Large Discrete Structures

Statistical Inference, Combinatorics and Limits

Dissertation

zur Erlangung des Doktorgrades
der Naturwissenschaften

vorgelegt beim Fachbereich 12
der Johann Wolfgang Goethe-Universität
in Frankfurt am Main

von
Maximilian Grischa Hahn-Klimroth
aus Bad Soden

Frankfurt 2021
(D 30)

vom Fachbereich Informatik und Mathematik der

Johann Wolfgang Goethe-Universität als Dissertation angenommen.

Dekan:

Prof. Dr. Lars Hedrich

Gutachter:

Prof. Dr. Amin Coja-Oghlan

Prof. Dr. Uriel Feige

Datum der Disputation: 11.05.2021

Contents

Acknowledgement

Throughout my doctoral studies that ultimately led to the writing of this dissertation I have received a great deal of support.

At first, I would like to express my deep gratitude to my first supervisor, Prof. Dr. Coja-Oghlan, for his guidance and continued mentorship as well as fruitful collaboration on an equal footing. I am especially thankful that he gave me the opportunity to be part of the active research community by, for instance, letting me attend various conferences and workshops. Not to forget, he always had an open door and gave me helpful advice as well as honest feedback. Also, I appreciate his tireless efforts to sweeten our joint lunch breaks which always contained interesting and relaxed conversations. Furthermore, I am grateful for him making sure that all of us were able to enjoy the sights and landmarks around the aforementioned conference venues.

Second, I thank my second supervisor, Prof. Dr. Yury Person, for his guidance and mentorship. While being physically separated, he made sure that we stayed in contact, by, for instance, inviting me multiple times to Ilmenau. Moreover, it was he who initially excited me for the subject of probabilistic and extremal combinatorics during my Bachelor's studies and introduced me to the "Stiftung Polytechnische Gesellschaft" during my Master's degree. The latter institution played a key role in my academic career.

More precisely, this dissertation would not have been possible without the funding of Stiftung Polytechnische Gesellschaft Frankfurt am Main. Besides this financial support, I thank especially Tobias König for his never ending interest and support in a variety of ways. He meticulously organised some of the best seminar courses I could ever imagine which brought me a big step forward in terms of personal and professional skills.

Furthermore, I especially thank Prof. Dr. Uriel Feige whose very detailed comments led to numerous corrections of one of my publications and increased its quality significantly.

Additionally, I would like to thank all my esteemed colleagues in Frankfurt: Jean Bernoulli Ravelomanana, Joon Lee, Maurice Rolvien, Michèle Fellinghauer, Noëla Müller, Oliver Gebhard and Philipp Loick who made my time as a PhD student unforgettable. I thank you for our productive joint research as well as many stimulating conversations and welcome distractions to rest my mind. In especially, I thank my former office mate Philipp Loick that I could always count on his support. Furthermore, I am really glad having found two coffee lovers amongst my colleagues, Joon Lee and Oliver Gebhard, with whom I could enjoy short breaks on our joint office days. Last but not least, I especially thank Oliver Gebhard for his touristic guidance on our various joint trips. Moreover, I want to thank Olaf Parczyk for his patience in our joint projects and especially for giving me the possibility to visit him in London.

I am very grateful that my mother, Inge Hahn-Klimroth, gave me the opportunity to study mathematics in the first place and supported me throughout my whole academic training. Further, I thank Jennifer Gübert for her emotional support while I was writing this thesis.

Finally, I want to thank Jennifer Gübert, Inge Hahn-Klimroth, Joon Lee and Philipp Loick for proofreading parts of this thesis.

Abstract

Studying large discrete systems is of central interest in, non-exclusively, discrete mathematics, computer sciences and statistical physics. The study of *phase transitions*, e.g. points in the evolution of a large random system in which the behaviour of the system changes drastically, became of interest in the classical field of random graphs, the theory of spin glasses as well as in the analysis of algorithms [78, 82, 121].

It turns out that ideas from the statistical physics' point of view on spin glass systems can be used to study inherently combinatorial problems in discrete mathematics and theoretical computer sciences (for instance, satisfiability) or to analyse phase transitions occurring in inference problems (like the group testing problem) [68, 135, 168]. A mathematical flaw of this approach is that the physical methods only render mathematical conjectures as they are not known to be rigorous.

In this thesis, we will discuss the results of six contributions. For instance, we will explore how the theory of *diluted mean-field models* for spin glasses helps studying random constraint satisfaction problems through the example of the random 2−SAT problem. We will derive a formula for the number of satisfying assignments that a random 2−SAT formula typically possesses [2].

Furthermore, we will discuss how ideas from spin glass models (more precisely, from their *planted* versions) can be used to facilitate inference in the group testing problem. We will answer all major open questions with respect to non-adaptive group testing if the number of infected individuals scales sublinearly in the population size and draw a complete picture of phase transitions with respect to the complexity and solubility of this inference problem [41, 46].

Subsequently, we study the group testing problem under sparsity constrains and obtain a (not fully understood) phase diagram in which only small regions stay unexplored [88].

In all those cases, we will discover that important results can be achieved if one combines the rich theory of the statistical physics' approach towards spin glasses and inherent combinatorial properties of the underlying random graph.

Furthermore, based on partial results of Coja-Oghlan, Perkins and Skubch [42] and Coja-Oghlan et al. [49], we introduce a consistent limit theory for discrete probability measures akin to the graph limit theory [31, 32, 128] in [47]. This limit theory involves the extensive study of a special variant of the cut-distance and we obtain a continuous version of a very simple algorithm, the pinning operation, which allows to decompose the phase space of an underlying system into parts such that a probability measure, restricted to this decomposition, is close to a product measure under the cut-distance. We will see that this *pinning lemma* can be used to rigorise predictions, at least in some special cases, based on the physical idea of a *Bethe state decomposition* when applied to the Boltzmann distribution.

Finally, we study sufficient conditions for the existence of perfect matchings, Hamilton cycles and bounded degree trees in randomly perturbed graph models if the underlying deterministic graph is sparse [93].

1. Introduction

Large discrete systems play a central role in discrete mathematics, theoretical computer sciences, statistical physics as well as in statistics. Analysing such systems became of major interest in the last decades. Prominent examples are the study of phase transitions on classical random graphs, the modelling and analysis of spin glasses, creating and analysing limit theories of discrete structures (like the graph limit theory) as well as the asymptotic analysis of algorithms. While this list is far from being complete, researchers from different fields realised that the interdisciplinary application of certain methods can be used very profitably [135, 168].

This thesis gives examples in which methods inspired by the statistical physics analysis' of large spin glass systems are used to provide rigorous mathematical understanding of statistical inference problems as well as a classical random constraint satisfaction problem. The richness provided by this approach is due to combining ideas from physics like message passing with combinatorial arguments and interpretations. Different combinatorial insights will be used to study large perturbed graphs, thus deterministic graphs where a bit of randomness is added.

Finally, a rigorous mathematical approach to certain physics' intuitions is carried out by analysing a particular form of the cut-distance for discrete probability measures. We will create a theory of limiting objects comparable to the well known and rich theory of graph limits [31, 32, 128] and formalise the intuition of basic statistical physics' concepts like *pure states*. This limiting theory will provide an elegant algorithmic regularity lemma for probability measures which can be translated into an algorithmic version of the weak regularity lemma for graphs [84].

1.1. Message passing and a statistical physics' approach

In this section, we will briefly introduce some of the most basic concepts of statistical physics that will be used to describe mathematical problems in an elegant and uniform way. We will, as [135], describe statistical physics as a part of probability theory, and will only sometimes give a meaningful interpretation of the concepts in nature.

1.1.1. Physical systems and important quantities

Let Ω be a finite set and n be the size of a physical system. Then we say that this system contains n *particles* and call Ω^n the *configuration space*. Thus, a *configuration* $\sigma \in \Omega^n$ assigns each particle i a *spin* $\sigma_i \in \Omega$. Furthermore, for $k \geq 1$, we define the *Hamiltonian* (or *energy*) of a configuration $\sigma \in \Omega^n$ as a function

$$H(\sigma) = -\sum_{i_1,\dots,i_k} J_{i_1,\dots,i_k}(\sigma_{i_1},\dots,\sigma_{i_k}). \tag{1.1.1}$$

In the special case $k = 1$, the system is called a *non-interacting* system as the different particles do not interact. Conversely, if $k \geq 2$, the system is called *$k-$body interacting*.

The real numbers $J_{i_1,\dots,i_k}$ express the interactions of the particles. This setup is quite general and allows for a lot of modifications, i.e., for each choice of Ω, k and H a different system is described. Some systems of this family of systems are well known physical models, for example for magnetism or for glasses. Depending on the choice of the interaction between its particles, a model is called *ferromagnetic*, *anti-ferromagnetic* or a *spin glass*. Intuitively, a ferromagnetic system prefers the interaction of particles with the same spin while an anti-ferromagnetic system prefers the opposite. Finally, in a spin glass, we find ferromagnetic as well as anti-ferromagnetic interactions.

Some prominent models that fit into this generic setup are the Potts model [16] or the Edwards-Anderson model [69]. The former is frequently used to study phase transitions – we will come to phase transitions later – or to model systems with (easy) nearest neighbour interactions while the latter is a

widely accepted mathematical model for magnetism [135]. Both models are defined on the d-dimensional grid $L = (V, E)$ as an underlying graph structure. The configuration space of the Potts model is $\Omega_{\text{Potts}} = \{1, 2, ..., q\}$ and it is a 2−body interacting system with Hamiltonian

$$H_{\text{Potts}}(\sigma) = -\sum_{ij \in E} J\mathbf{1}\{\sigma_i = \sigma_j\}. \tag{1.1.2}$$

Therefore, if two neighboured particles have the same spin under a configuration σ, the correspondent summand vanishes. Clearly, for $J > 0$, this system is ferromagnetic and for $J < 0$ it is anti-ferromagnetic. On the other hand, the Edwards-Anderson model has only two spins (negative and positive), thus $\Omega_{\text{EA}} = \{-1, +1\}$. Furthermore, as it is a model for magnetism, it adds the possibility of the appearance of some external magnetic field of strength $B > 0$. Moreover, its Hamiltonian allows for different interactions between two neighboured spins, depending on where they are placed inside the system, therefore

$$H_{\text{EA}}(\sigma) = -\sum_{ij \in E} J_{ij}\sigma_i\sigma_j - B\sum_{i \in V} \sigma_i. \tag{1.1.3}$$

Again, depending on the choice of J_{ij} being positive or negative for all edges ij, the model is ferromagnetic or anti-ferromagnetic, and when allowing different signs, it is a spin glass model. In such cases, where the interaction between particles can be written as in (1.1.3), the interactions J_{ij} are called *coupling constants*. The most prominent variant of the Edwards-Anderson model is the spin glass case in which the coupling constants are chosen from a symmetric probability distribution, for instance as standard Gaussians. In contrast, if we set $J_{ij} > 0$ for all coupling constants, we get the ferromagnetic *Ising model* on the grid as a special case [100]. Analogously, setting all coupling constants to a negative value renders the anti-ferromagnetic Ising model.

Now, to capture the idea behind a spin glass, an important physical observation is that any system strives for being in a state of minimal energy. In the anti-ferromagnetic or ferromagnetic Ising model it is quite easy to construct the configuration σ minimising $H(\sigma)$. But in the spin glass situation of the general Edwards-Anderson model it happens that a particle receives contradicting constraints from its neighbours. If this happens, we call the system *frustrated* at this particle. It is computationally hard to find a configuration that minimises the energy of the system [22]. On the other hand, there are many configurations with comparable low energy that *almost* minimise the system's energy – those states are called *metastable* [167]. Besides being mathematically challenging, spin glasses are a very good model for many real world materials, e.g. window glass or polymers or even granular media. Recent models like spin glasses for these materials assume that certain atoms and molecules occur randomly at random positions, thus the single particles behave either anti-ferromagnetically or ferromagnetically or are frustrated. The emerging theory of spin glasses allows to analyse those materials although they are random when seen microscopically, because on a macroscopic scale, they show describable properties [167].

1.1.1.1. Mean field and diluted mean field models

The interactions in the Potts model as well as the Edwards-Anderson model are defined on the grid. Therefore, those models have direct physical interpretations. The Edwards-Anderson model, for instance, is a realistic mathematical model for magnetism [135]. Unfortunately, the interactions facilitating this geometric constraints are challenging to analyse mathematically. One approach of simplification are so-called *mean-field* models. A well known example is the Sherrington-Kirkpatrick model (SK-model) [158] in which all particles $x_1 \dots x_n$ of a system interact with each other. More precisely, with $\Omega = \{-1, 1\}$, and $(\boldsymbol{J}_{ij})_{i,j=1\dots n}$ being standard Gaussians, its Hamiltonian reads

$$H_{\text{SK}}(\sigma) = -\frac{1}{\sqrt{n}}\sum_{i,j=1}^{n} \boldsymbol{J}_{ij}\sigma_i\sigma_j. \tag{1.1.4}$$

An important feature of the SK-model (as well as of general mean-field models) is the fact that the (distribution) of the Hamiltonian is invariant under the permutation of the coordinates, thus the geometric constraints of similar models on the grid (like the Edwards-Anderson model) vanish in the mean-field theory. The way mean-field models are defined, they neglect local and long-range structure, thus they are not able to describe all physically necessary properties of a system properly [157]. On the other hand, the models are somewhat easy, thus studying mean-field models is physically as well as mathematically more accessible than analysing their corresponding grid-models. For instance, the physical predictions of the properties of the SK-model [143] were proven rigorously by Talagrand [162].

Diluted mean-field models try to overcome the weakness of mean-field models while staying accessible to mathematical analyses. We introduce the diluted version of the SK-model by Viana and Bray [166]. Let $\Omega = \{-1, 1\}$ and $\boldsymbol{k} \sim \mathbf{Po}(\alpha n)$ be a Poisson random variable. Given $\boldsymbol{k}$, let $\{\boldsymbol{i}_k, \boldsymbol{j}_k\}_{k=1\dots\boldsymbol{k}}$ be uniform samples from $\{1, \dots, n\}$ and choose $\{\boldsymbol{J}_k\}_{k=1\dots\boldsymbol{k}}$ from a symmetric distribution. All those random variables are supposed to be mutually independent. Then, the Hamiltonian of the Viana-Bray model reads

$$H_{\mathrm{VB}}(\sigma) = -\sum_{k=1}^{\boldsymbol{k}} \boldsymbol{J}_k \sigma_{\boldsymbol{i}_k} \sigma_{\boldsymbol{j}_k}. \tag{1.1.5}$$

Thus, the Viana-Bray interactions are defined on a sparse random graph where each particle interacts with a Poisson number of other particles. Therefore, the long-range structure of a grid-model is clearly missing, but in contrast to the mean-field approach, the random graph looks locally a bit more like a grid, i. e., a very important feature is the finite connectivity on the random graph [134]. It turns out that analysing appropriately chosen diluted mean-field models may yield exact solutions of spin glass like models [144]. Thus, diluted mean field models carry important features of realistic models but are still mathematically approachable.

Furthermore, it turns out that very important problems in theoretical computer sciences, like random satisfiability and other random constraint satisfaction problems, can by expressed in the framework of diluted mean field models [140]. We will introduce random constraint satisfaction problems in Section 1.1.2.

After this short excursion to some prominent examples of particle systems, we formalise the intuition behind finding a configuration of minimal energy within a large system. We may think of any physical system with a Hamiltonian defined as in (1.1.1), but we will focus on diluted mean-field models.

1.1.1.2. Boltzmann distributions and ground states

If a system is observed showing a configuration of minimal energy, we will call this state a *ground state*. We denote by $\Omega_0^n \subset \Omega^n$ the set of configurations of minimal energy. Let us define a probability measure on the configuration space. Intuitively, a suitable probability measure should output the probability of observing a certain configuration, preferring those configurations of low energy.

Therefore, let $\beta > 0$ be the *inverse temperature* of the system. Having fixed an energy function H and an inverse temperature β, we define the *Boltzmann distribution* as a probability distribution on Ω^n as

$$\mu_\beta(\sigma) = \frac{\exp\left(-\beta H(\sigma)\right)}{Z_\beta}, \qquad \text{where} \qquad Z_\beta = \sum_{\sigma \in \Omega^n} \exp\left(-\beta H(\sigma)\right). \tag{1.1.6}$$

In (1.1.6), the normalising constant Z_β is known as the *partition function* of the system and we will see in due course that the partition function itself carries a lot of information about the system. Clearly, the lower the Hamiltonian $H(\sigma)$, the more probable is σ under μ_β. This effect decreases with β being small (the temperature of the system is large) and increases with β being large (the system's energy is small). More precisely, with $\beta \to 0$ (high-temperature limit), the Boltzmann distribution becomes the uniform distribution on Ω^n. On the other hand, in the low-temperature limit, the Boltzmann distribution corresponds to the uniform distribution on the ground states, thus on the configurations minimising the

system's energy. Formally,

$$\lim_{\beta\to 0}\mu_\beta(\sigma)=\frac{1}{|\Omega^n|},\qquad\text{and}\qquad\lim_{\beta\to\infty}\mu_\beta(\sigma)=\frac{\mathbf{1}\{\sigma\in\Omega_0^n\}}{\left|\Omega_0^n\right|}. \tag{1.1.7}$$

As statistical physics' main intention is to study the macroscopic behaviour of very large systems consisting of single particles and their interactions, various properties of a statistical system are studied in the *thermodynamical limit* ($n\to\infty$) [135]. Such a physical system can be described by various thermodynamic quantities. In the following, we will introduce the most important ones for our purposes. As already mentioned, Z_β is the partition function of the system. We denote by $\phi_{n,\beta}=\ln(Z_\beta)$ the *free entropy* and by $\phi_\beta=\lim_{n\to\infty}\frac{\ln(Z_\beta)}{n}$ the *free entropy density*. For any $\beta>0$, ϕ_β is convex, therefore it is continuous in every point in which it exists. We call the non-analytic points of ϕ_β *phase transitions*. Of course, non-analytical points are interesting mathematical objects, but physically speaking, those phase transition points indicate qualitative changes in the physical system [135].

In particular, those phase transitions corresponding to the so-called *replica symmetry breaking* are of deeper interest and will be studied later.

1.1.2. Phase transitions in special systems: Random Constraint Satisfaction

In the very general setup of a physical system with a given energy function (1.1.1), it is possible to describe constraint satisfaction problems (CSPs) which are prominently studied in computer sciences and mathematics. One of the most important constraint satisfaction problem is the $k-SAT$ problem.

1.1.2.1. $k-$SAT and factor graphs

A $k-SAT$ formula Φ is a conjunction of m clauses

$$\Phi=\Phi_1\wedge\ldots\wedge\Phi_m$$

such that each clause is a disjunction of exactly k literals out of n variables $x_1\ldots x_n$. One of the most intriguing questions is, obviously, whether there is a mapping $\sigma:\{x_1,\ldots,x_n\}\to\{-1,+1\}^n$ that assigns the Boolean values TRUE (+1) and FALSE (-1) to each variable such that each clause (and therefore the formula Φ) is satisfied. Without loss of generality, we suppose that a variable appears only once in a clause (i.e., x and $\neg x$ do never belong to one clause). Let us describe a $k-$SAT formula as a *factor graph*.

A factor graph $G=(V\cup F,E)$ is a bipartite graph with vertex classes V (variable nodes) and F (factor nodes) and edges E [81]. Sometimes, it is convenient to call the factor nodes *constraints*. We follow Mézard and Montanari [135] for construction of a factor graph G^Φ corresponding to a $k-$SAT formula Φ. Define

$$V=\{x_1,\ldots,x_n\}\qquad\text{and}\qquad F=\left\{a_1^\Phi,\ldots,a_m^\Phi\right\}.$$

Furthermore, we introduce two different types of edges $E=E^+\cup E^-$ such that $x_i a_j^\Phi\in E^+$ if and only if x_i appears in Φ_j and $x_i a_j^\Phi\in E^-$ if and only if $\neg x_i$ appears in Φ_j. The resulting factor graph has a fixed degree of k at each factor node while the variable node degree is not fixed in general.

While deciding whether a given $k-$SAT formula is satisfiable is known to be $\boldsymbol{NP}$ hard for $k\geq 3$ [109], it is easily possible to describe the problem as a physical system. Similarly as Kirkpatrick et al. [111], we interpret the n variables as particles of a system and assign each particle a spin from $\Omega=\{-1,+1\}$. A specific assignment is expressed by a configuration $\sigma\in\Omega^n$. The Hamiltonian turns out to be

$$H_{k-SAT}(\sigma)=\sum_{a_j^\Phi\in F}\mathbf{1}\left\{a_j^\Phi\text{ is not satisfied under }\sigma\right\}=\sum_{a_j^\Phi\in F}\mathbf{1}\left\{\sum_{x_i a_j^\Phi\in E^+}\sigma_i-\sum_{x_i a_j^\Phi\in E^-}\sigma_i=0\right\}.$$

Therefore, $H_{k-SAT}(\sigma)$ equals the number of unsatisfied clauses of Φ under σ. Let us now take the low-temperature limit and observe that the resulting Boltzmann distribution $\mu_\infty=\lim_{\beta\to\infty}\mu_\beta$ is the uniform

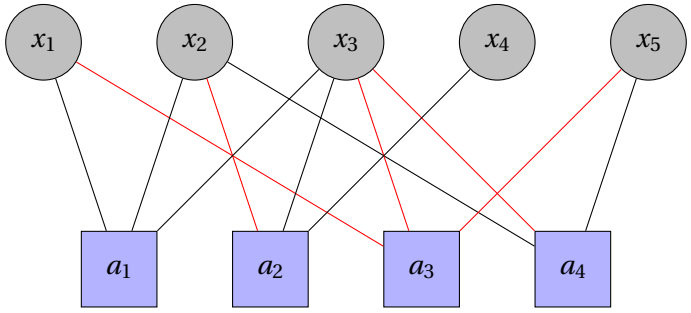

Figure 1.1.: The factor graph G^Φ corresponding to the 3−SAT formula $\Phi : (x_1 \vee x_2 \vee x_3) \wedge (\neg x_2 \vee x_3 \vee x_4) \wedge (\neg x_1 \vee \neg x_3 \vee \neg x_5) \wedge (x_2 \vee \neg x_3 \vee x_5)$. The $n = 5$ variable nodes are represented as circles while the $m = 4$ factor nodes are drawn as rectangles. The edge color represents the sign of the literal x_i in clause a_j. Thus, $s_{ij} = -1$ if edge $x_i a_j$ is red and $s_{ij} = 1$ otherwise. At each factor node a_j we have a local function $\Psi_{a_j} : \{-1,+1\}^{|\partial a_j|} \to \{-1,+1\}$ such that we have $\Psi_{a_j}(\sigma_{\partial a_j}) = \mathbf{1}\left\{\max_{x_i \in \partial a_j} \{\sigma_i s_{ij} = 1\}\right\}$ for an assignment $\sigma \in \{-1,+1\}^5$.

distribution on all configurations that satisfy the most clauses of Φ. Now let $\boldsymbol{\sigma} \sim \mu_\infty$ be a random sample, then the formula Φ is satisfiable if and only if $H_{k-SAT}(\boldsymbol{\sigma}) = 0$. The ***NP*** hardness of k−SAT for $k \geq 3$ immediately confirms the fact that it is in general computationally hard to find configurations of minimal energy in a physical system. We observe that

$$\mu_\infty(\sigma) = \lim_{\beta \to \infty} \frac{\prod_{a_j^\Phi \in F} \exp\left(-\beta \mathbf{1}\left\{a_j^\Phi \text{ is not satisfied under } \sigma\right\}\right)}{Z(\Phi)}. \tag{1.1.8}$$

The possibility to write the Boltzmann distribution in a factorised form as in (1.1.8) yields a few insights that directly generalise to other constraint satisfaction problems.

- Each factor of (1.1.8) corresponds to a factor node in the factor graph G^Φ, therefore, the factor graph perfectly describes the factorisation of the Boltzmann distribution into local constraints.
- For each $\beta > 0$ a non-satisfied clause gives a penalty of $\exp(-\beta)$ to the probability of observing a certain configuration. Given that Φ is satisfiable, in the low-temperature limit (or sometimes called *at zero temperature*), the Boltzmann distribution is supported on the satisfying assignments.
- Therefore, given that a formula is satisfiable, the partition function of the Boltzmann distribution in the low-temperature limit of the k-SAT problem ($Z_0(\Phi) = \left|\Omega_0^n\right|$) just equals the number of satisfying assignments and is an object of major interest in solving CSPs.

1.1.2.2. Random CSPs

In the general context of CSPs *random* constraint satisfaction problems gained a lot of attention [156]. Let us briefly sketch, what *random* means in this context. As we already discussed, a CSP can be expressed by a factor graph. Thus, given n variables and $\boldsymbol{m}$ factors (where the number of factors might be random itself), we randomly connect variable nodes and factor nodes. Depending on the problem itself, we can have arbitrary distributions of degree sequences for both, factor nodes as well as variable nodes. Given specific degree sequences, a random factor graph is chosen uniformly at random among all possible factor graphs satisfying the degree sequences.

Let us explain this very general description as before with the example of a random k-SAT formula. Given n variables and m clauses, initialise a factor graph on n variable nodes $x_1, \ldots, x_n$ and m factor nodes $a_1, \ldots, a_m$. Each factor node has degree k and variable x_i has a random degree $\boldsymbol{d}_i \sim \mathbf{Po}(mk/n)$.

Given the event that $\sum_{i=1}^{n} \boldsymbol{d}_i = mk$, select one (simple) graph with the given degree sequences uniformly at random.

In this setting, a very natural question arises: Is a formula obtained by this process satisfiable? Obviously, such questions can only be answered with high probability, thus with a probability tending to 1 with $n \to \infty$. This question has been studied for a long time and various tools of statistical physics have been applied to this problem [37, 111] that led to a (non-rigorously proven) prediction for a critical ratio $\alpha_s = m_s/n$, such that in the thermodynamic limit, each random $k-SAT$ formula with a smaller clause-to-variable ratio than α_s is satisfiable with high probability, whilst each formula with a larger ratio is not satisfiable with high probability. Such a phenomenon is called a *phase transition* and will be further discussed in a moment. This conjecture attained a lot of attention within the mathematical community and a lot of important steps towards proving this conjecture were done [4, 51, 54, 91] until Ding, Sly and Sun [64] managed to prove the existence of the phase transition for large enough k at

$$\alpha_s = 2^k \ln 2 - \frac{1+\ln 2}{2} + O(2^{-k}).$$

The existence of such a *satisfiability threshold* is not limited to the random k-SAT but is a genuine phenomenon of random constraint satisfaction problems [136]. The satisfiability threshold, however, is not the only interesting threshold regarding random CSPs. Indeed, let $\mathscr{S}$ be the solution space of a random constraint satisfaction problem (we may think about $\mathscr{S}$ being the set of all configurations satisfying a random k-SAT formula). We say that a pair of solutions is connected if its Hamming distance equals 1 and call a subset of Ω^n consisting of connected solutions a *cluster*. It turns out that the geometry of $\mathscr{S}$ has a highly complicated structure, but fortunately, the so-called *1-Replica Symmetry Breaking (1-RSB) Ansatz* from statistical physics draws a non-rigorously proven but fine grained picture. We will discuss this Ansatz in a moment. It conjectures that, with growing clause-to-variable ratio α, the solution space $\mathscr{S}$ undergoes four phase transitions (see Figure 1.2). While being a non-rigorous tool, the existence of the predicted phase transitions could be already proven rigorously for some models [161]. We let α start at 0 and let it increase continuously, then we observe four critical values $\alpha_u \le \alpha_{\text{clus}} \le \alpha_{\text{cond}} \le \alpha_s$ at which $\mathscr{S}$ changes dramatically [121, 138, 169, 170].

1. For $\alpha < \alpha_u$, there is exactly one cluster of solutions. Therefore, this phase is called the *unique* phase.

2. Once α exceeds α_u, the system is in the *extremal* phase. (Very) few and exponentially small clusters of satisfying configurations appear besides one cluster containing almost all solutions.

3. At the *clustering* threshold α_{clus}, the set of solutions shatters into exponentially many exponentially small clusters. Withing this phase, the size and and number of clusters reduces further with α increasing.

4. Subsequently, when α passes the *condensation* threshold, the solutions condense into a bounded number of clusters.

5. Finally, at the satisfiability threshold α_s, $\mathscr{S}$ becomes empty.

While an intuitive explanation what a phase transition is was already given, until now, we lacked a formal definition. We distinguish between *strict* and *coarse* phase transitions in random discrete systems. Let $\mathscr{P}$ be a property, and α be a parametrisation of the system. For instance, $\mathscr{P}$ could be the property that a random k-SAT formula is unsatisfiable and α is the clause-to-variable ratio. Then the system (the random factor graph) $\mathscr{G}$ undergoes a strict phase transition at a threshold α^* if for any $\varepsilon > 0$ we have

$$\mathbb{P}\left(\mathscr{G} \in \mathscr{P} \mid \alpha \le (1-\varepsilon)\alpha^*\right) = o(1) \qquad \text{and} \qquad \mathbb{P}\left(\mathscr{G} \in \mathscr{P} \mid \alpha \ge (1+\varepsilon)\alpha^*\right) = 1 - o(1).$$

Similarly, a coarse phase transition occurs if

$$\mathbb{P}\left(\mathscr{G} \in \mathscr{P} \mid \alpha = o(\alpha^*)\right) = o(1) \qquad \text{and} \qquad \mathbb{P}\left(\mathscr{G} \in \mathscr{P} \mid \alpha = \omega(\alpha^*)\right) = 1 - o(1).$$

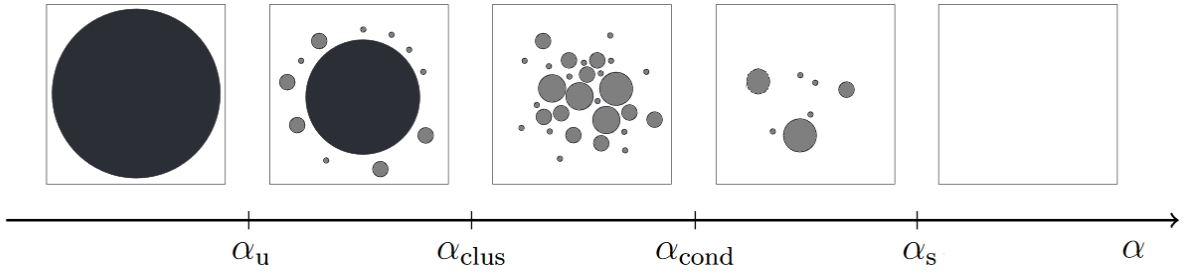

Figure 1.2.: Overview over the geometry of the solution space in random constraint satisfaction problems undergoing four important phase transitions obtained by the physics' 1-RSB Ansatz. The graphic is modified after [138, 169, 170].

(Clearly, the definition above requires $\mathscr{P}$ to become more probable for increasing α, it can analogously be defined for a property whose probability increases with decreasing α.)

Due to Friedgut and Bourgain [82] it is very well understood that non-uniform strict phase transitions exist in random structures if we look at monotonously increasing or decreasing properties. For example, the property for a random $k-$SAT formula to be unsatisfiable is clearly of this kind, as adding more clauses only increases the probability to be unsatisfiable. Of course, Friedgut and Bourgain's result only tells us that there *are* phase transitions, but it cannot tell us, *where* they are. For many graph properties regarding the random graph $\mathscr{G}(n,p)$[1] the exact position of the phase transitions is known [103]. For random constraint satisfaction problems, things are much more unexplored, even if recently the satisfiability thresholds could be pinned down rigorously for many random CSPs [25, 43, 55, 64]. The large temporal gap between the beginnings of studying random CSPs and the discovery of the exact satisfiability thresholds might be due to the reason that a simple first-moment calculation overestimates the number of satisfying assignments in many random CSPs dramatically, thus $\mathbb{E}[Z(\mathbf{\Phi})] \gg Z(\mathbf{\Phi})$ w.h.p. [55]. In the statistical physics' interpretation, this is due to the existence of the condensed phase in which the solution space is dominated by few large clusters. Therefore, the expected number of solutions is blown up by solutions that do not occur in typical formulas frequently and therefore, obviously, the first moment method that calculates the critical value α_{avg} such that $\mathbb{E}[Z(\mathbf{\Phi})] = 0$, fails.

Dealing with such effects is non-trivial, but the physics' 1-RSB Ansatz is a very promising (non-rigorous) approach towards understanding the condensed phase.

1.1.2.3. Phase transitions in physical systems and replica symmetry

Let us start with explaining major concepts of this Ansatz using a fairly easy model. The Curie-Weiss model [71] is a simple model for ferromagnetism. Let $\Omega = \{-1,1\}$ and define the corresponding Hamiltonian as

$$H_{CW}(\sigma) = -\frac{1}{n}\sum_{i\neq j\in[n]}\sigma_i\sigma_j - B\sum_{i=1}^{n}\sigma_i. \tag{1.1.9}$$

We directly observe by above's discussion that the Curie-Weiss model is a prime example of a *mean-field* model as all pairs of variables interact with each other. Again, B is the strength of an external magnetic field and the first sum in (1.1.9) reduces the energy of such configurations that do not exhibit many different spins. Following Mézard and Montanari [135], we are going to compute the partition function of the Curie-Weiss model non-rigorously, but each step can in principle be turned into a rigorous argument. We emphasise that all formulas (and their derivations) can be found in [135]. First, we need a

[1] We define the binomial random graph as Gilbert [90], thus every edge is present independently of the rest with probability p.

simple global property of a given configuration σ: the magnetisation. Let

$$m(\sigma) = \frac{1}{n}\sum_{i=1}^{n}\sigma_i$$

be the magnetisation of σ, then the Hamiltonian (1.1.9) can be written as

$$H_{CW}(\sigma) = n\left(\frac{1}{2} - \frac{1}{2}m(\sigma)^2 - Bm(\sigma)\right), \tag{1.1.10}$$

thus can be expressed in terms of the magnetisation. Therefore, in the partition function, we can cluster the summands by their magnetisation $m \in \mathscr{M} = \{-1+2\ell/n : \ell = 0\dots n\}$ and find [135, Eq. (2.75)]

$$Z_{CW}(\beta,B) = \exp\left(-n\beta/2\right)\sum_{m\in\mathscr{M}}\binom{n}{n(1+m)/2}\exp\left(\frac{n\beta}{2}m^2 + n\beta Bm\right). \tag{1.1.11}$$

Using the standard asympotical behaviour of the binomial coefficient $\binom{n}{\alpha n} \sim \exp\left(nH(\alpha)\right)$ where $H(\alpha)$ is the entropy of a $\mathbf{Be}(\alpha)$ variable and the definition

$$\phi_m(\beta,B) = -\frac{\beta}{2}(1-m^2) + \beta Bm + H\left(\frac{1+m}{2}\right),$$

(1.1.11) can be simplified to [135, Eq. (2.77)]

$$Z_{CW}(\beta,B) \sim \int_{-1}^{1}\exp\left(n\phi_m(\beta,B)\right)\mathrm{d}m. \tag{1.1.12}$$

Here and subsequently, we use $\sim$ for asymptotic equality. Then the Laplace method applied to (1.1.12) yields [135, Eq. (2.79)]

$$\frac{1}{n}\ln\left(Z_{CW}(\beta,B)\right) \sim -\beta\max_{m\in[-1,1]}\phi_m(\beta,B). \tag{1.1.13}$$

The maximum turns out to be achieved away from the boundary and is dependent on the inverse temperature β as well as the external magnetic field B. Let us, for the moment, suppose that there is no external magnetic field, thus $B = 0$. Then for $0 \le \beta \le 1$, $\phi_m(\beta,0)$ is concave and takes its unique maximum at $m = 0$. Therefore, for $n \to \infty$, if we sample a configuration from the underlying Boltzmann distribution μ_{CW} defined as in (1.1.6), we do not know the exact configuration, but in almost all cases, its magnetisation will turn out to be 0, as all other configurations have exponentially smaller probability mass under μ_{CW}. In other words, almost all created instances of the system will exhibit the same property - thus we find a high degree of symmetry between different instances.

If now, on the other hand, β exceeds the critical value of 1, $\phi_m(\beta,0)$ is symmetric with respect to the y-axis and exhibits two global maxima $m^*(\beta) > 0$ and $-m^*(\beta) < 0$ whose peaks become larger with increasing β. Therefore, if we sample sufficiently often from the Boltzmann distribution, half of the configurations will have magnetisation $m^*(\beta)$ and the other half will show a magnetisation of $-m^*(\beta)$. In this sense, the symmetry is *broken*.

Therefore, in the thermodynamic limit, the space of all configurations of minimal energy Ω_0^n can be decomposed into two *pure states* $\Omega_0^n = \Omega_+^n \cup \Omega_-^n$ where

$$\Omega_+^n = \left\{\sigma \in \Omega_0^n : m(\sigma) = m^*(\beta)\right\} \qquad \text{and} \qquad \Omega_-^n = \left\{\sigma \in \Omega_0^n : m(\sigma) = -m^*(\beta)\right\}.$$

In the setting of Figure 1.2, this corresponds to a condensed phase, where the configurations of one cluster have significantly more particles of spin -1 than the configurations in the second cluster.

This observation corresponds intuitively to a fairly important feature of condensation: the marginal probabilities of the single particles under the Boltzmann distribution are dependent. Indeed, if σ is a random element of Ω_0^n and we get the information that particle x_1 has spin $\sigma_1 = -1$, it is more likely for

σ to be an element of Ω_-^n and therefore, in expectation, the other particles prefer spin -1 as well.

Let us at this point shortly discuss the so-called *replica symmetry* and *replica symmetry breaking*. Two replicas can be just seen as identically distributed instances of the same system. If such replicas are symmetric, the following effect takes place. Suppose we sample $\boldsymbol{\sigma}, \boldsymbol{\sigma}'$ from the Boltzmann distribution of a physical system and calculate the *overlap* of $\boldsymbol{\sigma}$ and $\boldsymbol{\sigma}'$ which is, for the Curie-Weiss model, defined as

$$\langle \boldsymbol{\sigma}, \boldsymbol{\sigma}' \rangle = n^{-1} \sum_{i=1}^{n} \boldsymbol{\sigma}_i \boldsymbol{\sigma}'_i.$$

Now, if the system is replica symmetric, we suppose that the overlap between two randomly sampled configurations is concentrated around exactly one value. If this is not the case, we say that *replica symmetry breaking* is present. The last formulation has to be read very cautiously because its definition is actually a bit more complicated. We say that a system is still replica symmetric if the pure states are related by *global symmetries* [131, 135]. The latter is clearly the case in the Curie-Weiss model as Ω_-^n can be turned into Ω_+^n (and vice versa) by flipping the spin of each particle.

Overall, we learned the intuition behind replica symmetry and replica symmetry breaking as well as that we defined pure states. In the following, we will formalise this intuition a bit.

1.1.2.4. Pure state decomposition, replica symmetry and the cut-distance

Let us suppose that we analyse the solution space and the Boltzmann distribution of a random constraint satisfaction problem (or a diluted mean-field model). Thus, the underlying graph corresponding to the interactions of the n particles is a sparse random graph. The physics' prediction suggests that the Boltzmann distribution has a fairly comfortable property before condensation (thus, in the replica symmetric phase): there are no long-range correlations present. More specifically, suppose we know the prevalence for a specific spin of the $i-$th particle x_i. If particle x_j is far away (in the sparse random graph) from x_i, then we do not gain any information about the spin of x_j from this knowledge. Formally, given a probability measure μ on Ω^n, we denote by μ_i the marginal of μ on spin i. More generally, for a set $I \subset [n]$, μ_I is the marginal probability on the particles in I. Then, we say that μ is ε*-symmetric* if

$$\mathbb{E}_{\boldsymbol{i},\boldsymbol{j}} \left[\left\| \mu_{\{\boldsymbol{i},\boldsymbol{j}\}} - \mu_{\boldsymbol{i}} \otimes \mu_{\boldsymbol{j}} \right\|_{\mathrm{TV}} \right] \leq \varepsilon.$$

Thus, ε-symmetry formalises the intuition that far-apart particles do not influence each other.

Now, after the condensation threshold, such long-range correlations start to appear. Even if the graph still looks locally tree-like, like every sparse random graph, the knowledge about a spin of a far-away particle might influence the probability of observing a specific spin on another particle. Recall that in the condensed phase of a random constrained satisfaction problem, the solution space is predicted to consist of clusters of solutions. Knowing the spin of a single particle might contain information on which of those clusters the configuration at hand lies in and therefore, this information pushes the probability to observe a certain spin on a different particle. But, if $1-$RSB is present, it is predicted that within one of those clusters, those long-range correlations disappear. More precisely, if $\Xi_1, \ldots, \Xi_\ell$ is the partition into clusters of Ω^n in the condensation phase and the physicist's prediction of $1-$RSB takes place, the Boltzmann distribution conditioned on one of these *pure states*, is symmetric. Formally, if μ is the Boltzmann distribution, we have for all $1 \leq k \leq \ell$

$$\mathbb{E}_{\boldsymbol{i},\boldsymbol{j}} \left[\left\| \mu_{\{\boldsymbol{i},\boldsymbol{j}\}}[\cdot|\Xi_k] - \mu_{\boldsymbol{i}}[\cdot|\Xi_k] \otimes \mu_{\boldsymbol{j}}[\cdot|\Xi_k] \right\|_{\mathrm{TV}} \right] = o(1).$$

We stress at this point that we only described the so-called $1-$RSB Ansatz. In physics' literature, the replica symmetry breaking might be of different orders. Intuitively, the clusters of solutions described here might themselves decompose into clusters of solutions ($2-$RSB) iteratively (until $\infty-$RSB) before the pure states (and thus absence of long-range correlations) appear. If the higher order RSB is interpreted via the overlap distribution as previously 1-RSB, we find that $k-$RSB corresponds to the concentration of the overlap onto $k+1$ values up to symmetry. Of course, this perfectly fits into the picture drawn before with clusters of clusters of clusters... Indeed, suppose that $2-$RSB is present and suppose we have *big*

clusters C_j and contained *small* clusters C_{jk}. If we draw two configurations from different big clusters C_j and $C_{j'}$ (independent of the small clusters), we observe a certain overlap value concentrated around q_0 (this are two configurations which are very far apart). If, on the other hand, we sample configurations within the same small cluster C_{ij}, the overlap will be concentrated around q_1 and finally, if the configurations come from the same big cluster but from different small clusters, we find an overlap concentrated around a third value q_2.

Thus, the geometrical interpretation directly shows that the overlap distribution facilitates a kind of *ultra-metricity*. What is an ultra-metric? We say that a metric d is an ultra-metric if the triangle inequality holds in a stronger version, e.g. $d(a,b) \leq \max\{d(a,c), d(b,c)\}$. This analytical property has one advantage if compared to the geometric interpretation of clusters of clusters of... Indeed, in the case of $\infty-$RSB, the single clusters would not be well separated but the analytical property of ultra-metricity still holds and describes the relevant behaviour [135, Section 8.2.2]. The interested reader can find a detailed discussion of (approximate) ultra-metricity and the replica symmetry breaking in an article of Jagannath [101].

Let us leave the interpretation of replica symmetry breaking for the moment, we will come back to it in a moment. Mathematically spoken, there is a decomposition of the Boltzmann distribution which kind of resembles this pure state decomposition. It comes along with a very simple algorithm: the pinning operation [49]. Extending and analysing results in context of the pure-state decomposition is part of one contribution of this thesis (c.f. Section 2.3). To be more precise, for an arbitrary probability measure on Ω^n, we can express the property of being $o(1)$-symmetric in terms of being close to a specific measure under a carefully chosen metric - the *cut-distance*. For two probability measures η, ν on a finite set Ξ, let $\Gamma(\eta,\nu)$ denote the set of all couplings γ of η and ν, thus γ is a probability measure on $\Xi \times \Xi$ with marginals η and ν. Further, let $\mathbb{S}_n$ denote the set of permutations $\phi : [n] \to [n]$. Then, the cut-distance is defined as

$$\Delta_{\boxtimes}(\mu,\nu) = \inf_{\substack{\gamma\in\Gamma(\mu,\nu),\\ \phi\in\mathbb{S}_n}} \sup_{\substack{S\subset\Omega^n\times\Omega^n,\\ X\subset[n],\\ \omega\in\Omega}} \left| \sum_{\substack{(\sigma,\tau)\in S,\\ x\in X}} \gamma(\sigma,\tau)\left(\mathbf{1}\{\sigma_x=\omega\} - \mathbf{1}\left\{\tau_{\phi(x)}=\omega\right\}\right)\right|. \tag{1.1.14}$$

The definition (1.1.14) can be interpreted as a kind of a two-player game. The first player chooses a coupling and a permutation of the particles under which μ and η look as much alike as possible. Now, given the choices γ and ϕ, the second player tries to find a subset of coordinates and configurations, on which the two measures differ as much as possible. We stress that the cut-distance is indeed a very weak metric [49]. But it suffices to gather the idea of $\varepsilon-$symmetry. Let μ be a probability measure on Ω^n and define

$$\bar{\mu}(\sigma) = \prod_{i=1}^{n} \mu_i(\sigma_i) \tag{1.1.15}$$

as the product measure on the marginals of μ. Clearly, $\bar{\mu}$ is a probability measure on Ω^n as well and if μ is a Boltzmann distribution such that the spins on all particles are completely independent, we have $\mu \equiv \bar{\mu}$. Thus, if most pairs of particles do not influence each other, we would expect that μ and $\bar{\mu}$ are close to each other under a certain distance. The cut-distance turns out to exactly express this connection. As proven by Coja-Oghlan and Perkins [53, Proposition 2.5], we have

$$\Delta_{\boxtimes}(\mu,\bar{\mu}) \leq \varepsilon^3 \Longrightarrow \mathbb{E}_{\boldsymbol{i},\boldsymbol{j}}\left[\left\|\mu_{\{\boldsymbol{i},\boldsymbol{j}\}} - \mu_{\boldsymbol{i}}\otimes\mu_{\boldsymbol{j}}\right\|_{\mathrm{TV}}\right] \leq O(\varepsilon)$$

and

$$\mathbb{E}_{\boldsymbol{i},\boldsymbol{j}}\left[\left\|\mu_{\{\boldsymbol{i},\boldsymbol{j}\}} - \mu_{\boldsymbol{i}}\otimes\mu_{\boldsymbol{j}}\right\|_{\mathrm{TV}}\right] \leq \varepsilon^3 \Longrightarrow \Delta_{\boxtimes}(\mu,\bar{\mu}) \leq O(\varepsilon).$$

Thus, the cut-distance is a perfectly suiting mathematical tool to formalise the idea of pure states and the absence of long range correlations. Furthermore, as already mentioned, the cut-distance can be used to find a decomposition of any configuration space Ω^n for a large enough system ($n \geq n_0 \in \mathbb{N}$) into

finitely many parts $E_1, \dots, E_\ell$ such that any probability measure conditioned on a specific part $\mu[\cdot \mid E_j]$ is close to the product of its marginals under the cut-distance, hence $\Delta_{\boxtimes}\left(\mu[\cdot \mid E_j], \bar{\mu}[\cdot \mid E_j]\right) \leq \varepsilon$ (for any $\varepsilon > 0$). This result is a type of a *regularity lemma* and will be discussed in more detail in Section 1.3.1.

The cautious reader might ask how this finite decomposition fits into the predictions of replica theory. Indeed, the physics' prediction is that the phase space decomposes into infinitely many pure states (if $n \to \infty$) if some kind of replica symmetry breaking is present [131] while the pinning operation – and regularity lemmas which will be discussed later – yield a finite decomposition. But the pinning operation (as well as the regularity lemmas) only guarantee the existence of states $E_1, \dots, E_\ell$ with above's property such that $\mu(E_1 \cup \dots \cup E_\ell) \geq 1 - \varepsilon$ for any small but constant $\varepsilon > 0$. Therefore, there can be infinitely many clusters which carry very little probability mass under the Boltzmann distribution.

After having described how the solution space of random constraint satisfaction problems (respectively, diluted mean field models) might look like, we will study a very specific random constraint satisfaction problem in more detail which plays the leading role in one of this thesis's contributions.

1.1.3. The random 2-SAT problem

While we already introduced the $k-$SAT problem for arbitrary k, we will focus on the very special case $k = 2$. This setup plays a very specific role, as it is the only k-SAT problem in which deciding whether a given formula is satisfiable and, if so, finding a satisfying assignment is computationally easy [120]. For the random $2-$SAT problem, the satisfiability threshold is well known since the early 1990's by independent works of Chvatal and Reed [40] and Goerdt [91], which both link the problem to the well studied percolation phase transition in a random directed graph. Subsequently, Bollobás et al. [30] managed to analyse the scaling window of the satisfiability threshold in more detail which turns out to correspond to the scaling window of the giant component in the binomial random graph [29].

Even if the satisfiability threshold itself is well understood, things were completely different for a question that seems quite innocuous at first. If a random $2-$SAT formula is satisfiable w.h.p., how many satisfying assignments exist? Indeed, this question was posed prominently by Fernandez de la Vega [78] and it actually turns out that counting the number of satisfying assignments is computationally hard, this counting problem lies in $\#\mathbf{P}$ [165].

Describing and analysing the partition function of a random $2-$SAT formula would answer this question as the partition function exactly counts the number of satisfying assignments, at least, if the formula is satisfiable in the first place. Luckily, there is a non-rigorous heuristics from statistical physics which leads to a precise prediction of how this partition function looks like. One of this thesis's contribution verifies this prediction rigorously as will be seen in Section 2.2. Intuitively, the non-rigorous physics' approach calculates the marginals of the Boltzmann distribution, thus for each variable node x they get the probability that x is assigned the value 0 or 1 under a randomly chosen satisfying assignment $\boldsymbol{\Phi}$. Now, it is possible to connect $Z(\boldsymbol{\Phi})$ with those marginal probabilities through an operator called *Bethe functional*. We will subsequently describe how to (non-rigorously) calculate marginals of the Boltzmann distribution in Section 1.1.4 and how to link them to the partition function in Section 1.1.5. We start by finding the marginal probabilities of the Boltzmann distribution. In physics' literature, specific message passing algorithms are conjectured to be able to do so.

1.1.4. Getting the marginals through message passing

The key idea behind message passing algorithms is the following. Given a factor graph $\mathcal{G}$, an algorithm computes *messages* for each edge of $\mathcal{G}$ such that in each round of the algorithm, a set of messages is sent on every edge of the graph in parallel. Those messages can be sent from the variables to the factors and vice versa. A computationally intriguing advantage of message passing algorithms is that all messages are computed through local functions at the vertices of $\mathcal{G}$. Probably the most prominent message passing algorithm is Belief Propagation BP. It was discovered under various names in different fields of research, for instance, in statistical physics it is known as the *Bethe-Peierls approximation* [27], while coding theorists developed the *sum-product algorithm* [122]. The terminology *Belief Propagation* which we will use subsequently, has its origins in the research towards artificial-intelligence [145].

1.1.4.1. Belief Propagation

Let $\mathcal{G} = (V \cup F, E)$ be a given factor graph. BP is used to estimate the marginal distribution for each variable, for instance, if $\mathcal{G}$ represents a $k-$SAT formula, this corresponds to the probability of a specific variable being set to 1 under a random satisfying assignment. While its first occurrence can be traced back to the 1930's in statistical physics [27], it was Pearl [145] who proved that BP correctly calculates the marginals if $\mathcal{G}$ is a tree. Furthermore, it is a widespread conjecture that BP performs well if the graph is at least locally tree-like and there are almost no long-range correlations as discussed in Section 1.1.2.4 [146]. Probably due to its performance as well as being an efficient algorithm, BP finds its applications in artificial intelligence and information theory. Empirically it was shown that BP can be used in various applications including ldpc-codes and turbo-codes in coding theory, free energy approximation in statistical physics and satisfiability in theoretical computer science [38].

Let us now introduce the messages sent by BP formally. In order to do so, we require a few definitions. Let Ω be a finite set (the set of spins as before) and $\mathcal{G} = (V \cup F, E)$ a factor graph. Denote by $x_1, \dots, x_n \in V$ the n variables and by $a_1, \dots, a_m \in F$ the m factors of $\mathcal{G}$. Furthermore, let $\sigma \in \Omega^n$ be a configuration that assigns each variable a specific spin from Ω. Further, for $I = \{i_1, \dots, i_k\} \subset [n]$, we set $\sigma_I = (\sigma_{i_1}, \dots, \sigma_{i_k}) \in \Omega^k$ and denote by Ω^I the subspace of Ω^n given by the coordinates of I. Moreover, for a vertex $v \in V \cup F$, we denote by $\partial_{\mathcal{G}} v \subset V \cup F$ the set of neighbours of v in $\mathcal{G}$. If the context clarifies what the underlying graph structure is, we write ∂v for the sake of brevity. Finally, we use $\propto$ to express equality up to a normalisation constant and write $f(n) \sim g(n)$ if $f(n)/g(n) \to 1$ for $n \to \infty$.

Before stating the BP messages, let us revisit the definition of a factor graph. As already introduced, one merit of a factor graph is that the Hamiltonian of the corresponding physical system factorises such that the Boltzmann distribution can be written as

$$\mu(\sigma) \propto \prod_{i=1}^{m} \psi_{a_i}(\sigma_{\partial a_i}). \tag{1.1.16}$$

Here, $\psi_{a_1}, \dots, \psi_{a_m}$, are the local contributions of the single factor nodes to the system's energy. Of course, given a factor graph $\mathcal{G} = (V \cup F, E, \Psi)$ with *weight functions* $\Psi = \{\psi_{a_1}, \dots, \psi_{a_m}\}$, we can associate a corresponding Hamiltonian such that the system's energy is calculated via (1.1.16). It turns out that sometimes it is more convenient to use the factor graph notation and talk about weight functions instead of referring to the physical interpretations.

We are now in position to state the BP messages on such a factor graph $\mathcal{G} = (V \cup F, E, \Psi)$. For $y \in \Omega$, a variable $x \in V$ and a test $a \in F$ they read

$$\nu^{(t+1)}_{x \to a}(y) \propto \prod_{b \in \partial x \setminus a} \nu^{(t)}_{b \to x}(y) \qquad \text{and} \qquad \nu^{(t)}_{a \to x}(y) \propto \sum_{\sigma_{\partial a \setminus x}} \psi_a(\sigma_{\partial a}) \prod_{x' \in \partial a \setminus x} \nu^{(t)}_{x' \to a}(y). \tag{1.1.17}$$

More precisely, the set $\left\{\nu^{(t)}_{x \to a}\right\}_{x \in V, a \in F}$ is called the set of *variable-to-factor messages* at time t while the *factor-to-variable messages* are $\left\{\nu^{(t)}_{a \to x}\right\}_{x \in V, a \in F}$. Clearly, each message itself is a probability distribution on Ω. Intuitively speaking, $\nu^{(t)}_{x \to a}$ represents the marginal distribution of spins on variable x in a model that does not contain factor a and analogously, $\nu^{(t)}_{a \to x}$ is the marginal distribution of x in a model where all factors in which x is contained except of a have been deleted.

Observing those messages, various natural questions arise. Probably the most intriguing are the following.

- Do these messages converge to a fixed-point?
- If so, is the fixed-point unique?
- If these messages indeed converge to a unique fixed-point, how are those messages related to the marginals of the Boltzmann distribution?

These questions can be answered rigorously for tree-factor graphs, thus for graphs $\mathcal{G}$ that do not contain cycles. Furthermore, it is conjectured that similar results hold for locally tree-like graphs if the solution

space exhibits certain properties. This conjecture was proven rigorously in some special cases [121, 50]. We will discuss those questions in Section 1.1.5. BP is designed to be applied to any distribution that can be written like (1.1.16), but in many CSPs, the weight functions have a very specific form, as the Hamiltonian just counts the number of unsatisfied constraints (clauses, in above's terminology on satsifiability). Therefore,

$$\psi_a(\sigma_{\partial a}) = 1 - \mathbf{1}\{a \text{ is satisfied under } \sigma\}. \tag{1.1.18}$$

In this special case, the messages (1.1.17) simplify remarkably, i.e., if the set of messages at time zero is initialised with values in $\{0,1\}$, then this condition still holds after an arbitrary number of updates [135, Proposition 14.5]. The resulting update rules are known as *Warning Propagation* WP [75].

1.1.4.2. Warning Propagation

The semantic interpretation of those integer-valued BP messages is the reason why WP holds its name. More precisely, for a spin $y \in \Omega$, a variable x and a factor a, we have

$$\begin{aligned} \nu_{x\to a}(y) = 1 \quad &\leftrightarrow \quad \text{according to all factors } b \in \partial x \setminus a \text{ variable } x \\ &\qquad \text{should not have spin } y \text{ under a satisfying assignment.} \\ \nu_{x\to a}(y) = 0 \quad &\leftrightarrow \quad \text{according to all factors } b \in \partial x \setminus a \text{ variable } x \\ &\qquad \text{may take spin } y \text{ under a satisfying assignment.} \end{aligned} \tag{1.1.19}$$

Thus, the variable-to-factor messages $\{\nu_{x\to a}\}$ *warn* a factor node a if a variable can probably not be used to satisfy a taking certain values. WP itself is a broadly used algorithm in random CSPs and it is known that it finds all direct implications of a partial assignment of the variables, based on the local structure [135]. Furthermore, it is known to find satisfying assignments of different satisfiability problems under mild conditions, for instance, it can be used to solve certain instances of random 3-SAT [75].

Let us suppose for the moment that BP or WP are exact on a given problem. We will now discuss what we actually understand by *being exact*.

1.1.5. Boltzmann marginals and the partition function

Let, as above, $\left\{\nu^{(t)}_{x\to a}\right\}$ and $\left\{\nu^{(t)}_{a\to x}\right\}$ be the BP messages at time t. Then we define the *BP estimation of the marginals at time t* as

$$\nu^{(t)}_x(y) \propto \prod_{a\in\partial x} \nu^{(t)}_{a\to x}(y). \tag{1.1.20}$$

Hence, $\nu^{(t)}_x$ is the product measure over all incoming messages at variable x. If μ is the corresponding Boltzmann distribution, we have that $\nu^{(t)}_x$ is an estimation for μ_x. If we say that BP is exact (for instance, on factor graphs that are acyclic), then it is rigorously proven that

$$\nu^{(t)}_x = \mu_x$$

if t is large enough. This, of course, already implies that the BP messages converge to a fixed-point. We call the set of fixed-point messages $\{\nu^*_{x\to a}\}$ and $\{\nu^*_{a\to x}\}$ respectively.

Now, let us discuss, how this helps understanding the partition function of the underlying system. Given a factor graph $\mathcal{G} = (V \cup F, E, \Psi)$ and a set of messages $\nu = (\{\nu_{x\to a}\}, \{\nu_{a\to x}\})$, we define the *Bethe free*

entropy as follows. Recall that those messages are probability distributions on Ω and define

$$\Xi_a(\nu) = \ln\left(\sum_{\sigma_{\partial a}\in\Omega^{\partial a}} \psi_a(\sigma_{\partial a}) \prod_{x\in\partial a} \nu_{x\to a}\right), \qquad \Xi_x(\nu) = \ln\left(\sum_{\omega\in\Omega} \prod_{b\in\partial x} \nu_{b\to x}\right),$$
$$\Xi_{x,a} = \ln\left(\sum_{\omega\in\Omega} \nu_{x\to a}(\omega)\nu_{a\to x}(\omega).\right) \tag{1.1.21}$$

Thus $(\Xi_a)_{a\in F}$ and $(\Xi_x)_{x\in V}$ describe the entropy on the factor nodes and on the variables respectively, while $\Xi_{x,a}$ measures their interaction. If we put in the set of fixed-point messages ν^*, we get the Bethe free entropy $\mathbf{\Phi}$ as

$$\mathbf{\Phi} = \sum_{a\in F} \Xi_a(\nu^*) + \sum_{x\in V} \Xi_x(\nu^*) - \sum_{ax\in E} \Xi_{x,a}(\nu^*). \tag{1.1.22}$$

If BP is exact on a model, then we find

$$\ln Z = \mathbf{\Phi}. \tag{1.1.23}$$

Thus, (1.1.23) gives a recipe, how to calculate the partition function of a physical system, for instance, a random constraint satisfaction problem, by using an easy to implement efficient algorithm. Due to the ***NP***-hardness of calculating $\ln Z$ in general, as discussed earlier, this can of course only be true for very specific instances of random CSPs (if we suppose $\boldsymbol{P} \neq \boldsymbol{NP}$).

Indeed, it turns out that the *Bethe approximation by Belief Propagation*, which we will call the procedure above, yields the correct value of the partition function sometimes [50, 52, 145] and sometimes it does not [126]. The latter is assumed to happen, if the system's solution space had undergone a phase transition at which replica symmetry breaking occurred. A statistical physics' approach suggests to run BP on a modified problem in this case.

1.1.5.1. The (1-RSB) Cavity Method

Before going deeper into the *1-RSB cavity approach*, we shortly stress that the method itself is a highly non-rigorous technique. Furthermore, this presentation's focus is solely to draw an intuitive idea behind the 1-RSB cavity method and not to carry out technical details.

Intuitively, BP depends on a central assumption, namely, whenever we delete a factor node, the spins on the affected variables become roughly independent. This assumption is clearly violated when the factor graph either contains short cycles, or whenever long-range correlations appear [135]. As (sparse) random factor graphs which we deal with in this thesis (or: diluted mean-field models) look locally tree-like, the failure of BP needs to be due to the appearance of such long-range correlations. In the replica symmetric phase, those are supposed to be negligible, thus we study systems in a phase in which replica symmetry is broken.

We recall that in the presence of 1-RSB, the configuration space Ω^n is supposed to decompose into pure states (c.f. Section 1.1.2.4). Let us take a different look on those pure states. Recall that BP is known to render the correct marginals of the Boltzmann distribution on some replica symmetric factor graphs and recall that given $\mathcal{G} = (V \cup F, E, \Psi)$, we can write the Boltzmann distribution as

$$\mu(\sigma) \propto \prod_{a\in F} \psi_a(\sigma_{\partial a}).$$

Let $U \subset V$ be a subset of the variables, then we define $\mathcal{G}[U] = (U \cup F[U], E[U], \Psi)$ as the induced factor graph on U as follows. It contains

- all variables of U,
- all factor nodes $a \in F$ such that $\partial a \subset U$,
- all edges a, x with $a \in F[U], x \in U$,

- all half-edges x, a with $a \notin F[U]$ but $x \in U$.

The half-edges play an important role. Let us denote the set of those half-edges by $\mathscr{H}(U)$. Suppose that BP renders the correct marginals, then the messages $\{\nu_{a\to x}\}_{(x,a)\in\mathscr{H}(U)}$ on the half-edges can be seen as *boundary conditions* which represent the influence of the factors and variables outside of U. We will call such a set U coming with $\mathscr{G}[U]$ a *cavity*. Observe that some variables in U miss factor nodes (constraints) compared to their distribution in $\mathscr{G}$. If the system is replica symmetric, we would suppose that we can express the Boltzmann marginal on U – at least approximately – by the weight functions at the factor nodes inside of U and the boundary condition, thus

$$\mu_U(\sigma_U) \propto \prod_{a\in F[U]} \psi_a(\sigma_{\partial a}) \prod_{ax\in\mathscr{H}(U)} \nu_{a\to x}(\sigma_x) + \varepsilon_n. \tag{1.1.24}$$

Led by this intuition, we call a probability measure μ on Ω^n a *Bethe measure* or *Bethe state* if in the thermodynamic limit, $n\to\infty$, there is a set of messages $\{\nu_{a\to x}\}$ such that μ satisfies (1.1.24) for almost all finite subsets U [135, Definition 19.1]. We highlight that there has to be one set of BP messages satisfying (1.1.24) for (almost) all finite subsets of variables and in fact it turns out that messages satisfying the equation for almost all U are actually very close to valid Belief Propagation messages (1.1.17). To be a bit more precise, an *almost-solution* of the BP equations is a set of messages that satisfy almost all BP equations up to a vanishing error term of $o(1)$.

What does this imply? Any set of messages satisfying (1.1.24) for almost all U, thus a Bethe measure, corresponds to an almost-solution of the Belief Propagation messages. It is far from true that the converse is correct as well [135, Example 19.2] in every situation, but a core assumption of the 1-RSB cavity method is that this is indeed correct in the problem at hand. Suppose we have a set of almost-solutions of the BP equations $\{\nu^i = (\nu^i_{x\to a}, \nu^i_{a\to x})\}_{i=1\dots\ell}$, then we can associate with each of those (almost) fixed-points a corresponding Bethe measure μ^i (1.1.24) as well as the corresponding Bethe free entropy $\mathbf{\Phi}_i$ via (1.1.22). We call a Bethe measure *extremal* if it does not exhibit long-range correlations. We denote the set of all extremal Bethe measures out of $\{\mu^i\}_{i=1\dots\ell}$ as $\{\tilde\mu^i\}_{i=1\dots\tilde\ell}$ corresponding to BP messages $\{\tilde\nu^i\}_{i=1\dots\tilde\ell}$ Now, there are three basic assumptions on a model that make the 1-RSB method work heuristically. To this end, let $\mathbf{\Sigma}:\mathbb{R}\to\mathbb{R}_+$ be a function, which is called *complexity function* in physics' language.

1. For any interval $[\phi,\phi']$, the number of almost-solutions with Bethe free entropy $\mathbf{\Phi}_i \in [n\phi, n\phi']$ equals $\exp(n\mathbf{\Sigma}^* + o(n))$ where $\mathbf{\Sigma}^* = \sup_{\sigma^*\in[\phi,\phi']}\{\mathbf{\Sigma}(\phi^*)\}$. Thus, roughly speaking, the number of almost-solutions with Bethe free entropy of approximately $n\phi$ is given by $\exp\big(n\mathbf{\Sigma}(\phi)\big)$.

2. The Boltzmann distribution can be expressed as a convex combination of extremal Bethe measures
$$\mu(\sigma) = \sum_{i=1}^{\tilde\ell} \omega_i \tilde\mu^i(\sigma)$$
with weights $\omega_i = \exp(\mathbf{\Phi}_i)/\mathscr{Z}$ such that $\mathscr{Z} = \sum_{i=1}^{\ell}\exp(\mathbf{\Phi}_i)$.

3. The number of extremal Bethe measures $\tilde\ell$ equals, up to the leading order, the number of almost-solutions to the BP equations, thus the number of extremal Bethe measures with Bethe free entropy $\sim n\phi$ equals approximately $\exp\big(n\mathbf{\Sigma}(\phi)\big)$.

Now suppose that the three assumptions are satisfied. Then we build a new physical system (an auxiliary model) from the original system we started with. More precisely, we interpret the BP messages as variables of the auxiliary model and the corresponding Bethe measures become the configurations. To this end, let ζ be the set of the extremal Bethe measures and let Λ be a probability measure on ζ defined as follows. Let κ be the inverse temperature, sometimes called *Parisi 1-RSB parameter*, of the auxiliary model. Then we define

$$\Lambda(\tilde\mu^i) = \omega_i(\kappa) = \frac{\exp(\kappa\mathbf{\Phi}_i)}{\mathscr{Z}(\kappa)}. \tag{1.1.25}$$

At a first glance, (1.1.25) strongly resembles the definition of the Boltzmann distribution (1.1.6) and indeed, as it turns out, this resemblance has a deeper meaning. But first observe that for $\kappa = 1$, the auxiliary model equals the original system. Nevertheless, studying $\mathscr{Z}(\kappa)$ for general κ turns out to be the key for calculating $\boldsymbol{\Sigma}$, which ultimately allows understanding the original model. Without going too much into detail, we will shortly describe how to translate a physical model with an underlying sparse random factor graph which undergoes 1-RSB, into an auxiliary model which can then, itself, be solved using Belief Propagation. We stress at this point that the Boltzmann distribution (1.1.25) of the auxiliary model is a probability distribution on probability distributions, thus a difficult object to analyse.

Let us assume that we start with a problem described by the factor graph $\mathscr{G} = (V \cup F, E, \Psi)$. Then, following Mézard and Montanari [135, Section 19.2.1], we create an auxiliary model in three straightforward steps. Suppose that $\{\nu_{x\to a}\}$ and $\{\nu_{a\to x}\}$ are the BP messages on $\mathscr{G}$. We highlight that we leave out technical details (for instance, we suppose implicitly that those messages were discrete measures) as we are only interested in sketching the main idea of the 1-RSB cavity method. Recall Ξ_a, Ξ_x, Ξ_{ax} from (1.1.21). Then, we proceed as follows in the construction of the factor graph $\boldsymbol{G} = (\boldsymbol{V} \cup \boldsymbol{F}, \boldsymbol{E}, \boldsymbol{\Psi})$ representing the auxiliary model.

1. For $ax \in E$, we create a variable node $\boldsymbol{xa}$ representing $(\nu_{x\to a}, \nu_{a\to x})$ and a factor node connected to this variable with weight function $\boldsymbol{\Psi}_{xa} = \exp(-\kappa \Xi_{xa}(\nu))$.

2. For any factor $a \in F$, we introduce a factor node $\boldsymbol{a}$ in the auxiliary model and connect it to all variable nodes $\boldsymbol{xa}$ with $x \in \partial a$ in the original model. The corresponding weight function reads

$$\boldsymbol{\Psi}_a = \prod_{x\in\partial a} \mathbf{1}\left\{\nu_{a\to x} \propto \sum_{\sigma_{\partial a}\in\Omega^{\partial a}} \Psi_a(\sigma_{\partial a}) \prod_{x'\in\partial a\setminus x} \nu_{x'\to a}\right\} \exp\left(-\kappa \Xi_a\left(\left\{\nu_{y\to a}\right\}_{y\in\partial a}\right)\right).$$

   (Here, we write $\Xi_a\left(\left\{\nu_{y\to a}\right\}_{y\in\partial a}\right)$ instead of $\Xi_a(\nu)$ in order to highlight the local dependencies.) The purpose of the weight function is two-fold. First, it makes sure that the correct BP equations are observed and second, it weighs the contribution by a factor of $\exp(\kappa \Xi_a)$.

3. Each variable $x \in V$ produces a factor node $\boldsymbol{x} \in \boldsymbol{F}$ in the auxiliary model. This factor connects to all variables $\boldsymbol{xa}$ if $a \in \partial x$ in the original model. The corresponding weight function is defined as

$$\boldsymbol{\Psi}_x = \prod_{a\in\partial x} \mathbf{1}\left\{\nu_{x\to a} \propto \prod_{b\in\partial x\setminus a} \nu_{b\to x}\right\} \exp\left(\kappa \Xi_x\left(\{\nu_{b\to x}\}_{b\in\partial x}\right)\right).$$

   Again, this weight function guarantees observing the valid BP messages as well as it weighs the factor node's contribution.

Now it is possible to run BP on this auxiliary model. If the physics' intuition is correct, this auxiliary model is replica symmetric and BP yields a valid estimate of $\ln \mathscr{Z}(\kappa)$ via the Bethe free entropy. Recall that we have by definition

$$\mathscr{Z}(\kappa) = \sum_{i=1}^{\tilde{\ell}} \exp(\kappa \boldsymbol{\Phi}_i).$$

Leaving out technical details and justifications, we obtain

$$\mathscr{Z}(\kappa) \sim \int \exp\left(n\left(x\phi + \boldsymbol{\Sigma}(\phi)\right)\right) \mathrm{d}\phi.$$

Then, if $\mathscr{Z}(\kappa) \sim \exp\left(n f(\kappa)\right)$ for some function f, we find $\boldsymbol{\Sigma}$ through a Legendre transformation [135, Section 19.2] as

$$f(\kappa) = x\kappa + \boldsymbol{\Sigma}(\phi) \qquad \text{such that} \qquad \frac{\partial \boldsymbol{\Sigma}}{\boldsymbol{\phi}} = -\kappa.$$

In summary, if a model shows 1-RSB, Belief Propagation will not render the correct estimate for the partition function, i.e. it will have various (almost) fixed-points. In this case, the physics' intuition

proposes to create a statistical auxiliary model with the almost-solutions being variables and analysing the partition function of the auxiliary model by BP yielding to a (hopefully) unique fixed-point.

To put this section into context, we recall the already discussed phase diagram of random CSPs. Suppose that the Boltzmann distribution was concentrated on finitely many pure states (which is called *static 1-RSB*). Then the solution space has undergone the condensation phase transition. In the case that the Boltzmann distribution is not concentrated on a small amount of pure states but there are exponentially many pure states each with exponentially small probability mass (*dynamical 1-RSB*), the solution space finds itself in the clustering phase. We furthermore stress that one basic assumption of the 1-RSB Ansatz is that the Boltzmann distribution can be written as a convex combination of *extremal Bethe measures*, thus probability distributions which lack long-range correlations and which describe the local behaviour of the Boltzmann distribution based only on a finite neighbourhood with a boundary condition given by (almost)-solutions to the BP messages. This strongly resembles the finding of a partition of the configuration space Ω^n into clusters $\mathscr{E}_1,\dots,\mathscr{E}_\ell$ such that the Boltzmann distribution conditioned on those clusters is $\varepsilon-$extremal (c.f. Section 1.1.2.4). And indeed, the cut-distance formalism turned out to be a key tool in verifying the results predicted by the cavity method in some special problems [21, 48, 49].

Until now, we studied sparse random CSPs from a specific point of view which ultimately results in calculating the number of satisfying assignments. Thus, given a degree distribution of the factor nodes $(\deg(a_1),\dots,\deg(a_m))$ and a degree distribution of the variable nodes $(\deg(x_1),\dots,\deg(x_n))$ (which might be obtained by a distribution of choice), we first draw the set of weight functions $\Psi_1,\dots,\Psi_m$ from a distribution (which might, in principle, turn out to be deterministic choices like in the random $k-$SAT problem). Then, given the degree sequences as well as the weight functions, a factor graph is drawn uniformly at random from all factor graphs on those degree distributions. We can now ask the question, whether this random CSP is satisfiable w.h.p., or determine the number of solutions or study the geometry of the solution space. There is yet another interesting model closely related, which will be discussed in the next section.

1.2. Statistical Inference

The task of statistical inference can be modelled by the so-called *teacher-student scheme* quite intuitively. The scheme itself was introduced by Gardner and Derrida [87] in the context of studying the perceptron, a fairly easy binary classifier. In this section, we follow an introduction into statistical inference based on the study of physical systems by Zdeborová and Krzakala [168].

1.2.1. A statistical physics' approach

1.2.1.1. The teacher-student scheme

As a first step, the teacher generates some *ground-truth* $\boldsymbol{\sigma}$ from an arbitrary probability distribution μ^{TP} - this is the *teacher's* prior. Now he generates some observable data $\hat{\boldsymbol{\sigma}}$ from $\boldsymbol{\sigma}$ by a statistical model. This model is characterised by a distribution $\mu^{TM}(\hat{\boldsymbol{\sigma}} \mid \boldsymbol{\sigma})$, which expresses the likelihood of observing $\hat{\boldsymbol{\sigma}}$ given that the ground-truth was $\boldsymbol{\sigma}$ - this likelihood distribution is the *teacher's model*. And finally, the teacher conveys the data $\hat{\boldsymbol{\sigma}}$ and some information about μ^{TP} as well as μ^{TM} to the student.

The student's ultimate goal is to infer as much information as possible about $\boldsymbol{\sigma}$ from the given data and the (probably limited) information about the teacher's prior as well as the teacher's model. If the teacher gives the full information about the prior as well as the model to the student, we call this setting *Bayes optimal*. Let us focus on statistical inference problems which exhibit quite convenient properties. More precisely, we suppose the following. Let Ω be a finite set and sample the ground-truth $\boldsymbol{\sigma} \in \Omega^n$ from μ^{TP}. Furthermore, suppose that $\hat{\boldsymbol{\sigma}}$ is an m-dimensional vector with entries in a finite set χ and that the following assumptions hold.

- The prior distribution factorises, thus

$$\mu^{TP}(\sigma) = \prod_{i=1}^{n} \mu_i^{TP}(\sigma_i).$$

- The single observable data points $\hat{\boldsymbol{\sigma}}_1, \dots \hat{\boldsymbol{\sigma}}_m$ are independently generated based on a subset of variables $\partial\hat{\boldsymbol{\sigma}}_1, \dots, \partial\hat{\boldsymbol{\sigma}}_m$, hence

$$\mu^{TM}(\hat{\boldsymbol{\sigma}} \mid \boldsymbol{\sigma}) = \prod_{j=1}^{m} \mu_{\partial\hat{\boldsymbol{\sigma}}_j}(\hat{\boldsymbol{\sigma}}_j \mid \partial\hat{\boldsymbol{\sigma}}_j).$$

Let us now view the whole scenario through the student's eyes. As the student might not have full access to the teacher's prior as well as the model's statistics, we suppose that the known prior is described by a distribution ν on Ω^n and the information about data generation is modelled by a distribution $\tilde{\nu}$ on χ^m. If ν and $\tilde{\nu}$ satisfy above's assumptions, the student can employ Bayes' theorem and write the posterior distribution as

$$\nu(\boldsymbol{\sigma} \mid \hat{\boldsymbol{\sigma}}) \propto \prod_{i=1}^{n} \nu_i(\boldsymbol{\sigma}_i) \prod_{j=1}^{m} \tilde{\nu}_{\partial\hat{\boldsymbol{\sigma}}_j}(\hat{\boldsymbol{\sigma}}_j \mid \partial\hat{\boldsymbol{\sigma}}_j). \tag{1.2.1}$$

Now, the best the student can do in order to infer $\boldsymbol{\sigma}$ from $\hat{\boldsymbol{\sigma}}$, ν and $\tilde{\nu}$, is to sample a configuration $\tilde{\sigma}$ from $\nu(\boldsymbol{\sigma} \mid \hat{\boldsymbol{\sigma}})$. If $\tilde{\sigma} = \boldsymbol{\sigma}$, we say that the student succeeded in inferring the ground-truth completely. If we are in the convenient situation of Bayes optimality, thus the student gets the full information about the teacher's prior and the data generation, we can observe the following. Denote by $\boldsymbol{\tau}, \boldsymbol{\tau}', \boldsymbol{\tau}''$ three independent uniform samples from $\nu(\cdot \mid \hat{\boldsymbol{\sigma}})$. Furthermore, let $f : \Omega^n \times \Omega^n \to \mathbb{R}$ be some arbitrary function, then we have [168, Eq. (15)]

$$\mathbb{E}\left[f(\boldsymbol{\tau}', \boldsymbol{\tau}'')\right] = \mathbb{E}\left[f(\boldsymbol{\tau}, \boldsymbol{\sigma})\right].$$

Thus, at least with respect to the expectation, there is no difference between the ground-truth $\boldsymbol{\sigma}$ and uniform samples from the posterior distribution. This result in the Bayes optimal setting is called the *Nishimori property* and from now on, we will tacitly assume to satisfy this Nishimori property, thus being in the Bayes optimal setting. Next, we express the problem of inference in the spin glass language from the previous sections.

1.2.1.2. Planted models

Probably the first connection between statistical inference problems and statistical physics was observed by Jaynes [104] while the expression of statistical inference problems in terms of spin glass language was strongly influenced by pioneering work of Kirkpatrick, Gelatt and Vecchi [110] on simulated annealing. We first observe that (1.2.1) can be written as [168, Eq. (25)]

$$\nu(\boldsymbol{\sigma} \mid \hat{\boldsymbol{\sigma}}) \propto \exp\left(\sum_{i=1}^{n} \ln\left(\nu_i(\boldsymbol{\sigma}_i)\right) + \sum_{j=1}^{m} \ln\left(\tilde{\nu}_{\partial\hat{\boldsymbol{\sigma}}_j}\left(\hat{\boldsymbol{\sigma}}_j \mid \partial\hat{\boldsymbol{\sigma}}_j\right)\right)\right). \tag{1.2.2}$$

Now, we introduce an inverse temperature β, such that $\beta = 1$ recovers the original posterior distribution and introduce $\ln\left(\nu_i(\boldsymbol{\sigma}_i)\right)$ as an external magnetic field at particle i while $\ln\left(\tilde{\nu}_{\partial\hat{\boldsymbol{\sigma}}_j}\left(\hat{\boldsymbol{\sigma}}_j \mid \partial\hat{\boldsymbol{\sigma}}_j\right)\right)$ expresses the interaction between particles. Thus, as given through (1.1.1), we find a Hamiltonian $H(\boldsymbol{\sigma}, \hat{\boldsymbol{\sigma}})$ such that (1.2.2) becomes

$$\nu(\boldsymbol{\sigma} \mid \hat{\boldsymbol{\sigma}}) = \frac{\exp\left(-\beta H(\boldsymbol{\sigma}, \hat{\boldsymbol{\sigma}})\right)}{Z(\hat{\boldsymbol{\sigma}})} \tag{1.2.3}$$

with $Z(\hat{\boldsymbol{\sigma}})$ being the partition function of the described physical system. Such a physical system is called a *planted model*. Let us elaborate on this shortly. While the particle interactions $J_{i_1,\dots,i_k}(\sigma_{i_1}, \dots, \sigma_{i_k})$

in (1.1.1) can be any function on k spins, spin glasses exhibit positive as well as negative interactions between particles. One possibility to generate such glasses is to choose the interactions randomly from a (often symmetric) probability distribution, like in the case of Gaussian spin glass models [23]. In contrast, the planted model is a very special system, as the interactions between particles are random but mutually correlated as they are all generated given the ground-truth $\boldsymbol{\sigma}$. This is why we call this ground-truth $\boldsymbol{\sigma}$ the *planted configuration*. Following [168], we will introduce a short example on random linear equations that shows why this correlations of the particle interactions can influence the system's behaviour drastically. Suppose that $\boldsymbol{y} \in \{0,1\}^m$ is a random vector and $\boldsymbol{A}$ is a random $m \times n$ matrix over $\mathbb{F}_2$. If $m > n$, the random system of linear equations $\boldsymbol{A}x = \boldsymbol{y}$ has no solution x with high probability [18]. But if we sample a uniform binary vector $\boldsymbol{x}$, calculated $\boldsymbol{y}$ as $\boldsymbol{Ax}$, then the system of linear equations will always feature at least one solution, namely $\boldsymbol{x}$. In our setting, $\boldsymbol{x}$ corresponds to the planted ground-truth while the teacher shows the student $\boldsymbol{A}$ and $\boldsymbol{y}$ (and the information that $\boldsymbol{x}$ and $\boldsymbol{A}$ are independently and uniformly chosen). Now the student's inference task is to infer $\boldsymbol{x}$ from $(\boldsymbol{A}, \boldsymbol{y})$.

The principle of *planting* did not only appear in statistical physics. For instance, it has been used to prove a hardness result on the planted clique problem [74, 105] or within proofs on the geometry of the solution space of random CSPs [1]. Probably one of the most studied planted models is the *stochastic block model* which is a widely used model for community detection on networks [62, 85, 95]. To describe it in the framework at hand, given n individuals $V = \{x_1 \dots x_n\}$, a teacher generates q communities, thus a coloring $\boldsymbol{\sigma} \in [q]^n$ assigning each individual a community. Now, a random graph $\boldsymbol{G} = (V, E)$ is generated as follows. Each edge $x_i x_j$ is present with probability p_1 if $\boldsymbol{\sigma}_i = \boldsymbol{\sigma}_j$ and with probability p_2 if $\boldsymbol{\sigma}_i \neq \boldsymbol{\sigma}_j$. Clearly, if $p_1 = p_2$, this is just a binomial random graph containing no information about the communities. If $p_1 \gg p_2$, it is more likely to observe edges within one community and if $p_2 \gg p_1$, most edges are expected between different communities. Of course, there are various generalisations to this problem. Nevertheless, the student's task is, given $(\boldsymbol{G}, p_1, p_2, q)$ to infer a $q-$coloring $\tilde{\sigma}$ that has the highest possible overlap with $\boldsymbol{\sigma}$.

As one can conclude, planted models are a fairly common used technique in statistical inference. Studying the corresponding physical system whose Boltzmann distribution equals the posterior distribution obtained by the student might bring together powerful tools from different fields of research. For the sake of convenience and as this thesis's contributions on statistical inference require this setup, we tacitly assume that the Boltzmann distribution of the planted model can be expressed by a sparse random factor graph, such that it has the form (1.1.16). As we previously studied different phase transitions, mostly corresponding to the geometry of the solution space of such random CSPs, it is not very surprising that there are important phase transitions in statistical inference problems as well.

1.2.1.3. Phase transitions in statistical inference

Suppose being in the Bayes optimal setting, thus, the student's guess ν is exactly the teacher's prior and the student's model knowledge $\tilde{\nu}$ equals the teacher's model generating distribution. Furthermore, suppose that we observe the underlying physical system in the thermodynamic limit ($n \to \infty$). Moreover, let $\boldsymbol{\sigma} \in \Omega^n$ for some finite set Ω. In principle, the following discussion can be extended to more general ground-truth domains. For two configurations $\sigma, \tau \in \Omega^n$, we denote the overlap $\langle \sigma, \tau \rangle$ as the number of coordinates in which σ and τ coincide. Let q_0 denote the expected overlap between a uniformly at random chosen $\boldsymbol{\tau}$ from ν and $\boldsymbol{\sigma}$, formally

$$q_0 = \mathbb{E}_{\boldsymbol{\tau} \sim \nu} \langle \boldsymbol{\tau}, \boldsymbol{\sigma} \rangle.$$

Given data $\hat{\boldsymbol{\sigma}}$ as well as ν and $\tilde{\nu}$, the student's task might be evaluated with respect to two levels of reconstruction.

- Is the student able to guess $\tilde{\sigma}$ such that $\langle \tilde{\sigma}, \boldsymbol{\sigma} \rangle > q_0$? *(weak reconstruction)*
- Is the student able to infer $\boldsymbol{\sigma}$? *(strong reconstruction)*

With the statistical inference tasks being part of this thesis, we are only interested in the strong reconstruction scenario. Now, this question can be answered under two different sets of restrictions.

- Reconstruction is *information-theoretically* possible, if the student can infer $\boldsymbol{\sigma}$ from $(\hat{\boldsymbol{\sigma}}, \nu, \tilde{\nu})$ given unlimited computational power.
- Reconstruction is *algorithmically* possible, if there is a polynomial-time algorithm $\mathscr{A}$ that outputs $\boldsymbol{\sigma}$ on input $(\hat{\boldsymbol{\sigma}}, \nu, \tilde{\nu})$.

Clearly, the planted model undergoes phase transitions with respect to this questions. We suppose that the planted model comes as a factor graph $\mathscr{G}$ with n variables and m factor nodes and that the student has access to this graph. With a slight misuse of notation, we suppose that the student gains knowledge about $\tilde{\nu}$ as well, if she has access to the factor graph. Furthermore, on each factor node a we find a weight function ψ_a such that $\psi_a(\partial a) = \hat{\boldsymbol{\sigma}}_a$. As before, let $\alpha = m/n$ be the factor-to-variable ratio. In this case, the more factors there are (as α getting large), the more measurements of the ground-truth are available for gathering information and thus, the easier the task seems to become. We denote by α_{inf} the information-theoretic threshold and by α_{alg} the algorithmic threshold. Then we have the following, assuming tacitly that all thresholds might be either strict or coarse phase transitions as already discussed.

- For $\alpha < \alpha_{\mathrm{inf}}$, there is no algorithm (efficient or not) that is able to infer $\boldsymbol{\sigma}$ from $(\mathscr{G}, \hat{\boldsymbol{\sigma}}, \nu)$.
- For $\alpha_{\mathrm{inf}} < \alpha < \alpha_{\mathrm{alg}}$, there is no efficient algorithm that is able to infer $\boldsymbol{\sigma}$ from $(\mathscr{G}, \hat{\boldsymbol{\sigma}}, \nu)$.
- If $\alpha > \alpha_{\mathrm{alg}}$, there is a polynomial-time algorithm $\mathscr{A}$ that outputs $\boldsymbol{\sigma}$ on input $(\mathscr{G}, \hat{\boldsymbol{\sigma}}, \nu)$.

If α^* is some threshold, we will subsequently refer to negative results (*for $\alpha < \alpha^*$ inference is not possible*) as *converse statements* and to positive results (*for $\alpha > \alpha^*$ inference is possible*) as *achievability statements* respectively. When mentioning algorithmic achievability, a natural question arises: which class of algorithms is supposed to perform well on inference problems? We make use of the observation that the Boltzmann distribution of the planted model can be tackled by the message passing algorithms of Section 1.1.4. While Belief Propagation is indeed an efficient algorithm in a complexity theoretical way, in each iteration at each factor a, we need to compute roughly $\deg(a)$ messages. If the underlying graph is not too sparse, this might not be feasible computationally on large instances. Suppose we have weight functions that are not too sensitive to single messages' contributions. Then a family of algorithms called *approximate message passing algorithms* is supposed to perform well.

1.2.1.4. Approximate Message Passing AMP

Let us first discuss what a sensitive weight function is. A prime example would be the weight functions occurring in the random $k-$SAT problem where a single message can turn the evaluation from nearly zero to almost 1. On the other hand, a weight function that, for instance, counts the adjacent variables with spin 1, would be very insensitive – at least if the factor node degree is large. Thus, suppose the latter is the case and suppose further that the average degree of a factor $k = \omega(1)$ is large. Instead of calculating k different messages (depending on which variable is removed) at each factor as in Belief Propagation, one could compute one message where no variable is removed and send it to all neighbours. Donoho, Maleki and Montanari [66] introduced a family of message passing algorithms which is, intuitively speaking, an approximate version of Belief Propagation.

More precisely, let $\mathscr{G} = (V \cup F, E)$ be a factor graph representing a statistical inference problem on n variables and m factors with $\alpha = m/n$. For a vector $\tau \in \mathbb{R}^k$ we write $\langle \tau \rangle = k^{-1} \sum_{i=1}^{k} \tau_i$ for the average over all entries of τ. Denote by $\{\eta_t : \mathbb{R}^n \to \mathbb{R}^n\}_{t \geq 1}$ a family of coordinate-wise applied (non-linear) threshold functions and let $A \in \mathbb{R}^{m \times n}$ be the (normalised) adjacency matrix of the factor graph, thus normalise the columns to ℓ_2 norm 1. Then, approximate message passing starts with an initial guess $\sigma^{(0)}$ of the ground-truth $\boldsymbol{\sigma}$ and computes iteratively

$$\sigma^{(t+1)} = \eta_t\left(A^T z^{(t)} + \sigma^{(t)}\right) \qquad \text{and} \qquad z^{(t)} = \hat{\boldsymbol{\sigma}} - A\sigma^{(t)} + \alpha^{-1} z^{(t-1)} \left\langle \eta_t'\left(A^T z^{(t-1)} + \sigma^{(t-1)}\right)\right\rangle. \tag{1.2.4}$$

Let us only briefly sketch the meaning of the single parts of (1.2.4), a complete introduction and a formal justification of those equations is provided by [66, 129].

- $\sigma^{(t)}$ is the current estimate for the ground-truth $\boldsymbol{\sigma}$.
- $z^{(t)}$ can be interpreted as a current residual.
- The threshold function η pushes $\sigma^{(t)}$ towards the sparsest solution. In its absence, the algorithm would converge to a solution of $\hat{\boldsymbol{\sigma}} = A\boldsymbol{\sigma}$ of least ℓ_2 norm.
- $\alpha^{-1} z^{(t-1)} \left\langle \eta'\left(A^T z^{(t-1)} + \sigma^{(t-1)}\right)\right\rangle$ is derived from the Belief Propagation update rules on the corresponding factor graph. It improves the convergence towards sparse solutions even further.

While AMP is clearly fast to run and easy to implement, it also achieves the best algorithmic performance presently known in some of the most prominent inference problems like compressed sensing [67] or the pooled data problem [70].

After this excursion into the statistical physics' foundations of statistical inference, we will introduce the *group testing* problem in the next section. Group testing is a prime example of a statistical inference problem and is the protagonist of multiple contributions of this thesis.

1.2.2. Group Testing

In the group testing problem, one is given n individuals $x_1, \dots, x_n$ out of which a small number k is infected. We may employ a testing procedure that allows to pool various individuals into one *group test* that renders a positive result if and only if at least one infected individual is contained in the test. Given probes of those n individuals and the prevalence k/n, the ultimate goal is to find a testing strategy (we will refer to this as a *pooling scheme*) that is able to infer the infection status of each individual with the minimum number of tests possible. Group testing itself found its first appearance in literature in the early 1940's when Dorfman [68] proposed the following, fairly simple, inference algorithm.

(D1) Assign a group of Γ individuals to a test, such that each individual gets tested once.

(D2) If a test renders a negative result, all of the contained individuals are uninfected. If a test renders a positive result, test all individuals individually.

Supposing that the prevalence is k/n and each individual is infected independently of all other individuals, it is straightforward to calculate the expected number of tests $\boldsymbol{m}$ required in this testing scheme as

$$\mathbb{E}[\boldsymbol{m}] = \frac{n}{\Gamma} + n\left(1 - \left(1 - \frac{k}{n}\right)^{\Gamma}\right).$$

Given an estimation of the prevalence (this is the teacher's prior), it is possible to optimise the test size Γ in order to minimise the expected number of tests.

Even if the Dorfman scheme is, as we will see, a suboptimal design, it finds its applications in various medical applications which might be due to the very simple inference algorithm and the fact that it recovers the infection status of each individual correctly (supposing that each test outputs the correct result) [130].

Since its first appearance, group testing gained a lot of attention in various installments. In the first decades, the focus was lying on *combinatorial* group testing, where one aims to construct a pooling scheme that successfully recovers every possible ground-truth from the teacher's prior distribution. This problem was studied intensively, amongst others, by D'yachkov et al. [58], Erdős and Rényi [73], Fischer, Klasner and Wegener [80], Hwang [98] and Ungar [164]. But in the early 2000's, the focus changed to so-called *probabilistic* test-designs in which inference is only required with high probability with respect to the random choice of the ground-truth. Some of the most influential contributions are due to Aldridge, Baldassini and Johnson [8], Aldridge, Johnson and Scarlett [9], Damaschke [59], Gandikota et al. [86], Johnson, Aldridge and Scarlett [108], Mezard and Toninelli [132], Mézard, Tarzia and Toninelli [133] and Scarlett and Cevher [154]. Let us describe more systematically which kind of group testing problems exist in literature. The exact problem description may vary with respect to ...

- ... the teacher's prior $\boldsymbol{\sigma} \in \{0,1\}^n$:

  - $\boldsymbol{\sigma}$ can be a uniformly sampled configuration out of all configurations with exactly k non-zero entries *(hypergeometric group testing model)*,
  - Alternatively, $\boldsymbol{\sigma}$ might be a binomial random vector, such that each entry equals 1 independently of all other entries with probability k/n *(i.i.d. group testing model)*.
  - Finally, each individual x_i might be (independent of the other individuals) infected with probability p_i such that p_i and p_j do not need to be equal necessarily. This choice of the teacher's prior is called *group testing with priors*.

- ... the number of subsequent rounds a test-design may contain:
  - In the *non-adaptive* group testing problem all tests need to be conducted in parallel, thus one cannot use information gained in previous stages.
  - The *adaptive* group testing problem allows to design subsequent stages of tests based on the outcome of previous stages.

- ... the level of required certainty:
  - In *combinatorial group testing*, a testing scheme needs to output the correct infection status of all individuals on *any* ground-truth $\boldsymbol{\sigma}$.
  - On the other hand, in *probabilistic group testing*, it suffices to recover $\boldsymbol{\sigma}$ with high probability (with respect to the randomly generated ground-truth).

- ... the type of recovery:
  - If we demand *exact recovery*, each individual has to be assigned the correct infection status.
  - Otherwise, if we can tolerate a small number of falsely classified individuals, we call this criterion *partial recovery*.

- ... the correctness of each test:
  - In *noiseless group testing*, each test outputs the correct result.
  - *Noisy group testing* instances are characterised by a random flip of each test-result. This noise might be uniformly at random (binary symmetric channel), dependent of the tests' correct result (e.g. $Z-$channel and reverse $Z-$channel) or correlated with the number of infected and uninfected individuals in the test (diluted noise models).
  - In *threshold group testing*, a test-result is negative if the number of contained infected individuals is below a given threshold t_1 and positive if it exceeds a second threshold $t_2 \geq t_1$. In the range (t_1, t_2), the test-result might be randomly chosen.

- ... the constraints on individuals-per-test and tests-per-individual:
  - In the *unrestricted group testing problem* each individual might take place in an arbitrary number of tests. Furthermore, each test might contain between one and all individuals. For the sake of brevity, we will refer to the unrestricted group testing problem as the *group testing problem*.
  - In $\Delta-$*divisible sparsity constrained group testing*, each individual might take place in a maximum of Δ tests.
  - Analogously, in the $\Gamma-$*sparse group testing problem*, each tests may contain at most Γ individuals.

- ... the prevalence of infected individuals:
  - If the number of infected individuals k satisfies $k = \Theta(1)$, thus is independent of the number of individuals n, we call the setting the *ultra-sparse* regime.
  - If, on the other hand, $k \sim n^{\theta} (\theta \in (0,1))$, the regime is called *sublinear*.

– Finally, if we suppose $k = \alpha n$ for some constant $\alpha \in (0,1)$, the setting is denoted as the *linear* regime.

In this thesis's contributions, we will analyse phase transitions in **noiseless probabilistic hypergeometric group testing** instances with a focus on non-adaptive group testing. Nevertheless, some contributions contain results on adaptive group testing as well. We will discuss this thesis's results in detail in Section 2.1. Let us briefly express a non-adaptive group testing instance in terms of the statistical physics' framework. A pooling scheme for such an instance can be represented as a factor graph $\mathcal{G} = (V \cup F, E)$. We denote with $V = \{x_1, \dots, x_n\}$ the n individuals and the m tests are given by $F = \{a_1, \dots, a_m\}$. Furthermore, an edge $x_i a_j$ exists if and only if individual x_i takes part in test a_j. As we are in the setting of inference, suppose that the ground-truth $\boldsymbol{\sigma} \in \{0,1\}^n$ assigns each individual its infection status. Furthermore, let $\hat{\boldsymbol{\sigma}} \in \{0,1\}^m$ denote the sequence of test-results, hence

$$\hat{\boldsymbol{\sigma}}_a = \max_{x \in \partial a} \boldsymbol{\sigma}_x.$$

This completely describes the planted model introduced via (1.2.3), as we, as usual in group testing, suppose that we have complete knowledge about the model's generation and the teacher's prior (Bayes optimal setting). A visualisation can be found in Figure 1.3. In the following, we will use a slightly different notation as in the previous sections on phase transitions. While we previously denoted by α the factor-to-variable ratio and analysed the system's behaviour with respect to the size of α, it is usual in the group testing community to express phase transitions in terms of the required number of tests $m = m(n,k)$ in the large-system limit $n \to \infty$. Of course, those statements can directly be translated into a statement about the factor-to-variable ratio.

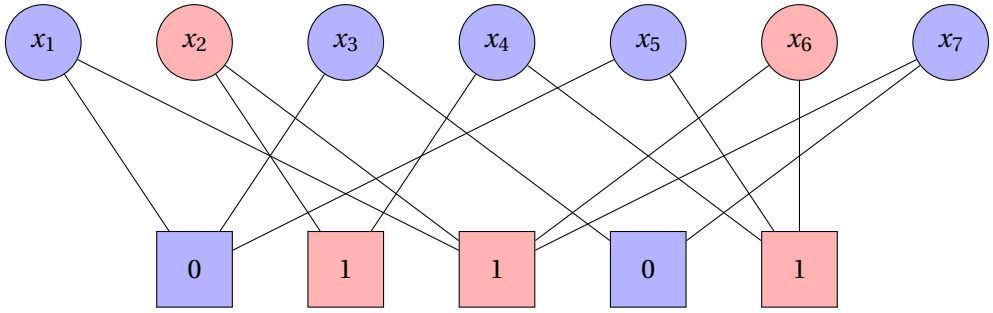

Figure 1.3.: The factor graph representation $G = (V \cup F, E)$ of a group testing instance with $n = 7$ individuals out of which $k = 2$ are infected on $m = 5$ tests. Blue individuals are uninfected while red individuals are infected. Furthermore, a test renders a positive result if and only if at least one infected individual is contained.

Clearly, if the problem was to be studied for finite n, the different levels of prevalence would not be well defined. It is, for instance, not possible to distinguish in a population of $n = 10^3$ individuals whether the occurrence of $k = 20$ infected individuals should be described as $k = 20, k \approx n^{0.43}$ or as $k = 0.02n$, thus results obtained in the large-system limit need to be verified empirically for small n if they should be applied in real laboratories.

Next, we will shortly describe known results and open problems ahead of this thesis's contributions. As this overview is with respect to noiseless probabilistic group testing, we refer the interested reader to the detailed overview article of Aldridge, Johnson and Scarlett [10] for an overview over the state of the play in different models.

1.2.2.1. Prior results on noiseless probabilistic hypergeometric group testing

We will tacitly assume throughout this section that all results are meant to be read with respect to exact recovery if not stated differently. Let us start this section by a folklore counting argument which is a good indication for how many tests we need to use to have a chance of exact recovery. Namely, independent of the choice of the pooling scheme (adaptive or non-adaptive), by making m tests, we can

generate 2^m different sequences of test-results. Clearly, this number needs to exceed the number of possible ground-truth values $\binom{n}{k}$, therefore, if m_{inf} is the minimum number of tests necessary to solve the problem information-theoretically, we find

$$2^{m_{\text{inf}}} \geq \binom{n}{k} \quad \Leftrightarrow \quad m_{\text{inf}} \gtrsim \frac{n\ln n - (n-k)\ln(n-k) - k\ln k}{\ln 2}. \tag{1.2.5}$$

Therefore, with $H(\alpha)$ denoting the entropy of $\mathbf{Be}(\alpha)$ and $\alpha, \theta \in (0,1)$, the counting bound (1.2.5) yields

$$m_{\text{inf}} \gtrsim \begin{cases} \frac{H(\alpha)}{\ln 2} n, & k = \alpha n \\ \frac{1-\theta}{\ln 2} k \ln n, & k \sim n^\theta. \end{cases}$$

Thus, if the prevalence is constant, we require linearly many group tests but if the spread of the disease scales sublinearly in the population size, we can do much better. The counting bound (1.2.5) gives a lower bound on the number of tests required in any testing scheme with respect to a hypergeometric problem setup, thus it applies for adaptive pooling schemes as well as for non-adaptive pooling schemes. Clearly, it should be easier to come along with m_{inf} tests adaptively rather than non-adaptively. Indeed, it turns out that this intuition is mostly correct.

Adaptive group testing Pretty soon after group testing found its way into mathematical literature, there were some negative results on group testing. More precisely, Ungar [164] proved that under the i.i.d. prior there is a phase transition at prevalence $p = \frac{3-\sqrt{5}}{2} \approx 0.38$. More precisely, there is no test design succeeding at inference of $\boldsymbol{\sigma}$ on less than n tests if the problem's prevalence is larger and on the other hand, if the prevalence becomes smaller, there is a test design achieving inference on at most $n-1$ tests. In the hypergeometric group testing problem, it is conversely conjectured that this phase transition (of course with n replaced by $n-1$) occurs at a prevalence of $1/3$ [80, 97] but until now it was only proven that for prevalence $p > \log^{-1.5}(3) \approx 0.369$ performing $n-1$ individual tests is optimal [149]. Clearly, this number of required tests is far from the counting bound. Assuming there are αn infected individuals, Hwang [98] provides a generalised binary splitting algorithm which was later improved by Allemann [13] such that inference of $\boldsymbol{\sigma}$ is possible with

$$m_{\text{Hwang}} \sim m_{\text{inf}} + \alpha n \qquad \text{and, respectively} \qquad m_{\text{Allemann}} \sim m_{\text{inf}} + 0.255\alpha n \tag{1.2.6}$$

tests.

We stress that in the linear regime it turns out that m_{Allemann} exceeds n at roughly $\alpha = 1/3$, underlying the conjecture of Hu, Hwang and Wang [97] and that in the *sublinear regime*, we find

$$m_{\text{Hwang}} \sim m_{\text{Allemann}} \sim m_{\text{inf}},$$

as the deviations are of lower order.

Unfortunately, the binary splitting approach comes with a practical flaw - in order to guarantee successful inference, one requires $\Omega(\ln n)$ rounds of adaptive tests. A natural question is if there might be algorithms succeeding at inference of $\boldsymbol{\sigma}$ with a bounded number of testing rounds. We underline that, with a slight misuse of wording, we tacitly assume that an algorithm is always efficient, thus runs in polynomial time, whenever we do not explicitly state differently. Up to the best of our knowledge there is no algorithm known which succeeds on an arbitrary group testing instance in less than $\Omega(\ln n)$ rounds on m_{Allemann} tests. The situation becomes much more convenient if the problem gets sparser.

More precisely, suppose we have a vanishing prevalence, thus $k \sim n^\theta$ for some $\theta \in (0,1)$. The binary splitting algorithm by Allemann clearly performs asymptotically optimal in this sublinear regime, as already discussed. But now, there are well known efficient algorithms achieving the same asymptotic performance in much less rounds, the probably best known algorithms are due to Damaschke and Muhammad [60] whose algorithm achieves inference in not more than 4 rounds and Scarlett [153] whose algorithm needs only two rounds in the described setting with no more than $(1+o(1))m_{\text{inf}}$ tests,

improving on the 2-stage algorithm of Mezard and Toninelli [132] requiring $\frac{(1+o(1))m_{\inf}}{\ln 2}$ tests.

Therefore, the adaptive group testing problem is, up to the exact phase transition point in hypergeometric linear group testing, well understood. Things turn out to be completely different for non-adaptive group testing.

Non-adaptive group testing Let us again start with discussing known results for a constant prevalence $k = \alpha n$. It turns out that this case is, from a mathematical viewpoint, completely uninteresting. Due to Bay, Price and Scarlett [24] it is known that recovery of the ground-truth is impossible with fewer than $(1-\varepsilon)n$ tests for any $\varepsilon, \alpha \in (0,1)$. The authors built up on work by Aldridge [7] who established a coarse phase transition, thus proved that recovery under the given circumstances fails with positive probability.

Therefore, we let our focus be the sublinear regime in which we suppose throughout that the prevalence is given by $k/n = n^{-(1-\theta)}$ for some density parameter θ. Thus, if θ becomes larger, the prevalence gets higher and the group testing problem is said to become denser.

We will first introduce two specific non-adaptive pooling schemes. Given the teacher's prior in the hypergeometric model, we might construct a random graph, thus the teacher's model, as follows.

- *Bernoulli testing:* Each individual takes place in any test with probability p independently.
- *Random (almost) regular testing:* Each individual chooses Δ tests uniformly at random without (or with, respectively) replacement.

It turns out that both model choices have some similarities but that the first is inferior to the second. We will verify this fact in Section 2.1, which is basically due to high fluctuations of the individual degree. Indeed, the information-theoretically optimal designs require each test to be positive with probability $1/2$ implying that $\Delta = \Theta(\ln n)$, or $p = \Theta\left(k^{-1}\right)$ respectively. The choice that any test needs to be positive with probability $1/2$ is, intuitively speaking, due to the fact that this choice maximises the system's entropy, thus the gain on information per test is maximised.

While there was no universal converse statement sharpening $m_{\inf}$ known prior to this thesis's contributions, there are several algorithmic and information-theoretical achievability as well converse statements on those two random models. We will discuss the most influential ones. Regarding Bernoulli testing, it were Scarlett and Cevher [152] who proved that it is information-theoretically possible to infer the ground-truth with $(1+o(1))m_{\inf}$ tests if the group testing instance is fairly sparse, thus $\theta \leq 1/3$. This result was strengthened by Aldridge [11] who pinned down the information-theoretic strict phase transition of the Bernoulli group testing design at

$$m_{\text{Bernoulli}} = \frac{1}{c_{\text{Ber}} \ln 2} k \ln \frac{n}{k} \qquad \text{where} \qquad c_{\text{Ber}} = \max_{\nu>0} \min \left\{ \frac{(1-\theta)\nu \exp(-\nu)}{\theta \ln 2}, \frac{H\left(\exp(-\nu)\right)}{\ln 2} \right\} k \ln \frac{n}{k} \tag{1.2.7}$$

which is strictly worse than (1.2.5) for all $\theta > 1/3$. Subsequently, Aldridge, Johnson and Scarlett [9] provided an information-theoretical converse statement (though, no achievability result) in the random regular model. More precisely, inference of $\boldsymbol{\sigma}$ fails with positive probability, if the test-design contains less than

$$m_{\text{rand-reg}} = \max \left\{ \frac{\theta}{(1-\theta)\ln^2 2}, \frac{1}{\ln 2} \right\} k \ln \frac{n}{k} \tag{1.2.8}$$

tests. A short calculation verifies that $m_{\text{rand-reg}} < m_{\text{Bernoulli}}$ for $\theta \in \left(\frac{1}{3}, 1\right)$ and $m_{\text{rand-reg}} = m_{\inf}$ for $\theta \leq \frac{\ln 2}{1+\ln 2} \approx 0.409$ but nevertheless, $m_{\text{rand-reg}} > m_{\inf}$ for larger θ. Thus, the random regular model might outperform the Bernoulli testing but it is clearly far from the counting bound for a high prevalence. Such an information-theoretic gap might be due to the model itself or it might be the case that non-adaptive pooling schemes cannot perform at $m_{\inf}$. We will discuss in Section 2.1 that the latter is the case and actually $m_{\text{rand-reg}}$ is a universal information-theoretic converse, independent of the pooling scheme, for any non-adaptive design. We will see furthermore that the converse statement (1.2.8) of [9] actually marks an information-theoretic phase transition point.

At this point, we make a very short excursion into the setting of partial recovery. The already mentioned paper of Scarlett and Cevher [152] actually proves that m_{inf} is an important threshold for partial recovery models. More precisely, the simple Bernoulli test design suffices to recover all but γk individuals correctly with m_{inf} tests. On the negative side, no test-design can come along with less than $(1-\gamma)m_{\text{inf}}$ tests when trying to recover all but γk individuals correctly.

Let us come back to the problem of exact recovery. While we already discovered the state of the play prior to this thesis's contributions with respect to information-theoretic aspects, we will now introduce three of the most prominent non-adaptive group testing algorithms and state known results about their performances. In detail, we will introduce the COMP algorithm as well as the DD algorithm and its greedy extension called SCOMP.

The two probably most basic algorithms are COMP and DD and their descriptions can be found in Algorithms 1 – 2.

Input: Pooling scheme $\mathcal{G} = (V \cup F, E)$, test-results $\hat{\boldsymbol{\sigma}} \in \{0,1\}^m$
Output: Estimate $\tilde{\sigma}$ of $\boldsymbol{\sigma}$
1 Mark all individuals occurring in a negative test as uninfected.
2 Declare all other individuals as infected.

Algorithm 1: The COMP algorithm as first introduced by Chan et al. [39].

Input: Pooling scheme $\mathcal{G} = (V \cup F, E)$, test-results $\hat{\boldsymbol{\sigma}} \in \{0,1\}^m$
Output: Estimate $\tilde{\sigma}$ of $\boldsymbol{\sigma}$
1 Mark all individuals occurring in a negative test as uninfected and remove them and the corresponding negative test from the graph.
2 Mark an individual as infected if it appears as the only individual in a positive test in this reduced graph.
3 Declare all other individuals as uninfected.

Algorithm 2: DD algorithm as defined by Aldridge, Baldassini and Johnson [8].

While COMP cannot produce any false negatives, thus $\tilde{\sigma}_i^{\texttt{COMP}} = 0 \Rightarrow \boldsymbol{\sigma}_i = 0$, DD guarantees that all declared infected individuals are indeed infected, hence $\tilde{\sigma}_i^{\texttt{DD}} = 1 \Rightarrow \boldsymbol{\sigma}_i = 1$. Nevertheless, it might happen that the estimate $\tilde{\sigma}$ does not even *explain* the test-results $\hat{\boldsymbol{\sigma}}$. In this context, we say that an individual $x \in \partial a$ explains test a under $\tilde{\sigma}$ if $\hat{\boldsymbol{\sigma}}_a = 1$ and $\tilde{\sigma}_x = 1$. Conversely, a positive test is called explained by $\tilde{\sigma}$ if there is at least one individual $x \in \partial a$ which explains a. With COMP or DD it might furthermore happen that $\tilde{\sigma}$ contains less than k infected individuals (if DD was applied). In terms of inference, we observe that the estimates of COMP and DD do not necessarily belong to the solution space of the underlying random CSP. If we look a bit closer into the DD-algorithm, we find that the first two steps do not misclassify any individual. We furthermore observe that the estimate was correct if after the first step of DD, there is no individual left that does not belong to at least one (positive) test of degree one. We will formalise this observation in Section 2.1. Suppose that this does not hold and thus, after the second step of DD, we are left with some unexplained tests. As the prevalence is small, it might be a natural idea to declare greedily those individuals as infected that explain the most unexplained tests. This is exactly what the SCOMP-algorithm does.

Positive news about SCOMP is clearly that it produces an estimate $\tilde{\sigma}$ which explains the test-results $\hat{\boldsymbol{\sigma}}$. The flaw is, of course, that it might produce false positive as well as false negative predictions. Nevertheless, it was conjectured based on simulations that SCOMP performs better than DD does [8], thus requires less tests to succeed at inference of $\boldsymbol{\sigma}$. We will see in Section 2.1 that this conjecture turned out to be actually false.

Algorithms 1 – 3 can be applied to any arbitrary (non-adaptive) pooling scheme. Nevertheless, they were studied on the Bernoulli model [8] as well as the random regular model [108]. As it turns out that those algorithms require less tests on the random regular model in order to infer $\boldsymbol{\sigma}$ with high probability, we focus on those results from [108]. More precisely, the authors prove a strict phase transition of the

Input: Pooling scheme $\mathcal{G} = (V \cup F, E)$, test-results $\hat{\boldsymbol{\sigma}} \in \{0,1\}^m$
Output: Estimate $\tilde{\sigma}$ of $\boldsymbol{\sigma}$
1 Mark all individuals occurring in a negative test as uninfected and remove them and the corresponding negative test from the graph.
2 Mark all individuals that are the sole individual in a test in this reduced graph as infected.
3 **while** *there is an unexplained test* a **do**
4 Take the individual of highest degree, breaking ties arbitrarily, and declare it as infected.
5 Remove all adjacent tests from the graph.
6 Declare all left individuals (now, isolated in the reduced graph) as uninfected.

Algorithm 3: The SCOMP algorithm by Aldridge, Baldassini and Johnson [8] can be seen as a greedy extension of DD.

COMP-algorithm on the random regular model at $m_{\mathtt{COMP}}$ as well as an achievability result for DD at $m_{\mathtt{DD}}$ where

$$m_{\mathtt{COMP}} = \frac{1}{(1-\theta)\ln^2 2} k \ln \frac{n}{k} \qquad \text{and} \qquad m_{\mathtt{DD}} \leq \max\left\{\frac{\theta}{(1-\theta)\ln^2 2}, \frac{1}{\ln^2}\right\} k \ln \frac{n}{k}. \tag{1.2.9}$$

Clearly, if DD achieves inference at $m_{\mathtt{DD}}$, so does SCOMP as it performs the two first steps of DD. Comparing the performance of the DD algorithm on the random regular model with its information-theoretic converse (1.2.8), we find that DD is an optimal inference algorithm on the random regular model for $\theta \geq 1/2$ while there remains a gap for smaller θ. This might have been due to a weakness in the achievability proof of DD, a weakness in the information-theoretic converse or because DD does not perform best possible in this regime. As we will see in Section 2.1, the latter is the case.

Let us, at this point, briefly discuss the DD algorithm itself. While it was first stated in its comfortable and easy to digest version in [8], it turns out that the estimate $\tilde{\sigma}$ of $\boldsymbol{\sigma}$ coincides with the estimate computed by the well known Warning Propagation algorithm. Such message passing algorithms were already applied by Mézard, Tarzia and Toninelli [133] to the group testing problem. Indeed it turns out that the estimates of DD and WP coincide. If an individual x is part of a negative test, WP sends the warning to x that it may not take value 1, thus we just need to analyse the messages at positives test. If now there is a positive test a of size Γ_a containing $\Gamma_a - 1$ individuals being part of a (different) negative test, the message sent to the last individual warns this individual not to take value 0 as well, as otherwise the test was unexplained. Finally, if there were two individuals in a not being warned about taking the value 0, the test won't send a warning. The possibility to write down (equivalent) forms of the statistical physics' message passing algorithms seems to be a key feature of group testing and similar problems as we will see in due course more often.

Until now, all presented (non-adaptive) designs and algorithms have one thing in common. If we restrict ourselves to the sublinear regime, each individual takes part in $\Omega(\ln n)$ tests and there are tests containing at least $\Omega(n/k)$ tests. Understanding the group testing problem under given restrictions on the maximum degrees of individuals and tests is not only a challenging problem but it might influence the group testing schemes used in real-world laboratories. Thus, let us introduce *sparsity constrained group testing*.

Sparsity constrained group testing As before, we restrict ourselves to the noiseless case. We distinguish between $\Gamma-$sparse group testing in which each test may contain at most Γ individuals and $\Delta-$divisible group testing whose restriction is that each individual may only be tested at most Δ times. Clearly, these models try to built up real world conditions in the sense that the test's sensitivity might decrease with very large pools ($\Gamma-$sparse) or that it is not possible to duplicate the patient's samples arbitrarily often.

Probably the most influential prior work to this thesis's contributions is the one of Gandikota et al. [86]. They stated (universal) information-theoretic converse results in both restriction models and achievability results using the COMP algorithm on the random regular pooling scheme. Let us denote

by $m_{\inf,\mathrm{G}}(\Gamma)$ and $m_{\inf,\mathrm{G}}(\Delta)$ the information-theoretic converse bounds and by $m_{\mathtt{COMP},\mathrm{G}}(\Gamma)$ and $m_{\mathtt{COMP},\mathrm{G}}(\Delta)$ the achievability bounds of COMP in the $\Gamma-$sparse and $\Delta-$divisible setting respectively obtained in [86]. More precisely, the authors find for $\Gamma = \Theta\left(\left(\frac{n}{k}\right)^{\beta}\right)$ for some $\beta \in [0,1)$ that

$$m_{\inf,\mathrm{G}}(\Gamma) = \frac{1}{1-\beta}\frac{n}{\Gamma} \qquad \text{and} \qquad m_{\mathtt{COMP},\mathrm{G}}(\Gamma) = \left\lceil \frac{1}{(1-\theta)(1-\beta)} \right\rceil \left\lceil \frac{n}{\Gamma} \right\rceil . \tag{1.2.10}$$

Furthermore, with $\Delta = o(\ln n)$ they prove weak converse respectively achievability statements at

$$m_{\inf,\mathrm{G}}(\Delta) = \Delta k \left(\frac{n}{k}\right)^{1/\Delta} \qquad \text{and} \qquad m_{\mathtt{COMP},\mathrm{G}}(\Delta) = \left(\mathrm{e}\Delta k n^{1/\Delta}\right). \tag{1.2.11}$$

Comparing the achievability result with the converse statement, we observe a sizeable gap in the $\Delta-$divisible setting whilst in the $\Gamma-$sparse case the gap is only a multiplicative constant factor. We will improve on the converse statements as well as provide a rigorous analysis of the DD algorithm in a tailor-made pooling scheme which improves the achievability bounds in Section 2.1.

After having presented a prime example of a statistical inference problem in large planted versions of statistical physics' models, let us return to the question of how to express the physics' intuition behind the handling of such large random CSPs in a rigorous way. This might help studying random CSPs as well as their planted versions, hence to study statistical inference problems.

1.3. Large discrete systems and their limits

The purpose of this section is two-fold. First, we will dive deeper into the already defined cut-distance for probability measures and give an overview of recent results prior to this thesis's contributions, for instance we will present a regularity lemma for such measures. We will see that this notion is highly inspired by graph regularity and the cut-distance used in graph limit theory and give a very gentle and short introduction into this field as well. Second, we will slightly change the point of view on large graphs from diluted mean-field models (random graphs) to deterministic graphs which are perturbed slightly with a little bit of randomness. The latter ones are useful structures in order to study the expected behaviour of algorithms or to obtain structural results on real world occurrences of large graphs.

1.3.1. Approaching pure states of spin glass systems: the cut-distance

We already learned about the cut-distance $\Delta_{\boxtimes}$ in Section 1.1.2.4. For the sake of the reading flow, recall that it was defined in (1.1.14) for two probability measures μ, ν on some finite set Ω^n as

$$\Delta_{\boxtimes}(\mu,\nu) = \inf_{\substack{\gamma\in\Gamma(\mu,\nu),\\ \phi\in\mathbb{S}_n}} \sup_{\substack{S\subset\Omega^n\times\Omega^n,\\ X\subset[n],\\ \omega\in\Omega}} \left| \sum_{\substack{(\sigma,\tau)\in S,\\ x\in X}} \gamma(\sigma,\tau)\left(\mathbf{1}\{\sigma_x=\omega\} - \mathbf{1}\{\tau_{\phi(x)}=\omega\}\right)\right|,$$

where $\Gamma(\mu,\nu)$ is the set of couplings of μ and ν and $\mathbb{S}_n$ is the set of permutations on $[n]$. Let us write $\mathscr{P}(\Omega^n)$ for the set of probability measures over Ω^n. It can be easily verified that $\Delta_{\boxtimes}(\cdot,\cdot)$ satisfies the triangle-inequality and is symmetric on $\mathscr{P}(\Omega^n)$ but it might happen that $\Delta_{\boxtimes}(\mu,\nu) = 0$ even if $\mu \neq \nu$. Therefore, let $\mathscr{L}_n(\Omega)$ be the set of equivalence classes over $\mathscr{P}(\Omega^n)$ such that one class consists of those measures with cut-distance zero. With a slight misuse of notation, we say that $\mu \in \mathscr{L}_n(\Omega)$ is a probability measure on Ω^n rather than representing some equivalence class and now, $\Delta_{\boxtimes}(\cdot,\cdot)$ defines a metric on $\mathscr{L}_n(\Omega)$. As n is supposed to grow to ∞ in typical applications, it might be tempting to introduce some kind of limit theory.

1.3.1.1. The cut-distance in the thermodynamic limit

The cut-distance for probability measures was introduced by Coja-Oghlan, Perkins and Skubch [42] and the authors provided an idea of how to get meaningful limit objects of discrete probability measures

by using this cut-distance. Let us introduce some notation in order to grasp this idea. If $\sigma \in \Omega^n$ is a configuration, we can translate this configuration into a measurable function from $[0,1)$ into the set of probability measures over Ω. Denote by Σ_Ω the space of all measurable functions from $[0,1)$ to $\mathscr{P}(\Omega)$ up to equality almost everywhere. Then, express σ as

$$\hat{\sigma} : [0,1) \to \mathscr{P}(\Omega) \qquad \text{s.t.} \qquad x \mapsto \sum_{i=1}^{n} \delta_{\sigma_i} \mathbf{1}\left\{ x \in \left[\frac{i-1}{n}, \frac{i}{n} \right) \right\}.$$

If now $\mu \in \mathscr{P}(\Omega^n)$ is a probability measure on Ω^n, Coja-Oghlan, Perkins and Skubch [42] define

$$\hat{\mu} = \sum_{\sigma \in \Omega^n} \mu(\sigma) \delta_{\hat{\sigma}} \qquad \text{s.t.} \qquad \hat{\mu} \in \mathscr{P}(\Sigma_\Omega). \tag{1.3.1}$$

Thus, μ and $\hat{\mu}$ are in 1-to-1-correspondence. Now it is possible to equip $\mathscr{P}(\Sigma_\Omega)$ with a corresponding continuous version of the cut-distance. To this end, let $\mathbb{S}_{[0,1)}$ denote the set of all measure-preserving bijections on $[0,1)$ whose inverse is measure-preserving as well, then the cut-distance of two measures $\mu, \nu \in \mathscr{P}(\Sigma_\Omega)$ is defined as

$$D_{\boxtimes}(\mu,\nu) = \inf_{\substack{\gamma \in \Gamma(\mu,\nu), \\ \varphi \in \mathbb{S}_{[0,1)}}} \sup_{\substack{B \subset \Sigma_\Omega^2, \\ U \subset [0,1), \\ \omega \in \Omega}} \left| \int_B \int_U \sigma_x(\omega) - \tau_{\varphi(x)}(\omega) \mathrm{d}x \mathrm{d}\gamma(\sigma,\tau) \right|. \tag{1.3.2}$$

Again, we identify two measures μ, ν if their cut-distance is zero and obtain $\mathscr{L}_\Omega$ as the space of all such equivalence classes. It can be proven that $\mathscr{L}_\Omega$ is a compact metric space [42, Corollary 2.5]. The authors prove this by the compactness of (a special version) of the so-called graphon space [127] and the existence of a homeomorphism from $\mathscr{L}_\Omega$ into the latter due to [102, Theorem 7.1]. Nevertheless, prior to this thesis's contributions, there was no rigorous limit theory for such probability measures. Especially the limit objects (we will call them $\Omega-$*laws* later on) were not described explicitly and the connection between the discrete and continuous cut-distance was not understood for embedded measures constructed via (1.3.1). Furthermore, the exact connection to the graph limit theory was not carried out and some important key features of graph limit theory, like reconstruction of a finite object by *sampling* from a limit object were not known. We will discuss those missing pieces in Section 2.3. As we referred to graph limit theory a couple of times, let us briefly describe this beautiful theory on a fairly short and intuitive level. Lovász [127] provides a very detailed introduction into graph limit theory for the interested reader.

1.3.1.2. A spark of graph limit theory

In a series of seminal papers, Borgs et al. and Lovász and Szegedy [31, 32, 128] introduced a very powerful theory which connects large discrete objects - growing sequences of dense graphs - with analytical functions called graphons which are measurable functions from the unit-square into $[0,1]$. We give a little insight into this theory following Lovász [127]. In this context, we call a weighted graph $G = (V = \{1,\dots,n\}, E, \Psi)$ on $|V| = n$ vertices *dense*, if there is a constant $\varepsilon > 0$ such that $|E| \geq \varepsilon n^2$, thus a positive proportion of all edges is present. Furthermore, $\Psi : V \times V \to [0,1]$ describes the weight of each edge (observe that a $\{0,1\}-$valued function Ψ makes G a simple graph). Without going much into detail, one way of obtaining such limit objects is by using their version of the cut-distance for graphs. We begin by introducing a very intuitive version of this distance, namely for graphs on the same vertex set of size n. We identify a graph G with its adjacency matrix and denote by G_{ij} its entry, then the cut-distance reads

$$\delta_{\boxtimes}(G,H) = n^{-2} \min_{\varphi \in \mathbb{S}_n} \max_{S,T \subset V} \left| \sum_{s \in S, t \in T} G_{s,t} - H_{\phi(s),\phi(t)} \right|. \tag{1.3.3}$$

Intuitively speaking, the cut-distance measures the largest deviation of edges on subsets of vertices under a best possible re-labelling of the vertices. In particular, isomorphic graphs are those graphs with

cut-distance zero. One can imagine that this definition gets quite involved if it is defined for graphs on a different number of vertices. Roughly said, in this case the cut-distance is defined as the minimum over all values of the cut-distance between blow-ups of the graphs that have the same number of vertices [127, Section 8.1.4].

Let us now embed such finite graphs into the space of *kernels*, thus measurable functions $W : [0,1]^2 \to [0,1]$. Denote the space of all such kernels as $\mathscr{K}$. In order to embed a graph we can, figuratively speaking, shrink the adjacency matrix of a graph G from size $n \times n$ onto the unit-square $[0,1]^2$ such that the built function takes a constant value on each little square corresponding to a matrix entry. Formally, if $G = (V,E)$ is a graph on n vertices, we construct W_G as the corresponding $G-$kernel as follows. Let $S_1, \dots, S_\ell$ be a partition of $[0,1]$ into n intervals such that $S_i = [(i-1)/n, i/n)$. Then let

$$W_G(x,y) = \sum_{i,j=1}^{n} \mathbf{1}\{x \in S_i, y \in S_j\} G_{i,j}$$

be the step-function which takes value 1 on square $S_i \times S_j$ if and only if edge i,j is present in G. Of course, this definition easily generalises to weighted graphs as well.

Thus, we can embed any finite graph into the space of kernels $\mathscr{K}$. But clearly, there are much more functions in $\mathscr{K}$ as those that correspond to finite graphs. Following [127, Section 8.2.1], we define the continuous version of the cut-distance for two kernels $W, W' \in \mathscr{K}$ as

$$\mathscr{D}_{\boxtimes}(W,W') = \inf_{\varphi \in \mathbb{S}_{[0,1]}} \sup_{S,T \subset [0,1]} \left| \int_{S \times T} W(x,y) - W'(\varphi(x),\varphi(y)) \mathrm{d}y \mathrm{d}x \right|. \tag{1.3.4}$$

As in the previous section, the continuous version seems to be the intuitive generalisation of the discrete cut-distance (1.3.3) by replacing sums with integrals and permutations with measure preserving bijections. On th other hand, we are in a good position that such intuition can be formalised rigorously as for two graphs G, H and their corresponding $G-$kernel and $H-$kernel W_G, W_H, we have [127, Lemma 8.9]

$$\delta_{\boxtimes}(G,H) = \mathscr{D}_{\boxtimes}(W_G, W_H).$$

We will see in Section 2.3 that we are not that lucky in the case of limit objects of probability measures but it is possible to achieve similar but slightly weaker results.

Sampling A very elegant aspect about graph limits is that sampling from a kernel yields a finite graph which is close to the kernel in the cut-distance. This fact is often referred to as a *sampling lemma*. More precisely, given a kernel W, let $k > 1$ be an integer and sample $x_1, \dots, x_k$ uniformly and independently from $[0,1]$. Define a random graph $\boldsymbol{G} = \mathscr{G}(k,W)$ on k vertices such that edge ij is present with probability $W(x_i, x_j)$. Then we have the following sampling lemma [127, Lemma 10.16].

> **Sampling Lemma.** Let W be a kernel and $\boldsymbol{G} = \mathscr{G}(k,W)$ defined as above. Then we have with probability at least $1 - \exp(-k/(2\ln k))$ that
>
> $$\mathscr{D}_{\boxtimes}(W_{\boldsymbol{G}}, W) < \frac{22}{\sqrt{\ln k}}.$$

It will turn out in Section 2.3 that a similar fact holds for $\Omega-$laws as well.

Subgraph counts and homomorphism densities A further fairly important aspect about graph limit theory is that the space of kernels really consists of meaningful limit object for sequences of graphs. To this end, we need to briefly mention that graph convergence is actually defined as the convergence of all homomorphism densities of finite graphs into the graph sequence. Therefore, a series of graphs $(G_n)_n$ converges, by definition, if all series of homomorphism densities of finite graphs $(t(H,G_n))_H$ converge in $\mathbb{R}$. Let us define the homomorphism density. A graph homomorphism f from a graph $F = (V(F), E(F))$

into a graph $G = (V(G), E(G))$ is a function $f : V(F) \to V(G)$ that maps edges on edges, thus $ij \in V(F) \Rightarrow f(i)f(j) \in V(G)$. Now, $\hom(F, G_n)$ counts the number of homomorphisms from a given $k-$vertex graph F into G_n and we define the homomorphism density as

$$t(F, G_n) = \frac{\hom(F, G_n)}{n^k}.$$

With this notation at hand, convergence of graphs should be understand as follows. If two graph sequences produce the same convergent series of homomorphism densities, they should have much in common (as they contain equally many copies of all finite graphs) and therefore converge to the same limit object. Fortunately, it turns out that the kernels are good limit objects as we find that for any convergent sequence $(G_n)_n$ of graphs there is a kernel W such that $t(F, G_n)$ converges to $t(F, W)$ for every finite graph F [127, Theorem 11.22]. Such a kernel is called the limit of the graph sequence $(G_n)_n$ and we say that $(G_n)_n$ converges to W. Now it could in principle happen that multiple kernels satisfy this condition and clearly, kernels representing isomorphic graphs should definitely do so. To this end, denote by $\mathcal{W}$ the space of kernels such that we identify kernels of cut-distance zero. We have that $\mathcal{W}$ is a compact Polish space [127, Theorem 9.23]. It turns out that we are in good shape as [127, Theorem 11.22] proves that a sequence of graphs $(G_n)_n$ with a diverging number of vertices converges to a kernel W if and only if $\mathcal{D}_\boxtimes\left(W_{G_n}, W\right) \to 0$ and as $\mathcal{W}$ is a compact metric space, the uniqueness up to equivalent kernels W' - kernels that satisfy $\mathcal{D}_\boxtimes\left(W, W'\right) = 0$ - follows directly.

After discussing some aspects of the kernel representation of a graph limit, let us shortly describe a second important possibility to describe a limit object.

Aldous-Hoover representation The Aldous-Hoover representation theorem for exchangeable arrays of random variables is closely connected to the graph limit theory [6, 96, 102]. More precisely, with W being a kernel, we recall that $\mathcal{G}(k, W)$ is the random graph obtained as follows.

- Draw $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_k$ uniformly at random and independently from $[0, 1]$.
- Let each edge ij be present in $\mathcal{G}(k, W)$ with probability $W(\boldsymbol{X}_i, \boldsymbol{X}_j)$.

This can be naturally extended to an infinite random graph model $\mathcal{G}(\infty, W)$ by sampling infinitely many points $\boldsymbol{X}_i$. Of course, $\mathcal{G}(\infty, W)$ contains every finite random graph $\mathcal{G}(k, W)$ as an induced subgraph [31]. We observe that $\mathcal{G}(\infty, W)$ is an *exchangeable* random graph. Indeed, its distribution is invariant under permutation of the vertices.

Thus, the Aldous-Hoover representation theorem says that any such infinite random graph can be written as a mixture of random variables $A_{ij} = \tilde{W}_k(\boldsymbol{Y}_i, \boldsymbol{Y}_j)$ where $\{\boldsymbol{Y}_i\}_{i\in\mathbb{N}}$ is a family of mutually independent $[0, 1]-$valued random variables and $\left\{\tilde{W}_k\right\}_k$ is a family of symmetric functions $\tilde{W}_k : [0, 1]^2 \to [0, 1]$. This matrix $A = (A_{ij})_{i,j}$ is clearly an infinite random exchangeable array in $[0, 1]^{\mathbb{N}\times\mathbb{N}}$ and one would suggest that it is in 1-to-1 correspondence with the adjacency matrix of $\mathcal{G}(\infty, W)$. And this intuition is indeed correct. It is possible to proof that the set of exchangeable infinite arrays corresponding to such an infinite graph equals the set of extremal points in the space of all exchangeable random arrays. Therefore, the mixture is no real mixture but there is exactly one function $W : [0, 1]^2 \to [0, 1]$ such that $A_{ij} = W(\boldsymbol{Y}_i, \boldsymbol{Y}_j)$ for all i, j. This is, of course, exactly the kernel W.

After having learned a spark of graph limit theory, let us use it to tackle one of the most influential concepts in graph theory, the graph regularity.

1.3.1.3. Regularity of graphs, graph limits and probability measures

Since the first occurrence of a regularity lemma for graphs in 1975, graph regularity and its generalisation have attracted a lot of attention. We refer to two very detailed overview articles by Rödl and Schacht [151] and Komlos et al. [112] describing and analysing various variants of graph regularity and their applications. In this thesis, we will only sketch two types of regularity, one which is just called *regularity* as defined by Szemerédi [160] and a much weaker version called *weak regularity* by Frieze and Kannan

[84]. Besides belonging clearly to the most studied types of regularity, we will see that those concepts are closely related to the cut-distance for probability distributions.

Throughout this section we suppose that $G = (V, E)$ is a graph on n vertices with at least εn^2 edges. The regularity lemma says, intuitively speaking, that the vertex set of each such graph can be partitioned into finitely many classes such that the edges between (almost all) those classes look random. Let us describe this in more detail.

For two subsets $X, Y \subset V$ of the vertex set we denote by $E(X, Y)$ the set of edges with one endpoint in X and one endpoint in Y. Then the edge density between X and Y is defined as

$$d(X, Y) = \frac{|E(X, Y)|}{|X||Y|}.$$

If the edges between X and Y were randomly chosen, we would expect that the edge density between not too small subsets of X and Y roughly equals the overall edge density. Thus, let us call the pair (X, Y) $\varepsilon-$regular if for all $X' \subset X$, $Y' \subset Y$ with $\left|X'\right| \geq \varepsilon |X|$ and $\left|Y'\right| \geq \varepsilon |Y|$ we have

$$\left|d(X, Y) - d(X', Y')\right| \leq \varepsilon.$$

If we now have a partition $\boldsymbol{S} = (S_0, S_1, \ldots, S_\ell)$ of the vertex set V, we say that $\boldsymbol{S}$ is $\varepsilon-$regular if

- the *exceptional set* S_0 satisfies $|S_0| \leq \varepsilon n$,
- for all $1 \leq i < j \leq \ell$ we have that the pair (S_i, S_j) is $\varepsilon-$regular.

Now, the famous regularity lemma of Szemerédi [160] guarantees that every large enough dense graph has such a partition of its vertex set.

> **Regularity Lemma**. For all $\varepsilon > 0$ and every $t \in \mathbb{N}$ there exists an integer $T = T(\varepsilon, t)$ such that each graph G on at least T vertices has an $\varepsilon-$regular partition $\boldsymbol{S} = (S_0, \ldots, S_\ell)$ of its vertex set where $t \leq \ell \leq T$.

A specific feature of this theorem is that T is independent of the graph G and its size. Nevertheless, it turns out that T is lower bounded by a tower of 2s of height proportional to $\ln(1/\varepsilon)$ [92]. One might ask, do we get smaller partitions if we only want to have the property of being regular on average? Indeed, Frieze and Kannan [84] answer this question positively.

Observe that given a regular partition $\boldsymbol{S} = (S_0, \ldots, S_\ell)$ we find that the number of edges between two disjoint subsets of vertices A, B is within $\pm\varepsilon n^2$ of

$$\sum_{i,j=1}^{\ell} d(S_i, S_j) |A \cap V_i| \left|B \cap V_j\right|,$$

thus the latter expression measures somehow the deviation from being regular. Therefore, we say that a partition $\boldsymbol{S} = (S_1, \ldots, S_k)$ of the vertex set of a graph $G = (V, E)$ is *weakly $\varepsilon-$regular* if we have for all disjoint $A, B \subset V$

$$\left| |E(A, B)| - \sum_{i,j=1}^{\ell} d(S_i, S_j) |A \cap S_i| \left|B \cap S_j\right| \right| < \varepsilon. \tag{1.3.5}$$

Now, the weak regularity lemma [84] guarantees the existence of a weakly regular partition for every graph.

> **Weak Regularity Lemma**. For all $\varepsilon > 0$ and every graph G there is a weakly $\varepsilon-$regular partition of its vertex set into k sets such that $k \leq \exp\left(O\left(\varepsilon^{-2}\right)\right)$.

Thus, the weak regularity lemma provides a partition which is on average regular but consists of considerably less parts. Interestingly, if G_S is the weighted graph on the same vertices as G with edge-weight

$d(S_i, S_j)$ for edge uv with $u \in S_i, v \in S_j$, we have that (1.3.5) implies

$$\delta_{\boxtimes}(G, G_S) < 2\varepsilon. \tag{1.3.6}$$

Therefore, the cut-distance is closely related to the concept of (weak) regularity. Of course, a similar result does hold for kernels as well. If W is a kernel and $\boldsymbol{S} = (S_1, \dots, S_\ell)$ is a partition of $[0,1]$ into measurable sets, we define following [127, Section 9.2.1][2]

$$W_S(x,y) = \frac{1}{\lambda(S_i)\lambda(S_j)} \int_{S_i \times S_j} W(x,y) \mathrm{d}y \mathrm{d}x \qquad \text{for } (x \in S_i, y \in S_j). \tag{1.3.7}$$

Thus, W_S is obtained by averaging over each *step* $S_i \times S_j$. Now the weak regularity lemma states that given a kernel W we find a partition $\boldsymbol{S}$ of the unit interval into $k \leq \exp\left(O\left(\varepsilon^{-2}\right)\right)$ sets such that $\mathscr{D}_{\boxtimes}(W, W_S) < \varepsilon$.

Let us now come back to discrete probability measures on Ω^n for some finite set Ω. More precisely, we look at their continuous embeddings as Ω-laws on Σ_Ω. It is possible to define a similar concept of regularity for those objects inspired by the concept of regularity in graph theory [21, 42]. In order to do so, we need to introduce some notation.

For a set $X \subset [0,1)$ and a configuration $\sigma \in \Sigma_\Omega$ as well as a spin $\omega \in \Omega$ define

$$\sigma[\omega \mid X] = \int_X \sigma_x(\omega) \mathrm{d}x.$$

Thus, $\sigma[\cdot \mid X] : [0,1) \to \mathscr{P}(\Omega)$ is a probability distribution on Ω and more precisely, it can be seen as a continuous valued analogue of the empirical distribution of σ on X.

If we now have a partition $\boldsymbol{V} = (V_0, V_1, \dots, V_\ell)$ of $[0,1)$, and a partition $\boldsymbol{S} = (S_0, S_1, \dots, S_k)$ of Σ_Ω we say that μ is $\varepsilon-$regular with respect to $(\boldsymbol{V}, \boldsymbol{S})$, if

(i) the non-exceptional sets $V_1, \dots, V_\ell$ and $S_1, \dots, S_k$ satisfy

$$\lambda(V_i)\,\mu\left(S_j\right) > 0 \qquad \text{and} \qquad \sum_{i=1}^{\ell} \sum_{j=1}^{k} \lambda(V_i)\mu(S_j) \geq 1 - \varepsilon,$$

(ii) for all $1 \leq i \leq \ell$, $1 \leq j \leq k$ we have

$$\max_{\sigma, \sigma' \in S_i} \left|\left|\sigma[\cdot|V_j] - \sigma'[\cdot|V_j]\right|\right|_1 < \varepsilon,$$

(iii) for all $1 \leq i \leq \ell$ and $1 \leq j \leq k$ we have for $U \subset V_i$ with $\lambda(U) \geq \varepsilon\lambda(V_i)$ and $T \subset S_j$ with $\mu(T) \geq \varepsilon\mu(S_j)$ that

$$\left|\left|\langle \boldsymbol{\sigma}[\cdot \mid U]\rangle_{\mu[\cdot|T]} - \langle \boldsymbol{\sigma}[\cdot \mid V_i]\rangle_{\mu[\cdot|S_j]}\right|\right|_1 < \varepsilon.$$

This definition deviates slightly from the one in [42] with respect to the exceptional sets but is clearly equivalent. As stated in this publication, (ii) guarantees that the averages $\sigma[\cdot \mid V_i]$ and $\sigma'[\cdot \mid V_i]$ over V_i of any two configurations from one cluster S_j are close. More importantly, (iii) requires that the average $\langle \boldsymbol{\sigma}[\cdot \mid U]\rangle_{\mu[\cdot|T]}$ over a large *sub-square* does roughly equal the mean over the square given by the partition $V_i \times S_j$.

As in the case of graph regularity, this can be expressed via the cut-distance. Given an $\varepsilon-$regular partition $(\boldsymbol{V}, \boldsymbol{S})$, we define

$$\sigma_x[\omega|\boldsymbol{V}] = \sum_{i=0}^{\ell} \mathbf{1}\{x \in V_i\}\, \sigma_x[\omega \mid V_i].$$

Furthermore, we let $\mu[\cdot \mid \boldsymbol{V}, \boldsymbol{S}]$ be the conditional expectation of μ with respect to this partition, thus

$$\mu[\cdot \mid \boldsymbol{V}, \boldsymbol{S}] = \sum_{j=0}^{k} \delta_{\int_{S_j} \sigma[\cdot|\boldsymbol{V}] \mathrm{d}\mu(\sigma)}.$$

[2] We denote by $\lambda(\cdot)$ the Lebesgue-measure.

Therefore, $\mu[\cdot \mid \boldsymbol{V}, \boldsymbol{S}] \in \mathscr{L}_\Omega$ is an $\Omega-$law which is supported on a discrete set of configurations $\sigma : [0,1) \to \mathscr{P}(\Omega)$ which themselves are constant on each of the partition classes $\boldsymbol{V}$. Comparing $\mu[\cdot \mid \boldsymbol{V}, \boldsymbol{S}]$ with the step-kernel W_S given via (1.3.7) we see that the two structures express very similar ideas, for graphs on the one hand and probability measures on the other hand.

Thus, it might be no surprise that we find for an Ω-law μ and an $\varepsilon-$regular partition $(\boldsymbol{V}, \boldsymbol{S})$ that [42, Proposition 2.14]

$$D_{\boxtimes}\left(\mu, \mu[\cdot \mid \boldsymbol{V}, \boldsymbol{S}]\right) < 2\varepsilon.$$

Moreover, the authors provide a regularity lemma for such $\Omega-$laws [42, Corollary 2.15].

> **Regularity Lemma for $\Omega-$laws.** For any $\varepsilon > 0$ there is a natural number $N = N(\varepsilon)$ such that for any Ω-law μ there are $\sigma_1, \dots, \sigma_N : [0,1) \to \mathscr{P}(\Omega)$ and $\omega = (\omega_1, \dots, \omega_N) \in \mathscr{P}([N])$ with $D_{\boxtimes}(\mu, \sum_{i=1}^N \omega_i \delta_{\sigma_i}) < \varepsilon$.

A similar statement is known for probability measures μ on Ω^n [21, Theorem 2.1]. Let us bring this notion of regularity together with the idea of $\varepsilon-$symmetry discussed previously. First we observe that any *refinement* of the classes $V_1, \dots, V_\ell$ only increases the cut-distance between μ and $\mu[\cdot \mid \boldsymbol{V}, \boldsymbol{S}]$ by a constant factor. Therefore, we can refine a regular partition $(\boldsymbol{V}, \boldsymbol{S})$ into singletons $V_i = \{i\}$ and observe that in this case $\mu[\cdot \mid \boldsymbol{V}, \boldsymbol{S}]$ becomes a convex combination over (on S_i conditioned) product measures on the marginals of $\mu[\cdot \mid S_i]$. Recall that for a probability measure $\nu \in \mathscr{P}(\Omega^n)$ we denote by $\bar{\nu}$ the corresponding product measure on the same marginals, therefore, on each partition class S_i we find $\mu[\cdot \mid \boldsymbol{V}, \boldsymbol{S}] = \bar{\mu}[\cdot \mid S_i]$.

A direct consequence [21, Corollary 2.2] of the regularity lemma is the following. If $\boldsymbol{S}$ is an ε^3-regular partition of (Ω^n) (we drop the singleton decomposition of $\boldsymbol{V}$ for the sake of readability from now on) w.r.t. $\mu \in \mathscr{P}(\Omega^n)$ then we have:

> **Regularity and symmetry.** For any $\varepsilon > 0$ there is $\eta = \eta(\varepsilon, \Omega) > 0$ such that for all $n \geq \eta^{-1}$ and any probability measure μ on Ω^n we have for $j = 1, \dots, k$ that $\Delta_{\boxtimes}(\mu[\cdot \mid S_j], \bar{\mu}[\cdot \mid S_j]) < O(\varepsilon)$.

Thus, finding an $\varepsilon-$regular partition of the phase space guarantees to express a probability measure as an convex combination of measures conditioned on the partition classes which look like a product measure under the cut-distance. Clearly, it would be interesting to have algorithms that construct such partitions efficiently. Here comes the flaw of all described regularity lemmas. While they guarantee the existence of regular partitions, it is not clear how to generate them. A fairly elegant way of obtaining a partition $\boldsymbol{S}$ of Ω^n for some measure $\mu \in \mathscr{P}(\Omega^n)$ is the so-called *pinning operation* introduced by Coja-Oghlan et al. [49].

1.3.1.4. Pinning

Suppose we have a measure $\mu \in \mathscr{P}(\Omega^n)$ and we want to obtain a fairly related measure which is $\varepsilon-$symmetric. In this case, the *pinning lemma* [49, Lemma 3.5] provides a very simple way of achieving this goal.

> **Pinning Lemma.** For any $\varepsilon > 0$ there is a natural number $T > 0$ such that for every $n > T$ and every $\mu \in \mathscr{P}(\Omega^n)$ the following holds. Construct a (random) probability measure $\boldsymbol{\nu} = \boldsymbol{\nu}(\mu) \in \mathscr{P}(\Omega^n)$ as follows.
>
> - Sample $\tilde{\boldsymbol{\sigma}} \sim \mu$.
> - Choose independently an integer $\boldsymbol{\theta} \in [1, T)$ uniformly at random.
> - Create a random subset $U \subset [n]$ by including each coordinate with probability $\boldsymbol{\theta}/n$ independently.
> - Define
>
> $$\boldsymbol{\nu}(\sigma) = \mu(\sigma) \frac{\mathbf{1}\{\forall i \in U : \sigma_i = \tilde{\boldsymbol{\sigma}}_i\}}{\mu(\{\tau : \forall i \in U : \tau_i = \tilde{\boldsymbol{\sigma}}_i\})}.$$
>
> Then we have $\Delta_{\boxtimes}(\boldsymbol{\nu}, \bar{\boldsymbol{\nu}}) < O\left(\varepsilon^{1/3}\right)$ with probability at least $1 - \varepsilon$.

Therefore, we draw just a single sample from μ and *pin* a relatively small number of coordinates to their spins under this sample. This reweighed measure is now likely to be $\varepsilon-$symmetric. Fairly related versions of such lemmas for probability measures were previously obtained by Montanari [137, Lemma 3.1] and Raghavendra and Tan [148].

It actually turns out that the pinning operation does not only yield such a reweighed measure but that we can actually obtain a partition of the phase space and a corresponding family of reweighed measures, when we just define the partition as given by all $|\Omega|^{|U|}$ possibilities of spins those variables in U can take. While this operation is clearly a mighty tool for producing regular partitions, it was not known prior to this thesis's contributions whether and how similar approaches might work for Ω-laws. We will discuss this in Section 2.3.

Let us for the sake of completeness mention that similar approaches of obtaining regular partitions for graphs have been studied as well by (non-exclusively) Tao [163] and Fischer, Matsliah and Shapira [79] by choosing partitions according to the adjacency of randomly selected vertices. Up to our knowledge, a complete understanding on how the procedure of sampling and reweighing can be explicitly applied in this context is not yet gained.

In this section we discussed the cut-distance and its connection to regularity for probability measures and graphs. We saw that regularity implies that (dense) large graphs cannot *look arbitrarily wild* but do contain a lot of structure. Structural results of (sparse) graphs will be the topic of the next section which introduces the concept of *random perturbation* of graphs.

1.4. Perturbing sparse graphs: when randomness meets determinism

The origin of randomly perturbing deterministic structures can be traced back to a contribution of Spielman and Teng [159] who introduced the *smoothed analysis of algorithms*. The key idea is fairly simple. While it is known for many algorithms that their worst-case running time exceeds the average case running time significantly, it can be observed in real world applications that this worst-case does not usually occur. A prime example might be the simplex algorithm [61] whose worst-case complexity is exponential but nevertheless, the method is used quite frequently in applications. Of course, one could analyse the *average case* running time but this might not give enough performance guarantees in production systems, as it is not unlikely to observe an input deviating from the average case. The smoothed analysis of algorithms overcomes this flaw by analysing a worst-case input on whom random changes have been applied. If the number of changes is fairly high, the instance clearly becomes average case by definition but if the number of manipulations is small, the modified input is fairly close to a worst-case instance which is reasonable to be observed in real applications as hitting the absolute worst-case is very unlikely.

Subsequently, this idea of randomly perturbing deterministic structures became of interest in the study of random graphs. For instance, one of the most famous results of extremal combinatorics might be Dirac's theorem [65] on Hamilton cycles. It states that whenever a graph on n vertices has minimum degree at least $\lceil n/2 \rceil$, this graph contains a Hamilton cycle, thus a closed loop through the graph visiting each vertex exactly once. This result is optimal in the sense that there are graphs with minimum degree $\lfloor n/2 \rfloor$ which do not contain such a cycle, i.e. let $n = 2m-1$ and take two copies of the complete graph on m vertices K_m which share exactly one common vertex. The latter graph has clearly no Hamilton cycle but has minimum degree $\lfloor n/2 \rfloor$. But it is of course very unlikely to observe such an extreme graph in a real world network, thus many graphs with smaller minimum degree contain a Hamilton cycle (clearly, the necessary minimum degree is 2 achieved by the cycle on n vertices C_n). Let us look at the average case, which is the binomial random graph. It is well understood that $\mathcal{G}(n,p)$ undergoes a strict phase transition with respect to containing a Hamilton cycle at $p = n^{-1} \ln n$ [116, 117, 147]. What happens if we combine the probabilistic and deterministic objects? Denote by $G_\alpha = (V,E)$ an arbitrary (probably adversely chosen) graph on $V = [n]$ with minimum degree αn and denote furthermore by $F = G_\alpha \cup \mathcal{G}(n,p)$ the union of this graph and an instance of the binomial random graph with edge probability p. Here we define the union as follows: For each pair of vertices $i, j \in \binom{V}{2}$, we add an edge of $\mathcal{G}(n,p)$ independently

of everything else with probability p. When does F contain a Hamilton cycle with high probability?

Clearly, if $\alpha \geq 1/2$, the existence follows solely from Dirac's theorem applied to G_α. If on the other hand $p \geq (1+\varepsilon)\ln n/n$, we find the cycle inside of the edges of $\mathscr{G}(n,p)$ with high probability. But what happens in between?

This model of randomly perturbed graphs with given minimum degree was introduced by Bohman, Frieze and Martin [28] for $\alpha = \Theta(1)$, thus for dense graphs. In the aforementioned contribution the authors explicitly find the trade-off between α and p. More precisely, they proved that for every constant $0 < \alpha < 1/2$ there are graphs G_α such that $G_\alpha \cup \boldsymbol{G}(n,p)$ does not contain a Hamilton cycle with high probability if $p = o(1/n)$. On the other hand, if $p = \omega(1/n)$, any such union of graphs contains a Hamilton cycle with high probability. It is important to observe that $p = 1/n$ is the phase transition point in (solely) $\mathscr{G}(n,p)$ of containing a cycle on all but εn vertices. Thus the transition point for the spanning structure in the perturbed model equals (in this case) the transition point of the existence of an almost spanning structure in the random graph which is due to the existence of isolated vertices below $p \sim \ln n/n$.

Of course, there is not only interest for Hamilton cycles but also for various different spanning structures. On the side of finding Dirac-like theorems for the existence of specific spanning structures in G_α solely, there are for instance results for spanning trees [115], factors [94] as well as powers of Hamilton cycles [113, 114]. Finally, there are even fairly generic results for the existence of a copy of any bounded degree graph in G_α by Böttcher, Schacht and Taraz [36]. And clearly, the existence of all of those structures is known to undergo phase transitions in the random graph $\mathscr{G}(n,p)$. To briefly name a few important contributions, there are results on the existence of matchings [72], spanning trees [118, 139], factors [107] and powers of Hamilton cycles [125, 141]. Finally, there are also generic results on the phase transitions with respect to general bounded degree graphs [14, 76, 77, 150]. We refer to a recent overview article of Böttcher [33] for a more detailed presentation.

It is not very surprising that since the first discussion of the existence of Hamilton cycles in randomly perturbed graphs various contributions obtained results with respect to the aforementioned spanning structures. Just to name a few, there are results on spanning trees [34, 119], factors [20] as well as powers of Hamilton cycles [26]. Ultimately, there are recent results on the existence of general bounded degree graphs in the perturbed model by Böttcher et al. [35]. Interestingly, in most of the results, the obtained phase transition for p is a multiplicative factor of order $\ln n$ smaller than in $\mathscr{G}(n,p)$.

The knowledge of things becomes completely different if one allows $\alpha = o(1)$, thus one has a deterministic sparse graph G_α and needs more edges from the random graph. We will investigate the existence of perfect matchings, Hamilton cycles and bounded degree trees in this model of sparse perturbed graphs in Section 2.4.

2. Results

This chapter summarises the main results obtained in this thesis's contributions. Those results will be presented and very short proof sketches will be given showing the most important steps in order to achieve those results. We emphasise that those sketches make simplifying assumptions and leave out all technical details because they are just meant to grasp the main idea of how to prove a result. For complete and rigorous proofs we refer the reader to the contributions in the appendix.

We start by discussing results with respect to the group testing problem. Subsequently, we will answer the question of how many satisfying assignments a random 2-SAT formula has and state results in context with a limit theory for discrete probability measures and the role of the cut-distance. Finally, we discuss the existence of spanning structures in randomly perturbed sparse graphs.

2.1. Group Testing

As already discussed in the introduction, all our results are within the framework of Bayes optimal sublinear hypergeometric probabilistic group testing, thus we try to achieve inference with high probability and the ground-truth $\boldsymbol{\sigma}$ is supposed to be chosen uniformly at random from all possible configurations in $\{0,1\}^n$ of Hamming weight $k \sim n^\theta$ for some $\theta \in (0,1)$. The results of this section were obtained in the following contributions which can be found in the appendix:

- *Information-Theoretic and algorithmic thresholds for group testing* [41],
- *Optimal group testing* [46],
- *Near optimal sparsity-constrained group testing: improved bounds and algorithms* [88].

We start by discussing results with respect to non-adaptive group testing schemes. A summary of the obtained results can be found at the end of the section.

2.1.1. Non-adaptive Group Testing

Almost all obtained results make extensive use of the occurrence of different types of individuals in a group testing instance. Following [88], we will shortly describe the combinatorial meaning of those types.

Throughout the section we suppose that we have individuals $V = \{x_1, \dots, x_n\}$ and tests $F = \{a_1, \dots, a_m\}$ and that each individual's infection status is given by the underlying ground-truth $\boldsymbol{\sigma}$ and all test-results are given by $\hat{\boldsymbol{\sigma}}$.

2.1.1.1. Combinatorial properties of individuals

Suppose that some (non-adaptive) pooling scheme is given through the factor graph $\mathcal{G} = (V \cup F, E)$. We abbreviate the set of uninfected individuals to V_0 and the set of infected individuals to V_1, thus

$$V_0(\mathcal{G}) = \{x \in V(\mathcal{G}) : \boldsymbol{\sigma}_x = 0\} \qquad \text{and} \qquad V_1(\mathcal{G}) = \{x \in V(\mathcal{G}) : \boldsymbol{\sigma}_x = 1\}.$$

Those uninfected individuals appearing in a negative test can be classified immediately and play therefore a special role. We define this set of individuals as V_{0-}, formally

$$V_{0-}(\mathcal{G}) = \{x \in V_0(\mathcal{G}) : \exists a \in \partial_{\mathcal{G}} x : \hat{\boldsymbol{\sigma}}_a = 0\}.$$

Furthermore, with respect to the `DD` algorithm, it intuitively makes sense to denote the set of infected individuals which appear in at least one test with only elements of V_{0-}. Thus, upon classifying the latter

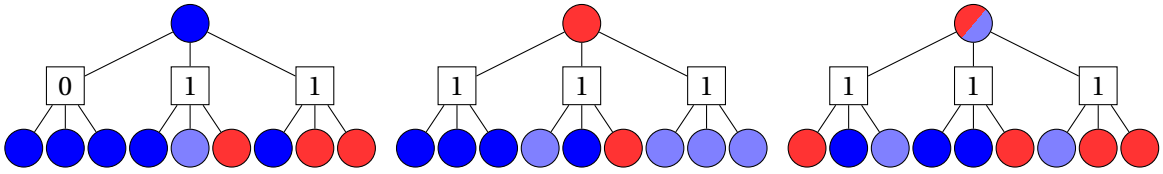

Figure 2.1.: The local structure given by the underlying factor graph in the non-adaptive group testing problem modified after [88]. Blue individuals are uninfected under the ground-truth $\boldsymbol{\sigma}$ while red individuals are infected. More precisely, we suppose that light blue individuals belong to V_{0+} and dark blue individuals to V_{0-}. From left to right, the upper individuals are elements of V_{0-}, V_{1--} and V_+ respectively.

individuals, those infected individuals are easy to identify. More precisely, we define

$$V_{1--}(\mathcal{G}) = \{x \in V_1(\mathcal{G}) : \exists a \in \partial_{\mathcal{G}} x : \partial_{\mathcal{G}} a \setminus \{x\} \subset V_{0-}(\mathcal{G})\}.$$

Moreover, there might be *totally disguised* individuals. Following Aldridge, Johnson and Scarlett [10] and Mézard, Tarzia and Toninelli [133], we say that an individual x is *disguised* in a test a if there is an infected individual $y \in \partial a \setminus \{x\}$. A totally disguised individual is disguised in all of its tests. Clearly, the infection status of those individuals cannot be inferred directly, but using the prior, it is possible to declare any of those individuals as uninfected as long as there are not too many totally disguised individuals [8, 41]. Formally, we let

$$V_+(\mathcal{G}) = \{x \in V(\mathcal{G}) : \forall a \in \partial_{\mathcal{G}} x : (\partial_{\mathcal{G}} a \setminus \{x\}) \cap V_1(\mathcal{G}) \neq \emptyset\}.$$

For the sake of completeness, we furthermore define the set of totally disguised infected and uninfected individuals, thus

$$V_{0+}(\mathcal{G}) = V_+(\mathcal{G}) \cap V_0(\mathcal{G}) \qquad \text{and} \qquad V_{1+}(\mathcal{G}) = V_+(\mathcal{G}) \cap V_1(\mathcal{G}).$$

If it is clear from the context which pooling scheme $\mathcal{G}$ is the matter of discussion, we will write $V_{0+} = V_{0+}(\mathcal{G})$ and equivalently abbreviate the other sets. A graphical visualisation of those types of individuals is given in Figure 2.1.

The sizes of the sets V_{1+}, V_{0+}, V_{1--} have direct impact on the algorithmic and information-theoretic feasibility of a group testing instance. For a configuration $\tau \in \{0,1\}^n$ we denote by $\hat{\tau} = \hat{\tau}(\mathcal{G})$ the corresponding test-results on a pooling scheme $\mathcal{G}$. Now we define

$$S_k = S_k(\mathcal{G}, \boldsymbol{\sigma}) = \left\{\tau \in \{0,1\}^n : ||\tau||_1 = k \quad \text{and} \quad \hat{\tau} = \hat{\boldsymbol{\sigma}}\right\} \qquad \text{as well as} \qquad Z_k = Z_k(\mathcal{G}, \boldsymbol{\sigma}) = |S_k|.$$

This notation enables us to find the following assertion which holds as the infection status of totally disguised individuals can be swapped arbitrarily due to the supposed Bayes optimality ([41, Corollary 2.2] and [88, Claim 2.3]).

Lemma 2.1.1. *Let $Z_k(\mathcal{G}, \boldsymbol{\sigma})$ be defined as above, then the following holds.*

- *If $Z_k(\mathcal{G}, \boldsymbol{\sigma}) = 1$, there exists a (not necessarily efficient) algorithm which infers $\boldsymbol{\sigma}$ from $(\mathcal{G}, \hat{\boldsymbol{\sigma}})$ with high probability.*
- *If $Z_k(\mathcal{G}, \boldsymbol{\sigma}) = \ell$, any algorithm (efficient or not) fails at inference of $\boldsymbol{\sigma}$ from $(\mathcal{G}, \hat{\boldsymbol{\sigma}})$ with probability at least $1 - \ell^{-1}$.*
- *For any test design $\mathcal{G}$, we have $Z_k(\mathcal{G}, \boldsymbol{\sigma}) \geq |V_{1+}(\mathcal{G}) \times V_{0+}(\mathcal{G})|$.*

A similar statement follows for the `DD` algorithm but here we need to take the set V_{1--} into account because it consists of exactly those individuals which will be misclassified by `DD`.

Lemma 2.1.2 (Corollary 2.4 of [88])**.** *The `DD` algorithm recovers $\boldsymbol{\sigma}$ from $(\mathcal{G}, \hat{\boldsymbol{\sigma}})$ if and only if $V_1(\mathcal{G}) = V_{1--}(\mathcal{G})$.*

Let us subsequently state achieved information-theoretic bounds for non-adaptive hypergeometric probabilistic group testing.

2.1.1.2. Information-theoretical results

Recall m_{inf} and $m_{\text{rand-reg}}$ from (1.2.5) and (1.2.8) as

$$m_{\text{inf}} = \frac{1}{\ln 2} k \ln \frac{n}{k} \qquad \text{and} \qquad m_{\text{rand-reg}} = \max\left\{ \frac{\theta}{(1-\theta)\ln^2 2}, \frac{1}{\ln 2} \right\} k \ln \frac{n}{k}.$$

As discussed earlier, the random regular model was known to fail with positive probability below $m_{\text{rand-reg}}$ due to [9]. We strengthen this converse statement and establish an achievability statement at the same value, thus obtaining a strict phase transition in the random regular model. With a slight misuse of notation we refer to a model as the *random (almost) regular model* if each individual chooses Δ tests uniformly at random with or without replacement. While the technical delicacies of the proofs change depending on the exact model formulation the results themselves stay the same.

Theorem 2.1.3 (Theorem 1.1 of [41])**.** *Let $\mathcal{G}$ be the random almost regular pooling scheme with n individuals and m tests, $\varepsilon > 0$ and $k \sim n^\theta$. Then the following holds.*

- *If $m > (1+\varepsilon) m_{\text{rand-reg}}$, there is an (exponential time) algorithm inferring $\boldsymbol{\sigma}$ from $(\mathcal{G}, \hat{\boldsymbol{\sigma}})$ with high probability.*
- *If $m < (1-\varepsilon) m_{\text{rand-reg}}$, no algorithm (efficient or not) outputs $\boldsymbol{\sigma}$ given $(\mathcal{G}, \hat{\boldsymbol{\sigma}})$ with a non-vanishing probability.*

Proof sketch of Theorem 2.1.3. We need to establish two directions of the theorem, let us start with the achievability result. We do this by employing Markov's inequality and an idea from statistical physics. More precisely, we denote by $Z_{k,\ell}$ the number of configurations $\tau \neq \boldsymbol{\sigma}$ satisfying all test-results that have overlap ℓ with $\boldsymbol{\sigma}$ where the overlap is defined as the number of individuals which are infected under τ and $\boldsymbol{\sigma}$. If we can show that the expected number of such individuals (for the sum over $\ell = 0 \dots k-1$) is $o(1)$, Markov's inequality shows that there is, w.h.p., only one satisfying assignment (namely $\boldsymbol{\sigma}$) which can be found by exhaustive search.

Thus, let us bound the expected number of such configurations. First, we suppose $\ell < (1 - 1/\ln n)k$ and calculate the expected number of individuals quite directly as

$$\mathbb{E}[Z_{k,\ell}(\mathcal{G}, \hat{\boldsymbol{\sigma}})] \leq O(1) \binom{k}{\ell} \binom{n-k}{k-\ell} \left(1 - 2(1-k/n)^{\bar{\Gamma}} + 2(1-2k/n+\ell/n)^{\bar{\Gamma}}\right)^m, \tag{2.1.1}$$

where $\bar{\Gamma} = n\Delta/m$ is the average test size. The combinatorial meaning is immediate. While the product of the two binomial coefficients counts the number of possible assignments τ that have overlap ℓ with $\boldsymbol{\sigma}$, the last factor is the probability that an average test renders the same result under $\boldsymbol{\sigma}$ and τ. Of course, we made many simplifying assumptions as, for instance, we supposed that all tests are independent and the single test degrees are sufficiently concentrated. But it turns out that this can be turned into a rigorous argument. Now we find – as long as $\ell < (1 - 1/\ln n)k$ – that an easy calculation provides

$$\sum_{\ell=0}^{(1-1/\ln n)k} \mathbb{E}[Z_{k,\ell}(\mathcal{G}, \hat{\boldsymbol{\sigma}})] = o(1)$$

for the choice of $m = (1+\varepsilon) m_{\text{rand-reg}}$.

Unfortunately, this argument fails for very large overlaps as the r.h.s. of (2.1.1) gets too large, thus the expectation overshoots the value of the random variable dramatically. This is a kind of a *lottery effect* and is well known in the random CSP literature as for big overlap values rare but very solution-rich clusters dominate the expectation [3]. But fortunately, the random CSP literature enables us to cope with such phenomena [1]. In short, we only need to show that the underlying random graph simply does not

allow solutions having an overlap close to k as $\boldsymbol{\sigma}$ is locally rigid. More precisely, we can proof that above $m_{\text{rand-reg}}$, each individual is part of $\Theta(\Delta) = \Theta(\ln n)$ tests such that all other contained individuals belong to V_{0-}. Therefore, upon changing the infection status of one individual from $1 \to 0$, we directly need to change the status of $\sim \ln n$ different individuals from $0 \to 1$ to compensate for those tests. But now, the same applies for those individuals, thus we have to change the status of another $\ln n$ individuals from $1 \to 0$. This argument goes on for a couple of rounds until we proved that ℓ needs to be smaller than $(1 - 1/\ln n)k$.

We observe at this point that the both expressions in the $\max(\cdot,\cdot)$ of $m_{\text{rand-reg}}$ exactly account for those two arguments. While the $\ln^{-1} 2$ from the universal counting bound suffices to guarantee that the expected number of alternative satisfying configurations is small, the non-adaptivity part accounts for the local rigidity. This perfectly fits into the previous discussion.

Let us subsequently sketch how to achieve the converse statement, namely that any inference algorithm fails below $m_{\text{rand-reg}}$ on the random regular model. As $m_{\text{rand-reg}}$ coincides with m_{inf} for $\theta < \frac{\ln 2}{1+\ln 2}$, we only need to consider larger values of θ. The proof idea is fairly simple, as by Lemma 2.1.1, we just need to calculate that we find $\omega(1)$ totally disguised infected and uninfected individuals with high probability. For this proof sketch, we suppose that all tests are stochastically independent but we remark that the actual proof is technically challenging in order to cope with the delicate dependencies in the random regular model. I.e., it is possible to describe the number of infected individuals in each test by a family of independent but conditioned binomial random variables. As a first step, we prove that in the random regular model on $m = ck\ln\frac{n}{k}$ tests the choice $\Delta = c\ln 2 \ln\frac{n}{k}$ is optimal. Intuitively, this maximises the entropy of the system. Having established this, the probability for an individual x to be totally disguised turns out to be roughly

$$(1-(1-k/n)^{\tilde{\Gamma}-1})^{\Delta} \sim 2^{-\Delta} \sim n^{-(1-\theta)c\ln^2 2}.$$

Thus, the expected number of individuals in V_{1+} is $n^{\theta-(1-\theta)c\ln^2 2}$ and for V_{0+} we find $\mathbb{E}[|V_{0+}|] \gg \mathbb{E}[|V_{1+}|]$. Clearly, $\mathbb{E}[|V_{1+}|]$ diverges as fast as a polynomial in n if $m = (1-\varepsilon)m_{\text{rand-reg}}$ and luckily it turns out that Chebyshev's inequality suffices to guarantee enough concentration. □

Actually, it turns out that Theorem 2.1.3 is not the strongest result we can achieve. While it answers all questions regarding information-theoretical inference in the random regular model, it might be the case that there are different pooling schemes facilitating better. This is, indeed, not the case.

Theorem 2.1.4 (Theorem 2 of [46])**.** *Let $\mathcal{G}$ be any arbitrary non-adaptive pooling scheme with n individuals and m tests, $\varepsilon > 0$ and $k \sim n^{\theta}$. If $m < (1-\varepsilon)m_{\text{rand-reg}}$, no algorithm (efficient or not) outputs $\boldsymbol{\sigma}$ given $(\mathcal{G}, \hat{\boldsymbol{\sigma}})$ with a non-vanishing probability.*

Theorem 2.1.4 implies two important facts. First of all, the random regular design is information-theoretically optimal as there cannot be any better designs for inference. Second, as

$$m_{\text{rand-reg}} > m_{\text{inf}} \qquad \text{for} \qquad \theta > \frac{\ln 2}{1+\ln 2},$$

we proved the existence of an *adaptivity-gap*. Thus, performing multiple stages of testing might decrease the number of required tests. Formerly, the question about the existence of such an adaptivity-gap has been raised prominently [10, 108]. In light of the second part of Theorem 2.1.3 and Theorem 2.1.4, we will write

$$m_{\text{non-ada}} = m_{\text{rand-reg}}$$

from now on as this marks a strict phase transition point for non-adaptive group testing. Let us now give a proof sketch of Theorem 2.1.4.

Proof sketch of Theorem 2.1.4. As before, we only need to establish the assertion for $\theta > \frac{\ln 2}{1+\ln 2}$ as the universal counting bound already proves the theorem for smaller values of θ.

This proof comes in three steps. First, we slightly change the model from a hypergeometric group testing ground-truth $\boldsymbol{\sigma}$ to an i.i.d. ground-truth $\tilde{\boldsymbol{\sigma}}$ where each entry is set to one with probability $p \sim$

$\frac{k-\sqrt{k}\ln n}{n}$. Thus, with high probability, we find a coupling of $(\boldsymbol{\sigma},\tilde{\boldsymbol{\sigma}})$ such that turning few uninfected individuals in $\tilde{\boldsymbol{\sigma}}$ to infected creates $\boldsymbol{\sigma}$. It turns out that finding totally disguised infected and uninfected individuals in a testing scheme under $\tilde{\boldsymbol{\sigma}}$ implies to find some of them also under $\boldsymbol{\sigma}$. This change of the model simplifies proving converse statements. As the underlying graph is deterministic, this enables us to create the only source of independent randomness within the setup.

Second, we show that the number of totally disguised infected and uninfected individuals below m_{inf} is large for θ really close to one. We first observe that even in an arbitrary test-design no test contains more than $n\ln n/k$ individuals as otherwise a union bound over all tests shows that these tests render a positive result anyways and could also be left out. But this already implies due to the handshaking lemma that there are very few individuals participating in more than $\ln^3 n$ many tests. Thus, the underlying testing scheme is, besides not knowing how it looks like exactly, fairly sparse if θ is large. Next, we strengthen an argument based on the FKG-inequality and the probabilistic method of Aldridge [7] and Mézard, Tarzia and Toninelli [133] to identify one individual y whose probability of being element of V_+ is not too small, say $\geq \exp\left(-m\ln^2(2)/k\right)$. Of course, this probability tends to zero (making it much more difficult to find enough such individuals compared to Aldridge's case). Next, we delete this individual and all tests and individuals in its first, second, third and fourth neighbourhood. Clearly, the left-over individuals are stochastically independent of y with respect to the property of belonging to V_+. Furthermore, upon removal of the individuals, chances of being totally disguised only increase. Let us denote by $\mathcal{G}_1$ the graph after removal of y and its described neighbourhoods. Due to the sparsity of the underlying graph we can show that $\frac{m}{n} \sim \frac{|F(\mathcal{G}_1)|}{|V(\mathcal{G}_1)|}$, thus the ratio between tests and individuals stays roughly the same. We repeat this procedure $n^{1-\delta}$ times where $\delta = \delta(\theta) > 0$ is a small constant depending on the prevalence and identify $n^{1-\delta}$ individuals which all are independently totally disguised with probability at least $q = \exp\left(-m\ln^2(2)/k\right)$. As the infection status is independent of being totally disguised, we find that the number of infected totally disguised individuals is dominated by a $\mathbf{Bin}(n^{1-\delta}, pq)$ random variable while the number of totally disguised uninfected individuals is dominated by a $\mathbf{Bin}(n^{1-\delta},(1-p)q)$ variable. Both binomial random variables turn out to have expectation $n^{\Omega(1)}$ for a suitable choice of δ if $m = (1-\varepsilon)m_{\text{non-ada}}$ and thus the Chernoff bound guarantees the existence of many infected and uninfected totally disguised individuals.

As a third step, we need to show that if we could solve a group testing instance with low prevalence given through θ with $(1-\varepsilon)m_{\text{non-ada}}(\theta)$ tests we were also able to solve an instance of larger prevalence given via θ' with $(1-\eta)m_{\text{non-ada}}(\theta')$ tests. To this end, let $\frac{\ln 2}{1+\ln 2} < \theta < \theta' < 1$ and suppose that $\mathcal{G} = (V \cup F, E)$ is a pooling graph satisfying

$$|V(\mathcal{G})| = n, \qquad \text{and} \qquad |F(\mathcal{G})| = m = (1-\varepsilon)m_{\text{non-ada}}(n,\theta).$$

We construct a pooling graph $\mathcal{G}'$ for $n' \approx n^{\theta/\theta'}$ individuals out of which $k' \sim n'^{\theta'} \sim k$ are infected, thus an instance with the same number of infected individuals but those are found within a much smaller population $n' \ll n$.

- Select n' individuals uniformly at random out of all individuals and define $\mathcal{G}'$ on those individuals but on the same tests as $\mathcal{G}$.
- Select an adjusted ground-truth $\boldsymbol{\sigma}' \in \{0,1\}^{n'}$ u.a.r. with Hamming weight k and denote by $\hat{\boldsymbol{\sigma}}'$ the corresponding test-results.

Now it is easy to proof that the probability of having multiple elements in the solution space of $(\mathcal{G},\hat{\boldsymbol{\sigma}})$ is at least as high as observing that many in the solution space of $(\mathcal{G}',\hat{\boldsymbol{\sigma}}')$. Indeed, by construction there is a coupling of $\boldsymbol{\sigma}$ and $\boldsymbol{\sigma}'$ such that all infected individuals coincide. This is true because we first choose n' individuals uniformly at random. But this implies $\hat{\boldsymbol{\sigma}} = \hat{\boldsymbol{\sigma}}'$, thus whenever a configuration τ explains $\hat{\boldsymbol{\sigma}}'$ we can construct a configuration explaining $\hat{\boldsymbol{\sigma}}$ by setting the infection status of the not in $\boldsymbol{\sigma}'$ contained individuals to 0. Finally, the assertion follows from the fact that the choice $n' \sim n^{\theta/\theta'}$ allows to calculate

$$m_{\text{non-ada}}(n',\theta') = \frac{\theta'}{(1-\theta')\ln^2 2}k'\ln\frac{n'}{k'} = \frac{\theta'}{\ln^2 2}k'\ln n' \sim \frac{\theta'\theta}{\theta'\ln^2 2}k\ln n = \frac{\theta}{(1-\theta)\ln^2 2}k\ln\frac{n}{k} = m_{\text{non-ada}}(n,\theta).$$

□

While Theorems 2.1.3 – 2.1.4 answer all questions with respect to information-theoretical phase transitions in non-adaptive group testing, we expect to require significantly more tests if we restrict the design choices. More precisely, as the random regular model has tests of size $\Theta(n/k)$ and each individual takes part in $\Theta(\ln n)$ tests, the number of tests required for inference might increase in sparsity constrained settings. As explained earlier, the work of Gandikota et al. [86] provided some information-theoretical converse statements, i.e. for the maximum test-degree $\Gamma = O\left((n/k)^{\beta}\right)$ and the maximum individual degree $\Delta = o(\ln n)$ they provide information-theoretical converse statements at $m_{\inf,\mathrm{G}}(\Gamma)$ and $m_{\inf,\mathrm{G}}(\Delta)$ respectively given in Equations (1.2.10) – (1.2.11) as

$$m_{\inf,\mathrm{G}}(\Gamma) = \frac{1}{1-\beta}\frac{n}{\Gamma} \qquad \text{and} \qquad m_{\inf,\mathrm{G}}(\Delta) = \Delta k\left(\frac{n}{k}\right)^{1/\Delta}.$$

We need to stress that the converse result with respect to $\Delta-$divisible group testing is of a weak nature, thus they prove that any testing scheme with success probability at least $1-\varepsilon$ requires at least $\Delta k\left(\frac{n}{k}\right)^{\frac{1-5\varepsilon}{\Delta}}$ tests.

Let us begin with the $\Delta-$divisible case. To this end define

$$m_{\text{non-ada}}(\Delta) = \max\left\{\exp(-1)\Delta k^{1+\frac{(1-\theta)}{\Delta\theta}}, \Delta k^{1+\frac{1}{\Delta}}\right\}. \tag{2.1.2}$$

Then we find that no testing-scheme with individual degree at most Δ can be used to infer $\boldsymbol{\sigma}$ from the test-results below $m_{\text{non-ada}}(\Delta)$. We tacitly suppose that

$$\theta/(1-\theta) < \Delta$$

as otherwise $m_{\text{non-ada}}(\Delta)$ would exceed n. Observe that we reduced the exponential dependency on the number of tests provided by [86] to a constant factor of $\exp(-1)$.

Theorem 2.1.5 (Theorem 3.1 and Theorem 3.2 of [88])**.** *Let $\Delta = \ln^{1-\delta} n$ for some $\delta \in (0,1]$ and suppose $k = n^{\theta}$ with $\theta \in (0,1)$. Let furthermore $\mathcal{G}$ be an arbitrary non-adaptive pooling scheme in which each individual gets tested at most Δ times and let $\varepsilon > 0$. Then the following holds.*

- *If $m \leq (1-\varepsilon)\Delta k^{1+1/\Delta}$, any non-adaptive pooling scheme with any algorithm (efficient or not) fails at inferring $\boldsymbol{\sigma}$ from $(\mathcal{G}, \hat{\boldsymbol{\sigma}})$ with probability at least $\max\left\{\Omega(\varepsilon^2), 1 - O\left((1-\varepsilon/2)^{\Delta}\right)\right\}$.*
- *If $m \leq (1-\varepsilon)\exp(-1)\,\Delta k^{1+(1-\theta)/(\theta\Delta)}$, any pooling scheme with any algorithm (efficient or not) fails at inferring $\boldsymbol{\sigma}$ from $(\mathcal{G}, \hat{\boldsymbol{\sigma}})$ with high probability.*

We remark that the second part of Theorem 2.1.5 does not only hold for non-adaptive pooling schemes but for adaptive ones as well. Therefore, we will prove this part later. Let us, at this point, introduce a technique which we will be using for proving the first part of the theorem. We make extensive use of the so-called 2−round exposure technique [103] which will help us to obtain further results later on.

Without going too much into detail, the technique reads as follows. Suppose we want to find a subgraph in $\mathcal{G}(n,p)$. It might be helpful to avoid stochastic dependencies by obtaining one part of the subgraph in $\mathcal{G}(n,p_1)$ (with $p_1 < p$) and afterwards expose the missing edges of $\mathcal{G}(n,p-p_1)$. It is known, that if a subgraph is found in $\mathcal{G}(n,p-p_1)\cup\mathcal{G}(n,p_1)$ with high probability, it is contained in $\mathcal{G}(n,p)$ as well with high probability. We will use this technique in order to expose the infected individuals in two rounds, more precisely, we will find a set of infected individuals with certain properties of size roughly αk and analyse their neighbourhood. Afterwards, we will infect each individual in this neighbourhood with probability $\sim (1-2\alpha)k/n$ independently which will yield totally disguised infected individuals. We will start with a proof sketch of Theorem 2.1.5.

Proof sketch of (the first part of) Theorem 2.1.5. In a first step we argue again that we may employ the i.i.d. model with $p \sim \frac{k+\sqrt{k}\ln n}{n}$. Furthermore, we prove that it suffices to find infected totally disguised individuals in the problem at hand as there will be at least as many uninfected totally disguised individuals with high probability. Before starting the actual proof, we modify the graph such that it does not

contain tests with less than constantly many individuals for a suitable constant and call the resulting pooling graph $\mathcal{G} = (V \cup F, E)$ such that $|V| = n$ and $|F| = (1-\varepsilon)\, m_{\text{non-ada}}(\Delta)$.

A key insight, again provided by the FKG-inequality, is that the probability for an individual x to belong to $V_+(\mathcal{G})$ is at least as high as in a model in which all tests are disjoint, thus

$$\mathbb{P}\left(x \in V_+(\mathcal{G})\right) \geq \prod_{a \in \partial x} \left(1 - \left(1-p\right)^{\Gamma_a - 1}\right)$$

if Γ_a denotes the degree of test a.

Now we employ the 2−round exposure technique and infect every individual with probability αp ending up with a set of infected individuals K_1 of size at least $\alpha k/2$. We observe that the expected number of totally disguised individuals can only increase if those individuals share tests. Therefore, we built an auxiliary model $\mathcal{G}'$ as follows. We include every element of K_1, and for each such individual we include as many disjoint tests of size corresponding to the tests it belongs to in G. Now we mark each individual added with those tests with probability $q = (1-2\alpha)\, p$ as infected.

Let

$$\boldsymbol{X}_u = \prod_{a \in \partial u} \left(1 - \left(1-q\right)^{\Gamma_a - 1}\right)$$

denote the probability that u is totally disguised in this auxiliary model. Jensen's inequality and the inequality of arithmetic and geometric means allow to calculate

$$\mathbb{E}\left[\boldsymbol{X}_u\right] \geq (1-\varepsilon/2)^{-\Delta}\, k^{-1}.$$

Therefore, with $\boldsymbol{X} = \sum_{u \in K_1} \boldsymbol{X}_u$ we have $\mathbb{E}[\boldsymbol{X}] \geq \alpha\,(1-\varepsilon/2)^{-\Delta}/2$ and as $\boldsymbol{X}_u$ and $\boldsymbol{X}_v$ are independent by construction a generalised Chernoff bound yields

$$\mathbb{P}\left(\boldsymbol{X} < \alpha/4\,(1-\varepsilon/2)^{-\Delta}\right) < \exp\left(-\Theta\left(\alpha\,(1-\varepsilon/2)^{-\Delta}\right)\right).$$

Finally, we observe that $\sum_{x \in K_1} \mathbb{P}\left(x \in V_{1+}(\mathcal{G})\right) \geq \boldsymbol{X}$ and Markov's inequality suffices to show that we observe at least one totally disguised infected individual with probability at least $\boldsymbol{X}/(1+\boldsymbol{X})$. Thus with the bound on $\boldsymbol{X}$ and a suitable choice of $\alpha = \alpha(\varepsilon)$ we obtain the result. □

The suspicious reader might ask why the statement (and the end of the proof sketch) require that much detailed calculation and explicit statements of probabilities. This is due to the fact that Δ might be a constant. Indeed, if Δ was diverging, the Chernoff-like bound on $\boldsymbol{X}$ would clearly suffice as $(1-\varepsilon/2)^{-\Delta} \to \infty$ but if Δ is a constant, the calculations become much harder. This is even more challenging under the $\Gamma = \Theta(1)$ restriction as rounding errors need to be taken into account.

In the Γ−sparse case, we strengthen the converse statement at $m_{\text{inf,G}}(\Gamma)$ for the special case of $\Gamma = \Theta(1)$ being a constant independent of the number of individuals. We define

$$m_{\text{non-ada}}(\Gamma) = \max\left\{\left(1 + \left\lfloor \frac{\theta}{1-\theta} \right\rfloor\right) \frac{n}{\Gamma}, 2\frac{n}{\Gamma+1}\right\}. \tag{2.1.3}$$

Again, we tacitly suppose that $\Gamma > \left(1 + \left\lfloor \frac{\theta}{1-\theta} \right\rfloor\right)$ as otherwise individual testing would be superior. Then we find that no non-adaptive pooling scheme can infer the ground-truth using less than $m_{\text{non-ada}}(\Gamma)$ tests.

Theorem 2.1.6 (Theorem 4.1 of [88])**.** *Let $\mathcal{G}$ be any non-adaptive pooling scheme with tests of size at most $\Gamma = \Theta(1)$. Suppose $\mathcal{G}$ contains at most $m = (1-\varepsilon)\, m_{\text{non-ada}}(\Gamma)$ tests for some $\varepsilon > 0$. Then any inference algorithm fails at recovering $\boldsymbol{\sigma}$ from $(\mathcal{G}, \hat{\boldsymbol{\sigma}})$ with high probability if $\theta/(1-\theta)$ is no integer and with probability $\Omega(1)$ if $\theta/(1-\theta)$ is an integer.*

Interestingly, for very few density levels θ, the phase transition could only be proven to be coarse rather than strict. This arises from technical reasons as, for instance, counting the number of nodes with degree at most $\theta/(1-\theta)$ which is tight in the integer-case.

Proof sketch of Theorem 2.1.6. We start by defining

$$d^+ = 1 + \left\lfloor \frac{\theta}{1-\theta} \right\rfloor \qquad \text{and} \qquad d^- = \left\lfloor \frac{\theta}{1-\theta} \right\rfloor$$

and need to distinguish between low prevalence and high prevalence.

Let us begin to sketch the case $1/2 \leq \theta < 1$. Again, we employ the i.i.d. model with $p = \frac{k-\sqrt{k}\ln n}{n}$. A first observation is that whenever $\mathscr{G}'$ is a pooling scheme with maximum test degree Γ on

$$m = (1-2\varepsilon)d^+ n/\Gamma = (1-2\varepsilon)m_{\text{non-ada}}(\Gamma)$$

tests achieving inference, there is a pooling scheme $\mathscr{G}$ on the same set of individuals with

$$m' = (1-\varepsilon)d^+ n/\Gamma$$

tests which achieves inference as well and which has the additional property that each individual gets tested at most $\Delta = \Theta(1)$ times. This follows immediately from a plain counting argument as there can only be n/C individuals of degree $\geq C'$ as m is linear in n. Thus, testing each individual of degree $> C'$ individually causes n/C additional tests. Clearly, inference in $\mathscr{G}'$ implies inference in $\mathscr{G}$ and we get the claim by choosing $C = \Gamma/(\varepsilon d^+)$. By a similar token, one can achieve that we might suppose that the minimum test-degree is at least 2 (if $d^+ \geq 3$).

A second observation is that $m = (1-\varepsilon)d^+ n/\Gamma$ implies that there are at least αn individuals of degree at most d^- by the handshaking lemma. Now we claim that there are also βn individuals of degree at most d^- and distance at least 6 in $\mathscr{G}$ which follows from simple counting as all variables have bounded degree. Let B denote a set of individuals satisfying those two properties.

Then clearly, the property of being disguised is independent for $x, x' \in B$ and the probability turns out to be, as before, at least

$$\mathbb{P}(x \in V_+) \geq \prod_{a \in \partial x} \left(1 - (1-p)^{\Gamma_a - 1}\right) = \Theta\left(p^{d^-}\right).$$

By the independence we directly find that a binomial random variable $\mathbf{Bin}\left(\beta n, \Theta\left(p^{d^-+1}\right)\right)$ dominates $|V_{1+}(\mathscr{G})|$. Therefore, the expected number of totally disguised infected individuals is $\Theta\left(n^{\theta-(1-\theta)d^-}\right)$.

The failure with high probability if $\theta/(1-\theta)$ is no integer and with positive probability if it is follows directly from the Chernoff bound.

Hence, we are only left to prove the result for $0 < \theta < 1/2$. The first fact which follows from double counting the edges of the pooling graph $\mathscr{G} = (V, E)$ is that there are at least εn individuals of degree one. A second observation is that there cannot be many tests containing two (or more) individuals of degree one. Assume there were $n/\sqrt{k}$ such tests, then the Chernoff bound guarantees that we have $\sqrt{k}\ln n$ tests containing two individuals of degree one out of which one is infected. Thus any inference algorithm has to guess the infection status of those individuals and therefore the chance of correct inference is $2^{-\omega(1)} = o(1)$. But those two simple observations already suffice as they imply directly that

$$(2-\varepsilon)n/\Gamma = m \geq \varepsilon n - o(n)$$

needs to hold. Solving for ε implies

$$m \geq 2\frac{n}{\Gamma+1} - o(n).$$

Therefore, the theorem follows from combining the arguments for small and large θ. □

Let us briefly mention that those sparsity constrained results do explicitly not converge to $m_{\text{rand-reg}}$ for $\Delta \to \ln n$ and $\Gamma \to n/k$ what might at the first glance be a surprise. But actually, it is not very surprising. First, the bounds are given only with respect to the first order and lower order terms might get relevant in the limit and second, the proofs extensively make use of the sparsity, thus many partial results do not carry over to the unconstrained setting.

We observed the splitting of the information-theoretic bound into two parts (thus a maximum over

two terms) in the unrestricted group testing in $m_{\text{non-ada}}$ as well as in the $\Delta-$sparse case in $m_{\text{non-ada}}(\Delta)$. This is actually not very surprising. The first part comes in both cases from an information-theoretic argument (for instance, the counting bound) which applies for adaptive as well as non-adaptive pooling schemes, as the amount of information provided by the test-results must exceed certain bounds. On the other hand, the non-adaptivity strikes with respect to the different types of individuals. If the set of totally disguised infected individuals is non-empty, non-adaptive pooling schemes are doomed to fail while adaptive pooling schemes might classify those individuals in subsequent stages of tests. This is why the second term in the maximum actually corresponds to the phase transition point from which on there are totally disguised infected individuals in a pooling scheme.

After having discussed the information-theoretic side, let us now present algorithmic results in the aforementioned models.

2.1.1.3. Algorithms

We will first discuss two major results obtained in [41] and [46] with respect to unrestricted non-adaptive group testing algorithms. The first result shows that DD and SCOMP fail at exactly the same threshold in the random regular model, thus we obtain a fitting converse statement to the achievability result of DD obtained by [108] and settle a strict phase transition. Furthermore, we reject the conjecture of Aldridge, Baldassini and Johnson [8] that SCOMP could outperform DD asymptotically. Recall

$$m_{\text{DD}} = \max\left\{\frac{\theta}{(1-\theta)\ln^2 2}, \frac{1}{\ln^2 2}\right\} k \ln \frac{n}{k}$$

from (1.2.9). Then our result reads as follows.

Theorem 2.1.7 (Theorem 1.2 of [41])**.** *Let $\varepsilon > 0$ be a constant. Then, if $\mathcal{G}$ is an instance of the random regular model with $m \leq (1-\varepsilon) m_{DD}$ tests, both SCOMP and DD fail at inferring $\boldsymbol{\sigma}$ from $(\mathcal{G}, \hat{\boldsymbol{\sigma}})$ with high probability.*

Therefore, Theorem 2.1.7 does not only establish a phase transition but it also shows that there is a gap of a factor of $\ln 2$ between the information-theoretic achievability bound and the best known achieving efficient algorithm on the random regular model. Let us sketch how to prove this converse statement.

Proof sketch of Theorem 2.1.7. In a first step we need to show that DD fails on the random regular model with $m = (1-\varepsilon) m_{\text{DD}}$ tests with high probability. To this end let $m = ck \ln \frac{n}{k}$ and recall that any individual chooses $\Delta = d \ln \frac{n}{k}$ tests uniformly at random. In this contribution we chose the tests with replacement, thus an individual is part of one test twice from time to time. Thus, the average test-degree $\bar{\Gamma}$ turns out to be $\frac{dn}{ck}$. Analogously as before, we suppose for gathering the main idea of the proof that there were no stochastic dependencies (which is clearly false). In this simplified version, the probability for an individual to be disguised is roughly given by

$$\left(1 - \left(1 - \frac{k}{n}\right)^{\bar{\Gamma}}\right)^{\Delta} \sim \left(1 - \exp\left(-\frac{d}{c}\right)\right)^{\Delta}.$$

Therefore, the expected number of totally disguised uninfected individuals is

$$\mathbb{E}\left[|V_{0+}|\right] \sim (n-k)\left(1 - \exp\left(-\frac{d}{c}\right)\right)^{\Delta}.$$

Actually, this argument as well as the fact that the value is concentrated, can be proven rigorously by describing the number of infected and uninfected individuals per test as a family of independent binomials conditioned on a (not too unlikely) event. Furthermore, it is possible to prove that DD achieves its best results for the choice $d = c \ln 2$, therefore $|V_{0+}| \sim n 2^{-\Delta} \sim n^{1-(1-\theta)c\ln^2 2}$.

However, it is not clear at all how to find a rigorous argument which enables us to calculate the size of V_{1--}. Let us change our point of view and observe that DD identifies an infected individual in a positive

test a if and only if, besides x, there are no elements of V_1 nor V_{0+} in a. Therefore, let

$$\boldsymbol{W} = \text{number of positive tests containing exactly one element of } V_1 \cup V_{0+}.$$

We are left with calculating the expectation of $\boldsymbol{W}$ which turns out to be a mildly delicate calculation involving the description of the number of uninfected disguised, uninfected non-disguised and infected individuals per test as a family of independent conditioned multinomial-variables. Nevertheless, it can be shown that

$$\mathbb{E}[\boldsymbol{W}] \sim \frac{k\Delta}{2}\exp\left(-\ln 2 n^{(1-\theta)(1-c\ln^2 2)}\right)$$

which tends to zero below the DD threshold. Thus, by Markov's inequality we already established the converse statement. Let us furthermore show how to calculate the size of V_{1--} from that point. It is possible to show that $\boldsymbol{W}$ is actually tightly concentrated around its mean, thus

$$\boldsymbol{W} \sim \frac{k\Delta}{2}\exp\left(-\ln 2 n^{(1-\theta)(1-c\ln^2 2)}\right).$$

Now, we can rigorously calculate the probability that a given infected individual x does not belong to V_{1--}, as

$$\mathbb{P}(x \notin V_{1--} \mid x \in V_1) \sim \binom{k\Delta - \boldsymbol{W}}{\Delta}\binom{k\Delta}{\Delta}^{-1}.$$

Indeed, such an infected individual would need to choose all its Δ edges out of the $k\Delta - \boldsymbol{W}$ edges belonging to tests containing either a totally disguised individual or a second infected one. By plugging in $m = (1-\varepsilon)m_{\mathtt{DD}}$ Markov's inequality shows that DD fails with high probability. Nevertheless, we are still left to prove that the greedy extension SCOMP fails as well.

To this end, let us observe that there are $\Omega(|V_{0+}|) \gg k$ totally disguised uninfected individuals which are contained in exactly Δ tests. Indeed, with high probability, each individual is contained in at least $\Delta - \ell$ ($\ell = O(1)$) different tests which can be easily verified. Furthermore, the probability of being in exactly $\Delta - \ell'$ tests for $0 \le \ell' \le \ell$ is $\Omega(1)$, thus a positive fraction of all totally disguised uninfected individuals is in Δ different tests. Therefore, already the first individual taken by SCOMP is only infected with probability $\frac{k}{\Omega(|V_{0+}|)+k} = o(1)$. Thus, SCOMP fails with high probability. □

While the previous result is a converse statement about the performance of specific algorithms, we could not answer the question whether there is an efficient algorithm achieving at $m_{\text{non-ada}}$ on the random regular model. Fortunately, we could prove that this gap between information-theoretic and algorithmic achievability is not due to the group testing problem itself. More precisely, we can introduce a different (random) pooling scheme coming with an efficient decoding algorithm called *Spatial Inference Vertex Cover*-algorithm (SPIV) which succeeds with $m = (1+\varepsilon)m_{\text{non-ada}}$ many tests at inference of $\boldsymbol{\sigma}$ with high probability.

Theorem 2.1.8 (Theorem 1.2 of [46])**.** *Let $\varepsilon > 0$ be a constant. Then there is a pooling scheme called* spatially coupled random regular model $\mathcal{G}_{sc}$ *coming with an efficient inference algorithm* SPIV *which succeeds at inferring $\boldsymbol{\sigma}$ from $(\mathcal{G}_{sc}, \hat{\boldsymbol{\sigma}})$ with $m = (1+\varepsilon)m_{\text{non-ada}}$ many tests.*

In the specific case of Theorem 2.1.8 we will not give an explicit proof sketch as the explanation of the pooling scheme and the inference algorithm allow presenting the proof idea on the fly whereby we follow the contribution [46].

We start by defining the pooling scheme. The main idea of the *spatial coupling* has its origins in coding theory [15, 123, 124]. It was applied, for instance, to LDPC-codes. The key idea is to add some kind of geometry to the random regular graph such that, for a specific individual, the neighbourhood looks fairly similar in both models (as the random regular model is known to be information-theoretic optimal). But this geometry constraint allows a trick at inference. Let us partition the set of individuals into $\ell \sim \sqrt{\ln n}$ *compartments* $V[1], \dots, V[\ell]$ of size n/ℓ and the m tests into compartments $F[1], \dots, F[\ell]$ as well. Let $s \sim \ln\ln n$ be the size of the *sliding window*. Then we furthermore add $10ks\ln n/\ell$ additional tests into a larger compartment $F[0]$ whose purpose will become clear in due course.

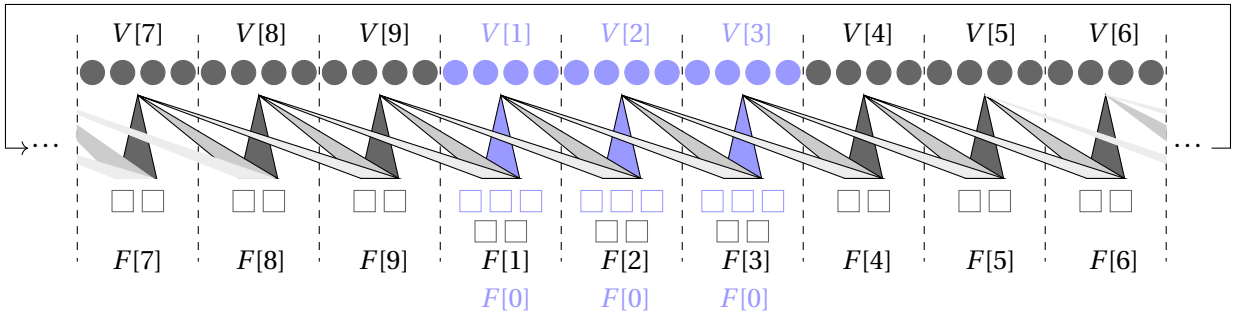

Figure 2.2.: The spatially coupled test design with $n = 36, \ell = 9, s = 3$. The graphic is modified after [41]. The individuals in the seed (blue) join additional tests from $F[0]$.

As in the random regular graph, we let $\Delta = c\ln 2 \ln \frac{n}{k}$ denote the individual degree if $m = ck\ln\frac{n}{k}$ and let the individuals choose their tests as follows.

- For $i = 1,\dots,\ell$ and $j = 1,\dots,s$ every individual $x \in V[i]$ chooses independently from all other randomness Δ/s tests from $F[i+j-1]$ without replacement.
- Those individuals from $V[1] \cup \dots \cup V[s]$ independently choose $10\ln(2)\ln n$ additional tests from $F[0]$ uniformly at random without replacement.

All indices of compartments need to be read such that $V[\ell+i] = V[i]$ and $F[i+\ell] = F[i]$ for $i = 1\dots\ell$. Thus the pooling graph has a ring structure such that, locally seen, all edges from an individual are going to its *right* whereas all incoming edges in a test come from the *left*. A visualisation can be found in Figure 2.2.

Why should such a graph perform better? Suppose we already classified all individuals in compartments $V[1],\dots,V[h]$ successfully. If we now are about to infer the infection status of those individuals in $V[h+1]$ we can use a lot of information. Indeed, in the tests in $F[h+1]$ there are just the individuals of $V[h+1]$ unclassified whilst in the tests in $F[h+1+j]$ at least proportion $(s-j)/s$ is already known. Therefore, if an *early* test emphasises that an individual was infected or uninfected, this information might be more confidential than an information gained in a test far to the right. This is exactly the idea behind the `SPIV` algorithm.

The main challenge of inference is to distinguish the k infected individuals from those $\omega(k)$ totally disguised uninfected individuals. Therefore, we introduce a random variable $\boldsymbol{W}_{x,j}$ whose distribution differs for real infected individuals x and uninfected but disguised individuals. By choice of Δ, the number of infected individuals per test a is (approximately) $\mathbf{Po}(\ln 2)$ distributed. But it turns out that this number's distribution changes when conditioning on specific events.

We define for an individual $x \in V[i+1]$ and a compartment $F[i+j]$ for $j = 1\dots s$

$$\begin{aligned}\boldsymbol{W}_{x,j} = {}& \text{number of tests in compartment } F[i+j] \text{ containing } x \\ & \text{and no infected individual from the preceding compartments.}\end{aligned}$$

Because the number of infected individuals in a test from a specific compartment is $\mathbf{Po}(\ln 2/s)$ distributed, the probability that a specific test contains no infected individual from any of the compartments $V[i+j-s+1],\dots,V[i]$ turns out to be $2^{-(s-j)/s}$, therefore

$$\mathbb{E}\left[\boldsymbol{W}_{x,j} \mid x \in V[i+1] \cap V_1\right] \sim \frac{\Delta}{s} 2^{j/s-1}.$$

Equivalently, we can calculate

$$\mathbb{E}\left[\boldsymbol{W}_{x,j} \mid x \in V[i+1] \cap V_{0+}\right] \sim \frac{\Delta}{s}\left(2^{j/s}-1\right)$$

as the sole difference is that a specific test a needs to contain at least one infected individual which is not x. Therefore, the number of infected individuals in a, conditioned on containing an element of V_{0+}, turns out to be a conditioned Poisson distributed variable $\mathbf{Po}_{\geq 1}(\ln 2)$. We clearly find

$$\mathbb{E}\left[\boldsymbol{W}_{x,j} \mid x \in V[i+1] \cap V_{0+}\right] < \mathbb{E}\left[\boldsymbol{W}_{x,j} \mid x \in V[i+1] \cap V_1\right].$$

Unfortunately, it turns out that trying to distinguish the infection status solely on the sum of those random variables does not suffice as there will be too much fluctuation [46, Section 4.3]. Therefore, we should try to incorporate the intuition that tests in a compartment close to the individual might contain more valuable information. We define

$$\boldsymbol{W}^*_x = \sum_{i=1}^{s-1} w_j \boldsymbol{W}_{x,j} \qquad \text{where} \qquad w_j \sim -\ln\left(1-2^{-j/s-1}\right). \tag{2.1.4}$$

Thus, W^*_x weighs information from closer tests higher. The optimal value of the weighs w_j was obtained by a delicate large deviation analysis using a Lagrange optimisation [46, Sections 4.7 and 4.8].

If we call the individuals in $V[1] \cup \ldots \cup V[s] = V_{seed}$ the *seed* (those individuals which are connected to tests in $F[0]$) the idea is now to classify the individuals subsequently starting at the seed. As we clearly cannot compute $\boldsymbol{W}^*_x$ as this would require knowledge about $\boldsymbol{\sigma}$, we have to rely on the current estimate of the ground-truth, thus let

$$W_{x,j}(\tau) = \left|\left\{a \in \partial x \cap F[i+j-1] : \max_{y \in \partial a \cap (V[1] \cup \ldots \cup V[i])} \tau_y = 0\right\}\right|$$

and analogously

$$W^*_x(\tau) = \sum_{i=1}^{s-1} w_j W_{x,j}(\tau).$$

We apply `DD` to the seed individuals as the graph on the seed individuals and the tests of $F[0]$ is an instance of the random regular model which is dense enough to allow classification by `DD`. We let τ be the current estimate of $\boldsymbol{\sigma}$ which we initialise to the all zero vector outside of the seed. Then we proceed with the individuals in the next compartment and

- declare an individual as uninfected if it is in at least one negative test,
- declare an individual x as uninfected, if $W^*_x(\tau)$ is smaller than the expected value for infected individuals under $\boldsymbol{\sigma}$,
- declare an individual as infected otherwise.

We iterate with this procedure through the graph and stop when all individuals are classified. We call the estimate of the ground-truth produced by this algorithm τ. Unfortunately, it turns out that τ does probably not coincide with $\boldsymbol{\sigma}$. But fortunately, if $|F[1] \cup \ldots F[\ell]| = (1+\varepsilon)m_{\inf}$, we have at least [46, Proposition 4.6]

$$|\{x : \tau_x \neq \boldsymbol{\sigma}_x\}| = kn^{-\Omega(1)}.$$

Thus, using only as many tests as given by the universal counting bound, we achieve already *partial recovery* of $\boldsymbol{\sigma}$ with high probability. To this end observe that $|F[0]| = o(m)$ is of lower order.

Corollary 2.1.9. *The `SPIV` algorithm on $\mathcal{G}_{sc}$ with $(1+\varepsilon)m_{\inf}$ tests succeeds at $(1-o(1))-$partial recovery of $\boldsymbol{\sigma}$ from $(\mathcal{G}_{sc}, \hat{\boldsymbol{\sigma}})$ with high probability.*

But of course we want to achieve *exact* recovery. It turns out that a rigidity argument helps in establishing an exact recovery statement. If $m = (1+\varepsilon)m_{\text{non-ada}}$, each infected individual is contained in at least $\Theta(\ln n)$ tests in which no second infected individual appears. Let $S_x(\tau)$ denote the number of (positive) tests in which an individual would be the only infected individual under τ if it was infected. Then a combinatorial clean-up step could read as follows.

For $\ln n$ steps repeat thresholding $S_x(\tau)$ with respect to the current estimate τ in order to classify individual x as infected or uninfected. By the expansion properties of the random graph it is possible to prove that this reduces the number of misclassified individuals by a factor of at least 3 each step. Thus, the final estimate τ coincides with $\boldsymbol{\sigma}$ w.h.p.. This is again an example of the splitting of the phase transition point into two parts. The non-adaptive tail of the phase transition point corresponds to the combinatorial observation of having some kind of local rigidity, similarly as we already saw in the proof sketch of Theorem 2.1.3.

We can now state the SPIV algorithm completely and formally.

Input: Spatially coupled pooling scheme $\mathcal{G}_{sc} = (V \cup F, E)$, test-results $\hat{\boldsymbol{\sigma}} \in \{0,1\}^m$
Output: Estimate $\tilde{\sigma}$ of $\boldsymbol{\sigma}$

Infer the infection status for all $x \in V_{seed}$ by DD and obtain $\tilde{\sigma}_{V_{seed}}$.
Initialise $\tilde{\sigma}_x = 0$ for all $x \notin V_{seed}$.
for $i = s, \dots, \ell - 1$ **do**
  for $x \in V[i+1]$ **do**
    if *x is in at least one negative test* **then**
      $\tilde{\sigma}_x = 0$ // set infection status to uninfected
    else if $W_x^*(\tilde{\sigma}) < (1-\zeta)\sum_{j=1}^{s-1} \frac{\Delta}{s} w_j 2^{j/s-1}$ **then**
      $\tilde{\sigma}_x = 0$ // set infection status to uninfected
    else
      $\tilde{\sigma}_x = 1$ // classify as infected
Let $\tilde{\sigma}^{(1)} = \tilde{\sigma}$.
for $i = 1, \dots, \ln n$ **do**
  For all $x \in V[s+1] \cup \dots \cup V[\ell]$ calculate

$$S_x\left(\tilde{\sigma}^{(i)}\right) = \sum_{a \in \partial x : \hat{\boldsymbol{\sigma}}_a = 1} \mathbf{1}\left\{\forall y \in \partial a \setminus \{x\} : \tilde{\sigma}_y^{(i)} = 0\right\}$$

  Let $\tilde{\sigma}_x^{(i+1)} = \begin{cases} \tilde{\sigma}_x^{(i)} & \text{if } x \in V[1] \cup \dots \cup V[s], \\ \mathbf{1}\left\{S_x\left(\tilde{\sigma}^{(i)}\right) > \ln^{1/4} n\right\} & \text{otherwise} \end{cases}$
return $\tilde{\sigma}^{(\lceil \ln n \rceil)}$

Algorithm 4: The SPIV algorithm by Coja-Oghlan et al. [46] using a spatially coupled pooling scheme.

We should emphasise that the decoding algorithms in prior applications on spatially coupled inference graphs like coding or compressed sensing were approximate message passing or BP like algorithms [106, 123, 124]. On the first glance, those algorithms are clearly much more sophisticated than the combinatorial and easy to digest SPIV algorithm. But based on the discussion about WP and DD, we should not be too surprised that actually the decisions based on (a normalised version of) W_x^* correspond to the estimate after one round of BP.

Therefore, we have proven that the group testing problem is very special with respect to its solvability. Either, it is completely impossible or it is easy (there is an efficient algorithm for inference) but there is no hard phase. Such phenomena are known as *impossible-easy* phase transitions. Furthermore, we discussed that there is indeed a gap between adaptive algorithms and non-adaptive algorithms, thus multiple stages of testing can decrease the number of tests required in total, at least if many individuals are supposed to be infected.

As we will discuss next, we observe similar phenomena even if we restrict the maximum capacity of a test or the number of tests an individual can be part of.

Let us first present the results we obtained with respect to the $\Delta-$divisible restricted group testing problem. We recall that we already provided a result showing that each non-adaptive group testing scheme which tests each individual at most Δ times fails with high probability if it contains less than $m_{\text{non-ada}}(\Delta)$ tests where

$$m_{\text{non-ada}}(\Delta) = \max\left\{\exp(-1)\Delta k^{1+\frac{(1-\theta)}{\Delta\theta}}, \Delta k^{1+\frac{1}{\Delta}}\right\}.$$

We will show that the DD algorithm on the random regular model solves the group testing problem almost at this bound, thus let $m_{\mathtt{DD}}(\Delta)$ denote the phase transition point of DD in the random (almost) regular model, then we have

$$m_{\mathtt{DD}}(\Delta) = \max\left\{\Delta k^{1+\frac{(1-\theta)}{\Delta\theta}}, \Delta k^{1+\frac{1}{\Delta}}\right\} \tag{2.1.5}$$

which is a factor of e away from the converse statement for $\theta < 1/2$ and matches it for $\theta \geq 1/2$. We define the random almost regular model in a way that each individual chooses Δ tests uniformly at random with replacement, thus the test-degree sequence is random (but sufficiently concentrated). Then we obtain the following theorem.

Theorem 2.1.10 (Theorem 3.3 [88])**.** *Let $\mathcal{G}_\Delta$ be the random regular model where each individual joins $\Delta = O(\ln^{1-\delta} n)$ $(0 < \delta \leq 1)$ tests on m tests. Then , if $m \geq (1+\varepsilon)m_{DD}(\Delta)$, DD succeeds at inference of $\boldsymbol{\sigma}$ from $(\mathcal{G}_\Delta, \hat{\boldsymbol{\sigma}})$ with probability $\Omega(1)$ if $\Delta = O(1)$ and with probability $(1-o(1))$ if $\Delta = \omega(1)$.*

Thus, Theorem 2.1.10 shows in combination with the information-theoretic converse at $m_{\text{non-ada}}(\Delta)$ that the random-regular model is information-theoretic optimal for large θ and that DD performs optimally in this regime.

Proof sketch of Theorem 2.1.10. Analogously as in the DD analysis in the unrestricted problem (Theorem 2.1.7), we bound the expected number of positive tests containing exactly one infected and no uninfected totally disguised individual $\mathbb{E}[\boldsymbol{W}]$ with the only difference that some probabilities have to be calculated slightly more carefully if Δ does not diverge. Again we require to describe some local properties of tests by a family of independent conditioned multinomial random variables. Of course, the resulting formulas differ such that

$$\mathbb{E}[\boldsymbol{W}] \sim \Delta k \cdot \left(1 - (1+\varepsilon)^{-1} k^{-1/\Delta}\right).$$

Nevertheless, we will omit any details here.

But in this case, as we want to proof that DD actually succeeds, we require a stronger result as Markov's inequality applied on the expectation of $\boldsymbol{W}$ does not suffice. Luckily, it turns out that $\boldsymbol{W}$ is fairly concentrated around its expectation. Therefore, with high enough probability, we find

$$\boldsymbol{W} \sim \Delta k \cdot \left(1 - (1+\varepsilon)^{-1} k^{-1/\Delta}\right).$$

Now, we can rigorously calculate the probability that a given infected individual x is classified falsely by DD, thus does not belong to V_{1--}, as

$$\mathbb{P}(x \notin V_{1--} \mid x \in V_1) \sim \binom{k\Delta - \boldsymbol{W}}{\Delta}\binom{k\Delta}{\Delta}^{-1} \sim \left((1+\varepsilon)^{-1} k^{-1/\Delta}\right)^{\Delta}.$$

Indeed, such an infected individual would need to choose all its Δ edges out of the $k\Delta - \boldsymbol{W}$ edges belonging to tests containing either a totally disguised individual or a second infected one. A standard calculation involving a case distinction between small and large θ suffices in order to obtain the expected number of individuals $x \in V_1 \setminus V_{1--}$ and to observe that this number is at most $(1+\varepsilon)^{-\Delta}$ if $m = (1+\varepsilon)m_{\mathtt{DD}}(\Delta)$ tests are carried out and the theorem follows by Markov's inequality. □

Of course, above's theorem does only give an achievability result for DD on a specific design. We can actually prove that DD is failing below $m_{\mathtt{DD}}(\Delta)$ on the random almost regular model. Clearly, we only need to show this for small θ as the assertion for large θ follows from the universal converse bound.

Theorem 2.1.11 (Theorem 3.4 of [88])**.** *Let $0 < \theta < 1/2$ and let $\mathcal{G}_\Delta$ be the random almost regular model on m tests. If $m \leq (1-\varepsilon)\Delta k^{1+(1-\theta)/(\Delta\theta)}$, DD fails at inference of $\boldsymbol{\sigma}$ from $(\mathcal{G}_\Delta, \hat{\boldsymbol{\sigma}})$ with probability $1-o(1)$ of $\Delta = \omega(1)$ and with probability $\Omega(1)$ if $\Delta = \Theta(1)$.*

Proof sketch of Theorem 2.1.11. Similarily as in the achievability proof, we find by analysing the number of positive tests containing exactly one infected and no uninfected totally disguised individual that

$$\mathbb{E}\left[|V_{1--}(\mathcal{G}_\Delta)|\right] \sim k\left(1-\left(1-\exp\left(-(1-\varepsilon)^{-\Delta}(1-1/\Delta)\right)\right)^{\Delta}\right).$$

But then Markov's inequality readily yields that the probability of having at least γk infected individuals outside of $V_{1--}(\mathcal{G})$ is at least

$$1-\frac{1-\left(1-\exp\left(-(1-\varepsilon)^{-\Delta}(1-1/\Delta)\right)\right)^{\Delta}}{1-\gamma}$$

and therefore, the assertion of the theorem follows. Indeed, this probability is $1-o(1)$ for $\Delta\to\infty$ for any $0<\gamma<1$ and if Δ is a constant, a suitable choice of γ makes the probability $\Omega(1)$. □

Thus, we understand DD on the random regular model completely.

We are left to discuss the $\Gamma-$sparse case. This is an especially interesting case as the underlying (almost regular) model has constant degrees on both sides, thus the construction of the graph requires a lot more attention. Ultimately, we will see that DD actually succeeds at the information-theoretic universal converse $m_{\text{non-ada}}(\Gamma)$ on a suitable chosen almost regular graph for all θ outside of a set of Lebesgue-measure zero. Recall that

$$m_{\text{non-ada}}(\Gamma)=\max\left\{\left(1+\left\lfloor\frac{\theta}{1-\theta}\right\rfloor\right)\frac{n}{\Gamma},2\frac{n}{\Gamma+1}\right\}$$

and define

$$m_{\mathtt{DD}}(\Gamma)=\begin{cases}\max\left\{\left(1+\left\lfloor\frac{\theta}{1-\theta}\right\rfloor\right)\frac{n}{\Gamma},2\frac{n}{\Gamma+1}\right\}, & \frac{\theta}{1-\theta}\notin\mathbb{Z}\\ \max\left\{\left(2+\left\lfloor\frac{\theta}{1-\theta}\right\rfloor\right)\frac{n}{\Gamma},2\frac{n}{\Gamma+1}\right\}, & \frac{\theta}{1-\theta}\in\mathbb{Z}.\end{cases}$$

Therefore, $m_{\mathtt{DD}}$ coincides with the universal converse if $\theta/(1-\theta)$ is no integer.

Let us first describe how to obtain the almost regular pooling scheme. If $\theta\geq 1/2$, we create the random pooling scheme by the configuration model as a random regular multi-graph $\mathcal{G}(\Gamma,\Delta)$. Thus, with $\Delta=m\Gamma/n$, each individual node gets Δ clones and each test node gets Γ clones and a perfect matching is chosen uniformly at random. On the other hand, if $\theta<1/2$, it turns out that this model is not optimal. In this case, we select $\gamma\leq\frac{2n}{\Gamma+1}$ individuals $X\subset\{x_1,\dots,x_n\}$ uniformly at random and put them apart. The exact value of γ is chosen such that the remaining individuals can be pooled by $\mathcal{G}(\Gamma-1,2)$. Now we select a uniform matching between the tests $F(\mathcal{G}(\Gamma-1,2))$ and the remaining individuals X. For a brief check of sanity, observe that this pooling is only possible for $m\geq 2\frac{n}{\Gamma+1}$ by comparing degrees. Furthermore observe that in the final graph any test has size $\Gamma-1$ or Γ. We will call this graph model $\mathcal{G}^*(\Gamma)$. Therefore, let us define

$$\mathcal{G}(\Gamma)=\begin{cases}\mathcal{G}(\Gamma,m\Gamma/n), & \theta\geq 1/2\\ \mathcal{G}^*(\Gamma), & \theta<1/2.\end{cases}.$$

With this model at hand, we can state the achievability result.

Theorem 2.1.12 (Theorems 4.10 and 4.18 of [88])**.** *If DD is applied on an instance of $\mathcal{G}(\Gamma)$ with $m\geq m_{DD}(\Delta)$, it succeeds at inference of $\boldsymbol{\sigma}$ from $(\mathcal{G}(\Gamma),\hat{\boldsymbol{\sigma}})$ with high probability.*

Observe that interestingly, the achievability bound is tight, thus we do not even need a multiplicative factor of $(1+\varepsilon)$.

Proof sketch of Theorem 2.1.12. The proof comes in two major steps. First, we give an achievability result of DD on the random regular model $\mathcal{G}(\Gamma,\Delta)$ for any choice of θ. More precisely, let

$$\Delta_{\mathtt{DD}}=\max\left\{2,1+\left\lfloor\frac{\theta}{1-\theta}\right\rfloor\right\}$$

denote the individual degree, then we have the following.

Achievability on the random regular model. DD recovers $\boldsymbol{\sigma}$ from $(\mathcal{G}_\Gamma(\Gamma,\Delta_{\mathtt{DD}}),\hat{\boldsymbol{\sigma}})$ correctly w.h.p..

Observe that by definition of the random regular graph we have $m \geq \Delta_{\mathtt{DD}}\frac{n}{\Gamma}$. It turns out that this part of the proof follows completely analogous ideas as the achievability proof of DD in the $\Delta-$divisible case and will thus be omitted. Of course, a technical challenge is to deal with the (higher) number of multi-edges. Fortunately, the heavy stochastic dependencies caused by the configuration model vanish as we can describe the important local properties of tests by a family of independent conditioned multinomial random variables as before. The sole technical challenge is to deal with the (higher) number of multi-edges.

As we have proven the success of DD with $m = \Delta_{\mathtt{DD}}\frac{n}{\Gamma}$ tests, the part of the theorem for $\theta \geq 1/2$ follows. But it turns out that we can indeed perform better if $\theta < 1/2$. The main idea is the following. We let $\mathcal{G}_\Gamma^{*,r}$ be the subgraph created in the first step of generating $\mathcal{G}(\Gamma)$, thus an instance of $\mathcal{G}(\Gamma-1,2)$ on $n' = n-\gamma$ individuals where $\gamma \leq \frac{2}{\Gamma+1}n$. Without loss of generality we suppose that $\gamma = \frac{2}{\Gamma+1}n$ as otherwise we could fill the instance with dummy-individuals which are known to be uninfected.

We furthermore let $\boldsymbol{\sigma}[\mathcal{G}_\Gamma^{*,r}]$ and $\hat{\boldsymbol{\sigma}}[\mathcal{G}_\Gamma^{*,r}]$ be the infection status vector and test-result vector on this induced subgraph. Observe that for $\theta < 1/2$ we have $\Delta_{\mathtt{DD}} = 2$. Now we obtain the result as follows.

(i) Let $\theta' = \theta'(\theta)$ describe the prevalence on $\boldsymbol{\sigma}[\mathcal{G}_\Gamma^{*,r}]$, then we have $\theta' \sim \theta$ with high probability.

(ii) We already know that $m = 2\frac{n-\gamma}{\Gamma-1} = 2\frac{n}{\Gamma+1}$ tests suffice for DD to infer $\boldsymbol{\sigma}[\mathcal{G}_\Gamma^{*,r}]$ from $\mathcal{G}(\Gamma-1,2)$ and $\hat{\boldsymbol{\sigma}}[\mathcal{G}_\Gamma^{*,r}]$.

(iii) We prove that adding the matching edges in order to generate $\mathcal{G}(\Gamma)$ from $\mathcal{G}(\Gamma-1,2)$ does, with high probability, enable DD to infer $\boldsymbol{\sigma}$ from $\mathcal{G}(\Gamma)$.

It is easy to see that (i) follows from the Chernoff bound as the number of infected individuals in $\boldsymbol{\sigma}[\mathcal{G}_\Gamma^{*,r}]$ is a hypergeometrically distributed random variable $\boldsymbol{k}' \sim H\left(n,k,n'\right)$ and thus concentrated around its mean $\boldsymbol{k}' \sim \frac{\Gamma-1}{\Gamma+1}k$. Furthermore, (ii) is a direct consequence of above's result with respect to achievability on the random regular model. Therefore, we only need to discuss (iii).

The key property of the proof is that for $k = o(\sqrt{n})$ we do not find two infected individuals in any bounded part of the graph. Suppose that an individual x gets connected to a negative test. If x is uninfected itself, the test stays uninfected an DD recovers x (and the other individuals) correctly. If x is infected, each uninfected individual in the test joins a second test which is negative as well with high probability due to the fact that there are no two infected individuals within a finite range within the random graph with high probability.

If on the other hand x connects to a positive test and is uninfected, a similar argument shows that the causing already contained positive individual can be identified by DD through both its tests with high probability. Thus the previously infected individual in $\boldsymbol{\sigma}[\mathcal{G}_\Gamma^{*,r}]$ can be still inferred by DDand x will be declared uninfected.

Finally, if x is infected, it will connect to a negative test with high probability. Indeed, as it is infected with probability $\sim k/n$ and its test is chosen uniformly at random, this test will contain a second infected individual with probability $\sim k^2/n^2$ and a union bound shows that therefore, with high probability, all infected individuals connect to negative tests in the matching process.

Therefore, DD infers $\boldsymbol{\sigma}$ from $\mathcal{G}(\Gamma)$ and $\hat{\boldsymbol{\sigma}}$ w.h.p., if it infers $\boldsymbol{\sigma}[\mathcal{G}_\Gamma^{*,r}]$ from $\mathcal{G}^{*,r}$ and $\hat{\boldsymbol{\sigma}}[\mathcal{G}_\Gamma^{*,r}]$ w.h.p..

Above's bounding of the probability of bad events can be done rigorously and hinges on precise but elementary calculations which make extensive use of the fact that a uniformly at random chosen subset of individuals was put apart before creating the regular part of the graph. □

Next, we will present the results we obtained with respect to adaptive group testing. Afterwards, we summarise the results on non-adaptive group testing as well as adaptive group testing shortly in Section 2.1.3.

2.1.2. Adaptive Group Testing

As in the section about non-adaptive group testing, we split our results into a section about information-theoretic aspects as well as algorithmic aspects respectively.

2.1.2.1. Information-theoretic results

With respect to the information-theoretic phase transitions, we achieved primarily results in the context of sparsity constrained group testing. More precisely, while the contribution of Gandikota et al. [86] only provided non-adaptive converse statements, we give information-theoretic converse results for all adaptive pooling schemes in the $\Delta-$divisible and the $\Gamma-$sparse case. To this end, let $\Delta = \ln^{1-\delta} n$ and $\Gamma = \left(\frac{n}{k}\right)^{\beta}$ with $0 < \delta \leq 1$ and $0 \leq \beta < 1$. Then define

$$m_{\text{inf}}(\Delta) = \exp(-1)\Delta k^{1+\frac{1-\theta}{\Delta\theta}} \qquad \text{and} \qquad m_{\text{inf}}(\Gamma) = \frac{n}{\Gamma}$$

as the information-theoretic threshold for any test-design in which individuals get tested at most Δ times and, respectively, each test has size at most Γ. We let $\boldsymbol{\sigma}$ denote the ground-truth and for an $\ell-$stage testing procedure $\mathcal{G} = (\mathcal{G}_1, \dots, \mathcal{G}_\ell)$ we let $\hat{\boldsymbol{\sigma}} = (\hat{\boldsymbol{\sigma}}^1, \dots, \hat{\boldsymbol{\sigma}}^\ell)$ denote the sequence of test-results.

Theorem 2.1.13 (Theorems 3.1 and 6.2 of [88])**.** *Suppose $\mathcal{G}(\Delta) = (\mathcal{G}_1(\Delta), \dots, \mathcal{G}_\ell(\Delta))$ is an $\ell-$stage testing procedure such that any individual gets tested at most Δ times in total and respectively that $\mathcal{G}(\Gamma) = (\mathcal{G}_1(\Gamma), \dots, \mathcal{G}_\ell(\Gamma))$ is an $\ell-$stage testing procedure such that any test contains at most Γ individuals. Then the following holds.*

- *If $\mathcal{G}(\Delta)$ contains at most $(1-\varepsilon)m_{\text{inf}}(\Delta)$ tests, inference of $\boldsymbol{\sigma}$ from $(\mathcal{G}(\Delta), \hat{\boldsymbol{\sigma}})$ fails with high probability.*
- *If $\mathcal{G}(\Gamma)$ uses at most $(1-\varepsilon)m_{\text{inf}}(\Gamma)$ tests, inference of $\boldsymbol{\sigma}$ from $(\mathcal{G}(\Gamma), \hat{\boldsymbol{\sigma}})$ fails with probability $\Omega(1)$.*

Proof sketch of Theorem 2.1.13. The proof of the first part of the theorem resembles the counting based proof of the universal counting bound in the unrestricted group testing problem.

We first show that any adaptive strategy with m tests succeeds with probability at most

$$\frac{\sum_{i=0}^{\Delta k}\binom{m}{i}}{\binom{n}{k}} \sim \frac{\exp\left(H(\Delta k/m)\right)}{\binom{n}{k}}$$

which is given by a short calculation using the Nishimori property that guarantees that choosing one possible solution is the best an inference algorithm can do. We directly find that there can be at most Δk positive tests in total, as each infected individual can be tested at most Δ times. Therefore, the summation accounts for all possible choices of positive tests. Plugging in $m = (1-\varepsilon)m_{\text{inf}}(\Delta)$ yields the assertion of the theorem.

The second part, namely the $\Gamma-$sparse case, might be actually called a folklore argument. Indeed, $\frac{n}{\Gamma}(1-o(1))$ tests of size at most Γ are clearly required to test $n-o(n)$ of all individuals at least once, which is a necessary requirement for inference. □

Overall it turns out that under both kinds of restrictions the adaptive converse statements are strictly below the non-adaptive converse results and non-adaptive achievability results. In the case of $\Delta-$ divisible group testing, this is only true for $\theta > 1/2$. This alone can clearly not answer the question whether there is an adaptivity-gap in restricted group testing or not. But the next section will show that such a gap really exists.

2.1.2.2. Algorithms

We introduce two new algorithms for the two restricted group testing models. Let us start with the $\Delta-$divisible case which can be solved by Algorithm 5.

Clearly, this algorithm strongly resembles the binary splitting approach of Hwang [98] and Allemann [13]. The only major difference is that we do not split groups into halves in order to guarantee the

Input: n, k, Δ
Output: Estimate $\tilde{K} \subset \{x_1, \dots, x_n\}$ of the infected individuals.
Set $\tilde{n} = \left(\frac{n}{k}\right)^{\frac{\Delta-1}{\Delta}}$.
Set $\tilde{K} = \emptyset$.
Arbitrarily divide the n individuals into $n/\tilde{n}$ groups of size $\tilde{n}$.
Test each of these groups and discard those with a negative result.
Denote the remaining groups by $A_j^{(0)}$.
for $i = 1$ *to* $\Delta - 1$ **do**
  for *each group* $A_j^{(i-1)}$ *from the previous stage* **do**
    Arbitrarily divide all individuals in $A_j^{(i-1)}$ into $\tilde{n}^{1/(\Delta-1)}$ sub-groups of size $\tilde{n}^{1-i/(\Delta-1)}$.
    Test each sub-group and discard any that returns a negative outcome.
    Label the remaining sub-groups as $A_j^{(i)}$.
Add the individuals from all of the remaining singleton groups $A_j^{(\Delta-1)}$ to $\tilde{K}$.

Algorithm 5: Splitting algorithm for the $\Delta-$divisible restricted group testing as posed in [88].

$\Delta-$restriction. As we will see, it does not perform tightly at the information-theoretic converse bound proven before. Let

$$m_{\text{adap-alg}}(\Delta) = \Delta k^{1+\frac{1-\theta}{\theta\Delta}}$$

be an algorithmic threshold. Then we find the following theorem.

Theorem 2.1.14 (Theorem 5.1 of [88])**.** *There is a choice of* $\tilde{n} = \tilde{n}(n, k, \Delta)$ *such that Algorithm 5 succeeds with at most* $(1+\varepsilon)m_{\text{adap-alg}}(\Delta)$ *tests.*

Therefore, the algorithm performs for all θ at exactly one of the bounds of non-adaptive group testing ($m_{\text{non-ada}}(\Delta)$) establishing an adaptivity-gap as in the unrestricted group testing problem.

Proof sketch of Theorem 2.1.14. It is clear from the definition of the algorithm that it recovers $\boldsymbol{\sigma}$ correctly as all infected individuals will be tested individually. The core idea is to optimise the choice of $\tilde{n}$ in such a way that tests do not get too large as otherwise they will be very likely positive. Starting initially with groups of size $\tilde{n}$, the size is continuously decreasing whenever the test-outcome is still positive.

Clearly, in the first stage we conduct $\tilde{n}$ tests and as there are k infected individuals, each subsequent stage of testing can produce at most $k\tilde{n}^{\frac{1}{\Delta-1}}$ additional (smaller) tests, therefore

$$m \leq \frac{n}{\tilde{n}} + (\Delta - 1)\, k\tilde{n}^{\frac{1}{\Delta-1}}.$$

The assertion of the theorem follows with $\tilde{n} = n^{(1-\theta)(\Delta-1)/\Delta}$. □

After having understood a splitting approach towards $\Delta-$divisible constrained instances of group testing, we will subsequently present an algorithm for $\Gamma-$sparse group testing. In this case, we will use a standard binary splitting approach as a sub-routine, thus let us shortly describe how it works in detail. Given a group of individuals, test the whole group and if the test is positive, split the group into two equal parts and tests each part. If the test is negative, we know that all individuals are uninfected and do not need to test further. Iterate this process until all positive tests contain exactly one individual. Now we can state Algorithm 6. We stress that the algorithm basically reduces to applying the first round of the Dorfman-algorithm followed by a binary splitting algorithm for all positive tests.

How many tests does this algorithm require? It turns out that at least its first order complexity coincides (up to rounding) with the information-theoretic converse if $\Gamma = O\left(\frac{n}{k}\right)^{\beta}$ for some $\beta \in [0, 1)$. Thus, let

$$m_{\text{adap-alg}}(\Gamma) = \left\lceil \frac{n}{\Gamma} \right\rceil + \frac{\ln \Gamma}{\ln 2} k$$

denote the algorithmic achievability bound.

Input: n, k, Γ
Output: Estimate $\tilde{K} \subset \{x_1, \dots, x_n\}$ of the infected individuals.
Set $T = \lceil n/\Gamma \rceil$ and $\tilde{K} = \emptyset$.
Choose one partition of all individuals into groups $G_1, \dots, G_T$ of size Γ.
for $i = 1 \dots T$ **do**
    Test group G_i.
    if *the outcome is positive* **then**
        Infer the infection status of all individuals in G_i by binary splitting.
    else
        Declare all individuals in G_i as uninfected.
Add the individuals from all of the remaining singleton groups $A_j^{(\Delta-1)}$ to $\tilde{K}$.

Algorithm 6: Splitting algorithm in the $\Gamma-$sparse restricted group testing. A similar version was introduced in [88].

Theorem 2.1.15. *Algorithm 6 succeeds at inference of $\boldsymbol{\sigma}$ requiring at most $(1+\varepsilon)m_{\mathrm{adap-alg}}(\Gamma)$ tests.*

Proof sketch of Theorem 2.1.15. Again, it is clear that the algorithm succeeds at inference. Thus, we only need to bound the number of tests required. At the first stage, we clearly conduct $T = \lceil \frac{n}{\Gamma} \rceil$ tests. Subsequently, we apply the binary splitting algorithm to groups of at most Γ individuals each. But due to Hwang [98] it is known that binary splitting on a group of Γ individuals out of which k_i are infected requires not more than

$$m_{\mathrm{Hwang}} \sim \frac{k_i \ln \frac{\Gamma}{k_i}}{\ln 2}$$

tests. As $\Gamma \geq 2$, we find $\frac{\ln \Gamma}{\ln 2} \geq 1$ and therefore, testing all those groups of size at most Γ requires at most

$$\sum_{i=1}^{T} \frac{k_i \ln \Gamma - k_i \ln k_i}{\ln 2} \leq k \frac{\ln \Gamma}{\ln 2}$$

tests. Therefore, we clearly find

$$m \leq \left\lceil \frac{n}{\Gamma} \right\rceil + k \frac{\ln \Gamma}{\ln 2}$$

yielding the assertion of the theorem. □

Finally, with respect to the unrestricted group testing problem, we find as a direct consequence of the previous discussion on `SPIV` that we can use `SPIV` to infer all individuals with m_{inf} tests within two rounds.

Corollary 2.1.16 (Theorem 1.3 of [46])**.** *There is a two-stage inference algorithm which achieves inference of $\boldsymbol{\sigma}$ with high probability requiring at most m_{inf} tests.*

Indeed, it turns out that applying `SPIV` with the spatially coupled testing strategy using m_{inf} tests does render an estimate $\tilde{\sigma}$ of $\boldsymbol{\sigma}$ where all but $kn^{-\Omega(1)} = o(k/\ln n)$ individuals are identified correctly. Instead of applying the previously described clean-up step based on a local rigidity argument (which requires $m_{\mathrm{non-ada}}$ tests), we proceed as follows.

(i) Test any individual which is infected under $\tilde{\sigma}$ individually.

(ii) Test all individuals which are uninfected under $\tilde{\sigma}$ with the random regular model using `DD`.

Clearly, (i) requires at most $(1+o(1))k = o(m_{\mathrm{inf}})$ tests and the prevalence in (ii) is $o(k/\ln n)$, thus `DD` requires $o(m_{\mathrm{inf}})$ tests in order to succeed.

2.1.3. Summary of phase transitions in group testing

Let us begin by drawing a picture of unrestricted group testing. In Figure 2.3 we present a phase diagram based on the results obtained in [41, 46]. While exact recovery in the red areas (below $m_{\text{non-ada}}$) is not possible within one round of testing, in the light red area adaptive algorithms (e.g. Alleman's algorithm) succeed. The existence of this adaptivity-gap was unknown prior to this thesis's contributions. Furthermore, recovery in the blue area is information-theoretically possible on the random-regular model which was previously only known for $\theta > \frac{\ln 2}{1+\ln 2}$ while the efficient DD algorithm was proven to fail in the light blue area. Finally, we introduced the spatially coupled test-design and the efficient SPIV algorithm which already succeeds in this light blue area. Observe that we also managed to proof that inference of all but $o(k)$ individuals is possible within one round and inference of all individuals within two rounds by SPIV outside of the dark-red area.

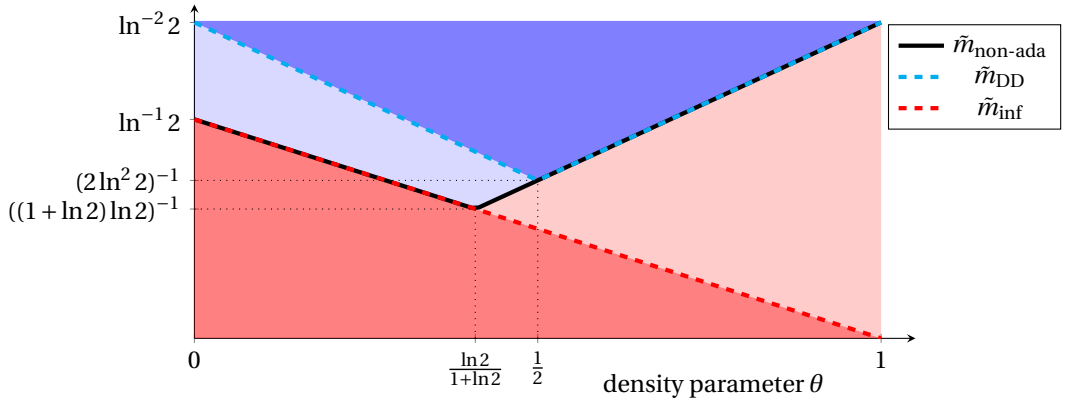

Figure 2.3.: The phase transitions in non-adaptive unrestricted hypergeometric group testing. The graphic is modified after [46, Figure 1]. We write $\tilde{m} = m \cdot (k \ln(n))^{-1}$.

With respect to the $\Delta-$divisible group testing problem, we present the (not completely understood) phase diagram in Figure 2.4. In the dark blue area, the simple COMP algorithm was known to perform on the random-regular model non-adaptively [86] while we proved that DD performs on the same model already in the light blue area. Furthermore, below the black line (red area), every testing scheme (even adaptive schemes) fail while below the red-dotted line all non-adaptive schemes do not succeed. Thus, the yellow area shows a regime where there might be non-adaptive schemes coming with efficient algorithms (e.g. SPIV-like ideas) but they are currently not known. Further, in the yellow and orange area there might be adaptive algorithms for inference of $\boldsymbol{\sigma}$ but they are also currently not known. Finally, the light red area marks a regime where we proved the existence of an adaptivity-gap, thus we found an efficient adaptive algorithm performing in this regime but there cannot be a non-adaptive pooling scheme facilitating inference.

Finally, the discussed phase transitions in the $\Gamma-$sparse group testing problem for $\Gamma = \Theta(1)$ are given in Figure 2.5. Above the blue line, the COMP algorithm studied by Gandikota et al. [86] succeeds at inference on the random regular model. In contrast, we can observe that DD (succeeding at $m_{\text{non-ada}}(\Gamma)$ on all θ outside of a set of measure zero) requires n/Γ tests less almost everywhere for $\theta > 1/2$ on the same model. Furthermore, DD performs slightly better in sparse instances on the matching model. Moreover, no algorithm (efficient or not) can succeed at inference below the red line on any non-adaptive pooling scheme. Finally, any adaptive testing scheme fails below the black line while we presented an algorithm who succeeds at this point (up to rounding). We stress that the points on which the achievability bound of DD satisfies $\tilde{m}_{\text{DD}}(\Gamma) = \tilde{m}_{\text{non-ada}}(\Gamma) + 1$ rather than being equal correspond to the jumps in the phase diagram.

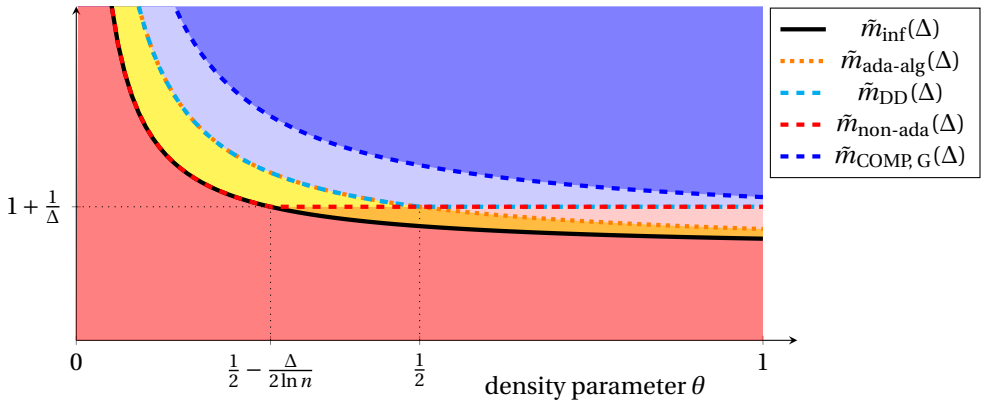

Figure 2.4.: Important phase transitions in $\Delta-$divisible hypergeometric group testing with $\Delta = O\left(\ln^{1-\delta} n\right)$ for some $\delta \in (0,1]$. The plot is with respect to the choice of parameters $\Delta = 5$, $n = 10^5$. The phase-transition lines correspond to the exponents of the actual phase transition points, thus we have $m = \Delta k^{\tilde{m}}$.

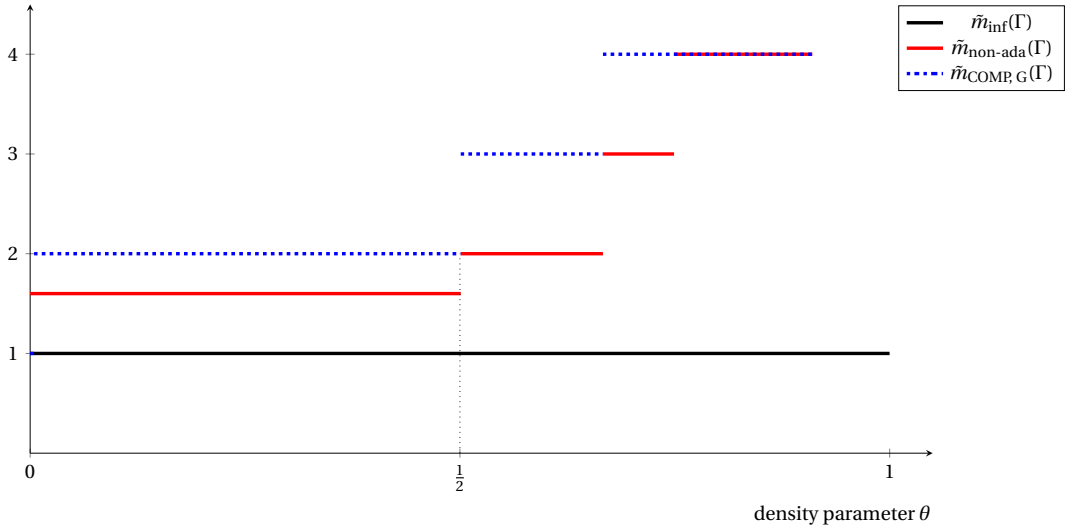

Figure 2.5.: The phase transitions in non-adaptive hypergeometric group testing under the $\Gamma-$sparse restriction for $\Gamma = \Theta(1)$. We define $m = \tilde{m}\frac{n}{\Gamma}$ and tacitly assume that $n/\Gamma \in \mathbb{Z}$. Furthermore, the plot is with respect to the choice $\Gamma = 4$.

In the next section, we will present how to achieve a formula that counts the number of solutions for a random $2-$SAT formula.

2.2. Counting solutions of a random 2-SAT formula

As already discussed in the introduction, it was a prominently posed open question, how many satisfying assignments a random $2-$SAT formula typically possesses [78]. Fortunately, the marginals obtained through Belief Propagation plugged into the Bethe functional yield a precise prediction. In the contribution

The number of satisfying assignments of random 2-SAT formulas [2]

we prove that this non-rigorous prediction is indeed correct. Before stating the main theorem, we require a bit of additional notation. Suppose we have n variables $x_1, \ldots, x_n$ taking spins in $\Omega = \{\pm 1\}$ and

a parameter $0 < d < 2$. Then we create $\boldsymbol{m} \sim \mathbf{Po}(dn/2)$ clauses and obtain a formula $\boldsymbol{\Phi}$ by choosing one uniformly at random from all possible formulas that have two distinct variables per clause. As before, $Z(\boldsymbol{\Phi})$ denotes the corresponding partition function. Let us introduce a Belief Propagation operator $\mathtt{BP}_d : \mathscr{P}((0,1)) \to \mathscr{P}((0,1))$ that maps a probability measure π onto π' as follows. Let $\boldsymbol{d}^+, \boldsymbol{d}^- \sim \mathbf{Po}(d/2)$ and $\left\{\boldsymbol{\mu}_{\pi,j}\right\}_j$ be a family of independent samples from π. Thus, $\boldsymbol{\mu}_{\pi,j}$ is a random probability measure on $(0,1)$. Then we define $\pi' = \mathtt{BP}_d(\pi)$ as the distribution of

$$\frac{\prod_{i=1}^{\boldsymbol{d}^-} \boldsymbol{\mu}_{\pi,j}}{\prod_{i=1}^{\boldsymbol{d}^-} \boldsymbol{\mu}_{\pi,j} + \prod_{i=1}^{\boldsymbol{d}^+} \boldsymbol{\mu}_{\pi,j+\boldsymbol{d}^-}}.$$

We will see in due course that this operator basically coincides with the Belief Propagation estimate of the marginals (1.1.20). We furthermore denote by $\mathtt{BP}_d^\ell$ the $\ell-$fold iteration of the $\mathtt{BP}_d$ operator. Finally, we let $\delta_x \in \mathscr{P}((0,1))$ denote the atom on x. Now we can finally state our main theorem.

Theorem 2.2.1 (Theorem 1.1 of [2])**.** *For any $0 < d < 2$ the limit $\pi_d = \lim_{\ell\to\infty} BP^\ell(\delta_{0.5})$ exists and*

$$\lim_{n\to\infty} \frac{1}{n} \ln Z(\boldsymbol{\Phi}) = \mathbb{E}\left[\ln\left(\prod_{i=1}^{\boldsymbol{d}^-} \boldsymbol{\mu}_{\pi_d,i} + \prod_{i=1}^{\boldsymbol{d}^+} \boldsymbol{\mu}_{\pi,i+\boldsymbol{d}^-}\right) - \frac{d}{2}\ln\left(1 - \boldsymbol{\mu}_{\pi_d,1}\boldsymbol{\mu}_{\pi_d,2}\right)\right]$$

in probability.

Theorem 2.2.1 implies that the free entropy density is really given through the Bethe functional (1.1.22). Let us stress that the paper's main contribution is actually a lower bound on $\ln Z(\boldsymbol{\Phi})$ as a tight upper bound could be computed by the so-called interpolation method [142]. Furthermore we stress that, of course, the result of the theorem might be hard to digest as it is far from obvious how to calculate a closed form. Nevertheless, it turns out that numerical approximations are easy to get.

How can we proof the statement of Theorem 2.2.1? To this end, let $\mu_{\boldsymbol{\Phi}} \in \mathscr{P}(\{\pm 1\}^n)$ be the Boltzmann distribution of the physical system given through the random factor graph corresponding to the random formula $\boldsymbol{\Phi}$. As we are in the zero-temperature limit, $\mu_{\boldsymbol{\Phi}}$ corresponds to the uniform distribution over all satisfying assignments (see (1.1.7)).

Supposing that the BP prediction is correct, it is not very surprising that π_d can actually be written in terms of the marginals of the Boltzmann distribution, more precisely as the random probability measure which is the $\mathbb{P}-$weak limit of

$$\pi_{\boldsymbol{\Phi}} = \frac{1}{n}\sum_{i=1}^n \delta_{\mu_{\boldsymbol{\Phi}}(\boldsymbol{\sigma}_{x_i}=1)}.$$

A core argument of the proof is that Belief Propagation is able to not only find the correct marginals but even the correct marginals for any boundary condition. Let $\boldsymbol{\sigma} \sim \mu_{\boldsymbol{\Phi}}$ be a sample from the Boltzmann distribution and τ be a second satisfying assignment. We define $\partial^{2\ell} x$ as the variables of distance exactly 2ℓ from variable x and ν^ℓ as the Belief Propagation estimation of the marginals. For the sake of the reading flow we recall the Belief Propagation messages and its estimation of the marginals from the introduction and plug in the specific setup at hand. A more detailed derivation can be found in [2]. We let $r \in \{\pm 1\}$ indicate whether variable x appears positively or negatively in clause a and let s be the corresponding sign of variable y. Then we let

$$\nu^{(\ell)}_{\boldsymbol{\Phi},a\to x}(t) = \frac{1 - \mathbf{1}\{r \neq t\}\nu^{(\ell-1)}_{\boldsymbol{\Phi},y\to a}(-s)}{1 + \nu^{(\ell-1)}_{\boldsymbol{\Phi},y\to a}(s)}, \qquad \nu^{(\ell)}_{\boldsymbol{\Phi},x\to a}(t) = \frac{\prod_{b\in\partial x\setminus\{a\}} \nu^{(\ell)}_{\boldsymbol{\Phi},b\to x}(t)}{\prod_{b\in\partial x\setminus\{a\}} \nu^{(\ell)}_{\boldsymbol{\Phi},b\to x}(1) + \prod_{b\in\partial x\setminus\{a\}} \nu^{(\ell)}_{\boldsymbol{\Phi},b\to x}(-1)}$$

be the Belief Propagation messages and define

$$\nu^{(\ell)}_{\boldsymbol{\Phi},x}(t) = \frac{\prod_{a\in\partial x} \nu^{(\ell)}_{\boldsymbol{\Phi},a\to x}(t)}{\prod_{a\in\partial x} \nu^{(\ell)}_{\boldsymbol{\Phi},a\to x}(1) + \prod_{a\in\partial x} \nu^{(\ell)}_{\boldsymbol{\Phi},a\to x}(-1)},$$

as the BP estimate of the marginals. Finally, we let

$$\mu_{\mathbf{\Phi}}\left(\cdot \mid \boldsymbol{\sigma}_{\partial^{2\ell} x_1} = \tau_{\partial^{2\ell} x_1}\right) = \mu_{\mathbf{\Phi}}\left(\cdot \mid \forall y \in \partial^{2\ell} x_1 : \boldsymbol{\sigma}_y = \tau_y\right)$$

be the Boltzmann distribution conditioned on having a specific boundary condition on variables of distance 2ℓ from x_1.

Then, if $\mathbf{\Phi}$ is satisfiable with high probability, we have [2, Theorem 1.2]

$$\lim_{\ell\to\infty} \limsup_{n\to\infty} \mathbb{E}\left[\left|\mu_{\mathbf{\Phi}}(\boldsymbol{\sigma}_{x_1} = \pm 1) - \nu^{(\ell)}_{\mathbf{\Phi},x_1}(\pm 1)\right| \mid Z(\mathbf{\Phi}) > 0\right] = 0$$

and

$$\lim_{\ell\to\infty} \limsup_{n\to\infty} \mathbb{E}\left[\max_{\tau}\left|\mu_{\mathbf{\Phi}}(\boldsymbol{\sigma}_{x_1} = 1 \mid \boldsymbol{\sigma}_{\partial^{2\ell} x_1} = \tau_{\partial^{2\ell} x_1}) - \nu^{(\ell)}_{\mathbf{\Phi},x_1}(1)\right| \mid Z(\mathbf{\Phi}) > 0\right] = 0.$$

We emphasise again that this observation is really strong and it is clearly one of the most important features of the contribution to prove this assertion. Indeed, it shows that Belief Propagation does render the correct marginals given any boundary condition. Why is this so important? It formally justifies core assumptions of the statistical physics' 1-RSB Ansatz (see Section 1.1.5.1) in the 2−SAT problem. Having the (non-rigorous) discussion of Section 1.1.5.1 in mind, let $S(\mathbf{\Phi})$ be the set of all satisfying assignments of $\mathbf{\Phi}$. Then we can observe by the triangle inequality that the Boltzmann distribution itself is a Bethe state as

$$\lim_{\ell\to\infty} \limsup_{n\to\infty} \mathbb{E}\left[\max_{\tau\in S(\mathbf{\Phi})}\left|\mu_{\mathbf{\Phi}}(\boldsymbol{\sigma}_{x_1} = 1 \mid \boldsymbol{\sigma}_{\partial^{2\ell} x_1} = \tau_{\partial^{2\ell} x_1}) - \mu_{\mathbf{\Phi}}(\boldsymbol{\sigma}_{x_1} = 1)\right| \mid Z(\mathbf{\Phi}) > 0\right] = 0.$$

Thus, as in Sections 1.3.1 and 1.1.5.1, we established being in a replica symmetric phase, hence the Boltzmann distribution does not exhibit long-range correlations but is $o(1)$−symmetric.

We recall, that $o(1)$−symmetry is defined as

$$\sum_{s,t\in\{\pm 1\}} \mathbb{E}\left[\left|\mu_{\mathbf{\Phi}}(\boldsymbol{\sigma}_{x_1} = s, \boldsymbol{\sigma}_{x_2} = t) - \mu_{\mathbf{\Phi}}(\boldsymbol{\sigma}_{x_1} = s)\cdot\mu_{\mathbf{\Phi}}(\boldsymbol{\sigma}_{x_2} = t)\right| \mid Z(\mathbf{\Phi}) > 0\right] = o(1).$$

Let us now sketch in four steps how to obtain the results, of course, as in the previous sections, this sketch only carries the main ideas of the proof and is not meant to be complete nor rigorous.

Existence of the limit. First, we need to verify that the limit π_d actually exists. This can be done rigorously by showing that the Belief Propagation operator BP is a contraction and actually converges quite fast towards a unique fixed-point. Furthermore, we can prove that the limit π_d does not only exist but satisfies a tail-bound of the form $\mathbb{E}\left[\ln^2\left(\boldsymbol{\mu}_{\pi_d}/\left(1-\boldsymbol{\mu}_{\pi_d}\right)\right)\right] < \infty$.

The Boltzmann distribution is a Bethe state. Second, we prove that BP renders the correct conditional marginals and that the marginals do not depend on the boundary conditions. We stress that the formula locally looks like a Galton-Watson tree, thus we suppose working on a tree. We make use of a feature which might be exclusive in the 2−SAT problem compared to general k−SAT, namely that we can actually construct the worst-case boundary conditions. Indeed, suppose we start at a root vertex x_0 and we want BP to output an as-high-as-possible marginal of the truth value +1 for x_0. Look at the clauses x_0 is part of, more precisely, partition them into those clauses A^- in which x_0 is negated and those clauses in which x_0 comes positively (A^+). As we want to nudge x_0 towards taking the value +1, we set all the variables being its partner in a clause of A^+ to a value that does not satisfy the clause, thus x_0 needs to be set to +1 in order to satisfy all clauses in A^+. Analogously, we set all its partner-variables in the tests of A^- to the value that satisfies the clauses already, thus x_0 does not need to satisfy them. This procedure can now be iterated until depth 2ℓ and a visualisation is given in Figure 2.6.

Of course, there are two different worst-case boundary conditions, one that nudges x_0 to +1 and one which nudges it to −1. Due to symmetry, it actually suffices to prove that the marginals of the +1-nudging configuration, call it $\boldsymbol{\sigma}^+$, coincide with the unconditional marginals. As the construction of $\boldsymbol{\sigma}^+$ clearly depends on the formula $\mathbf{\Phi}$, respectively on the Galton-Watson tree with root x_0, we have to

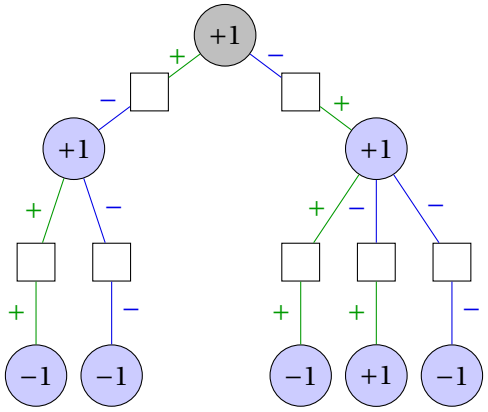

Figure 2.6.: Construction of a worst-case boundary condition. The value of the variables are chosen in the (unique) way that nudges the parent variables in the direction provided by setting x_0 (grey vertex) to +1. The graphic is modified after [2, Figure 2].

deal with delicate dependencies. More precisely, if we *really* went down the tree starting at x_0 to create $\boldsymbol{\sigma}^+$, we already would have revealed all randomness and it is far from clear how to work up this tree afterwards in order to compute the BP marginals. Fortunately, we can associate a random variable $\boldsymbol{\eta}_x \in \mathbb{R} \cup \{\pm\infty\}$ with each variable x of this tree which expresses how much x can nudge its grand-parent to the value it should take under $\boldsymbol{\sigma}^+$. These random variables $\boldsymbol{\eta}_x$ have a very comfortable (Markov-like) property which basically says that the value at vertices of distance $k \gg k'$ from x_0 does not depend on the vertices of distance at most k' from x_0.

How could such a random variable look like? To this end let $Z(\boldsymbol{T}_{x_0}, \boldsymbol{\sigma}^+, \boldsymbol{\sigma}^+_x)$ denote the number of satisfying assignments with boundary $\boldsymbol{\sigma}^+$ on the tree $\boldsymbol{T}_{x_0}$ of depth 2ℓ rooted at x_0 and truth value $\boldsymbol{\sigma}^+_x$ at x. Analogously, let $Z(\boldsymbol{T}_{x_0}, \boldsymbol{\sigma}^+, -\boldsymbol{\sigma}^+_x)$ count exactly those such assignments that have truth value $-\boldsymbol{\sigma}^+_x$ at x. Now we define $\boldsymbol{\eta}_x$ as the log-likelihood ratio between those quantities

$$\boldsymbol{\eta}_x = \ln \frac{Z(\boldsymbol{T}_{x_0}, \boldsymbol{\sigma}^+, \boldsymbol{\sigma}^+_x)}{Z(\boldsymbol{T}_{x_0}, \boldsymbol{\sigma}^+, -\boldsymbol{\sigma}^+_x)}.$$

Intuitively, as BP should calculate the marginal of x_0 with respect to a uniformly chosen satisfying assignment, the fraction between $Z(\boldsymbol{T}_{x_0}, \boldsymbol{\sigma}^+, \boldsymbol{\sigma}^+_x)$ and $Z(\boldsymbol{T}_{x_0}, \boldsymbol{\sigma}^+, -\boldsymbol{\sigma}^+_x)$ exactly expresses the extend to which x can be used to nudge its grand-parent into the correct direction. It turns out that the distribution of $\boldsymbol{\eta}_{x_0}$ can be expressed as the iterative application of a suitable operator which itself turns out to be a W_1-contraction. Analysing this operator is technically challenging, i.e. due to studying the problem at zero temperature which allows Z to decrease from an exponentially large number to zero by setting just a single variable to a certain truth value. But at least it is possible to analyse it and therefore prove that BP renders the correct marginals. The details can be found in [2, Section 5].

The Aizenman-Sims-Starr scheme. The third step is to prove that

$$\frac{1}{n}\mathbb{E}[\ln(Z(\boldsymbol{\Phi}) \vee 1)] \to \mathbb{E}\left[\ln\left(\prod_{i=1}^{\boldsymbol{d}^-} \boldsymbol{\mu}_{\pi_d,i} + \prod_{i=1}^{\boldsymbol{d}^+} \boldsymbol{\mu}_{\pi,i+\boldsymbol{d}^-}\right) - \frac{d}{2}\ln\left(1 - \boldsymbol{\mu}_{\pi_d,1}\boldsymbol{\mu}_{\pi_d,2}\right)\right] \tag{2.2.1}$$

in probability where we suppose that $\boldsymbol{d}^+, \boldsymbol{d}^- \sim \mathbf{Po}(d/2)$ denote the number of clauses in which a variable appears positively and negated respectively and where $\vee$ abbreviates the maximum. This truncation inside of the mean is actually necessary to deal with the case that, with very little probability, the formula could be unsatisfiable. The proof is done by the so-called *Aizenman-Sims-Starr* scheme [5]. To explain the main-idea, let Z_n denote the partition function of a particle system with n variables. Then we clearly find

$$\lim_{n\to\infty} \frac{1}{n}\mathbb{E}[\ln Z_n] = \lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n-1} (\mathbb{E}[\ln Z_{i+1}] - \mathbb{E}[\ln Z_i]) = \lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n-1} \mathbb{E}\left[\ln \frac{Z_{i+1}}{Z_i}\right].$$

Therefore, in order to calculate the partition function, we only need to understand its development when going from a system of size n to a system of size $n+1$. How can this be calculated? We need to couple a system with n particles and a corresponding system with $n+1$ particles by adding a particle and a few random clauses such that both systems follow the correct distribution. Let us discuss this in the setting of the 2−SAT problem. Suppose $\mathbf{\Phi}_n$ is a random formula with n variables and $m_n \sim \mathbf{Po}(dn/2)$ clauses while $\mathbf{\Phi}_{n+1}$ has $n+1$ variables and $m_{n+1} \sim \mathbf{Po}(d(n+1)/2)$ clauses. How could we possibly couple the formulas?

We start by a formula $\mathbf{\Phi}'$ on n variables that obtains some cavities, thus a few clauses less than it actually requires. More precisely, it has $m' \sim \mathbf{Po}(dn/2 - d/2)$ clauses. Now we obtain $\mathbf{\Phi}_n$ from $\mathbf{\Phi}'$ by adding $\mathbf{Po}(d/2)$ uniformly at random chosen clauses and $\mathbf{\Phi}_{n+1}$ by adding one variable x_{n+1} coming along with $\mathbf{Po}(d)$ clauses to $\mathbf{\Phi}'$, each of them taking a uniformly at random chosen second variable. Suppose we add from $\mathbf{\Phi}'$ to $\mathbf{\Phi}_{n+1}$ the clauses $a_1, \dots, a_{\boldsymbol{d}}$ such that the sign of x_{n+1} in clause a_i is randomly sampled as $\boldsymbol{s}_i$ from $\{-1,+1\}$. A visualisation is given in Figure 2.7.

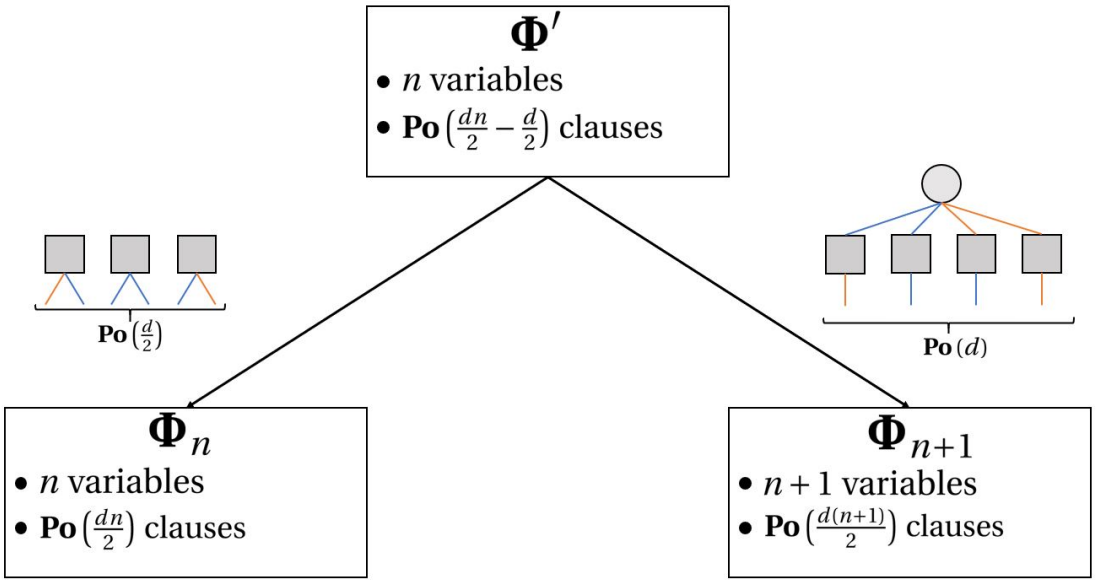

Figure 2.7.: A sketch of the Aizenmann-Sims-Starr scheme in the random 2−SAT problem. The formula $\mathbf{\Phi}'$ contains too few clauses and can be used to couple $\mathbf{\Phi}_n$ and $\mathbf{\Phi}_{n+1}$. The signs of the literals (red and blue edges) are drawn uniformly at random.

Then it suffices to prove that

$$\mathbb{E}\left[\ln \frac{Z(\mathbf{\Phi}_n)}{Z(\mathbf{\Phi}')}\right] \sim \frac{d}{2}\mathbb{E}\left[\ln\left(1-\boldsymbol{\mu}_{\pi_d,1}\boldsymbol{\mu}_{\pi_d,2}\right)\right] \quad \text{and} \quad \mathbb{E}\left[\ln \frac{Z(\mathbf{\Phi}_{n+1})}{Z(\mathbf{\Phi}')}\right] \sim \mathbb{E}\left[\ln\left(\sum_{\omega=\pm 1}\prod_{i=1}^{\boldsymbol{d}}\left(1-\mathbf{1}\{\omega \neq s_i\}\,\boldsymbol{\mu}_{\pi_d,i}\right)\right)\right]. \tag{2.2.2}$$

Indeed, suppose this is true. Then the second summand of (2.2.1) is already found as

$$\mathbb{E}\left[\ln \frac{Z(\mathbf{\Phi}_{n+1})}{Z(\mathbf{\Phi}_n)}\right] = \mathbb{E}\left[\ln \frac{Z(\mathbf{\Phi}_{n+1})}{Z(\mathbf{\Phi}')}\right] - \mathbb{E}\left[\ln \frac{Z(\mathbf{\Phi}_n)}{Z(\mathbf{\Phi}')}\right].$$

Furthermore, if we partition the newly added clauses with respect to the sign of x_{n+1} in the clause, we find that the last term can be written as

$$\mathbb{E}\left[\ln \frac{Z(\mathbf{\Phi}_n)}{Z(\mathbf{\Phi}')}\right] \sim \mathbb{E}\left[\ln\left(\sum_{\omega=\pm 1}\prod_{i=1}^{\boldsymbol{d}}\left(1-\mathbf{1}\{\omega \neq s_i\}\,\boldsymbol{\mu}_{\pi_d,i}\right)\right)\right] \sim \mathbb{E}\left[\ln\left(\prod_{i=1}^{\boldsymbol{d}^-}\left(1-\boldsymbol{\mu}_{\pi_d,i}\right) + \prod_{i=1}^{\boldsymbol{d}^+}\left(1-\boldsymbol{\mu}_{\pi_d,i+\boldsymbol{d}^-}\right)\right)\right].$$

This equals the first summand of (2.2.1) as $1-\boldsymbol{\mu}_{\pi,i}$ and $\boldsymbol{\mu}_{\pi,i}$ are equally distributed due to the symmetry of the signs of the single clauses. Thus, let as argue why (2.2.2) should intuitively hold.

Suppose we equipped $\mathbf{\Phi}'$ with a randomly chosen satisfying assignment $\boldsymbol{\tau}$, thus with a random sample from the Boltzmann distribution. Upon adding x_{n+1} with its connected clauses, there are two possibilities. Either, the truth value of x_{n+1} already satisfies a clause. In this case, the probability of extending $\boldsymbol{\tau}$ is one. If x_{n+1} does not satisfy clause a_i, then the second (randomly chosen) variable needs to satisfy it. This does happen with probability $(1-\boldsymbol{\mu}_{\pi_d,i})$. Altogether, this yields

$$\mathbb{E}\left[\ln\frac{Z(\mathbf{\Phi}_{n+1})}{Z(\mathbf{\Phi}')}\right] \sim \mathbb{E}\left[\ln\left(\sum_{\omega=\pm 1}\prod_{i=1}^{\boldsymbol{d}}\left(1-\mathbf{1}\left\{\omega\neq s_i\right\}\boldsymbol{\mu}_{\pi_d,i}\right)\right)\right].$$

On the other hand, if we add $\mathbf{Po}(d/2)$ clauses and connect them randomly to two variables and equip them with random signs, the probability that those two chosen variables both do not satisfy the clause under $\boldsymbol{\tau}$ is given by $\boldsymbol{\mu}_{\pi_d,1}\boldsymbol{\mu}_{\pi_d,2}$. Therefore, the proportion of satisfying assignments of $\mathbf{\Phi}'$ that are still satisfying for $\mathbf{\Phi}_n$ is expected to be $\left(1-\boldsymbol{\mu}_{\pi_d,1}\boldsymbol{\mu}_{\pi_d,2}\right)^{\frac{d}{2}}$ yielding

$$\mathbb{E}\left[\ln\frac{Z(\mathbf{\Phi}_n)}{Z(\mathbf{\Phi}')}\right] \sim \frac{d}{2}\mathbb{E}\left[\ln\left(1-\boldsymbol{\mu}_{\pi_d,1}\boldsymbol{\mu}_{\pi_d,2}\right)\right].$$

Of course, in a rigorous proof we need to deal with a lot of technical challenges. First, one needs to explicitly find a valid coupling $(\mathbf{\Phi}_n,\mathbf{\Phi}_{n+1})$. Second, randomly chosen variables might be almost and on average independent due to replica symmetry, but they are not completely independent. Third and most importantly, we are in the low-temperature limit. Therefore, adding one single clause could erase all satisfying assignments. A detailed proof how to handle those delicate technicalities can be found in [2, Section 6].

Concentration of the free entropy. Finally, as a last step, we need to show that $\ln(Z(\mathbf{\Phi})\vee 1)$ is concentrated around its mean. Proving this assertion seems on the first glance very challenging due to the huge fluctuations occurring in the low-temperature limit. Indeed, standard tools like the Azuma-Hoeffding inequality are doomed to fail as with little probability there might be exponentially large changes in the partition function. But fortunately, we are in the good shape that Panchenko and Talagrand [142] already proved that the partition function of the 2−SAT problem at positive temperature is concentrated applying the interpolation method of statistical physics. Let $Z_\beta(\mathbf{\Phi})$ be the partition function with respect to formula $\mathbf{\Phi}$ of the 2−SAT problem at inverse temperature β. It is clear by the definition that $Z_\beta(\mathbf{\Phi})\geq Z(\mathbf{\Phi})$ and it is known from [142] that $n^{-1}\ln Z_\beta(\mathbf{\Phi})$ does only exceed the value of the corresponding Bethe functional $\mathscr{B}_\beta$ by a factor of $(1+o(1))$ with high probability. Even more importantly, we find that the Bethe functional has a natural limit $\mathscr{B}_\infty(\pi_d)<\infty$ which coincides with the expectation of (2.2.1) and therefore, it is possible to show that $\ln Z(\mathbf{\Phi})$ does not exceed its expectation by more than $\pm\varepsilon n$ (for any $\varepsilon>0$). The details dealing with the exact functions, calculations and requirements for taking the limit can be found in [2, Section 7].

Therefore, the four described steps suffice to calculate the number of satisfying assignments of a random 2−SAT formula with high probability. In the next section we will discuss results with respect to a limit theory for discrete probability measures akin to the graph limit theory.

2.3. Limits of discrete probability measures and the cut-distance

All results of this section were obtained in

The cut metric for probability distributions [47]

and establish a consistent limit theory for discrete probability measures.

We recall from the introduction that the cut-distance of two probability measures $\mu,\nu\in\mathscr{P}(\Omega^n)$ is

defined as

$$\Delta_{\boxtimes}(\mu,\nu) = \inf_{\substack{\gamma\in\Gamma(\mu,\nu),\\ \phi\in\mathbb{S}_n}} \sup_{\substack{S\subset\Omega^n\times\Omega^n,\\ X\subset[n],\\ \omega\in\Omega}} \left| \sum_{\substack{(\sigma,\tau)\in S,\\ x\in X}} \gamma(\sigma,\tau)\left(\mathbf{1}\{\sigma_x=\omega\} - \mathbf{1}\left\{\tau_{\phi(x)}=\omega\right\}\right)\right|,$$

where $\Gamma(\mu,\nu)$ is the set of couplings of μ and ν and $\mathbb{S}_n$ is the set of permutations on $[n]$. Further, $\mathscr{L}_n(\Omega)$ is the set of equivalence classes over $\mathscr{P}(\Omega^n)$ identifying those measures with cut-distance zero.

Embedding discrete measures Moreover, we already learned about some kind of continuous embedding of configurations $\sigma\in\Omega^n$ into the space Σ_Ω of measurable functions from $[0,1)\to\mathscr{P}(\Omega)$ by

$$\hat{\sigma}:[0,1)\to\mathscr{P}(\Omega) \qquad \text{s.t.} \qquad x\mapsto\sum_{i=1}^{n}\delta_{\sigma_i}\mathbf{1}\left\{x\in\left[\frac{i-1}{n},\frac{i}{n}\right)\right\}$$

such that an associated probability measure $\mu\in\mathscr{P}(\Omega^n)$ can be expressed as

$$\hat{\mu}=\sum_{\sigma\in\Omega^n}\mu(\sigma)\delta_{\hat{\sigma}} \qquad \text{s.t.} \qquad \hat{\mu}\in\mathscr{P}(\Sigma_\Omega).$$

Finally, we recall that the cut-distance of two such measures $\mu,\nu\in\mathscr{P}(\Sigma_\Omega)$ is defined as

$$D_{\boxtimes}(\mu,\nu)=\inf_{\substack{\gamma\in\Gamma(\mu,\nu),\\ \varphi\in\mathbb{S}_{[0,1)}}}\sup_{\substack{B\subset\Sigma_\Omega^2,\\ U\subset[0,1),\\ \omega\in\Omega}}\left|\int_B\int_U\sigma_x(\omega)-\tau_{\varphi(x)}(\omega)\mathrm{d}x\mathrm{d}\gamma(\sigma,\tau)\right|$$

and identify measures with cut-distance zero. The resulting space of $\Omega-$laws is denoted by $\mathscr{L}=\mathscr{L}_\Omega$. From the definition it is immediate that $D_{\boxtimes}(\hat{\mu},\hat{\nu})\le\Delta_{\boxtimes}(\mu,\nu)$ but it is less clear if the other direction holds. We provide an argument that the embedding is, nevertheless, consistent as we have the following.

Theorem 2.3.1 (Theorem 1.2 of [47])**.** *There is a function $f:[0,1]\to[0,1]$ with $f^{-1}(0)=\{0\}$ such that for all $n\ge1$ and all $\mu,\nu\in\mathscr{P}(\Omega^n)$ we have*

$$f\left(\Delta_{\boxtimes}(\mu,\nu)\right)\le D_{\boxtimes}(\hat{\mu},\hat{\nu})\le\Delta_{\boxtimes}(\mu,\nu).$$

This assertion is far from being obvious. Indeed, equality would hold if the measure preserving bijection in the definition of $D_{\boxtimes}(\cdot,\cdot)$ was restricted to map intervals $I_i=[(i-1)/n,i/n)$ onto intervals I_j completely. But one can proof that the mass of one such interval does at least not split too much, more precisely, any reasonable such bijection maps at least mass n^{-3} from one interval into another one. Therefore, we get instantly $\Delta_{\boxtimes}(\mu,\nu)\le n^3D_{\boxtimes}(\hat{\mu},\hat{\nu})$. If n is small, this observation suffices clearly. If, on the other hand, n is sufficiently large, we partition the phase space Ω^n as well as the coordinates $[n]$ by the regularity lemma in such a way that on each little part of the partition, we can argue that the induced step functions $\hat{\mu},\hat{\nu}$ have to be very close to μ,ν under any permutation and obtain $\Delta_{\boxtimes}(\mu,\nu)\le D_{\boxtimes}(\hat{\mu},\hat{\nu})+o(1)$, establishing the theorem.

We will first show that there is a different, very elegant possibility to describe $\Omega-$laws. More precisely, such a description was already introduced in [42].

The kernel representation The kernel representation of $\Omega-$laws says briefly that it is possible to describe such a measure as something very similar to a graph limit if we define an appropriate distance. We will start discussion the other way round, thus we first define the space of kernels $\mathscr{K}$ by identifying all measurable functions $\kappa,\kappa':[0,1]^2\to\mathscr{P}(\Omega)$ with cut-distance zero where the cut-distance is defined as

$$D_{\boxtimes}(\kappa,\kappa')=\inf_{\phi,\phi'\in\mathbb{S}_{[0,1]}}\sup_{\substack{S,X\subset[0,1],\\ \omega\in\Omega}}\left|\int_S\int_X\kappa_{s,x}(\omega)-\kappa'_{\phi(s),\phi'(x)}(\omega)\mathrm{d}x\mathrm{d}s\right|.$$

Comparing $D_{\boxtimes}(\cdot,\cdot)$ with the usual cut-distance for kernels from the graph limit theory $\mathscr{D}_{\boxtimes}(\cdot,\cdot)$, we observe that (for a given ω) the major difference is that we may permute the entries on the two axis independently which clearly accounts for the fact that in the graph limit case both axes correspond to vertices of a graph while in the setting at hand, the $x-$axis represents coordinates while the $s-$axis represents configurations. Furthermore, we allow maximisation with respect to the spin ω while in the graph case we already have $\kappa_{s,x} \in \mathbb{R}$.

Now, suppose we have a kernel κ, then it induces a function $\kappa_s : [0,1] \to \mathscr{P}(\Omega) \in \Sigma_\Omega$ mapping $x \to \kappa_{s,x}$. We define $\mu^\kappa \in \mathscr{L}$ as the distribution of $\kappa_{\boldsymbol{s}}$ for an uniformly at random chosen $\boldsymbol{s} \in [0,1]$. Of course, each $\Omega-$law μ can be seen as the distribution of some $\kappa^\mu_{\boldsymbol{s}}$ as well, i.e. κ^μ coincides with $s : [0,1] \to \mathscr{P}(\Omega)$ on a set of measure $\mu(s)$.

With this picture in mind, it is actually not very surprising that those two objects are basically the same.

Theorem 2.3.2 (Theorem 1.4 of [47])**.** *The map $\mathscr{K} \to \mathscr{L}$ induced by $\kappa \mapsto \mu^\kappa$ is an isometric bijection.*

The actual proof which can be found in [47, Section 3] is quite technical but its main idea is to start with an arbitrary map $f : [0,1] \to \mathscr{P}(\Omega)$ that maps $s \mapsto f_s$ and to associate a kernel $\kappa^f : [0,1]^2 \to \mathscr{P}(\Omega)$ (by $\kappa^f_{s,x} = f_{s,x}$) with it. Analogously, the same map f can be used to transform the Lebesgue measure into a $\Omega-$law μ^f. All we need to show is that for two such functions f, g we find

$$D_{\boxtimes}(\kappa^f, \kappa^g) = D_{\boxtimes}\left(\mu^f, \mu^g\right).$$

Let us shortly come back to the connection between graph limits and our kernels. For the sake of clarity, we will refer to the graph limit kernels as graphons from now on. First, a graphon is a symmetric mapping and it turns out that we can make our kernels symmetric as follows. To this end, we define the *transpose* $\kappa^\dagger : (s,x) \mapsto \kappa_{x,s}$ of a kernel κ and observe that κ is symmetric if $\kappa = \kappa^\dagger$. For $\kappa \in \mathscr{K}$, define a family $\{\kappa^{(\omega)}\}_{\omega\in\Omega}$ of symmetric functions by

$$\kappa^{(\omega)}_{s/2,(1+x)/2} = \kappa_{s,x}(\omega), \qquad \kappa^{(\omega)}_{(1+s)/2,x/2} = \kappa_{x,s}(\omega), \qquad \kappa^{(\omega)}_{s/2,x/2} = \kappa^{(\omega)}_{(1+s)/2,(1+x)/2} = 0. \tag{2.3.1}$$

Intuitively, we squeeze the kernel from $[0,1]^2$ to $[0,\frac{1}{2}]^2$ and put it in the bottom left part of a unit square and add its transpose as the upper right part. The upper left as well as the upper right part will just equal zero. Clearly, each $\kappa^{(\omega)}$ is a symmetric function from $[0,1]^2 \to [0,1]$, thus a (bipartite) graphon where $\kappa^{(\omega)}_{s,x}$ is the corresponding edge weight.

Therefore, it is not surprising that various of the properties and results of graph limit theory carry over to the limit theory for discrete probability measures. We start by presenting one very important feature of graph limits, namely that it is possible to find a representation as an *exchangeable array*.

Exchangeable arrays We recall from the introduction that the graph limits were originally defined by the convergence of all series of homomorphism densities which basically expresses the density of how often which subgraph is present in the sequence of graphs. We will see that a similar equivalence is correct in the case of convergence of a sequence of probability measures $(\mu_n)_n$ to an $\Omega-$law μ with specific $\Omega^{n\times n}$-matrices replacing subgraphs.

To this end, we call a probability distribution $\Xi \in \mathscr{P}\left(\Omega^{\mathbb{N}\times\mathbb{N}}\right)$ *exchangeable* if the distribution of $\boldsymbol{X}^\Xi(i,j)$ coincides with the distribution of $\boldsymbol{X}^\Xi(\varphi(i),\psi(j))$ for any $\varphi,\psi \in \mathbb{S}_n$ where $i,j \in [n]$ and $\boldsymbol{X}^\Xi$ is a two-dimensional infinite array over Ω sampled from Ξ. The space of such exchangeable arrays has nice properties, for instance, it is compact and separable if equipped with the weak topology (Tychonoff's theorem). It is easy to generate such an infinite array from a kernel. Indeed, sample $\boldsymbol{s}_1, \boldsymbol{x}_1, \boldsymbol{s}_2, \boldsymbol{x}_2, \ldots \in [0,1]$ uniformly at random and independently and create an array $\boldsymbol{X}^\kappa$ such that each entry $\boldsymbol{X}^\kappa(i,j)$ is just an independent sample from $\kappa_{\boldsymbol{s}_i,\boldsymbol{x}_j} \in \mathscr{P}(\Omega)$. For the sake of brevity, we denote by $\boldsymbol{X}^\mu$ the infinite array obtained from κ^μ if κ^μ is the corresponding kernel to the $\Omega-$law μ.

Of course, if $\pi \in \mathscr{P}(\mathscr{K})$ is a distribution on such kernels, the same procedure induces a distribution Ξ^π on infinite arrays by first drawing $\boldsymbol{\kappa}$ from π and then creating $\boldsymbol{X}^{\boldsymbol{\kappa}}$. It turns out that this operation

is indeed a homeomorphism. If both probability spaces, $\mathscr{P}\left(\Omega^{\mathbb{N}\times\mathbb{N}}\right)$ and $\mathscr{P}(\mathscr{K})$ are equipped with the weak topology[1], we have the following theorem.

Theorem 2.3.3 (Theorem 1.8 of [47]). *The map $\pi \mapsto \Xi^\pi$ is a homeomorphism.*

This theorem is kind of an extension of a result in graph limit theory. Indeed, in the special case $\Omega = \{0,1\}$ it boils down to the the directed graph version of [63, Theorem 5.3]. Now we can see that a similar principle as the subgraph count is really an important feature in the theory of convergence of probability measures. Suppose that $(\mu_N)_{N\geq 1}$ is a sequence of Ω-laws that converges to $\mu \in \mathscr{L}$. Then we find an exchangeable array $\boldsymbol{X}^{\mu_N}$ such that for all $n \geq 1$ and all matrices $A \in \Omega^{n\times n}$ we have that

$$\lim_{N\to\infty} \mathbb{P}\left[\forall i,j \in [n] : \boldsymbol{X}^{\mu_N}(i,j) = A_{i,j}\right] = \mathbb{P}\left[\forall i,j \in [n] : \boldsymbol{X}^{\mu}(i,j) = A_{i,j}\right]. \tag{2.3.2}$$

If on the other hand (2.3.2) holds for all $n \geq 1$ and all $A \in \Omega^{n\times n}$, then the theorem implies

$$\lim_{N\to\infty} D_{\boxtimes}(\mu_N, \mu) = 0.$$

The existence of such a representation as an exchangeable array for an $\Omega-$law enables us to obtain a very basic result of graph limit theory in the context of $\Omega-$laws very elegantly: the sampling lemma.

Sampling from an $\Omega-$law A further feature of graph limit theory is the sampling operation. Given a large enough random graph obtained from a graphon by sampling, this finite graph will be very close to the original graphon under the cut-distance. We will use the possibility to express $\Omega-$laws as exchangeable infinite arrays to obtain a similar result for discrete probability measures. More precisely, given an $\Omega-$law μ and its representation as an exchangeable array $\boldsymbol{X}^\mu$, we define μ_n as the empirical distribution of the rows of the upper $n\times n-$submatrix of $(\boldsymbol{X}^\mu)$, thus formally

$$\mu_n(\sigma) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}\left\{\forall j \in [n] : \boldsymbol{X}^\mu(i,j) = \sigma_j\right\} \qquad (\sigma \in \Omega^n).$$

As $\boldsymbol{X}^\mu$ is clearly dependent on the random coordinates $(\boldsymbol{s}_i, \boldsymbol{x}_j)_{i,j\geq 1}$, the obtained measure μ_n is a random probability distribution. We obtain the following *sampling lemma*.

Theorem 2.3.4 (Theorem 1.9 of [47]). *There is $c > 0$ such that for all $n > 1$ and all $\mu \in \mathscr{L}$ we have $\mathbb{E}\left[D_{\boxtimes}(\mu,\mu_n)\right] \leq c/\sqrt{\ln n}$.*

It turns out that this dependence on n is actually best possible (besides a possibly optimised choice of the constant c).

Theorem 2.3.5 (Theorem 1.10 of [47]). *There is a constant $d > 0$ such that for any $\varepsilon > 0$ we find some $\Omega-$law μ such that $D_{\boxtimes}(\mu,\nu) \geq \varepsilon$ for every $\nu \in \mathscr{L}$ which is supported on at most $\exp(d/\varepsilon^2)$ configurations.*

The proofs of Theorems 2.3.4 and 2.3.5 make extensive use of the close connection to graph limit theory. The sampling lemma itself is proven analogously as its corresponding form in the case of graph limits (see, for instance, [127, Section 10]) and can be found in [47, Section 3]. While the proof idea is similar, technical challenges appear due to the missing symmetry of the kernels. Let κ be a kernel, then we begin by obtaining a (finite) kernel κ_n as the matrix $(\kappa_{\boldsymbol{s}_i,\boldsymbol{x}_j})_{i,j}$ which is obtained by sampling $\boldsymbol{x}_1, \boldsymbol{s}_1, \ldots, \boldsymbol{x}_n, \boldsymbol{s}_n \in [0,1]$ u.a.r.. Furthermore, $\hat{\kappa}_n$ is the $n\times n$ upper left sub-matrix of $\boldsymbol{X}^\kappa$. We first show that, with sufficiently high probability, the cut-distance between two kernels κ, κ' is sufficiently well approximated by the cut-distance between the corresponding finite kernels κ_n, κ'_n which follows from a carefully applied result from the graph limit theory. Afterwards, we use a standard large deviation bound to prove that $\mathbb{E}[D_{\boxtimes}(\kappa_n, \hat{\kappa}_n)] = o(1)$. To obtain the result itself, we require the the regularity lemma. More precisely, we find that the step-kernel $\hat{\kappa}$ coincides with the step-kernel guaranteed by the regularity

[1] We define the weak topology on a set with respect to a family of functions as the coarsest topology under which all those functions are continuous.

lemma, thus it is close to κ under the cut-distance and therefore κ_n and $\tilde{\kappa}_n$ are close as well. Finally, it suffices to see that the two step-kernels $\hat{\kappa}$ and $\hat{\kappa}_n$ are close to each other which can be done by comparing them on each step.

We furthermore want to show that the dependency on n in the sampling lemma is roughly optimal. In this case apply a result from graph theory directly. It is known that there are graphons W_G such that each partition P of the vertex set which induces a step-graphon satisfying $\mathscr{D}_{\boxtimes}(W_G, W_{G^P}) < \varepsilon$ (c.f. Section 1.3.1.2) has to consist of at least $\exp\left(\Theta(\varepsilon^{-2})\right)$ parts [56, Theorem 7.1]. Now, given such a graphon W_G, we create a kernel κ^G by (2.3.1) and apply the sampling lemma. If the assertion of Theorem 2.3.5 was false, we would obtain a step-kernel κ_n^G on fewer steps satisfying $D_{\boxtimes}(\kappa^G, \kappa_n^G) < \varepsilon/2$. But as $\Omega = \{0,1\}$, we have

$$\mathscr{D}_{\boxtimes}(\kappa^{G,1}, \kappa_n^{G,1}) \leq 2D_{\boxtimes}(\kappa^G, \kappa_n^G) < \varepsilon$$

which is a contradiction to the aforementioned result from graph limit theory [56, Theorem 7.1].

In the next paragraph we will discuss the pinning operation (c.f. Section 1.3.1.4) in the general context of $\Omega-$laws.

Pinning Recall from the introduction that the pinning operation says roughly that conditioned on pinning a few coordinates to specific values each probability measure over Ω^n becomes extremal. For the convenience of the reader, we recall that for $\mu \in \mathscr{P}(\Omega^n)$ we denote by

$$\bar{\mu}(\sigma) = \prod_{i=1}^{n} \mu_i(\sigma_i)$$

the corresponding product measure on the same marginals. Furthermore, we call a measure $\varepsilon-$*extremal* if we have $\Delta_{\boxtimes}(\mu, \bar{\mu}) < \varepsilon$.

It is possible to generalise this notion to $\Omega-$laws. Indeed, if $\mu \in \mathscr{L}$ is an $\Omega-$law, we define $\bar{\mu} \in \mathscr{L}$ as the generalised product measure on the same marginals, thus it is the atom concentrated on

$$[0,1] \to \mathscr{P}(\Omega), \qquad x \mapsto \int_{\Sigma_\Omega} \sigma_x \mathrm{d}\mu(\sigma). \tag{2.3.3}$$

We find clearly that $D_{\boxtimes}(\bar{\mu}, \bar{\nu}) = 0$ whenever $D_{\boxtimes}(\mu, \nu) = 0$, therefore, (2.3.3) induces a mapping from the space of $\Omega-$laws into itself by $\mu \mapsto \bar{\mu}$. It turns out that those generalised product measures carry a lot of information about the actual $\Omega-$laws.

Theorem 2.3.6 (Theorem 1.11 of [47])**.** *Let μ, ν be $\Omega-$laws and $\bar{\mu}, \bar{\nu}$ the corresponding generalised product measures given via* (2.3.3)*. Then we have*

$$D_{\boxtimes}(\bar{\mu}, \bar{\nu}) \leq D_{\boxtimes}(\mu, \nu) \quad \textit{and} \quad D_{\boxtimes}(\bar{\mu}, \bar{\nu}) \leq \max_{\omega \in \Omega} \int_0^1 \left| \int_{\Sigma_\Omega} \sigma_x \mathrm{d}\mu(\sigma) - \int_{\Sigma_\Omega} \sigma_x \mathrm{d}\nu(\sigma) \right| \mathrm{d}x \leq 2D_{\boxtimes}(\bar{\mu}, \bar{\nu}). \tag{2.3.4}$$

Furthermore, the set of extremal laws is a closed subset of all $\Omega-$laws.

The proof of this lemma follows from a fairly short technical calculation whose most important idea it is to partition the space of coordinates $[0,1]$ into $X^+(\omega)$ and $X^-(\omega)$ such that

$$X^+(\omega) = \left\{ x \in [0,1] : \int_{\Sigma_\Omega} \sigma_x(\omega) \mathrm{d}\mu(\sigma) - \int_{\Sigma_\Omega} \sigma_x(\omega) \mathrm{d}\nu(\sigma) \geq 0 \right\}.$$

This helps to cope with the possible cancelling out of contributions with different signs in the definition of the cut-distance.

Once we introduced $\varepsilon-$extremity also in the limit case of $\Omega-$laws, it is a natural question whether the pinning operation of Coja-Oghlan et al. [49] generalises and yields $\varepsilon-$extremal $\Omega-$laws. To this end, let us define the pinning operation for an $\Omega-$law μ. Given some integer $\theta \geq 1$ and θ coordinates $x_1, \ldots, x_\theta \in$

$[0,1]$ as well as a configuration $\tau \in \Omega^\theta$, we define a normalising constant

$$Z = Z_\mu(\tau, x_1, \dots, x_\theta) = \int_{\Sigma_\Omega} \prod_{i=1}^{\theta} \sigma_{x_i}(\tau_i) \mathrm{d}\mu(\sigma).$$

Furthermore, if $Z > 0$, let $\mu_{\tau\downarrow x_1,\dots,x_\theta}$ be defined as

$$\mathrm{d}\mu_{\tau\downarrow x_1,\dots,x_\theta}(\sigma) = \frac{1}{Z} \prod_{i=1}^{\theta} \sigma_{x_i}(\tau_i) \mathrm{d}\mu(\sigma) \tag{2.3.5}$$

and $\mu_{\tau\downarrow x_1,\dots,x_\theta} = \mu$ if $Z = 0$.

This pinning operation looks very similar to the discrete version and indeed, the discrete version is contained as a special case in which each of the factors on the r.h.s. of (2.3.5) is either one or zero.

As in the discrete case, it becomes interesting if we chose $\boldsymbol{x}_1, \dots, \boldsymbol{x}_\theta$ randomly. More specifically, for a given $\theta \geq 1$,

(i) let $\boldsymbol{x}_1, \boldsymbol{x}_2, \dots \in [0,1]$ be u.a.r. and mutually independent,

(ii) draw $\boldsymbol{\tau} \in \Sigma_\Omega$ from the distribution μ,

(iii) pick a reference configuration $\hat{\boldsymbol{\tau}}$ from $\boldsymbol{\tau}_{\boldsymbol{x}_1} \otimes \cdots \otimes \boldsymbol{\tau}_{\boldsymbol{x}_\theta} \in \mathscr{P}\left(\Omega^\theta\right)$,

(iv) obtain $\mu_{\hat{\boldsymbol{\tau}}\downarrow\theta} = \mu_{\hat{\boldsymbol{\tau}}\downarrow\hat{\boldsymbol{x}}_1,\dots,\hat{\boldsymbol{x}}_\theta}$ via (2.3.5).

By the choice of $\boldsymbol{\tau}$ being sampled from μ, we clearly find that $\boldsymbol{Z}_\mu(\hat{\boldsymbol{\tau}}) > 0$ almost surely. Finally, we let

$$\mu_{\downarrow\theta} = \mathbb{E}\left[\overline{\mu_{\hat{\boldsymbol{\tau}}\downarrow\theta}} \mid \boldsymbol{x}_1, \dots, \boldsymbol{x}_\theta\right] \in \mathscr{L}.$$

Intuitively spoken, $\mu_{\downarrow\theta}$ is a weighted probability measure such that each configuration's probability is weighted according to the probability of its reference configuration.

The pinning lemma itself in the continuous case does guarantee that we find with high probability an approximation of an Ω−law μ supported on few ε−extremal Ω−laws. More precisely, it reads as follows.

Theorem 2.3.7 (Theorem 1.12 of [47])**.** *Given $\varepsilon \in (0,1)$ and an Ω−law μ, draw $0 \leq \boldsymbol{\theta} = \boldsymbol{\theta}(\varepsilon) \leq 64\varepsilon^{-8}\ln|\Omega|$ uniformly and independently of everything else. Then we find with probability at least $1-\varepsilon$ that $\mu_{\hat{\boldsymbol{\tau}}\downarrow\boldsymbol{\theta}}$ is ε−extremal and $\mathbb{E}[D_\boxtimes(\mu, \mu_{\downarrow\boldsymbol{\theta}})] < \varepsilon$.*

While the proof of the pinning lemma is the technical main contribution of [47] and relies on delicate and technically challenging results, it is surprisingly easy to give a high level sketch. In the following sketch, let $\delta, \delta', \dots$ be small and suitable chosen constants.

First, we need to verify that the pinning operation is continuous with respect to the cut-distance. Second, given an Ω−law μ, we use the sampling lemma to obtain a discrete probability measure $\nu \in \mathscr{P}(\Omega^n)$ such that $D_\boxtimes(\mu, \nu) < \delta$. Now we apply the pinning operation to μ as well as to ν and obtain (written a bit shortly) $\mu_{\downarrow n}$ and $\nu_{\downarrow n}$. By the continuity of the pinning operation we have

$$D_\boxtimes(\mu_{\downarrow n}, \nu_{\downarrow n}) < \delta'$$

and as the pinning operation reduces to the discrete pinning regarding ν, the discrete pinning lemma guarantees that

$$\mathbb{E}\left[\Delta_\boxtimes\left(\overline{\nu_{\downarrow\boldsymbol{\theta}}}, \nu_{\downarrow\boldsymbol{\theta}}\right)\right] < \delta''.$$

Now, as the embedding of discrete probability measures into the Ω−laws respects the cut-distance (Theorem 2.3.1), we directly get

$$\mathbb{E}\left[D_\boxtimes\left(\overline{\nu_{\downarrow\boldsymbol{\theta}}}, \nu_{\downarrow\boldsymbol{\theta}}\right)\right] < \delta'''.$$

Now the results on how far the generalised product measures are from the actual measures in the cut-distance (Theorem 2.3.6) and the triangle inequality give

$$D_\boxtimes\left(\overline{\mu_{\downarrow\boldsymbol{\theta}}}, \mu_{\downarrow\boldsymbol{\theta}}\right) < \delta'''' + D_\boxtimes\left(\overline{\nu_{\downarrow\boldsymbol{\theta}}}, \nu_{\downarrow\boldsymbol{\theta}}\right).$$

Therefore, it is possible to reduce the continuous pinning to the discrete version and the assertion follows from Markov's inequality.

While the pinning operation stands at the heart of the contribution we obtained two more results. The first of those two is with respect to *multi-overlaps*.

Overlaps We will first describe two very basic operations on probability measures that turn out to be continuous with respect to the cut-distance. For the sake of simplicity, we will describe their meaning for discrete measures over Ω^n and will just shortly state the corresponding operation on $\Omega-$laws.

The first such operation is the construction of a product measure. More precisely, suppose we have two probability measures $\mu, \nu \in \mathscr{P}(\Omega^n)$. Then their product $\mu \otimes \nu$ is a probability distribution on $(\Omega \times \Omega)^n$ such that $(\mu \otimes \nu)(\sigma, \tau) = \mu(\sigma)\nu(\tau)$.

Analogously, we could define a tensor $\mu \otimes \nu$ such that for $\sigma_1, \tau_1, \ldots, \sigma_n, \tau_n \in \Omega$ we obtain

$$\mu \otimes \nu \left(\begin{pmatrix} \sigma_1 \\ \tau_1 \end{pmatrix}, \ldots, \begin{pmatrix} \sigma_n \\ \tau_n \end{pmatrix} \right) = \mu(\sigma_1, \ldots, \sigma_n)\nu(\tau_1, \ldots, \tau_n).$$

Clearly, both constructions are equivalent but the – perhaps less intuitive – tensor variant naturally extends to $\Omega-$laws [49]. More precisely, it is convenient to identify the $\Omega-$laws with their corresponding kernel. To be more precise, suppose that $\Lambda : [0,1] \to [0,1] \times [0,1]$, $x \mapsto (\Lambda_1(x), \Lambda_2(x))$ is any measurable bijection mapping the Lebesgue measure λ_1 on $[0,1]$ onto the Lebesgue measure λ_2 on $[0,1]^2$. Furthermore, Λ needs to satisfy that Λ^{-1} maps the Lebesgue measure on $[0,1]^2$ to the Lebesgue measure on $[0,1]$. Such transformations of the Lebesgue measure clearly exist, for instance, a proof is given by [102, Theorem A.7]. If now κ and κ' are the kernel representations of two $\Omega-$laws, we define their generalised product as

$$\kappa \otimes \kappa' : [0,1]^2 \to \mathscr{P}(\Omega^2), \qquad (s,x) \in [0,1] \times [0,1] \mapsto \kappa_{\Lambda_1(s),x} \otimes \kappa'_{\Lambda_2(s),x} \in \mathscr{P}(\Omega^2).$$

Thus, by re-translating the kernels into $\Omega-$laws we immediately find that, given Ω-laws μ, ν, above's procedure yields an Ω^2-law $\mu \otimes \nu$. As already pointed out, this operation is continuous.

Theorem 2.3.8 (Theorem 1.15 of [47])**.** *The map $(\mu, \nu) \in \mathscr{L}(\Omega) \mapsto \mu \otimes \nu \in \mathscr{L}(\Omega^2)$ is continuous with respect to the cut-distance.*

The second basic operation resembles the procedure which is used to generate an $n \times n$ (rank one) matrix over Ω from two vectors $\sigma, \tau \in \Omega^n$. More precisely, if $\sigma, \tau \in \Omega^n$ are two vectors we define $\sigma \oplus \tau \in (\Omega^2)^{n \times n}$ as the $n \times n$-matrix with entries $(\sigma \oplus \tau)_{ij} = (\sigma_i, \tau_j)$ for all $i, j \in [n]$. Given two probability distributions $\mu, \nu \in \mathscr{P}(\Omega^n)$, we define $\mu \oplus \nu$ as follows. First, sample $\boldsymbol{\sigma} \sim \mu$ and $\boldsymbol{\tau} \sim \nu$. Second, obtain $\mu \oplus \nu$ as $\boldsymbol{\sigma} \oplus \boldsymbol{\tau}$.

Of course, this operation can be generalised to $\Omega-$laws directly and interestingly, it can be expressed by the generalised product measure $\otimes$. Suppose that κ, κ' are kernel representations of $\Omega-$laws and define

$$\kappa \oplus \kappa' : [0,1]^2 \to \mathscr{P}(\Omega^2), \qquad (s,x) \mapsto \kappa_{s,\Lambda_1(x)} \otimes \kappa_{s,\Lambda_2(x)}.$$

As the $\oplus-$operation boils down to the $\otimes-$operation, it is not hard to guess that it is continuous as well.

Theorem 2.3.9. *The map $\mathscr{L}(\Omega) \to \mathscr{L}(\Omega^2)$, $(\mu, \nu) \mapsto \mu \oplus \nu$ is continuous with respect to the cut-distance.*

This two (very fundamental) operations are everything we need to express the quantity of multi-overlaps quite elegantly. Recall from the introduction that the overlap matrix of two configurations σ, τ was denoted by $\langle \sigma, \tau \rangle$ and expresses on how many particles the spins of σ and τ coincide. This can be naturally extended with regard to two aspects. First, instead of comparing discrete configurations we can calculate the overlap of two generalised configurations $\sigma, \tau \in \Sigma_\Omega$. Second, we can compute the overlap of more than two (generalised) configurations. Thus, let $\sigma_1, \ldots, \sigma_n \in \Sigma_\Omega$ and $\omega_1, \ldots, \omega_n \in \Omega$ and

define

$$R_{\omega_1,\dots,\omega_n}(\sigma_1,\dots,\sigma_n) = \int_0^1 \prod_{i=1}^n \sigma_{i,x}(\omega_i)\mathrm{d}x.$$

Given an $\Omega-$law μ and an integer $\ell \geq 1$ we calculate a similar quantity by averaging over the choice of the configuration through μ, thus

$$R_{\ell,\omega_1,\dots,\omega_n}(\mu) = \int_{\Sigma_\Omega} \cdots \int_{\Sigma_\Omega} R_{\omega_1,\dots,\omega_n}(\sigma_1,\dots,\sigma_n)^\ell \,\mathrm{d}\mu(\sigma_1)\cdots\mathrm{d}\mu(\sigma_n).$$

This finally enables us to define the *multi-overlap* of an $\Omega-$law μ as the array

$$R_{\ell,n}(\mu) = (R_{\ell,\omega_1,\dots,\omega_n}(\mu))_{\omega_1,\dots,\omega_n\in\Omega}.$$

As those multi-overlaps are built by concatenations of $\otimes$ and $\oplus$, it is no surprise that the following holds.

Corollary 2.3.10 (Corollary 1.15 of [47])**.** *The functions $\mu \in \mathscr{L} \mapsto R_{\ell,n}(\mu)$ with $\ell, n \geq 1$ are continuous.*

Let us now come to the last result obtained in [47] which can be seen as one of the most fundamental properties of a consistent limit theory.

Compactness of the space of $\Omega-$laws While [49] already provided an argument that the space $(\mathscr{L}, D_{\boxtimes})$ is a compact metric space by comparing it to the space of decorated graph limits, we give a self-contained argument which is based on previous results within the theory of limits of discrete probability measures.

We will first analyse the space of kernels $\mathscr{K}$. To this end, we define three different variants of the cut-distance. More precisely, we define

$$D_{\boxtimes}(\kappa,\kappa') = \inf_{\phi,\psi\in\mathbb{S}_{[0,1]}} \sup_{\substack{S,X\subset[0,1]\\ \omega\in\Omega}} \left|\int_S\int_X \left(\kappa_{s,x}(\omega) - \kappa'_{\phi(s),\psi(x)}(\omega)\right)\mathrm{d}x\mathrm{d}s\right|,$$

$$D_{\boxslash}(\kappa,\kappa') = \inf_{\phi\in\mathbb{S}_{[0,1]}} \sup_{\substack{S,X\subset[0,1]\\ \omega\in\Omega}} \left|\int_S\int_X \left(\kappa_{s,x}(\omega) - \kappa'_{\phi(s),x}(\omega)\right)\mathrm{d}x\mathrm{d}s\right|,$$

$$D_{\square}(\kappa,\kappa') = \sup_{\substack{S,X\subset[0,1]\\ \omega\in\Omega}} \left|\int_S\int_X \left(\kappa_{s,x}(\omega) - \kappa'_{s,x}(\omega)\right)\mathrm{d}x\mathrm{d}s\right|.$$

Therefore, the strongest version of the cut-distance is $D_{\square}(\cdot,\cdot)$ which does not allow for any measure preserving transformations while the slightly weaker variant $D_{\boxslash}(\cdot,\cdot)$ allows to permute the coordinates corresponding to the generalised configurations. Finally, $D_{\boxtimes}(\cdot,\cdot)$ which we previously studied, is the weakest form of the cut-distance which allows measure preserving transformations in both dimensions.

In a first step we verify that the set of kernels is complete with respect to $D_{\square}(\cdot,\cdot)$ which requires a delicate technical analysis employing the Riesz representation theorem as well as the Radon-Nikodym theorem. Then, the second and third step adapt the well known fact that the quotient space with respect to a linear subspace of each complete metric space is complete itself. Indeed, $(\mathscr{K}, D_{\boxslash})$ is a quotient space of $(\mathscr{K}, D_{\square})$ by identifying kernels κ,κ' with $D_{\boxslash}(\kappa,\kappa') = 0$. Analogously, $(\mathscr{K}, D_{\boxtimes})$ is a quotient space of $(\mathscr{K}, D_{\boxslash})$. By the kernel representation (Theorem 2.3.2) this immediately implies that the space of $\Omega-$laws is complete with respect to $D_{\boxtimes}(\cdot,\cdot)$ as well.

We are left to show that the space of kernels is separable. Fortunately, by the regularity lemma we know that the $\Omega-$laws with finite support are actually a dense subset of $\mathscr{L}$. Therefore, it suffices to prove that Σ_Ω is separable. But the latter is clear as the set of all finite linear combinations of indicator functions $x \mapsto \mathbf{1}\{a < x < b\}$ with $a, b \in \mathbb{Q}$ is a dense subset of all measurable continuous functions from $[0,1] \to \mathbb{R}$. Again employing the kernel representation, we find that this separability carries over to the space of kernels.

Up to now we found that $\mathscr{L}$ is complete and separable but of course this does not necessarily imply compactness. We use the representation of $\pi \in \mathscr{P}(\mathscr{L})$ as a distribution over exchangeable infinite arrays in $\mathscr{E} \subset \mathscr{P}\left(\Omega^{\mathbb{N}\times\mathbb{N}}\right)$ to prove compactness. As the latter space is known to be compact it suffices to find a continuous mapping from $\mathscr{E}$ to $\mathscr{P}(\mathscr{L})$.

Let us sketch how such a mapping could look like. Let $\xi \in \mathscr{E}$ be a distribution over infinite arrays, then we define a mapping $\xi \mapsto \rho^{\xi,n}$ as follows. First, we draw an array $\boldsymbol{X}^\xi$ from ξ. Subsequently, we define $\boldsymbol{\mu}^{\xi,n} \in \mathscr{P}(\Omega^n)$ as the empirical distribution of the rows of the top-left $n \times n$ submatrix of $\boldsymbol{X}^\xi$ and identify it with its embedding into $\mathscr{L}$. Finally, $\rho^{\xi,n}$ is the distribution of $\boldsymbol{\mu}^{\xi,n}$ w.r.t. the choice of $\boldsymbol{X}^\xi$. It turns out that we have for every $\xi \in \mathscr{E}$ that $\rho^\xi = \lim_{n\to\infty} \rho^{\xi,n}$ exists and $\xi \mapsto \rho^\xi$ is continuous [47, Lemma 3.13].

Into the other direction, it is also possible to associate a distribution $\xi^\mu \in \mathscr{E}$ with an $\Omega-$law μ. Indeed, as discussed in the paragraph about the representation as exchangeable arrays, we can simply define ξ^μ as the distribution of $\boldsymbol{X}^\mu$. This mapping actually shows that the space of $\Omega-$laws can be embedded into $\mathscr{E}$ as ρ^{ξ^μ} turns out to be the atom on μ (e.g. $\delta_\mu \in \mathscr{P}(\mathscr{L})$). Again, the corresponding mapping $\mu \mapsto \xi^\mu$ is continuous which directly implies that the mapping $\mathscr{E} \to \mathscr{P}(\mathscr{L})$, $\xi \mapsto \rho^\xi$ is surjective [47, Lemma 3.14, Corollary 3.15]. Altogether, this implies the compactness of the space of $\Omega-$laws.

2.3.1. Summary: the cut-distance for probability measures

We established a self-contained and consistent limit theory for discrete probability measures on Ω^n akin to the graph limit theory. We showed that the limit space $\mathscr{L} = \mathscr{L}(\Omega)$ of $\Omega-$laws is a compact space closely related to the space of graphons.

We furthermore analysed different representations for an $\Omega-$law μ. First, it is possible to define a corresponding kernel, thus a function from the unit square into the probability measures over Ω. Second, each μ is in correspondence with a (random) exchangeable two-dimensional infinite array $\Xi^\mu \in \Omega^{\mathbb{N}\times\mathbb{N}}$.

The latter representation enabled us to prove a sampling lemma comparable to the sampling lemma of graph limit theory. Moreover, we extended the pinning operation of [49] to the more general case of $\Omega-$laws. Finally, we proved that the operation of obtaining multi-overlaps is continuous with respect to the cut-distance which might be an important step towards rigorising some statistical physics' predictions as the (multi-)overlap is a frequently studied observable.

Let us in the next section leave the world of statistical physics once more and discuss results with respect to perturbed random graphs.

2.4. Spanning structures in randomly perturbed sparse graphs

Recall that in the setting of randomly perturbed graphs some arbitrary (possibly deterministic) graph $\mathscr{G}_\alpha = (V_\alpha, E_\alpha)$ with minimum degree αn is given. Furthermore, we take the edges from an instance $\boldsymbol{G}$ of $\mathscr{G}(n,p)$ and examine whether there are certain spanning structures present in $\mathscr{G}_\alpha \cup \boldsymbol{G}$ with high probability. We start by stating the obtained results.

Theorem 2.4.1 (Theorems 1.1 and 1.2 of [93])**.** *Let $\mathscr{G}_\alpha$ be a graph with minimum degree αn and $\boldsymbol{G}$ an instance of $\mathscr{G}(n,\beta/n)$.*

- *If $\beta \geq -(6+o(1))\ln\alpha$, $\mathscr{G}_\alpha \cup \boldsymbol{G}$ contains a Hamilton cycle with high probability.*
- *If $\beta \geq -(4+o(1))\ln\alpha$, $\mathscr{G}_\alpha \cup \boldsymbol{G}$ contains a perfect matching with high probability.*

Let us first observe that the theorem is tight (up to a constant factor). Indeed, if $\mathscr{G}_\alpha$ is the complete bipartite graph on classes $V_\alpha, V_{1-\alpha}$ of size αn and $(1-\alpha)n$, we cannot find a perfect matching if there is an independent set of size larger than αn in $V_{1-\alpha}$. It is known that the number of isolated vertices in the random graph is $\sim n\exp\left(-\beta\right) \gg \alpha n$ if $\beta = o(-\ln\alpha)$. Therefore, one requires $\beta = \Omega(-\ln\alpha)$, and of course, if there is no perfect matching, we cannot find a Hamilton cycle either.

Furthermore, we obtained results with respect to the existence of spanning bounded degree trees. More precisely, we state some kind of a *meta-theorem* which allows – given an almost spanning structure of sufficient size in $\mathscr{G}(n,\beta/n)$ – to find the spanning structure by the edges in the deterministic graph.

Theorem 2.4.2 (Theorem 1.6 of [93]). *Let $\Delta \geq 2$ be an integer and suppose that $\alpha, \beta, \varepsilon \colon \mathbb{N} \mapsto [0,1]$ are such that $4(\Delta+1)\varepsilon < \alpha^{\Delta+1}$. Furthermore suppose that $\mathcal{G}\left(n, \beta/n\right)$ contains a given tree with maximum degree Δ on $(1-\varepsilon)n$ vertices w.h.p. and that $\mathcal{G}_\alpha$ is an arbitrary graph with minimum degree αn.*

Then any tree with maximum degree Δ on n vertices can be found in $\mathcal{G}_\alpha \cup \mathcal{G}\left(n, \beta/n\right)$ with high probability.

The theorem solely does of course not answer the question whether we find spanning trees or not. But with a recent result of Balogh et al. [19], we find the following.

Corollary 2.4.3 (Corollary 1.8 of [93]). *For $\Delta \geq 2$ there exists $C > 0$ such that for $\alpha = \alpha(n) : \mathbb{N} \mapsto (0,1)$ and $\beta = \beta(\alpha) = -C\alpha^{-(\Delta+1)} \ln \alpha$ the following holds. Any n-vertex tree T with maximum degree Δ is a subgraph of $\mathcal{G}_\alpha \cup \mathcal{G}\left(n, \beta/n\right)$ with high probability.*

How can we proof such statements? We will only give an idea of the proof with respect to the Hamilton cycle as the other theorems follow fairly similar ideas. We need two major ingredients. First, we require a long cycle in the random graph. Fortunately, such a result already exists.

Lemma 2.4.4 (Frieze [83]). *Let $0 < \beta = \beta(n) \leq \ln n$. Then the random graph $\mathcal{G}(n, \beta/n)$ contains a cycle of length at least $\left(1 - (1-o(1))\,\beta \exp\left(-\beta\right)\right) n$ with high probability.*

Second, we will use the previously explained multiple round exposure technique with the difference that we in this case reveal edges of the random graph in multiple rounds instead of infected individuals like in the group testing problem. Now we already have everything at hand to sketch the proof of (the first part of) Theorem 2.4.1.

Proof sketch of the first part of Theorem 2.4.1. Observe that for very small $\alpha = O(n^{-1/6})$ our choice of β already guarantees that the random graph is known to contain a Hamilton cycle w.h.p..

Thus, suppose that $\alpha = \omega(n^{-1/6})$. We start by revealing almost all edges of the random graph. More precisely, we reveal the edges of $\mathcal{G}(n, (\beta-1)/n)$.

Within this random graph we find by the previous lemma a path P on all but at most $\beta \exp\left(-\beta\right) n$ vertices. Say that those left-over vertices are denoted by V'. We will subsequently *absorb* all but two such vertices onto the path P using edges from the deterministic graph $\mathcal{G}_\alpha$.

To this end, let $\boldsymbol{B}(u,v)$ denote the set of vertices x that lie on P and are a neighbour of u in $\mathcal{G}_\alpha$ such that the neighbours of u on the path P are also connected to v via edges in $\mathcal{G}_\alpha$. Formally,

$$\boldsymbol{B}(u,v) = \left\{x \in \partial_{\mathcal{G}_\alpha}(u) \cap P \mid \partial_P(x) \subset \partial_{\mathcal{G}_\alpha}(v)\right\}.$$

A visualisation is given in Figure 2.8. Suppose $P = p_1 \dots p_\ell$ is the path. Then clearly, if for a vertex

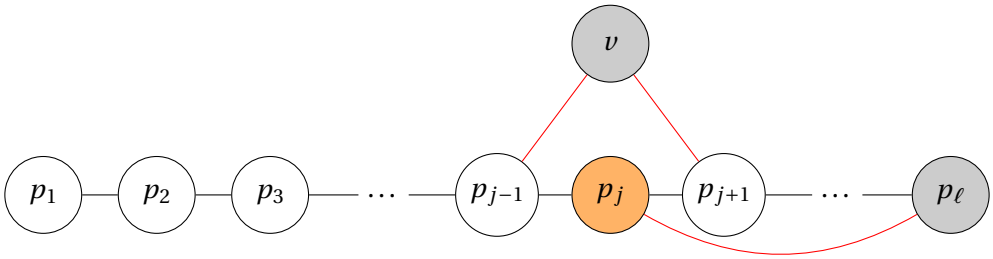

Figure 2.8.: Absorbing structure for vertex v onto the path $p_1 \dots p_\ell$ using edges from $\mathcal{G}_\alpha$ (red). The graphic is modified after [93].

$v \in V'$ there is some $p_j \in \boldsymbol{B}(p_\ell, v)$, we can create a longer path $p_1 \dots p_{j-1} v p_{j+1} \dots p_\ell p_j$ which contains v. It turns out that, up to some technicalities which guarantee that those absorbing structures do not overlap too much, we can follow this approach greedily until all but 2 vertices are absorbed. Indeed, as $\boldsymbol{B}(u,v)$ is a uniformly at random chosen set of vertices due to the randomness of $\mathcal{G}(n, (\beta-1)/n)$, we find

$$|\boldsymbol{B}(u,v)| \geq \frac{2}{5}\alpha^3 \beta \exp\left(-\beta\right) \geq \frac{\alpha^3 n}{4}$$

with high probability for the choice of β. While absorbing the left-over vertices greedily for

$$\left|V'\right| - 2 \leq \beta \exp\left(-\beta\right) n$$

rounds, we remove in every step all used vertices from all sets $\boldsymbol{B}(u,v)$ and the absorbed vertex from V'. As after $\beta \exp\left(-\beta\right) n$ rounds the size of $\boldsymbol{B}(u,v)$ only decreased by at most $\beta \exp\left(-\beta\right) n$ (for any pair of vertices u, v), we have after absorbing that

$$|\boldsymbol{B}(u,v)| \geq \frac{\alpha^3 n}{4} - \beta \exp\left(-\beta\right) n \geq \frac{\alpha^3 n}{8}$$

if $\alpha = \omega\left(n^{-1/6}\right)$. Thus, absorbing all but two vertices is indeed possible to be done greedily.

Therefore, all we need to do at this point is to close the cycle. Recall that we obtained a path $P = p_1 \dots p_{n-2}$ and are left with two additional vertices a, b. If there is an edge $p_i p_j$ between $\boldsymbol{B}(p_1, a)$ and $\boldsymbol{B}(p_{n-2}, b)$, we already found a Hamilton cycle $C = p_i p_1 \dots p_{i-1} a p_{i+1} \dots p_{j-1} p_{j+1} \dots p_{n-2} p_j p_i$. And indeed, by revealing the missing edges from $\mathcal{G}(n, 1/n)$, we find that such an edge is present with high probability. □

3. Outlook

This chapter will state open problems related with this thesis's contributions and aims for giving rise to possibly interesting further research directions within the respective fields. We start by stating such research questions with respect to the group testing problem.

3.1. Group testing

Non-adaptive group testing in the sublinear regime In the sublinear regime where the number of infected individuals scales as $k = n^\theta$, this thesis's contributions draw a somehow complete picture of non-adaptive noiseless hypergeometric group testing. While the analysis of SPIV and the proof of the universal information theoretic converse does not require exact knowledge about k but its order of magnitude suffices, it seems very likely that those results might carry over to i.i.d. models as well. Nevertheless, while SPIV is an efficient algorithm from a theoretical point of view, requiring $\ln\ln n \to \infty$ shows that the result is only of theoretical nature. Furthermore, the spatially coupled design is clearly much more complicated than the random regular model. Therefore, an intriguing question would be the following.

Question 3.1.1. *Is there an efficient algorithm succeeding at $m_{\text{non-ada}}$ on the random regular model? Furthermore, is there a deterministic graph G with $(1+\varepsilon)\, m_{\text{non-ada}}$ tests coming with an efficient algorithm for high probability recovery of $\boldsymbol{\sigma}$ from $(G, \hat{\boldsymbol{\sigma}})$?*

Besides searching for more practical algorithms, one could ask what happens when we leave the Bayes optimal setting. Suppose the student does only receive an estimate $\tilde{k}$ of k (thus, not the complete teacher's prior). If $\tilde{k} > k$, inference will still be possible by SPIV as we conduct only too many tests and each test is more likely to be negative. If, on the other hand, k exceeds the guess $\tilde{k}$, complete recovery is not possible under the studied designs. Therefore, the following question might arise.

Question 3.1.2. *Given an estimate $\tilde{k} < k$ of the number of infected individuals, how many tests suffice to infer the ground-truth completely or at least up to small errors depending on $\tilde{k}$ and k? How would an optimal design look like?*

Similarly, while we proved that $(1-o(1))-$recovery is possible at m_{inf} using the spatially coupled design with SPIV, it is also known that, information-theoretically, it is possible to achieve $(1-\gamma)-$recovery on a Bernoulli test-design with no more than m_{inf} tests. Up to our knowledge, there are only non-rigorous contributions which provide evidence that it might be possible algorithmically [99]. Further, there cannot be any design on less than $(1-\gamma)\, m_{\text{inf}}$ tests achieving $(1-\gamma)-$ recovery [152].

Question 3.1.3. *Is partial recovery of all but γk infected individuals possible algorithmically under Bernoulli group testing at the information-theoretic threshold?*

But nevertheless, this part of group testing is fairly well understood. Things are completely different for sparsity constrained group testing.

Sparsity constrained group testing As already seen in the previous chapter, the sparsity constrained group testing problem is not as well understood as the unrestricted case. A fairly natural question is if the phase diagram in the $\Delta-$divisible case does actually look similar to the one in unrestricted group testing.

Question 3.1.4. *Is there an adaptive algorithm testing each individual at most Δ times succeeding at $m_{\text{inf}}(\Delta)$ or can we proof that $m_{\text{inf}}(\Delta)$ is not tight? Furthermore, is there a spatially coupled design coming with an efficient algorithm performing as well as the binary splitting approach (or even better)?*

In the $\Gamma-$sparse case we completely understood the problem if Γ is a constant.

Question 3.1.5. *Does the analysis of the* $\Gamma = \Theta(1)$ *case extend to larger values of* Γ*? And, if so, does it strengthen existing bounds?*

Finally, one might be slightly irritated by the fact that the achievability and converse bounds in the $\Delta-$divisible case do not converge to their unrestricted counterpart.

Question 3.1.6. *How does the group testing problem behave in the critical regime* $\omega(\ln^{1-\delta} n) = \Delta = o(\ln n)$ *and what happens at the phase transition* $\Delta \to \ln n$*?*

Furthermore, there is already extensive work on noisy group testing.

Noisy group testing Without going too much in detail, the state of the play is comparable to the state of the art in noiseless non-adaptive group testing prior to this thesis's contributions. While simple algorithms are well understood [89, 108, 155], it is even not known under simplistic noise models like the binary symmetric channel if recovery of the ground-truth is possible efficiently at the Shannon capacity bound. It might be tempting to analyse whether a spatially coupled design could improve bounds in the noisy setting as well.

Question 3.1.7. *Does a spatially coupled design coming with a* `SPIV` *like algorithm perform better at inference under a noisy setting than currently known algorithms?*

Returning to the noiseless case, we find that even the linear case is not completely understood.

Linear group testing More precisely, while non-adaptive algorithms need to fail with high probability, it is possible to infer the ground-truth within multiple rounds of testing at the universal converse [13, 98]. Nevertheless, one important question is still open.

Question 3.1.8. *Are there any (potentially exponential-time) algorithms on less than* $n-1$ *tests which succeed at inference of the ground-truth if the prevalence is larger than* $1/3$ *in the hypergeometric group testing problem? Or conversely, can we proof that such algorithms do not exist?*

A bit more application-driven question is the following. Which influence do such asymptotic designs have on real world group testing?

Applications Suppose we have $k = 5$ infected individuals within a population of $n = 1000$. Then the following statement sounds correct.

> By the given prevalence of 0.5% we are clearly in the setting of linear group testing. Therefore, any non-adaptive group testing strategy fails due to Aldridge [7] and we need 1000 tests for inference within one round.

While this seems to be indeed true, let us provide a second statement which might also sound plausible.

> As the number of infected individuals k scales like $n^{0.233}$, we require $\frac{1}{\ln^2 2} \cdot 5 \cdot \ln(200) \approx 56$ tests such that `DD` infers the infected individuals correctly.

Finally, the folklore counting bound yields that we require $2^m > \binom{1000}{5}$, thus we need at least $m \geq 43$ tests.

Of course, above's statements do not only contradict each other, they are also false. All provided proofs in all contributions (besides the universal counting bound) are obtained under the assumption that n tends to ∞. Therefore, all we can say is that something in between 43 and 1000 tests will be the correct answer. But this is of course very unsatisfactory.

Question 3.1.9. *Given a real world instance on n individuals with an infection rate of* α *with a realistic false positive and false negative rate, how could an almost optimal non-adaptive or two-stage design look like in order to infer almost all individuals correctly?*

An answer to this question would be of high interest in applications. First intends are done by, for instance, Aldridge [12] and Cuturi et al. [57], but we are far from knowing precise statements. A natural suggestion is that an inference algorithm based on Belief Propagation might facilitate best-possible.

This question directly extends to different inference problems. It might be a fruitful research direction to apply message passing algorithms coming with a spatially coupled design to various inference problems. Unfortunately, in general it is much harder to find combinatorial descriptions of those message passing algorithms as in the group testing problem. Therefore, they might be highly challenging to analyse.

3.2. Random satisfiability

As described earlier, we could exactly pin down the number of satisfying assignments of a random 2−SAT formula in terms of the Bethe functional. A natural question would be if the result can be extended to general random k−SAT formulas. It is at least far from clear if this is possible. Indeed, the core of our proof technique is the possibility to construct worst-case conditions at the boundary of a random tree in order to prove that the Boltzmann distribution is a Bethe state itself. Even in a random 3−SAT formula it is not clear how such worst-case conditions might be constructed as the number of possibilities to nudge the marginal of a parent variable into a certain direction is huge and dependent on decisions in different sub-trees.

Within the setting of random 2−SAT we could furthermore ask if we can pin down the distribution of $\ln Z(\boldsymbol{\Phi})$ exactly.

Question 3.2.1. *Does* $n^{-1/2}(\ln Z(\boldsymbol{\Phi}) - \mathbb{E}\ln Z(\boldsymbol{\Phi}))$ *converge to a Gaussian random variable?*

3.3. The cut-distance, regularity and limits of probability measures

We managed to develop a consistent limit theory for discrete probability measures akin to the graph limit theory. Furthermore, we introduced the pinning operation on the limit objects (Ω−laws) as an elegant and easy algorithm to obtain a decomposition of the phase state on which the respective probability measures are extremal, thus close to product measures under the cut-distance. This is a kind of a regularity lemma which allows to write a probability measure as a convex combination of simple measures.

There is a second string of research, primarily by Austin [17], which develops similar regularity lemmas based on the so-called *dual total correlation (DTC)*. The latter can be seen as a generalisation of the classical mutual information from 2 to n random variables which reads as

$$DTC(\mu) = H(\boldsymbol{X}_1,\dots,\boldsymbol{X}_n) - \sum_{i=1}^{n} H(\boldsymbol{X}_i \mid \boldsymbol{X}_1,\dots,\boldsymbol{X}_{i-1},\boldsymbol{X}_{i+1},\dots,\boldsymbol{X}_n)$$

with $H(\cdot)$ being the entropy and $\boldsymbol{X}_1,\dots,\boldsymbol{X}_n$ random variables with joint distribution μ.

Described very briefly, it says that a small dual total correlation implies that a probability measure is close to a product measure with respect to a specific transportation metric [17, Theorem A]. Further, a regularity lemma is known which guarantees that there is a partition of the phase space which allows to write a probability measure as a convex combination of measures with low dual total correlation [17, Theorem 1.1]. An intriguing question is the following.

Question 3.3.1. *Is there a natural connection between the cut-distance of* μ *and* $\bar{\mu}$ *and the dual total correlation of* μ*?*

A second and a third question regarding regularity arise with respect to the pinning operation itself. First, we need to pin a random number of coordinates in order to achieve an extremal measure. It is not clear whether this part of the statement has a deeper reason or if it is just a relic of the proof technique which is originally based on a contribution of Raghavendra and Tan [148]. Furthermore, the pinning

lemma yields a sufficient condition on how many variables have to be pinned to specific spins. A lower bound on how many variables are necessary is currently unknown.

Question 3.3.2. *What is a lower bound on the number of variables which are necessary to be pinned in the pinning operation? Is it necessary that this number is random?*

3.4. Perturbed graphs

The note on spanning structures in randomly perturbed sparse graphs establishes sufficient results for Hamilton cycles and matchings in graphs that are optimal up to a constant. Furthermore, it gives a meta-theorem which proves that the existence of almost spanning bounded degree trees in $\mathcal{G}(n,\beta/n)$ suffices in order to gain the complete spanning structure in $\mathcal{G}(n,\beta/n)$ for $\beta = -\Omega((\Delta+1)\ln(\alpha))$ if Δ is the maximum degree of the tree. Thus, the first question which was also stated in the note reads as follows.

Question 3.4.1. *Is any given tree with maximum degree Δ on $(1-C\exp(-\beta))n$ vertices contained in $\mathcal{G}(n,\beta/n)$ for $0<\beta\leq\ln n$ and a suitable chosen constant C?*

Moreover, we already discussed that the sufficient conditions proven in the note are probably not tight but the proof technique yields to those expressions. It is likely to be true that the choice of other, more complicated, absorbing structures and a more carefully applied large deviations analysis might yield to better constants in the theorems. Therefore, it is an interesting question to pin down the exact phase transition points.

Question 3.4.2. *Can we establish strict phase transitions for the existence of a Hamilton cycle and a perfect matching in $\mathcal{G}(n,\beta/n)\cup\mathcal{G}_\alpha$?*

Finally, while we only studied certain spanning structures in graphs, the objects of interest can be extended. First, we can study different spanning structures in graphs like for instance triangle factors. Second, one can extend the analysis from perturbed sparse graphs to perturbed sparse hypergraphs.

4. Zusammenfassung

Die Analyse sehr großer diskreter Systeme ist ein wesentlicher Bestandteil aktueller Forschung in, unter anderem, der Diskreten Mathematik, der Informatik sowie der Statistischen Physik. Von besonders hoher Relevanz ist das Phänomen der *Phasenübergänge* [78, 82, 121]. Hierbei handelt es sich um Momente in der Evolution eines Systems, an welchen sich dessen Verhalten dramatisch verändert. Solche Phasenübergänge wurden und werden beispielsweise in zufälligen Graphen, Spin-Gläsern und bezüglich der Performance von Algorithmen untersucht.

Seit einigen Jahrzehnten kristallisiert sich immer weiter heraus, dass einige Ideen der Statistischen Physik auf die Untersuchung von eigentlich rein kombinatorischen Problemen übertragen werden können [168]. Zu solchen Problemen zählen zum Beispiel die Beantwortung der Frage nach Erfüllbarkeit gegebener zufälliger Formeln oder auch die Analyse der algorithmischen und informationstheoretischen Lösbarkeit von Inferenzproblemen wie dem *Group-Testing* Problem [68, 133]. Eine wesentliche Herausforderung aus Sicht der Mathematik besteht darin, die Ideen der physikalischen Heuristiken in rigorose mathematische Verfahren und Beweise zu übersetzen.

In dieser Dissertation untersuchen wir, wie Ideen aus der physikalischen Theorie der *diluted mean-field models* für Spin-Gläser dazu genutzt werden können, zufällige Erfüllbarkeitsprobleme zu analysieren. Insbesondere nutzen wir diese Ideen, um zu berechnen, wie viele Lösungen eine zufällige 2-SAT Formel in der Regel besitzt. Zudem werden wir die sogenannte *planted* Version von solchen zufälligen Erfüllbarkeitsproblemen untersuchen. Es stellt sich durch die Brille der Statistischen Physik betrachtet heraus, dass solche Modelle genutzt werden können, um statistische Inferenzprobleme auszudrücken [168].

Sowohl durch die Analyse von Problemen der statistischen Inferenz als auch durch das Untersuchen der zufälligen Erfüllbarkeit stellen wir fest, dass eine geschickte Kombination von Ideen der statistischen Physik mit ureigenen kombinatorischen Eigenschaften zufälliger Graphen zu rigorosen, neuen Resultaten führt. Wir beginnen mit einer knappen Einführung in wesentliche Begriffe der statistischen Physik.

Grundlagen der statistischen Physik Gegeben sei eine Menge $V = \{x_1, \dots, x_n\}$ von *n Partikeln*. Ein *Spin-System* mit diesen Partikeln besteht aus einer endlichen Menge Ω von möglichen Spins sowie einem $k-$uniformen dekorierten Hypergraphen $G = (V, E, J)$, welcher die Interaktionen zwischen den Partikeln beschreibt. Genauer gesagt bezeichnen wir mit $\sigma \in \Omega^n$ eine *Konfiguration*, welche jedem Partikel einen der möglichen Spins aus Ω zuordnet und mit $H : \Omega^n \to \mathbb{R}$ eine *Energie-Funktion*, die jeder Konfiguration ihre Energie zuweist. Formal definieren wir die Energie-Funktion H als

$$H(\sigma) = - \sum_{(i_1,\dots,i_k)\in E(G)} J_{i_1,\dots,i_k}(\sigma_{i_1}, \dots, \sigma_{i_k}).$$

Im Spezialfall $k = 1$ bezeichnen wir das System als ein nicht-interagierendes System, da die verschiedenen Partikel gar nicht miteinander interagieren, während höhere Werte von k ein sogenanntes $k-$*body interacting* System beschreiben. In dieser Dissertation werden wir nur den Fall $k = 2$ untersuchen, das heißt, der zugrunde liegende Hypergraph G ist ein einfacher Graph. Die Familie $\{J_{i,j}\}_{ij\in E(G)}$ beschreibt die Stärke und Art der Interaktion der interagierenden Partikel.

Jede Wahl der *Coupling-Konstanten* J_{ij} sowie der Energie-Funktion H und des Interaktionsgraphen G beschreibt ein spezifisches Spin-System. Klassische Systeme, wie das Potts-Modell oder das Edwards-Anderson-Modell sind Modelle, in denen G einem Gitter-Graphen (zum Beispiel $\mathbb{Z}^3$) entspricht. Diese Modelle mögen die naheliegensten Interaktionsmodelle für zum Beispiel Ferromagnetismus sein, allerdings sind diese auf Grund der geometrischen Beziehungen im Graphen mathematisch sehr schwer zu analysieren [135]. Eine mögliche Vereinfachung sind sogenannte *mean-field*-Modelle wie das SK-Modell [158], in denen der Interaktionsgraph dem vollständigen Graphen entspricht. Ferner wird eine Energie-Funktion gewählt, deren Wert invariant gegenüber Permutationen der Partikel ist. Auf diese Art und Weise werden einfachere Modelle definiert, welche die lokale und globale Struktur sowie Abhängig-

keit verschiedener Partikel ignorieren. Es stellt sich heraus, dass diese Modelle einfacher zu analysieren sind als ihre zugehörigen Gitter-Pendants [143, 162], allerdings wichtige physikalische Eigenschaften nicht korrekt beschreiben können [157].

Die sogenannten *diluted mean-field* Modelle, wie das Viana-Bray-Modell [166], versuchen die mathematische Einfachheit der mean-field-Theorie möglichst beizubehalten und dennoch wesentliche Eigenschaften realer Systeme möglichst gut zu modellieren. Der Kerngedanke ist, dass der zugrunde liegende Interaktionsgraph ein (zufälliger) dünner Graph ist, das heißt, dass zwar die globale Geometrie sicherlich nicht gegeben ist, der Graph lokal aber an einen Gittergraphen erinnert. Tatsächlich ist es so, dass die Analyse gut gewählter *diluted mean-field* Modelle zu exakten Lösungen von Problemen in (echten) Spin-Glas-Modellen führen kann [144], das heißt, diese Modelle tragen noch wesentliche Informationen von realitätsnahen Modellen in sich, während sie mathematisch analysierbar bleiben.

Außerdem können wesentliche Probleme der Informatik, wie das *zufällige Erfüllbarkeitsproblem* und weitere *Constraint Satisfaction*-Probleme (CSPs), als *diluted mean-field* Modell beschrieben werden.

Bevor wir uns diesem Punkt zuwenden, werden wir noch eine wichtige Wahrscheinlichkeitsverteilung, die sogenannte *Boltzmann-Verteilung*, einführen. Dazu erinnern wir uns an ein wesentliches Prinzip der Physik: Ein System versucht immer einen Zustand minimaler Energie einzunehmen. Auch Modelle von Partikelsystemen sollten diese Eigenschaft widerspiegeln. Wir sagen, dass das System in einem Zustand minimaler Energie ist, wenn $H(\sigma)$ durch die Konfiguration $\sigma \in \Omega^n$ minimiert wird. Außerdem bezeichnen wir mit $\Omega_0^n \subset \Omega^n$ die Menge der Konfigurationen, welche die Energie minimieren. Eine vernünftige und natürliche Wahrscheinlichkeitsverteilung sollte also Konfigurationen in Ω_0^n bevorzugen.

Dazu führen wir die *inverse Temperatur* $\beta > 0$ ein und definieren die Boltzmann-Verteilung auf Ω^n als

$$\mu_\beta(\sigma) = \frac{\exp\left(-\beta H(\sigma)\right)}{Z_\beta}, \qquad \text{wobei} \qquad Z_\beta = \sum_{\sigma \in \Omega^n} \exp\left(-\beta H(\sigma)\right).$$

Die Normalisierungskonstante Z_β der Boltzmann-Verteilung wird auch *Partitionsfunktion* genannt. Die Boltzmann-Verteilung spiegelt die Idee, dass das System einen Zustand minimaler Energie anstrebt, wider. Je geringer die Energie $H(\sigma)$, desto höher ist die Wahrscheinlichkeit σ unter μ_β zu beobachten. Steigt die Temperatur des Systems stark an (β wird klein), so verringert sich dieser Effekt und für $\beta \to 0$, also im *high-temperature limit*, wird die Boltzmann-Verteilung zur uniformen Verteilung auf Ω^n. Wächst β hingegen, wird das System also abgekühlt, so verstärkt sich o.g. Effekt, sodass die Boltzmann-Verteilung im *zero-temperature limit* die uniforme Verteilung auf allen Zuständen minimaler Energie wird. Formal ausgedrückt finden wir die folgenden Zusammenhänge:

$$\lim_{\beta \to 0} \mu_\beta(\sigma) = \frac{1}{|\Omega^n|} \qquad \text{und} \qquad \lim_{\beta \to \infty} \mu_\beta(\sigma) = \frac{\mathbf{1}\left\{\sigma \in \Omega_0^n\right\}}{\left|\Omega_0^n\right|}.$$

Da in der statistischen Physik (und auch in der theoretischen Informatik) oft das makroskopische Verhalten von sehr großen Systemen ($n \to \infty$) von Interesse ist, wird häufig der sogenannte *thermodynamische Grenzwert* eines Systems betrachtet [135]. Wir bezeichnen mit $\phi_{n,\beta} = \ln(Z_\beta)$ die *freie Entropie* und mit $\phi_\beta = \lim_{n\to\infty} \frac{\ln(Z_\beta)}{n}$ die *freie Entropiedichte* (*free entropy density*). Innerhalb eines physikalischen Systems definieren wir nun die nicht-analytischen Punkte von ϕ_β als *Phasenübergänge*. Das sind Punkte, an denen sich das qualitative Verhalten des Systems drastisch ändert [135]. Wir betrachten insbesondere Phasenübergänge von den bereits angesprochenen CSPs.

Constraint Satisfaction und Phasenübergänge Zunächst definieren wir, was wir unter CSPs verstehen. Ein besonders prominentes Beispiel ist das $k-$SAT Problem, also die Frage nach Erfüllbarkeit einer aussagenlogischen Formel in konjunktiver Normalform mit Klauseln der Größe k. Genauer gesagt ist eine $k-$SAT Formel Φ eine Konjunktion von m Klauseln

$$\Phi = \Phi_1 \wedge \ldots \wedge \Phi_m,$$

sodass jede Klausel selbst eine Disjunktion von exakt k Literalen der Variablen $x_1 \ldots x_n$ ist.

Ist eine Formel Φ gegeben, so lautet eine der wichtigsten Fragen offenbar, ob es eine Belegung $\sigma : \{x_1, \ldots, x_n\} \to \{-1, +1\}^n$ gibt, welche jeder der Variablen einen der Wahrheitswerte WAHR und FALSCH zuordnet, sodass jede Klausel Φ_j (und somit die gesamte Formel Φ) erfüllt ist. Wir interpretieren den Spin +1 als WAHR und den Spin -1 als FALSCH. Eine solche Formel Φ kann als *Faktorgraph* visualisiert werden.

Ein Faktorgraph $G = (V \cup F, E, \Psi)$ ist ein bipartiter Graph mit einer Menge von Variablenknoten V und einer Menge von Faktorknoten F, einer Kantenmenge E sowie einer Familie von Gewichtsfunktionen Ψ [81].

Wir beschreiben im Folgenden eine Konstruktionsanweisung für einen Faktorgraphen G^Φ, der die $k-$SAT Formel Φ repräsentiert [135]. Seien die Knotenmengen als

$$V = \{x_1, \ldots, x_n\} \qquad \text{sowie} \qquad F = \left\{a_1^\Phi, \ldots, a_m^\Phi\right\}$$

gegeben. Ferner besteht die Kantenmenge $E = \left\{a_1^\Phi, \ldots, a_m^\Phi\right\}$ aus zwei disjunkten Klassen E^+ und E^-, sodass die Kante $x_i a_j^\Phi$ genau dann in E^- liegt, wenn Variable x_i negiert in Klausel Φ_j vorkommt, und in E^+, falls x_i positiv in Φ_j enthalten ist. An jedem Faktorknoten a_j existiert eine lokale Funktion $\Psi_{a_j} : \{-1, +1\}^{|\partial a_j|} \to \{-1, +1\}$, sodass für eine Belegung $\sigma \in \{-1, +1\}^n$ das Folgende gilt:

$$\Psi_{a_j}(\sigma_{\partial a_j}) = \mathbf{1}\left\{\max_{x_i \in \partial a_j} \{\sigma_i s_{ij} = 1\}\right\}.$$

Eine Visualisierung eines solchen Faktorgraphen findet sich in Abbildung 4.1.

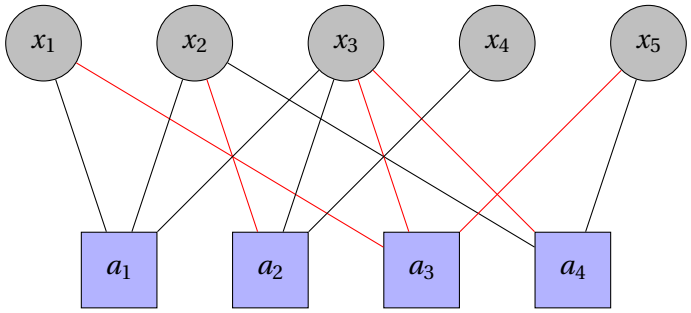

Abbildung 4.1.: Der Faktor-Graph G^Φ zur 3−SAT Formel $\Phi : (x_1 \vee x_2 \vee x_3) \wedge (\neg x_2 \vee x_3 \vee x_4) \wedge (\neg x_1 \vee \neg x_3 \vee \neg x_5) \wedge (x_2 \vee \neg x_3 \vee x_5)$. Die $n = 5$ Variablenknoten sind als Kreise dargestellt, die $m = 4$ Faktorknoten als Rechtecke und die Farbe der Kanten beschreibt, ob eine Variable negiert oder nicht negiert in einer Klausel vorkommt.

Während für $k \geq 3$ das Entscheidungsproblem, ob eine $k-$SAT Formel mindestens eine erfüllende Belegung besitzt, ***NP***-schwer ist [109], kann es leicht als physikalisches System aufgefasst werden [111]. Die n Variablen entsprechen den Partikeln des Systems und die Wahrheitswerte den Spins $\Omega = \{-1, +1\}$. Eine Belegung entspricht nun einer Konfiguration $\sigma \in \Omega^n$ und eine mögliche Energie-Funktion ist

$$H_{k-SAT}(\sigma) = \sum_{a_j^\Phi \in F} 1 - \Psi_{a_j}(\sigma_{\partial a_j}).$$

Also entspricht $H_{k-SAT}(\sigma)$ genau der Anzahl an nicht-erfüllten Klauseln unter einer gegebenen Konfiguration. Betrachten wir nun den *zero-temperature limit*, so ist die entsprechende Boltzmann-Verteilung $\mu_\infty = \lim_{\beta \to \infty} \mu_\beta$ die uniforme Verteilung auf solchen Konfigurationen, die am wenigsten Klauseln verletzen. Das heißt, wenn $\boldsymbol{\sigma} \sim \mu_\infty$ eine zufällige Konfiguration (gezogen von μ_∞) ist, so ist Φ genau dann erfüllbar, wenn $H_{k-SAT}(\boldsymbol{\sigma}) = 0$ gilt. Da wir wissen, dass $k-$SAT ***NP***-schwer ist, folgt direkt, dass es im Allgemeinen auch schwer ist, eine Konfiguration minimaler Energie in einem Spin-System zu finden.

An dieser Stelle fällt auf, dass sich die Boltzmann-Verteilung eines CSPs sehr elegant faktorisieren

lässt, das heißt,

$$\mu_\infty(\sigma) = \lim_{\beta\to\infty} \frac{\prod_{a_j^\Phi \in F} \exp\left(-\beta \mathbf{1}\left\{a_j^\Phi \text{ is not satisfied under } \sigma\right\}\right)}{Z(\Phi)}.$$

Das ist keine Eigenart spezieller CSPs, sondern eine sehr universelle Eigenschaft aller CSPs, die sich als Faktorgraph ausdrücken lassen. Wir bemerken, dass ...

- ... jeder Faktor in der Boltzmann-Verteilung zu einem Faktorknoten in G^Φ korrespondiert,
- ... für $\beta > 0$ jede nicht-erfüllte Klausel eine Art Strafe von $\exp(-\beta)$ auf die Wahrscheinlichkeit, diese Konfiguration zu beobachten, addiert.
- ... falls Φ erfüllbar ist, der Träger der Boltzmann-Verteilung im *low-temperature limit* den erfüllenden Belegungen einer Formel entspricht.
- ... falls Φ erfüllbar ist, die Partitionsfunktion im *low-temperature limit* der Anzahl der erfüllenden Belegungen entspricht.

Oft werden im Kontext von CSPs sogenannte zufällige CSPs betrachtet [156]. Was meint zufällig hierbei? Sind n Variablenknoten und $\boldsymbol{m}$ Faktorknoten (wobei $\boldsymbol{m}$ durchaus auch zufällig sein kann) sowie deren Grade gegeben, so bilden wir einen zufälligen bipartiten Graphen. Je nach CSP können die Gradsequenzen der Knoten selbst zufällig sein. Sind die Gradsequenzen gegeben, so ziehen wir uniform und unabhängig von allem anderen Zufall einen Faktorgraphen, der die entsprechenden Gradsequenzen hat.

Im Falle von $k-$SAT werden n Variablenknoten und m Faktorknoten gegeben. Jeder Faktorknoten hat Grad k und jede Variable hat Grad $\mathbf{Po}(mk/n)$. Wir wählen, gegeben $\left\{\sum_{i=1}^n \boldsymbol{d}_i = mk\right\}$, einen zufälligen einfachen Graphen mit den gegebenen Gradsequenzen. Ferner markieren wir jede Kante unabhängig und uniform mit $+1$ mit Wahrscheinlichkeit $1/2$ und mit -1 mit Wahrscheinlichkeit $1/2$. Eine sehr einfache Frage lautet: Ist für $n \to \infty$ die entstandene zufällige Formel mit hoher Wahrscheinlichkeit erfüllbar? Diese Frage wurde exzessiv untersucht, unter anderem auch mit Methoden der statistischen Physik [37, 111], und es wurde eine präzise Vermutung für ein kritisches Verhältnis $\alpha_s = m_s/n$ zwischen Klauseln und Variablen formuliert, sodass eine zufällige Formel mit geringerem Verhältnis erfüllbar und mit höherem Verhältnis nicht erfüllbar ist. Wir haben also einen Phasenübergang gefunden. Genauer gesagt wurde dieses Problem aufbauend auf vielfältigen Arbeiten [4, 51, 54, 91] schlussendlich von Ding, Sly und Sun [64] gelöst, die beweisen konnten, dass für k groß genug der *Erfüllbarkeits-Schwellenwert* bei

$$\alpha_s = 2^k \ln 2 - \frac{1+\ln 2}{2} + O(2^{-k})$$

liegt. Natürlich finden solche Phasenübergänge auch in allgemeinen zufälligen CSPs statt [136] und der Erfüllbarkeits-Schwellenwert ist nicht der einzige interessante Schwellenwert. Sei dazu $\mathscr{S}$ die Menge aller erfüllenden Belegungen einer zufälligen Formel. Wir sagen, dass zwei Lösungen verbunden sind, wenn ihr Hamming-Abstand 1 ist, und bezeichnen die Menge aller verbundenen Lösungen als Cluster. Es zeigt sich, dass die Geometrie von $\mathscr{S}$ hoch komplex ist, aber glücklicherweise liefert der 1-RSB-Ansatz der statistischen Physik ein nicht-rigoroses aber detailliertes Bild, wie sich $\mathscr{S}$ mit wachsendem Faktor-zu-Variablen-Verhältnis α entwickelt (siehe Abbildung 4.2).

Wir starten bei $\alpha = 0$ und lassen α stetig wachsen. Dann beobachten wir die Existenz von vier Schwellenwerten $\alpha_u \le \alpha_{\text{clus}} \le \alpha_{\text{cond}} \le \alpha_s$, an denen sich die Struktur von $\mathscr{S}$ drastisch verändert [121, 138, 169, 170]. Wir bezeichnen $\mathscr{S}$ manchmal als Lösungsraum und die erfüllenden Belegungen analog als Lösungen.

1. Ist $\alpha < \alpha_u$, so existiert exakt ein Cluster von Lösungen. Diese Phase wird als *unique phase* bezeichnet.

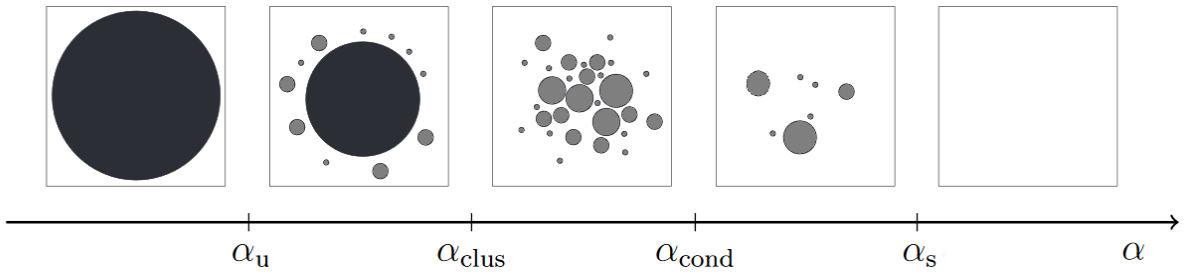

Abbildung 4.2.: Übersicht über die Geometrie des Lösungsraums in zufälligen CSPs. Der 1-RSB-Ansatz sagt die Existenz von vier wichtigen Phasenübergängen vorher. Die Abbildung ist modifiziert nach [138, 169, 170].

2. Sobald α größer wird als α_u, befindet sich das System in der *extremalen Phase*. Wenige und exponentiell kleine Cluster von Lösungen entstehen, die neben einem großen Cluster einige wenige Lösungen enthalten.

3. Am *clustering*-Schwellenwert α_{clus} zerfällt die Menge der Lösungen in exponentiell viele, exponentiell kleine Cluster.

4. Wenn α nun größer wird als der *condensation*-Schwellenwert, so befinden sich fast alle Lösungen in endlich vielen verschiedenen Clustern.

5. Schlussendlich, sobald α größer wird als der Erfüllbarkeits-Schwellenwert α_s, enthält $\mathscr{S}$ keine Lösungen mehr.

Formal betrachtet existieren strikte und schwache Schwellenwerte. Ist $\mathscr{P}$ irgendeine Eigenschaft und α eine Parametrisierung eines zufälligen Systems (im obigen Beispiel ist $\mathscr{P}$ die Menge der nicht erfüllbaren Formeln und α die Faktoren-zu-Variablen-Dichte), dann durchläuft das System $\mathscr{G}$ (der Faktorgraph) einen strikten Phasenübergang am strikten Schwellenwert α^*, falls für jedes $\varepsilon > 0$:

$$\mathbb{P}\left(\mathscr{G} \in \mathscr{P} \mid \alpha \leq (1-\varepsilon)\alpha^*\right) = o(1) \qquad \text{und} \qquad \mathbb{P}\left(\mathscr{G} \in \mathscr{P} \mid \alpha \geq (1+\varepsilon)\alpha^*\right) = 1 - o(1).$$

Analog findet ein schwacher Phasenübergang statt, falls

$$\mathbb{P}\left(\mathscr{G} \in \mathscr{P} \mid \alpha = o(\alpha^*)\right) = o(1) \qquad \text{sowie} \qquad \mathbb{P}\left(\mathscr{G} \in \mathscr{P} \mid \alpha = \omega(\alpha^*)\right) = 1 - o(1).$$

Die Definition verlangt offenbar, dass $\mathscr{P}$ mit wachsendem α wahrscheinlicher wird. Natürlich kann es analog auch für Eigenschaften definiert werden, die mit fallendem α wahrscheinlicher werden.

Friedgut und Bourgain [82] haben gezeigt, dass jedes zufällige System bezüglich einer monoton steigenden oder fallenden Eigenschaft einen nicht-uniformen strikten Phasenübergang durchläuft. Beispielsweise fällt die Frage, ob eine zufällige $k-$SAT Formel erfüllbar ist, darunter, da das Hinzufügen von Klauseln die Wahrscheinlichkeit reduziert, dass die Formel erfüllbar bleibt. Selbstverständlich liefert das Resultat nur eine Existenzaussage, das heißt, es bleibt unklar, was die entsprechenden Schwellenwerte sind. Für viele Eigenschaften bezüglich des zufälligen Graphen $\mathscr{G}(n,p)$[1] ist der Schwellenwert bekannt [103], aber in Bezug auf zufällige CSPs sieht die Welt anders aus, auch wenn in den letzten Jahren Schwellenwerte für manche zufälligen CSPs gefunden wurden [25, 43, 55, 64].

In dieser Dissertation beantworten wir eine offene Frage bezüglich des Lösungsraums des zufälligen $2-$SAT Problems. Das $2-$SAT Problem ist ein Spezialfall des bereits diskutierten $k-$SAT und auf eine gewisse Art und Weise eine Besonderheit. Es ist das einzige $k-$SAT Problem, in welchem es komplexitätstheoretisch einfach ist, eine erfüllende Belegung zu finden (sofern eine solche existiert) [120]. Auch

[1] Wir definieren $\mathscr{G}(n,p)$ wie Gilbert [90], das heißt, jede Kante ist existent mit Wahrscheinlichkeit p unabhängig von allem anderen Zufall.

der Erfüllbarkeits-Schwellenwert des zufälligen 2−SAT Problems ist seit den frühen 1990er Jahren bekannt [40, 91] und hängt eng mit dem Perkulationsphasenübergang im zufälligen Graphen zusammen [30]. Dennoch blieb eine sehr unschuldig vermutende, aber zentrale Frage offen [78]: Falls eine zufällige Formel erfüllbar ist, wie viele erfüllbare Belegungen gibt es? Es stellt sich heraus, dass dies schwer zu beantworten ist, insbesondere ist im (nicht-zufälligen) 2−SAT das Zählen der Lösungen komplexitätstheoretisch schwierig, es liegt in #**P** [165]. Die in der vorliegenden Dissertation enthaltene Publikation

The number of satisfying assignments of random 2-SAT formulas [2]

beantwortet diese Frage vollständig. Dies gelingt, indem wir zeigen, dass im Falle des zufälligen 2−SAT Problems bestimmte Heuristiken der statistischen Physik in einen rigorosen mathematischen Beweis verwandelt werden können. Genauer gesagt zeigen wir, dass die durch Belief Propagation berechnete Lösung der Marginale der zugehörigen Boltzmann-Verteilung korrekt ist.

Die Cut-Distanz Dies führt uns zum nächsten, in dieser Dissertation enthaltenen Beitrag mit dem Titel

The cut metric for probability distributions [47].

Die Boltzmann-Verteilung ist eine diskrete Wahrscheinlichkeitsverteilung auf Ω^n und in vielen zu studierenden Problemen betrachten wir den sogenannten *thermodynamischen Grenzwert* $n \to \infty$. Es liegt daher nahe, kontinuierliche Grenzobjekte von solchen diskreten Wahrscheinlichkeitsverteilungen zu betrachten. Die $\Omega-$*laws*, ursprünglich eingeführt durch Coja-Oghlan, Perkins und Skubch [42] beschreiben solche Grenzobjekte. Genauer gesagt verwandeln wir eine Konfiguration $\sigma \in \Omega^n$ in eine messbare Funktion $\hat{\sigma}$ von $[0,1)$ in die Menge der Wahrscheinlichkeitsmaße über Ω, wobei diese Menge als $\mathscr{P}(\Omega)$ bezeichnet wird. Es sei ferner Σ_Ω der Raum aller messbaren Funktionen $f : [0,1) \to \mathscr{P}(\Omega)$ bis auf Gleichheit fast überall. Nun definieren wir $\hat{\sigma}$ als

$$\hat{\sigma} : [0,1) \to \mathscr{P}(\Omega), \qquad \text{sodass} \qquad x \mapsto \sum_{i=1}^{n} \delta_{\sigma_i} \mathbf{1}\left\{x \in \left[\frac{i-1}{n}, \frac{i}{n}\right)\right\}.$$

Das zugehörige Wahrscheinlichkeitsmaß $\mu \in \mathscr{P}(\Omega^n)$ wird nun in den Raum der Wahrscheinlichkeitsmaße $\mathscr{P}(\Sigma_\Omega)$ wie folgt eingebettet. Wir definieren

$$\hat{\mu} = \sum_{\sigma \in \Omega^n} \mu(\sigma) \delta_{\hat{\sigma}}, \qquad \text{sodass} \qquad \hat{\mu} \in \mathscr{P}(\Sigma_\Omega).$$

Offenbar besteht eine 1-zu-1-Beziehung zwischen μ und $\hat{\mu}$. Es ist weiterhin möglich eine sehr schwache Metrik, die Cut-Distanz, auf $\mathscr{P}(\Sigma_\Omega)$ zu definieren. Dazu bezeichnen wir mit $\mathbb{S}_{[0,1)}$ die Menge der maßerhaltenden, invertierbaren Bijektionen auf $[0,1)$ sowie mit $\Gamma(\mu,\nu)$ die Menge der *Couplings* von μ, ν, also gemeinsamen Wahrscheinlichkeitsverteilungen mit Marginalen μ und ν. Die Cut-Distanz ist nun definiert als

$$D_{\boxtimes}(\mu,\nu) = \inf_{\substack{\gamma \in \Gamma(\mu,\nu),\\ \varphi \in \mathbb{S}_{[0,1)}}} \sup_{\substack{B \subset \Sigma_\Omega^2,\\ U \subset [0,1),\\ \omega \in \Omega}} \left| \int_B \int_U \sigma_x(\omega) - \tau_{\varphi(x)}(\omega) \mathrm{d}x \mathrm{d}\gamma(\sigma,\tau) \right|$$

und ist als eine Art 2-Spieler Spiel zu verstehen. Spieler 1 wählt ein mögliches Coupling von zwei Wahrscheinlichkeitsverteilungen, unter welchem sich die Verteilungen möglichst ähnlich sehen. Nun wählt Spieler 2 eine Menge von (verallgemeinerten) Koordinaten und (verallgemeinerten) Konfigurationen, an denen sich μ und ν stark unterscheiden. Selbstverständlich gibt es von der Cut-Distanz auch eine diskrete Variante, nämlich

$$\Delta_{\boxtimes}(\mu,\nu) = \inf_{\substack{\gamma \in \Gamma(\mu,\nu),\\ \phi \in \mathbb{S}_n}} \sup_{\substack{S \subset \Omega^n \times \Omega^n,\\ X \subset [n],\\ \omega \in \Omega}} \left| \sum_{\substack{(\sigma,\tau) \in S,\\ x \in X}} \gamma(\sigma,\tau) \left(\mathbf{1}\{\sigma_x = \omega\} - \mathbf{1}\{\tau_{\phi(x)} = \omega\} \right) \right|.$$

Hierbei bezeichnet $\mathbb{S}_n$ die Menge der Permutationen von $[n]$. In der o.g. Publikation zeigen wir unter anderem, dass für Maße $\mu, \nu \in \mathscr{P}(\Omega^n)$ die Abstände $\Delta_{\boxtimes}(\mu, \nu)$ und $D_{\boxtimes}(\hat{\mu}, \hat{\nu})$ eng zusammenhängen, wie man es intuitiv auch erwarten sollte.

Mit Hilfe der Cut-Distanz können wir $\mathscr{P}(\Sigma_\Omega)$ in einen kompakten metrischen Raum verwandeln. Genauer gesagt müssen wir Maße mit Cut-Distanz 0 identifizieren und definieren den Raum der $\Omega-$laws $\mathscr{L}(\Omega)$ als den Raum der Äquivalenzklassen unter dieser Identifikation.

Während $\mathscr{L}(\Omega)$, wie bereits erwähnt, durch Coja-Oghlan, Perkins und Skubch [42] eingeführt wurde, etablieren wir eine komplette und in sich konsistente Grenzwerttheorie für diskrete Wahrscheinlichkeitsmaße, die an die Theorie der Graph-Grenzwerte [31, 32, 128] angelehnt ist. Unter anderem beweisen wir, dass $(\mathscr{L}(\Omega), D_{\boxtimes}(\cdot,\cdot))$ ein kompakter metrischer Raum ist, und geben eine Art *schwaches Regularitätslemma* [84] für $\Omega-$laws, was ein zuvor bekanntes Resultat auf diskreten Maßen verallgemeinert [49]. Besonders elegant hierbei ist, dass das Regularitätslemma nicht nur eine Existenzaussage umfasst, sondern auch einen einfachen Algorithmus, das sogenannte *Pinning*, liefert, mit welchem eine entsprechende Partition des Phasenraums Σ_Ω bzw. Ω^n gefunden wird. Ohne zu sehr ins Detail zu gehen, verstehen wir hierbei unter Regularität, dass das Wahrscheinlichkeitsmaß - eingeschränkt auf die Partition - unter der Cut-Distanz sehr ähnlich wie das Produktmaß mit denselben Marginalen aussieht. Wir zeigen unter anderem, dass diese Eigenschaft, welche wir *Extremität* nennen, mit der schwachen Regularität der Graphentheorie eng verknüpft ist. Ferner zeigen wir, dass verschiedene wesentliche Operationen der statistischen Physik (wie zum Beispiel das Bilden von *Overlaps*) unter der Cut-Distanz stetig sind.

Insgesamt hilft die rigorose Analyse der Cut-Distanz in Spezialfällen dabei, Heuristiken der statistischen Physik in mathematische Beweise zu verwandeln. An dieser Stelle kehren wir zurück zur direkten Analyse von CSPs. Bislang haben wir Faktorgraphen von (zufälligen) CSPs betrachtet und uns deren Phasenraum angeschaut. Dieser Lösungsraum ändert sich drastisch, wenn wir die sogenannte *planted*-Version von CSPs betrachten.

Planted-Modelle und statistische Inferenz Wir betrachten dazu zunächst ein Beispiel. Nehmen wir an, wir haben n Variablenknoten gegeben sowie eine Färbung $\boldsymbol{\sigma}$ der Knoten mit q Farben. Nun erzeugen wir einen zufälligen Graphen derart, dass jede Kante ij mit Wahrscheinlichkeit p_1 existiert, falls $\boldsymbol{\sigma}_i = \boldsymbol{\sigma}_j$ (das heißt, ij ist monochromatisch) beziehungsweise mit Wahrscheinlichkeit p_2, falls $\boldsymbol{\sigma}_i \neq \boldsymbol{\sigma}_j$. Nach dem Einfügen der Kanten vergessen wir die zugrunde liegende Färbung $\boldsymbol{\sigma}$. Je nach Wahl von p_1 und p_2 sieht der Graph sehr unterschiedlich aus. Ist $p_1 = p_2$, so kann der entstandene Graph nicht von einem rein zufälligen $\mathscr{G}(n, p_1)$ unterschieden werden, ist hingegen $p_1 \ll p_2$ oder $p_1 \gg p_2$, so sollten wir - gegeben der zufällige Graph - in der Lage sein, eine Färbung $\tilde{\sigma}$ zu finden, die $\boldsymbol{\sigma}$ ähnelt.

Etwas formaler erklärt dieses Vorgehen das *Lehrer-Schüler-Modell* der statistischen Inferenz [168]. Im einfachsten Fall erzeugt ein Lehrer eine Grundwahrheit $\boldsymbol{\sigma}$, ein *planted*-Modell basierend auf dieser Grundwahrheit, und übermittelt einem Schüler das Modell und die Information, wie $\boldsymbol{\sigma}$ und das Modell erzeugt wurden. Die Aufgabe des Schülers ist es nun, eine Vermutung $\tilde{\sigma}$ zu formulieren, die möglichst nah an der Grundwahrheit liegt. Im obigen Beispiel, was eine simple Form des stochastischen Blockmodells [62, 85, 95] darstellt, muss der Schüler die Färbung $\boldsymbol{\sigma}$ möglichst genau aus dem zufälligen Graphen rekonstruieren.

Es stellt sich heraus, dass solche Inferenzprobleme als physikalisches Modell wie zuvor ausgedrückt werden können und sich somit Heuristiken zur Lösung von CSPs auch auf statistische Inferenzprobleme übertragen [168]. Daher ist es nicht überraschend, dass wir auch hier Phasenübergänge untersuchen können. Wir betrachten im Wesentlichen zwei Phasenübergänge. Sei $\mathscr{I}$ die Information, die der Schüler erhält (z.B. der zufällige Graph sowie p_1 und p_2). Die Menge der Information sei durch α parametrisiert. Beispielsweise kann α die Differenz von p_1 und p_2 sein oder die Anzahl von Messungen eines komprimierten Signals.

- Der *informationstheoretische* Schwellenwert bezeichnet den Moment, ab welchem der Schüler $\boldsymbol{\sigma}$ aus $\mathscr{I}(\alpha)$ rekonstruieren kann.
- Der *algorithmische* Schwellenwert bezeichnet die Menge an Information, die notwendig ist, damit ein effizienter Algorithmus bekannt ist, welcher $\boldsymbol{\sigma}$ aus $\mathscr{I}(\alpha)$ rekonstruieren kann.

In dieser Disseration betrachten wir ein spezielles Inferenzproblem – das Group-Testing – und studieren sowohl informationstheoretische Schwellenwerte als auch algorithmische Schwellenwerte.

Group Testing Das Group-Testing Problem fand in den 1940er Jahren den Weg in die mathematische Literatur [68] und wurde über die Jahrzehnte hinweg stetig untersucht [8, 9, 58, 59, 73, 80, 86, 98, 108, 132, 133, 154, 164]. In einer Population von $n \gg 1$ Individuen sind k mit einer Krankheit infiziert. Es ist möglich mehrere Individuen auf einmal zu testen und das Ergebnis eines solchen Gruppentests ist positiv, genau dann wenn mindestens ein infiziertes Individuum in dem Test enthalten ist. Gesucht ist nun eine Strategie, die möglichst wenige Tests benötigt um die infizierten Individuen (mit hoher Wahrscheinlichkeit) korrekt zu identifizieren. Eine Visualisierung findet sich in Abbildung 4.3.

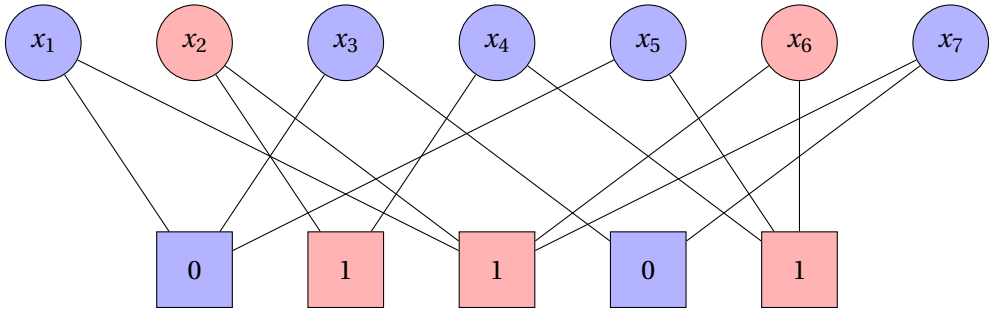

Abbildung 4.3.: Faktorgraph-Darstellung einer Group-Testing Strategie mit $n = 7$ Individuen von denen $k = 2$ infiziert sind. Blaue Individuen sind nicht infiziert und rote Individuen sind infiziert.

Das Group-Testing Problem kann unter vielen verschiedenen Modellen studiert werden. Beispielsweise können wir annehmen, dass jedes Individuum mit Wahrscheinlichkeit k/n infiziert ist oder dass es exakt k Infizierte gibt. Ferner können wir die Teststrategien insofern beschränken, als dass nur eine, zwei oder drei Runden Tests durchgeführt werden können, oder derart, dass Individuen nicht öfter als Δ mal getestet werden dürfen bzw. ein Gruppentest eine Maximalkapazität aufweist. Schlussendlich können die Tests immer ein korrektes Ergebnis liefern oder aber mit einer bestimmten Wahrscheinlichkeit ein falsches Ergebnis liefern. Die hier aufgelisteten Varianten sind weit entfernt davon vollständig zu sein und wir verweisen interessierte Leser auf einen Übersichtsartikel von Aldridge, Johnson und Scarlett [10].

In dieser Disseration beschäftigen wir uns mit dem sogenannten *hypergeometrischen sublinearen probabilistischen* Group-Testing. Das heißt, wir nehmen an, dass wir die Anzahl der infizierten Individuen k exakt kennen und sich k sublinear in n verhält, also $k = n^\theta (\theta \in (0,1))$. Ferner möchten wir die infizierten Individuen mit hoher Wahrscheinlichkeit rekonstruieren. Hierbei betrachten wir sowohl das *uneingeschränkte* Group-Testing Problem, in welchem Individuen beliebig oft getestet werden können und Tests beliebig groß werden dürfen, als auch das *eingeschränkte* Group-Testing Problem, in welchem dies nicht der Fall ist. Insbesondere sind wir an Schwellenwerten interessiert, welche die Anzahl der Tests $m = m(n,k)$ beschreiben, die notwendig bzw. hinreichend sind, um den Infektionsstatus aller Individuen zu rekonstruieren. Genauer gesagt enthält diese Dissertation drei Prublikationen, in denen wir uns mit dem Group-Testing Problem auseinandersetzen, nämlich

Information-Theoretic and Algorithmic Thresholds for Group Testing [41]

und sowohl

Optimal group testing [46]

als auch

Near optimal sparsity-constrained group testing: improved bounds and algorithms [88].

Im nicht-adaptiven Fall, das heißt, dass alle betrachteten Teststrategien alle Tests parallel ausführen müssen, war vor den Beiträgen der Dissertation die beste bekannte Strategie das sogenannte *zufällige reguläre Modell* [9]. In diesem Modell wählt jedes Individuum $\Delta = \Theta(\ln n)$ Tests zufällig aus und nimmt an ihnen teil. Zudem war der beste bekannte Algorithmus der DD-Algorithmus, in welchem zunächst alle Individuen in negativen Tests als gesund deklariert werden und alle Individuen, die nun alleine in einem positiven Test vorkommen, als infiziert. Alle übrigen Individuen werden ebenfalls als nicht-infiziert deklariert [108]. In Abbildung 4.4 sind die Ergebnisse der Publikationen [41, 46] zusammengefasst.

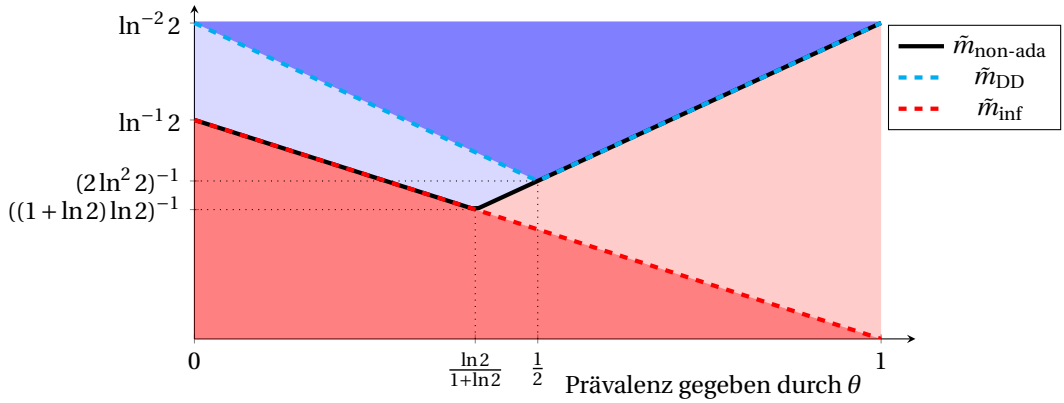

Abbildung 4.4.: Die Phasenübergänge im hypergeometrischen probabilistischen Group-Testing. Die Abbildung ist modifiziert nach [46, Abbildung 1]. Wir definieren $\tilde{m} = m \cdot (k \ln(n))^{-1}$.

Beginnen wir mit den bereits zuvor bekannten Tatsachen. Im dunkelroten Bereich, unter m_{inf}, ist es weder nicht-adaptiv noch adaptiv (in mehreren Runden) möglich, die infizierten Individuen zu identifizieren, wie aus dem Zählargument folgt, dass die Anzahl der möglichen Testergebnisse 2^m mindestens der Anzahl der Konfigurationen mit k infizierten Individuen $\binom{n}{k}$ entsprechen muss. Überhalb dieser Schwellenwertfunktion gibt es Algorithmen, die in mehreren Runden die infizierten Individuen rekonstruieren können [13, 60, 98, 153]. Ferner beschreibt die dunkelblaue Fläche (über m_{DD}) den Bereich, in welchem der DD-Algorithmus auf dem zufälligen regulären Modell funktioniert [108]. Es war ferner bekannt, dass es informationstheoretisch auf selbigem Modell unterhalb von $m_{\text{non-ada}}$ unmöglich ist, die Infizierten zu rekonstruieren [8]. Unsere Beiträge lassen sich wie folgt zusammenfassen, wobei alle angegebenen Schwellenwertfunktionen strikte Phasenübergänge ausdrücken.

- Unterhalb von m_{DD} ist keine Inferenz durch den DD-Algorithmus möglich.
- Unterhalb von $m_{\text{non-ada}}$ kann es keine nicht-adaptive Teststrategie geben, welche die Inferenz der infizierten Individuen ermöglicht. Das heißt, dass die hellrote Fläche einen Bereich darstellt, in welchem adaptive Algorithmen bekannt sind und funktionieren, während nicht-adaptive Strategien keinen Erfolg haben.
- Das zufällige reguläre Modell ermöglicht informationstheoretisch die Inferenz ab $m_{\text{non-ada}}$, ist also informationstheoretisch optimal.
- Wir definieren basierend auf Ideen der Coding-Theorie eine neuartige nicht-adaptive, *spatially-coupled* Teststrategie und einen effizienten Algorithmus, der ab $m_{\text{non-ada}}$, also auch bereits im hellblauen Bereich, funktioniert.
- Die Inferenz von allen bis auf $o(k)$ Individuen ist ab m_{inf} durch das eben genannte Modell mit demselben Algorithmus möglich. Ebenso kann der Algorithmus leicht zu einem zweistufigen Algorithmus verändert werden, der die Inferenz aller Individuen ab m_{inf} ermöglicht.

Zusammenfassend ist somit das hypergeometrische probabilistische Group-Testing im sublinearen Fall vollständig verstanden. Schränken wir die zulässige Anzahl an Tests pro Individuum ein, das heißt, jedes Individuum ist nur in maximal $\Delta = O\left(\ln^{1-\delta} n\right)$ ($\delta \in (0,1]$) Tests enthalten, so ist das Problem noch

nicht vollständig erforscht. Bekannte Resultate sowie unsere erzielten Ergebnisse sind in Abbildung 4.5 dargestellt.

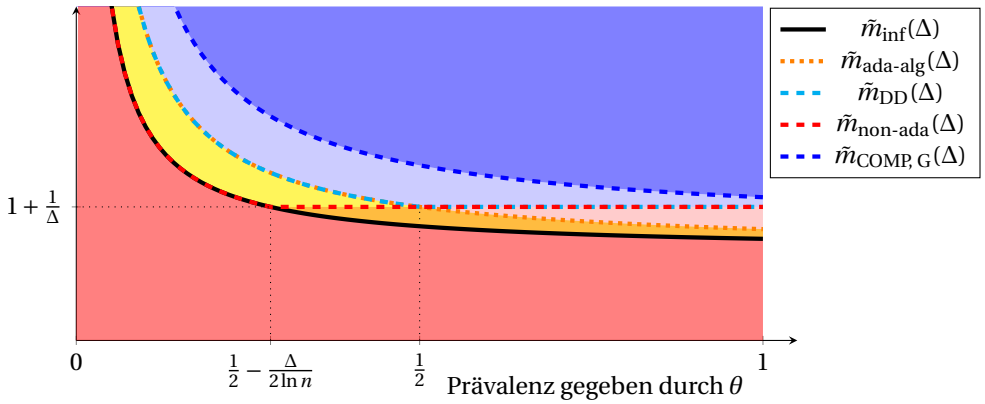

Abbildung 4.5.: Wichtige Phasenübergänge im hypergeometrischen probabilistischen Group-Testing Problem, falls jedes Individuum höchstens Δ-mal getestet werden darf mit $\Delta = O\left(\ln^{1-\delta} n\right)$ für ein $\delta \in (0,1]$. Die Abbildung zeigt die Phasenübergänge für die Wahl $\Delta = 5$ und $n = 10^5$. Wir parametrisieren die eigentliche Schwellenwertfunktion als $m = \Delta k^{\tilde{m}}$.

Die relevantesten zuvor bekannten Resultate gehen auf Gandikota et al. [86] zurück. Insbesondere analysieren die Autoren das zufällige reguläre Modell und einen einfachen Algorithmus `COMP`, der im dunkelblauen Bereich die Inferenz ermöglicht. Dieser sehr einfache Algorithmus deklariert alle Individuen, die in einem negativen Test vorkommen, als uninfiziert während alle weiteren Individuen als infiziert deklariert werden. Dieselben Autoren zeigen, dass jede nicht-adaptive Teststrategie mit höchstens $(\mathrm{e} m_{\text{inf}}(\Delta))^{1-\varepsilon}$ Tests scheitert. Wir erweitern diese Resultate wie folgt und nehmen wiederum implizit an, dass Schwellenwertfunktionen zu strikten Phasenübergängen korrespondieren.

- Der `DD`-Algorithmus auf dem zufälligen regulären Modell ermöglicht Inferenz im blauen Bereich, ist also strikt besser als `COMP`. Außerdem ermöglicht der `DD`-Algorithmus unterhalb von $\tilde{m}_{\text{DD}}$ mit hoher Wahrscheinlichkeit keine Inferenz auf dem zufälligen regulären Modell.
- Unterhalb von $m_{\text{inf}}(\Delta)$ kann keine – auch keine adaptive – Teststrategie erfolgreich sein.
- Unterhalb von $m_{\text{non-ada}}(\Delta)$ kann keine nicht-adaptive Teststrategie die Inferenz der infizierten Individuen ermöglichen.
- Überhalb von $m_{\text{adap-alg}}(\Delta)$ existiert ein effizienter adaptiver Algorithmus, der die Inferenz mit hoher Wahrscheinlichkeit ermöglicht. Das heißt insbesondere, dass adaptive Algorithmen im hellroten Bereich wie zuvor eine bessere Performance als nicht-adaptive Strategien liefern.

In diesem Setting bleibt somit offen, wo der informationstheoretische Phasenübergang für adaptive Algorithmen stattfindet (orangener Bereich) und ob es nicht-adaptive Teststrategien (ggf. mit effizienten Algorithmen) gibt, welche Inferenz im gelben Bereich ermöglichen.

Zuletzt haben wir uns ebenfalls mit einer anderen Art der Einschränkung im Group-Testing Problem beschäftigt. Falls jeder Test nur $\Gamma = \Theta(1)$ Individuen beinhalten darf, so war vor dem Beitrag der Dissertation nur wenig bekannt. Die Ergebnisse sind in Abbildung 4.6 visualisiert.

Gandikota et al. [86] analysieren wiederum den `COMP`-Algorithmus auf dem zufälligen regulären Modell, welcher ab der blauen Linie Inferenz ermöglicht. Ferner zeigt eine einfache Zählschranke, dass mindestens n/Γ Tests in jeder (adaptiven) Teststrategie benötigt werden (schwarze Linie). Wir erzielen die folgenden Resultate.

- Wir etablieren eine universelle informationstheoretische Schranke für alle nicht-adaptiven Teststrategien, das heißt, jede nicht-adaptive Teststrategie kann unterhalb der roten Linie mit hoher Wahrscheinlichkeit nicht die infizierten Individuen rekonstruieren.

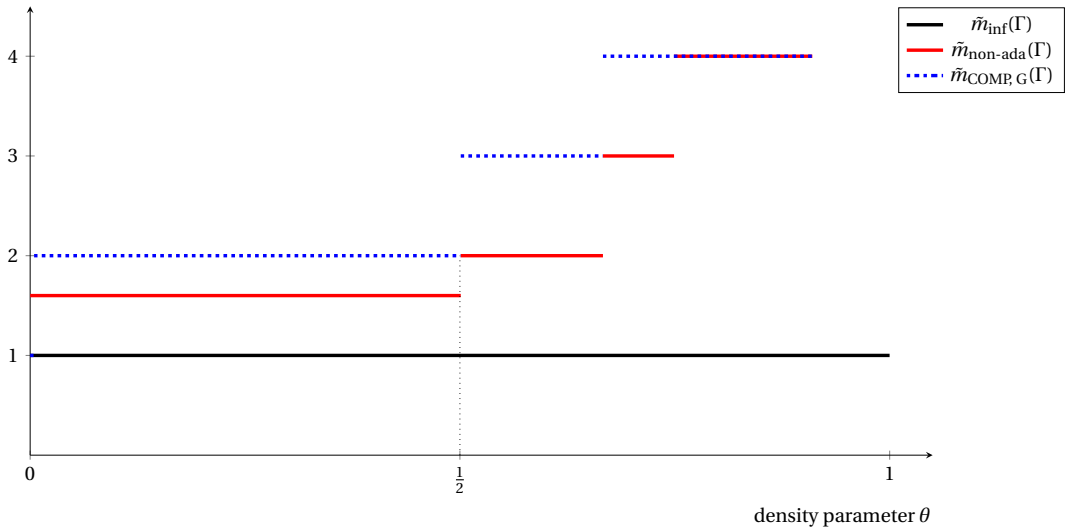

Abbildung 4.6.: Übersicht über die Phasenübergänge im hypergeometrischen probabilistischen Group-Testing, sofern jeder Test höchstens $\Gamma = 4$ Individuen beinhalten darf. Wir parametrisieren die Anzahl der Tests $m = \tilde{m}\frac{n}{\Gamma}$ und nehmen implizit an, dass $n/\Gamma \in \mathbb{Z}$.

- Wir definieren eine neue nicht-adaptive Teststrategie, die bei niedriger Prävalenz ($\theta < 1/2$) vom regulären Modell abweicht, und zeigen, dass der `DD`-Algorithmus für alle Werte von θ außerhalb einer Nullmenge auf diesem Modell ab der roten Linie die Infizierten mit hoher Wahrscheinlichkeit rekonstruiert. Dieses Modell mit dem `DD`-Algorithmus ist also optimal.
- Es existiert ein effizienter adaptiver Algorithmus, der – bis auf Rundung – oberhalb der schwarzen Linie die Infizierten erfolgreich rekonstruiert.

Entsprechend ist dieser Fall des Group-Testings auch nahezu vollständig verstanden. Es bleibt allerdings offen, ob ähnliche Analysen auch für $\omega(1) = \Gamma = o(n/k)$ durchgeführt werden können.

Schlussendlich befasst sich die vorliegende Dissertation mit sogenannten *zufällig perturbierten* Graphmodellen.

Zufällig perturbierte Graphen Ursprünglich stammt die Idee der zufälligen Perturbation deterministischer Systeme aus der *smooth analysis of algorithms*, also der *stufenlosen Analyse von Algorithmen* [159]. Während viele Algorithmen eine exponentielle Laufzeit im schlimmsten Fall (*Worst-Case-Analyse*) aufweisen, so zeigt sich in realen Anwendungen, dass sie meistens sehr effizient funktionieren. Das bekannteste Beispiel ist vermutlich der Simplex-Algorithmus [61] zum Lösen linearer Optimierungsprobleme. In Anwendungen scheinen somit oft die Worst-Case-Bedingungen nicht einzutreten, allerdings ist es wichtig zu verstehen, wie hoch die Laufzeit auf einer anwendungsspezifischen Eingabe vermutlich werden kann. Die Analyse der typischen Laufzeit auf einer zufälligen Eingabe (*Average-Case-Analyse*) beantwortet diese Frage nur ungenügend, da in einer Anwendung durchaus Konstellationen auftreten können, die weit entfernt von einer durchschnittlichen Eingabe sind. Allerdings ist es ebenso unwahrscheinlich, eine Worst-Case-Konfiguration zu beobachten. Aus diesem Grund werden perturbierte Modelle untersucht. Wir beginnen bei einer Worst-Case-Konfiguration, fügen (ein wenig) Zufall hinzu und möchten die Laufzeit des Algorithmus in Abhängigkeit der Menge des hinzugefügten Zufalls untersuchen [159].

Kurz nach Einführung dieser Betrachtungsweise von Algorithmen wurde das Prinzip auf die Existenz aufspannender Strukturen in Graphen übertragen. Eine zentrale aufspannende Struktur ist beispielweise der Hamiltonkreis, also ein Rundweg, der jeden Knoten exakt einmal besucht. Es sind zahlreiche hinreichende Bedingungen, wie Diracs Theorem [65] bekannt, welche zum Beispiel durch Minimalgradbedingungen die Existenz eines Hamiltonkreises in einem beliebigen deterministischen Graphen garantieren. Diese Art von Theoremen kann mit der Worst-Case-Analyse verglichen werden. Auf der anderen

Seite ist der exakte Schwellenwert $p = p(n)$ bekannt, ab welchem der zufällige Graph $\mathcal{G}(n.p)$ mit hoher Wahrscheinlichkeit einen solchen aufspannenden Kreis besitzt (Average-Case-Analyse) [116, 117, 147]. Selbstverständlich existieren solche Analysen nicht nur für Hamiltonkreise, sondern auch unter anderem für Spannbäume, Matchings, Potenzen von Hamiltonkreisen sowie für allgemeine aufspannende Graphen mit beschränktem Maximalgrad, sowohl in deterministischen Graphen [36, 94, 113, 114, 115] als auch in zufälligen Graphen [14, 33, 72, 76, 107, 118, 125, 139, 141, 150].

Das Modell der zufälligen perturbierten Graphen ist nun wie folgt zu verstehen. Gegeben sei ein beliebiger Graph $\mathcal{G}_\alpha$ mit Minimalgrad αn sowie ein zufälliger Graph $\mathcal{G}(n,p)$ mit $p = p(\alpha)$. Nun stellt sich die Frage, wann $\mathcal{G}_\alpha \cup \mathcal{G}(n,p)$ mit hoher Wahrscheinlichkeit einen Hamiltonkreis besitzt. Offenbar ist der Schwellenwert nicht nur abhängig von α, sondern auch von dem gegebenen Graphen $\mathcal{G}_\alpha$, das heißt, wir suchen einen Schwellenwert p^* im folgenden Sinne. Sofern $p > p^*$, so enthält $\mathcal{G}_\alpha \cup \mathcal{G}(n,p)$ einen Hamiltonkreis mit hoher Wahrscheinlichkeit unabhängig von der Wahl von $\mathcal{G}_\alpha$. Ist andererseits $p < p^*$, so existiert mindestens ein Graph $\mathcal{G}_\alpha$ mit Minimalgrad αn, sodass die Vereinigung mit dem zufälligen Graphen keinen Hamiltonkreis besitzt.

Dieses Modell wurde erstmalig von Bohman, Frieze und Martin [28] diskutiert. Es folgten mehrere Publikationen, welche hinreichende Bedingungen für die Existenz verschiedener aufspannender Strukturen in $\mathcal{G}_\alpha \cup \mathcal{G}(n,p)$ geben [20, 26, 35, 34, 119], allerdings liegt all diesen Beiträgen zu Grunde, dass der betrachtete deterministische Graph ein dichter Graph ist, das heißt, dass $\alpha = \Theta(1)$ eine Konstante ist. Das letzte in dieser Dissertation enthaltene Manuskript,

Random perturbation of sparse graphs [93],

beschäftigt sich mit dem Fall $\alpha = o(1)$, also mit dem Fall, dass der zugrunde liegende deterministische Graph dünn ist. In diesem Beitrag geben wir hinreichende Bedingungen für die Existenz von Matchings und Hamiltonkreisen sowie für die Existenz aufspannender Bäume mit beschränktem Maximalgrad, indem wir die Frage nach der Existenz einer aufspannenden Struktur in $\mathcal{G}_\alpha \cup \mathcal{G}(n,p)$ auf die Existenz nahezu-aufspannender Strukturen in $\mathcal{G}(n,p)$ zurückführen.

References

[1] D. Achlioptas and A. Coja-Oghlan. 'Algorithmic Barriers from Phase Transitions'. In: *2008 49th Annual IEEE Symposium on Foundations of Computer Science*. 2008, pp. 793–802.

[2] D. Achlioptas, A. Coja-Oghlan, M. Hahn-Klimroth, J. Lee, N. Müller, M. Penschuck and G. Zhou. 'The number of satisfying assignments of random 2-SAT formulas'. In: *Random Structures & Algorithms* (2021), pp. 1–39.

[3] D. Achlioptas and C. Moore. 'Random k-SAT: Two Moments Suffice to Cross a Sharp Threshold'. In: *SIAM Journal on Computing* 36.3 (2006), pp. 740–762.

[4] D. Achlioptas and Y. Peres. 'The threshold for random k-SAT is 2k (ln 2 - O(k))'. In: *Proceedings of the thirty-fifth ACM symposium on Theory of computing - STOC '03*. ACM Press, 2003.

[5] M. Aizenman, R. Sims and S. L. Starr. 'Extended variational principle for the Sherrington-Kirkpatrick spin-glass model'. In: *Phys. Rev. B* 68 (21 2003), p. 214403.

[6] D. J. Aldous. 'Representations for partially exchangeable arrays of random variables'. In: *Journal of Multivariate Analysis* 11.4 (1981), pp. 581–598.

[7] M. Aldridge. 'Individual Testing Is Optimal for Nonadaptive Group Testing in the Linear Regime'. In: *IEEE Transactions on Information Theory* 65.4 (2019), pp. 2058–2061.

[8] M. Aldridge, L. Baldassini and O. Johnson. 'Group Testing Algorithms: Bounds and Simulations'. In: *IEEE Transactions on Information Theory* 60.6 (2014), pp. 3671–3687.

[9] M. Aldridge, O. Johnson and J. Scarlett. 'Improved group testing rates with constant column weight designs'. In: *2016 IEEE International Symposium on Information Theory (ISIT)*. 2016, pp. 1381–1385.

[10] M. Aldridge, O. Johnson and J. Scarlett. *Group Testing: An Information Theory Perspective*. 2019.

[11] M. Aldridge. 'The Capacity of Bernoulli Nonadaptive Group Testing'. In: *IEEE Transactions on Information Theory* 63.11 (2017), pp. 7142–7148.

[12] M. Aldridge. *Conservative two-stage group testing*. 2020. arXiv: 2005.06617 [stat.AP].

[13] A. Allemann. 'An Efficient Algorithm for Combinatorial Group Testing'. In: *Information Theory, Combinatorics, and Search Theory*. Springer Berlin Heidelberg, 2013, pp. 569–596.

[14] N. Alon and Z. Füredi. 'Spanning subgraphs of random graphs'. In: *Graphs and Combinatorics* 8.1 (1992), pp. 91–94.

[15] E. Arikan. 'Channel Polarization: A Method for Constructing Capacity-Achieving Codes for Symmetric Binary-Input Memoryless Channels'. In: *IEEE Transactions on Information Theory* 55.7 (2009), pp. 3051–3073.

[16] J. Ashkin and E. Teller. 'Statistics of Two-Dimensional Lattices with Four Components'. In: *Physical Review* 64.5-6 (1943), pp. 178–184.

[17] T. Austin. 'Multi-variate correlation and mixtures of product measures'. In: *Kybernetika* (2020), pp. 459–499.

[18] P. Ayre, A. Coja-Oghlan, P. Gao and N. Müller. 'The Satisfiability Threshold For Random Linear Equations'. In: *Combinatorica* 40.2 (2020), pp. 179–235.

[19] J. Balogh, B. Csaba, M. Pei and W. Samotij. 'Large bounded degree trees in expanding graphs'. In: *the electronic journal of combinatorics* 17.1 (2010), p. 6.

[20] J. Balogh, A. Treglown and A. Z. Wagner. 'Tilings in randomly perturbed dense graphs'. In: *Combinatorics, Probability and Computing* 28.2 (2019), pp. 159–176.

[21] V. Bapst and A. Coja-Oghlan. 'Harnessing the Bethe free energy'. In: *Random Structures & Algorithms* 49.4 (2016), pp. 694–741.

[22] F. Barahona. 'On the computational complexity of Ising spin glass models'. In: *Journal of Physics A: Mathematical and General* 15.10 (1982), pp. 3241–3253.

[23] A. Barra, G. Genovese, F. Guerra and D. Tantari. 'About a solvable mean field model of a Gaussian spin glass'. In: *Journal of Physics A: Mathematical and Theoretical* 47.15 (2014), p. 155002.

[24] W. H. Bay, E. Price and J. Scarlett. *Optimal Non-Adaptive Probabilistic Group Testing Requires* $\Theta(\min\{k \ln n, n\})$ *Tests*. 2020. arXiv: 2006.01325 [cs.IT].

[25] M. Bayati, D. Gamarnik and P. Tetali. 'Combinatorial Approach to the Interpolation Method and Scaling Limits in Sparse Random Graphs'. In: *Proceedings of the Forty-Second ACM Symposium on Theory of Computing*. STOC '10. Cambridge, Massachusetts, USA: Association for Computing Machinery, 2010, pp. 105–114.

[26] W. Bedenknecht, J. Han, Y. Kohayakawa and G. O. Mota. 'Powers of tight Hamilton cycles in randomly perturbed hypergraphs'. In: *Random Structures & Algorithms* 55.4 (2019), pp. 795–807.

[27] H. A. Bethe and W. L. Bragg. 'Statistical theory of superlattices'. In: *Proceedings of the Royal Society of London. Series A - Mathematical and Physical Sciences* 150.871 (1935), pp. 552–575.

[28] T. Bohman, A. Frieze and R. Martin. 'How many random edges make a dense graph hamiltonian?' In: *Random Structures & Algorithms* 22.1 (2003), pp. 33–42.

[29] B. Bollobás. 'The Evolution of Random Graphs'. In: *Transactions of the American Mathematical Society* 286.1 (1984), pp. 257–274.

[30] B. Bollobás, C. Borgs, J. T. Chayes, J. H. Kim and D. B. Wilson. 'The Scaling Window of the 2-SAT Transition'. In: *Random Struct. Algorithms* 18.3 (2001), pp. 201–256.

[31] C. Borgs, J. Chayes, L. Lovász, V. Sós and K. Vesztergombi. 'Convergent sequences of dense graphs I: Subgraph frequencies, metric properties and testing'. In: *Advances in Mathematics* 219.6 (2008), pp. 1801–1851.

[32] C. Borgs, J. Chayes, L. Lovász, V. Sós and K. Vesztergombi. 'Convergent sequences of dense graphs II. Multiway cuts and statistical physics'. In: *Annals of Mathematics* 176.1 (2012), pp. 151–219.

[33] J. Böttcher. 'Large-scale structures in random graphs'. In: *Surveys in Combinatorics* 440 (2017), pp. 87–140.

[34] J. Böttcher, J. Han, Y. Kohayakawa, R. Montgomery, O. Parczyk and Y. Person. 'Universality for bounded degree spanning trees in randomly perturbed graphs'. In: *Random Structures & Algorithms* (2019).

[35] J. Böttcher, R. Montgomery, O. Parczyk and Y. Person. 'Embedding spanning bounded degree subgraphs in randomly perturbed graphs'. In: *Mathematika* (2019), pp. 1–25.

[36] J. Böttcher, M. Schacht and A. Taraz. 'Proof of the bandwidth conjecture of Bollobás and Komlós'. In: *Mathematische Annalen* 343.1 (2009), pp. 175–205.

[37] A. Braunstein, M. Mézard, M. Weigt and R. Zecchina. 'Constraint Satisfaction by Survey Propagation'. In: *Advances in Neural Information Processing Systems* 9 (2005), p. 424.

[38] A. Braunstein, M. Mézard and R. Zecchina. 'Survey propagation: An algorithm for satisfiability'. In: *Random Structures & Algorithms* 27.2 (2005), pp. 201–226.

[39] C. L. Chan, S. Jaggi, V. Saligrama and S. Agnihotri. 'Non-Adaptive Group Testing: Explicit Bounds and Novel Algorithms'. In: *IEEE Transactions on Information Theory* 60.5 (2014), pp. 3019–3035.

[40] V. Chvatal and B. Reed. 'Mick gets some (the odds are on his side) (satisfiability)'. In: *Proceedings., 33rd Annual Symposium on Foundations of Computer Science*. 1992, pp. 620–627.

[41] A. Coja-Oghlan, O. Gebhard, M. Hahn-Klimroth and P. Loick. 'Information-Theoretic and Algorithmic Thresholds for Group Testing'. In: *IEEE Transactions on Information Theory* 66.12 (2020), pp. 7911–7928.

[42] A. Coja-Oghlan, W. Perkins and K. Skubch. 'Limits of discrete distributions and Gibbs measures on random graphs'. In: *European Journal of Combinatorics* 66 (2017), pp. 37–59.

[43] A. Coja-Oghlan, C. Efthymiou and S. Hetterich. 'On the chromatic number of random regular graphs'. In: *Journal of Combinatorial Theory, Series B* 116 (2016), pp. 367–439.

[44] A. Coja-Oghlan, O. Gebhard, M. Hahn-Klimroth and P. Loick. 'Information-Theoretic and Algorithmic Thresholds for Group Testing'. In: *46th International Colloquium on Automata, Languages, and Programming (ICALP 2019)* (2019), 43:1–43:14.

[45] A. Coja-Oghlan, O. Gebhard, M. Hahn-Klimroth and P. Loick. 'Optimal Group Testing'. In: *Proceedings of 33rd Conference on Learning Theory* (2020), pp. 1374–1388.

[46] A. Coja-Oghlan, O. Gebhard, M. Hahn-Klimroth and P. Loick. 'Optimal group testing'. In: *Combinatorics, Probability and Computing* (2021), pp. 1–38.

[47] A. Coja-Oghlan and M. Hahn-Klimroth. *The cut metric for probability distributions*. 2020. arXiv: 1905.13619 [math.CO].

[48] A. Coja-Oghlan, M. Hahn-Klimroth, P. Loick, N. Müller, K. Panagiotou and M. Pasch. *Inference and mutual information on random factor graphs*. 2020. arXiv: 2007.07494 [cs.DM].

[49] A. Coja-Oghlan, F. Krzakala, W. Perkins and L. Zdeborová. 'Information-theoretic thresholds from the cavity method'. In: *Advances in Mathematics* 333 (2018), pp. 694–795.

[50] A. Coja-Oghlan, N. Müller and J. B. Ravelomanana. *Belief Propagation on the random k-SAT model*. 2020. arXiv: 2011.02303 [math.PR].

[51] A. Coja-Oghlan and K. Panagiotou. 'The asymptotic k-SAT threshold'. In: *Advances in Mathematics* 288 (2016), pp. 985–1068.

[52] A. Coja-Oghlan and W. Perkins. 'Belief propagation on replica symmetric random factor graph models'. In: *Annales de l'Institut Henri Poincaré D* 5.2 (2018), pp. 211–249.

[53] A. Coja-Oghlan and W. Perkins. 'Bethe States of Random Factor Graphs'. In: *Communications in Mathematical Physics* 366.1 (2019), pp. 173–201.

[54] A. Coja-Oghlan and N. Wormald. 'The Number of Satisfying Assignments of Random Regular k-SAT Formulas'. In: *Combinatorics, Probability and Computing* 27.4 (2018), pp. 496–530.

[55] A. Coja-Oghlan and L. Zdeborová. 'The condensation transition in random hypergraph 2-coloring'. In: *Proceedings of the Twenty-Third Annual ACM-SIAM Symposium on Discrete Algorithms* (2012).

[56] D. Conlon and J. Fox. 'Bounds for graph regularity and removal lemmas'. In: *Geometric and Functional Analysis* 22.5 (2012), pp. 1191–1256.

[57] M. Cuturi, O. Teboul, Q. Berthet, A. Doucet and J.-P. Vert. *Noisy Adaptive Group Testing using Bayesian Sequential Experimental Design*. 2020. arXiv: 2004.12508 [stat.ME].

[58] A. G. D'yachkov, I. V. Vorob'ev, N. A. Polyansky and V. Y. Shchukin. 'Bounds on the rate of disjunctive codes'. In: *Problems of Information Transmission* 50.1 (2014), pp. 27–56.

[59] P. Damaschke. 'Threshold Group Testing'. In: *Lecture Notes in Computer Science*. Springer Berlin Heidelberg, 2006, pp. 707–718.

[60] P. Damaschke and A. S. Muhammad. 'Randomized Group Testing Both Query-Optimal and Minimal Adaptive'. In: *SOFSEM 2012: Theory and Practice of Computer Science*. Springer Berlin Heidelberg, 2012, pp. 214–225.

[61] G. B. Dantzig. 'Maximization of a Linear Function of Variables Subject to Linear Inequalities'. In: *Activity Analysis of Production and Allocation, Cowles Commission Monograph* 13 (1951).

[62] R. David and U. Feige. 'On the effect of randomness on planted 3-coloring models'. In: *Proceedings of the 48th Annual ACM SIGACT Symposium on Theory of Computing, STOC 2016, Cambridge, MA, USA, June 18-21, 2016*. Ed. by D. Wichs and Y. Mansour. ACM, 2016, pp. 77–90.

[63] P. Diaconis and S. Janson. 'Graph limits and exchangeable random graphs'. In: *Rendiconti di Matematica e delle sue Applicazioni* 28 (2008).

[64] J. Ding, A. Sly and N. Sun. 'Proof of the Satisfiability Conjecture for Large k'. In: *Proceedings of the Forty-Seventh Annual ACM Symposium on Theory of Computing*. STOC '15. Portland, Oregon, USA: Association for Computing Machinery, 2015, pp. 59–68.

[65] G. A. Dirac. 'Some Theorems on Abstract Graphs'. In: *Proceedings of the London Mathematical Society* s3-2.1 (1952), pp. 69–81.

[66] D. L. Donoho, A. Maleki and A. Montanari. 'Message passing algorithms for compressed sensing: I. motivation and construction'. In: *2010 IEEE Information Theory Workshop on Information Theory (ITW 2010, Cairo)*. 2010, pp. 1–5.

[67] D. Donoho, A. Maleki and A. Montanari. 'Message Passing Algorithms for Compressed Sensing'. In: *Proceedings of the National Academy of Sciences of the United States of America* 106 (2009), pp. 18914–9.

[68] R. Dorfman. 'The Detection of Defective Members of Large Populations'. In: *Ann. Math. Statist.* 14.4 (1943), pp. 436–440.

[69] S. F. Edwards and P. W. Anderson. 'Theory of spin glasses'. In: *Journal of Physics F: Metal Physics* 5.5 (1975), pp. 965–974.

[70] A. El Alaoui, A. Ramdas, F. Krzakala, L. Zdeborová and M. I. Jordan. 'Decoding From Pooled Data: Phase Transitions of Message Passing'. In: *IEEE Transactions on Information Theory* 65.1 (2019), pp. 572–585.

[71] R. S. Ellis and C. M. Newman. 'The statistics of Curie-Weiss models'. In: *Journal of Statistical Physics* 19.2 (1978), pp. 149–161.

[72] P. Erdős and A. Rényi. 'On the existence of a factor of degree one of a connected random graph'. In: *Acta Mathematica Hungarica* 17.3-4 (1966), pp. 359–368.

[73] P. Erdős and A. Rényi. 'On Two Problems of Information Theory'. In: *Magyar Tud. Akad. Mat. Kutató Int. Közl* 8 (1963), pp. 229–243.

[74] U. Feige, S. Goldwasser, L. Lovász, S. Safra and M. Szegedy. 'Interactive Proofs and the Hardness of Approximating Cliques'. In: *J. ACM* 43.2 (1996), pp. 268–292.

[75] U. Feige, E. Mossel and D. Vilenchik. 'Complete Convergence of Message Passing Algorithms for Some Satisfiability Problems'. In: *Theory of Computing* 9.19 (2013), pp. 617–651.

[76] A. Ferber, K. Luh and O. Nguyen. 'Embedding large graphs into a random graph'. In: *Bulletin of the London Mathematical Society* 49.5 (2017), pp. 784–797.

[77] A. Ferber and R. Nenadov. 'Spanning universality in random graphs'. In: *Random Structures & Algorithms* 53.4 (2018), pp. 604–637.

[78] W. Fernandez de la Vega. 'Random 2-SAT: results and problems'. In: *Theoretical Computer Science* 265.1 (2001). Phase Transitions in Combinatorial Problems, pp. 131–146.

[79] E. Fischer, A. Matsliah and A. Shapira. 'Approximate Hypergraph Partitioning and Applications'. In: *48th Annual IEEE Symposium on Foundations of Computer Science (FOCS'07)*. 2007, pp. 579–589.

[80] P. Fischer, N. Klasner and I. Wegener. 'On the cut-off point for combinatorial group testing'. In: *Discrete Applied Mathematics* 91.1 (1999), pp. 83–92.

[81] B. J. Frey, F. R. Kschischang, H.-a. Loeliger and N. Wiberg. 'Factor Graphs and Algorithms'. In: *Proc. 35th Allerton Conf. on Communications, Control, and Computing, (Allerton)*. 1998, pp. 666–680.

[82] E. Friedgut and J. Bourgain. 'Sharp thresholds of graph properties, and the k-sat problem'. In: *Journal of the American Mathematical Society* 12.4 (1999), pp. 1017–1054.

[83] A. Frieze. 'On large matchings and cycles in sparse random graphs'. In: *Discrete Mathematics* 59.3 (1986), pp. 243–256.

[84] A. Frieze and R. Kannan. 'Quick Approximation to Matrices and Applications'. In: *Combinatorica* 19.2 (1999), pp. 175–220.

[85] T. Funke and T. Becker. 'Stochastic block models: A comparison of variants and inference methods'. In: *PLOS ONE* 14.4 (2019), pp. 1–40.

[86] V. Gandikota, E. Grigorescu, S. Jaggi and S. Zhou. 'Nearly Optimal Sparse Group Testing'. In: *IEEE Transactions on Information Theory* 65.5 (2019), pp. 2760–2773.

[87] E. Gardner and B. Derrida. 'Three unfinished works on the optimal storage capacity of networks'. In: *Journal of Physics A: Mathematical and General* 22.12 (1989), pp. 1983–1994.

[88] O. Gebhard, M. Hahn-Klimroth, O. Parczyk, M. Penschuck, M. Rolvien, J. Scarlett and N. Tan. *Near optimal sparsity-constrained group testing: improved bounds and algorithms*. 2020. arXiv: 2004.11860 [cs.DS].

[89] O. Gebhard, O. Johnson, P. Loick and M. Rolvien. *Improved bounds for noisy group testing with constant tests per item*. 2020. arXiv: 2007.01376 [cs.IT].

[90] E. N. Gilbert. 'Random Graphs'. In: *The Annals of Mathematical Statistics* 30.4 (1959), pp. 1141–1144.

[91] A. Goerdt. 'A Threshold for Unsatisfiability'. In: *Journal of Computer and System Sciences* 53.3 (1996), pp. 469–486.

[92] W. Gowers. 'Lower bounds of tower type for Szemerédi's uniformity lemma'. In: *Geometric and Functional Analysis* 7.2 (1997), pp. 322–337.

[93] M. Hahn-Klimroth, G. S. Maesaka, Y. Mogge, S. Mohr and O. Parczyk. *Random perturbation of sparse graphs*. 2020. arXiv: 2004.04672 [math.CO].

[94] A. Hajnal and E. Szemerédi. 'Proof of a conjecture of P. Erdős'. In: *Combinatorial theory and its applications* 2 (1970), pp. 601–623.

[95] P. W. Holland, K. B. Laskey and S. Leinhardt. 'Stochastic blockmodels: First steps'. In: *Social Networks* 5.2 (1983), pp. 109–137.

[96] D. Hoover. *Relations on Probability Spaces and Arrays of Random Variables*. 1979.

[97] M. C. Hu, F. K. Hwang and J. K. Wang. 'A Boundary Problem for Group Testing'. In: *SIAM Journal on Algebraic Discrete Methods* 2.2 (1981), pp. 81–87.

[98] F. K. Hwang. 'A Method for Detecting All Defective Members in a Population by Group Testing'. In: *Journal of the American Statistical Association* 67.339 (1972), pp. 605–608.

[99] F. Iliopoulos and I. Zadik. *Group testing and local search: is there a computational-statistical gap?* 2020. arXiv: 2011.05258 [math.ST].

[100] E. Ising. 'Beitrag zur Theorie des Ferromagnetismus'. In: *Zeitschrift für Physik* 31.1 (1925), pp. 253–258.

[101] A. Jagannath. 'Approximate Ultrametricity for Random Measures and Applications to Spin Glasses'. In: *Communications on Pure and Applied Mathematics* 70.4 (2017), pp. 611–664.

[102] S. Janson. *Graphons, cut norm and distance, couplings and rearrangements*. Vol. 4. NYJM Monographs. New York Journal of Mathematics, 2013.

[103] S. Janson, T. Luczak, A. Rucinski and R. Rucinski. *Random Graphs*. New York: Wiley, 2000.

[104] E. T. Jaynes. 'Information Theory and Statistical Mechanics'. In: *Phys. Rev.* 106 (4 1957), pp. 620–630.

[105] M. Jerrum. 'Large Cliques Elude the Metropolis Process'. In: *Random Structures and Algorithms* 3.4 (1992), pp. 347–359.

[106] A. Jimenez Felstrom and K. S. Zigangirov. 'Time-varying periodic convolutional codes with low-density parity-check matrix'. In: *IEEE Transactions on Information Theory* 45.6 (1999), pp. 2181–2191.

[107] A. Johansson, J. Kahn and V. Vu. 'Factors in random graphs'. In: *Random Structures & Algorithms* 33.1 (2008), pp. 1–28.

[108] O. Johnson, M. Aldridge and J. Scarlett. 'Performance of Group Testing Algorithms With Near-Constant Tests Per Item'. In: *IEEE Transactions on Information Theory* 65.2 (2019), pp. 707–723.

[109] R. M. Karp. 'Reducibility among Combinatorial Problems'. In: *Complexity of Computer Computations*. Springer US, 1972, pp. 85–103.

[110] S. Kirkpatrick, C. D. Gelatt and M. P. Vecchi. 'Optimization by Simulated Annealing'. In: *Science* 220.4598 (1983), pp. 671–680.

[111] S. Kirkpatrick, G. Györgyi, N. Tishby and L. Troyansky. 'The Statistical Mechanics of k-Satisfaction'. In: *Advances in Neural Information Processing Systems 6*. Ed. by J. D. Cowan, G. Tesauro and J. Alspector. Morgan-Kaufmann, 1994, pp. 439–446.

[112] J. Komlos, A. Shokoufandeh, M. Simonovits and E. Szemerédi. 'The Regularity Lemma and Its Applications in Graph Theory'. In: vol. 2292. 2000, pp. 84–112.

[113] J. Komlós, G. N. Sárközy and E. Szemerédi. 'On the Pósa-Seymour conjecture'. In: *Journal of Graph Theory* 29.3 (1998), pp. 167–176.

[114] J. Komlós, G. N. Sárközy and E. Szemerédi. 'Proof of the Seymour conjecture for large graphs'. In: *Annals of Combinatorics* 2.1 (1998), pp. 43–60.

[115] J. Komlós, G. N. Sárközy and E. Szemerédi. 'Spanning trees in dense graphs'. In: *Combinatorics, Probability and Computing* 10.5 (2001), pp. 397–416.

[116] J. Komlós and E. Szemerédi. 'Limit distribution for the existence of hamiltonian cycles in a random graph'. In: *Discrete Mathematics* 43.1 (1983), pp. 55–63.

[117] A. D. Korshunov. 'Solution of a problem of Erdős and Rényi on Hamiltonian cycles in nonoriented graphs'. In: *Dokl. Akad. Nauk SSSR* 228 (3 1976), pp. 529–532.

[118] M. Krivelevich. 'Embedding spanning trees in random graphs'. In: *SIAM Journal on Discrete Mathematics* 24.4 (2010), pp. 1495–1500.

[119] M. Krivelevich, M. Kwan and B. Sudakov. 'Bounded-degree spanning trees in randomly perturbed graphs'. In: *SIAM Journal on Discrete Mathematics* 31.1 (2017), pp. 155–171.

[120] M. R. Krom. 'The Decision Problem for a Class of First-Order Formulas in Which all Disjunctions are Binary'. In: *Mathematical Logic Quarterly* 13.1-2 (1967), pp. 15–20.

[121] F. Krzakala, A. Montanari, F. Ricci-Tersenghi, G. Semerjian and L. Zdeborová. 'Gibbs states and the set of solutions of random constraint satisfaction problems'. In: *Proceedings of the National Academy of Sciences* 104.25 (2007), pp. 10318–10323.

[122] F. R. Kschischang, B. J. Frey and H. Loeliger. 'Factor graphs and the sum-product algorithm'. In: *IEEE Transactions on Information Theory* 47.2 (2001), pp. 498–519.

[123] S. Kudekar, T. Richardson and R. L. Urbanke. 'Spatially Coupled Ensembles Universally Achieve Capacity Under Belief Propagation'. In: *IEEE Transactions on Information Theory* 59.12 (2013), pp. 7761–7813.

[124] S. Kudekar, T. J. Richardson and R. L. Urbanke. 'Threshold Saturation via Spatial Coupling: Why Convolutional LDPC Ensembles Perform So Well over the BEC'. In: *IEEE Transactions on Information Theory* 57.2 (2011), pp. 803–834.

[125] D. Kühn and D. Osthus. 'On Pósa's conjecture for random graphs'. In: *SIAM Journal on Discrete Mathematics* 26.3 (2012), pp. 1440–1457.

[126] M. Lelarge. 'Bypassing correlation decay for matchings with an application to XORSAT'. In: *2013 IEEE Information Theory Workshop (ITW)*. 2013, pp. 1–5.

[127] L. Lovász. *Large networks and graph limits*. Providence, Rhode Island: American Mathematical Society, 2012.

[128] L. Lovász and B. Szegedy. 'Limits of Dense Graph Sequences'. In: *J. Comb. Theory Ser. B* 96.6 (2006), pp. 933–957.

[129] A. Maleki and A. Montanari. 'Analysis of approximate message passing algorithm'. In: *2010 44th Annual Conference on Information Sciences and Systems (CISS)*. 2010, pp. 1–7.

[130] S. Mallapaty. 'The mathematical strategy that could transform coronavirus testing'. In: *Nature* 583.7817 (2020), pp. 504–505.

[131] E. Marinari, G. Parisi, F. Ricci-Tersenghi, J. J. Ruiz-Lorenzo and F. Zuliani. In: *Journal of Statistical Physics* 98.5/6 (2000), pp. 973–1074.

[132] M. Mezard and C. Toninelli. 'Group Testing With Random Pools: Optimal Two-Stage Algorithms'. In: *Information Theory, IEEE Transactions on* 57 (2011), pp. 1736–1745.

[133] M. Mézard, M. Tarzia and C. Toninelli. 'Statistical physics of group testing'. In: *Journal of Physics: Conference Series* 95 (2008), p. 012019.

[134] M. Mézard and G. Parisi. 'The Bethe lattice spin glass revisited'. In: *The European Physical Journal B* 20.2 (2001), pp. 217–233.

[135] M. Mézard and A. Montanari. *Information, Physics, and Computation*. Oxford University Press, Inc., 2009.

[136] M. Molloy. 'Models and thresholds for random Constraint Satisfaction Problems.' In: *Conference Proceedings of the Annual ACM Symposium on Theory of Computing* (2002).

[137] A. Montanari. 'Estimating random variables from random sparse observations'. In: *European Transactions on Telecommunications* 19.4 (2008), pp. 385–403.

[138] A. Montanari, F. Ricci-Tersenghi and G. Semerjian. 'Clusters of solutions and replica symmetry breaking in random k-satisfiability'. In: *Journal of Statistical Mechanics: Theory and Experiment* 2008.04 (2008), P04004.

[139] R. Montgomery. 'Spanning trees in random graphs'. In: *Advances in Mathematics* 356 (2019), p. 106793.

[140] T. Mora and L. Zdeborová. 'Random Subcubes as a Toy Model for Constraint Satisfaction Problems'. In: *Journal of Statistical Physics* 131.6 (2008), pp. 1121–1138.

[141] R. Nenadov and N. Škorić. 'Powers of Hamilton cycles in random graphs and tight Hamilton cycles in random hypergraphs'. In: *Random Structures & Algorithms* 54.1 (2019), pp. 187–208.

[142] D. Panchenko and M. Talagrand. 'Bounds for diluted mean-fields spin glass models'. In: *Probability Theory and Related Fields* 130.3 (2004), pp. 319–336.

[143] G. Parisi. 'A sequence of approximated solutions to the S-K model for spin glasses'. In: *Journal of Physics A: Mathematical and General* 13.4 (1980), pp. L115–L121.

[144] G. Parisi, F. Ricci-Tersenghi and T. Rizzo. 'Diluted Mean-Field Spin-Glass Models at Criticality'. In: *Journal of Statistical Mechanics: Theory and Experiment* 2014 (2014).

[145] J. Pearl. 'Reverend Bayes on Inference Engines: A Distributed Hierarchical Approach'. In: *Proceedings of the Second AAAI Conference on Artificial Intelligence*. AAAI'82. Pittsburgh, Pennsylvania: AAAI Press, 1982, pp. 133–136.

[146] J. Pearl. 'Chapter 4 - Belief Updating by Network Propagation'. In: *Probabilistic Reasoning in Intelligent Systems*. Ed. by J. Pearl. San Francisco (CA): Morgan Kaufmann, 1988, pp. 143–237.

[147] L. Pósa. 'Hamiltonian circuits in random graphs'. In: *Discrete Mathematics* 14.4 (1976), pp. 359–364.

[148] P. Raghavendra and N. Tan. 'Approximating CSPs with Global Cardinality Constraints Using SDP Hierarchies'. In: *Proceedings of the Annual ACM-SIAM Symposium on Discrete Algorithms* (2011).

[149] L. Riccio and C. J. Colbourn. 'Sharper Bounds in Adaptive Group Testing'. In: *Taiwanese Journal of Mathematics* 4.4 (2000), pp. 669–673.

[150] O. Riordan. 'Spanning subgraphs of random graphs'. In: *Combinatorics, Probability and Computing* 9.2 (2000), pp. 125–148.

[151] V. Rödl and M. Schacht. 'Regularity Lemmas for Graphs'. In: *Fete of Combinatorics and Computer Science*. Springer Berlin Heidelberg, 2010, pp. 287–325.

[152] J. Scarlett and V. Cevher. 'Limits on support recovery with probabilistic models: An information-theoretic framework'. In: *2015 IEEE International Symposium on Information Theory (ISIT)*. 2015, pp. 2331–2335.

[153] J. Scarlett. 'Noisy Adaptive Group Testing: Bounds and Algorithms'. In: *IEEE Transactions on Information Theory* 65.6 (2019), pp. 3646–3661.

[154] J. Scarlett and V. Cevher. 'Phase Transitions in Group Testing'. In: *Proceedings of the Twenty-Seventh Annual ACM-SIAM Symposium on Discrete Algorithms*. Society for Industrial and Applied Mathematics, 2015.

[155] J. Scarlett and V. Cevher. 'Phase Transitions in the Pooled Data Problem'. In: *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems*. 2017, pp. 376–384.

[156] G. Semerjian and R. Monasson. 'A Study of Pure Random Walk on Random Satisfiability Problems with "Physical" Methods'. In: *Theory and Applications of Satisfiability Testing*. Springer Berlin Heidelberg, 2004, pp. 120–134.

[157] J. M. L. Sengers. 'Mean-field theories, their weaknesses and strength'. In: *Fluid Phase Equilibria* 158-160 (1999), pp. 3–17.

[158] D. Sherrington and S. Kirkpatrick. 'Solvable Model of a Spin-Glass'. In: *Phys. Rev. Lett.* 35 (26 1975), pp. 1792–1796.

[159] D. A. Spielman and S.-H. Teng. 'Smoothed Analysis of Algorithms: Why the Simplex Algorithm Usually Takes Polynomial Time'. In: *J. ACM* 51.3 (2004), pp. 385–463.

[160] E. Szemerédi. 'On sets of integers containing no k elements in arithmetic progression.' In: *Acta Arithmetica* 27 (1975), pp. 199–245.

[161] M. Talagrand. 'Multiple levels of symmetry breaking'. In: *Probability Theory and Related Fields* 117.4 (2000), pp. 449–466.

[162] M. Talagrand. 'The Parisi Formula'. In: *Annals of Mathematics* 163.1 (2006), pp. 221–263.

[163] T. Tao. *Szemeredi's regularity lemma via random partitions*. 2019. URL: https://terrytao.wordpress.com/2009/04/26/szemeredis-regularity-lemma-via-random-partitions/ (visited on 30/11/2020).

[164] P. Ungar. 'The cutoff point for group testing'. In: *Communications on Pure and Applied Mathematics* 13.1 (1960), pp. 49–54.

[165] L. G. Valiant. 'The Complexity of Enumeration and Reliability Problems'. In: *SIAM Journal on Computing* 8.3 (1979), pp. 410–421.

[166] L. Viana and A. J. Bray. 'Phase diagrams for dilute spin glasses'. In: *Journal of Physics C: Solid State Physics* 18.15 (1985), pp. 3037–3051.

[167] E. Vincent and V. Dupuis. 'Spin Glasses: Experimental Signatures and Salient Outcomes'. In: *Frustrated Materials and Ferroic Glasses*. Springer International Publishing, 2018, pp. 31–56.

[168] L. Zdeborová and F. Krzakala. 'Statistical physics of inference: thresholds and algorithms'. In: *Advances in Physics* 65.5 (2016), pp. 453–552.

[169] L. Zdeborová and M. Mézard. ‘Constraint satisfaction problems with isolated solutions are hard’. In: *Journal of Statistical Mechanics: Theory and Experiment* 2008.12 (2008), P12004.

[170] Y. Zhang. *Phase Transitions of Random Constraints Satisfaction Problem*. UC Berkley, 2017.

A. Contained publications and the author's contributions

All contained publications of this dissertation are included in their current arXiv version. This section provides a detailed overview about the author's contributions as well as the publication status of each of those manuscripts. We will explicitly only state the contributions of this thesis's author (MHK). Therefore, the papers might contain results achieved by different authors whose contributions are not discussed below. For the sake of readability, we will abbreviate all author's names to their initials.

Information-theoretic and algorithmic thresholds for group testing This manuscript by A. Coja-Oghlan, O. Gebhard, M. Hahn-Klimroth and P. Loick appeared in the *IEEE Transactions on Information Theory* [41] and a short version appeared in the *Proceedings of the 46th ICALP* [44].

While the idea of this paper was found during the master's thesis of OG (supervised by MHK and PL), the further main contributions of MHK are the development and formalisation of Theorem 1.1 and its proof in joint work with PL as well as the formalisation of the proof of Theorem 1.2 based on an idea of PL.

Optimal Group Testing A short version of this article by A. Coja-Oghlan, O. Gebhard, M. Hahn-Klimroth and P. Loick appeared in the *Proceedings of 33rd Conference on Learning Theory (COLT)* [45] and the full version is accepted for publication at *Combinatorics, Probability and Computing*.

The algorithmic achievability result with respect to adaptive group testing (Theorem 1.3) was formally proven by MHK and PL while the corresponding non-adaptive result (Theorem 1.2) is joint work of all authors. The main technical contribution to Theorem 1.2, thus the derivation of the correct weights as well as the formal proof are due to ACO, MHK and PL. Furthermore, the proof idea of Theorem 1.1 based on a generalised argument of Aldridge was discussed by MHK and PL while the formal derivation, i.e. the reduction on a different prevalence is due to ACO, MHK and PL. Finally, all authors contributed to the writing of the manuscript.

Near-Optimal Sparsity-Constrained Group Testing: Improved Bounds and Algorithms This paper by O. Gebhard, M. Hahn-Klimroth, O. Parczyk, M. Penschuck, M. Rolvien, J. Scarlett and N. Tan is under review at *IEEE Transactions on Information Theory*.

To be more precise, OG, MHK, OP, MP, MR uploaded a draft on non-adaptive sparsity constrained group testing to arXiv and contemporaneously JS and NT published a preprint on adaptive methods. As those contributions could be perfectly merged all authors decided to extend their results and combine them.

While the writing of the manuscript was led by OG, MHK and JS, the development and formal elaboration of the non-adaptive converse bounds with respect to information theory (Theorem 3.2 and Theorem 4.1) are joint equal work of MHK and OP. A further contribution of MHK is the development of the proof idea of Theorem 3.4 (`DD` converse in the $\Delta-$divisible model) and its formal justification was joint work of MHK and OP. The corresponding achievability statement (Theorem 3.3) is joint work of OG and MHK. Moreover, the achievability statement for `DD` in the $\Gamma-$sparse model is joint work of OG and MHK (Theorem 4.10) and MHK and OP (Theorem 4.18) respectively.

The number of satisfying assignments of random 2-SAT formulas. This paper is joint work of D. Achlioptas, A. Coja-Oghlan, M. Hahn-Klimroth, J. Lee, N. Müller, M. Penschuck and G. Zhou and it appeared in *Random Structures & Algorithms*. The current arXiv version is entitled *The random 2-SAT partition function*. The main contribution of MHK is the intuitive description of how to obtain a worst-case boundary condition in order to prove that the Boltzmann distribution is a Bethe state while its formal justification is due to NM. A further contribution of MHK are some technical derivations in order

to extend the positive temperature concentration result to the zero temperature limit while the main idea of this approach and the detailed proof were provided by ACO.

The cut metric for probability distributions This article by A. Coja-Oghlan and M. Hahn-Klimroth is accepted for publication at the *SIAM Journal of Discrete Mathematics*. It is a joint project of MHK with his PhD advisor ACO. While the latter led the writing of the manuscript, the theorems and their proofs were obtained and discussed jointly.

Random perturbation of sparse graphs This manuscript by M. Hahn-Klimroth, G. Maesaka, Y. Mogge, S. Mohr and O. Parczyk is currently submitted to the *Electronic Journal of Combinatorics*.

While the idea for the paper was brought by OP to a workshop on extremal and probabilistic combinatorics, Theorems 1.1 and 1.2 which establish a sufficient condition for observing a Hamilton cycle and a perfect matching in $\mathcal{G}(n,\beta/n)\cup\mathcal{G}_\alpha$ were discussed intensively by OP and MHK and formally proven and written down by MHK.

B. Information-theoretic and algorithmic thresholds for group testing

INFORMATION-THEORETIC AND ALGORITHMIC THRESHOLDS FOR GROUP TESTING

AMIN COJA-OGHLAN, OLIVER GEBHARD, MAX HAHN-KLIMROTH, PHILIPP LOICK

ABSTRACT. In the group testing problem we aim to identify a small number of infected individuals within a large population. We avail ourselves to a procedure that can test a group of multiple individuals, with the test result coming out positive iff at least one individual in the group is infected. With all tests conducted in parallel, what is the least number of tests required to identify the status of all individuals? In a recent test design [Aldridge et al. 2016] the individuals are assigned to test groups randomly with replacement, with every individual joining an almost equal number of groups. We pinpoint the sharp threshold for the number of tests required in this randomised design so that it is information-theoretically possible to infer the infection status of every individual. Moreover, we analyse two efficient inference algorithms. These results settle conjectures from [Aldridge et al. 2014, Johnson et al. 2019].

1. INTRODUCTION

1.1. **Background and motivation.** The group testing problem goes back to the work of Dorfman from the 1940s [24]. Among a large population a few individuals are infected with a rare disease. The objective is to identify the infected individuals effectively. At our disposal we have a testing procedure capable of not merely testing one individual, but several. The test result will be positive if at least one individual in the test group is infected, and negative otherwise; all tests are conducted in parallel. We are at liberty to assign a single individual to several test groups. The aim is to devise a test design that identifies the status of every single individual correctly while requiring as small a number of tests as possible. A recently proposed test design allocates the individuals to tests randomly [10, 12, 13, 30, 33]. To be precise, given integers $n, m, \Delta > 0$ we create a random bipartite multi-graph by choosing independently for each of the n vertices $x_1, \ldots, x_n$ 'at the top' Δ neighbours among the m vertices $a_1, \ldots, a_m$ 'at the bottom' uniformly at random with replacement. The vertices $x_1, \ldots, x_n$ represent the individuals, the $a_1, \ldots, a_m$ represent the test groups and an individual joins a test group iff the corresponding vertices are adjacent (see Figure 1). The wisdom behind this construction is that the expansion properties of the random bipartite graph precipitate virtuous correlations, facilitating inference. Given n and (an estimate of) the number k of infected individuals, what is the least m for which, with a suitable choice of Δ, the status of every individual can be inferred correctly from the test results with high probability?Like in many other inference problems the answer comes in two instalments. First, we might ask for what m it is *information-theoretically* possible to detect the infected individuals. In other words, regardless of computational resources, do the test results contain enough information in principle to identify the infection status of every individual? Second, for what m does this problem admit *efficient algorithms*? The first main result of this paper resolves the information-theoretic question completely. Specifically, Aldridge, Johnson and Scarlett [13] obtained a function $m_{\mathrm{inf}} = m_{\mathrm{inf}}(n,k)$ such that for any fixed $\varepsilon > 0$ the inference problem is information-theoretically infeasible if $m < (1-\varepsilon)m_{\mathrm{inf}}$. They conjectured that this bound is tight, i.e., that for $m > (1+\varepsilon)m_{\mathrm{inf}}(n,k)$ there is an (exponential) algorithm that correctly identifies the infected individuals with high probability. We prove this conjecture. Furthermore, concerning the algorithmic question, Johnson, Aldridge and Scarlett [30] obtained a function $m_{\mathrm{alg}} = m_{\mathrm{alg}}(n,k)$ that exceeds m_{inf} by a constant factor for small k such that for $m > (1+\varepsilon)m_{\mathrm{alg}}$ certain efficient algorithms successfully identify the infected individuals with high probability. They conjectured that SCOMP, their most sophisticated algorithm, actually succeeds for smaller values of m. We refute this conjecture and show that SCOMP asymptotically fails to outperform a much simpler algorithm called DD. A technical novelty of the present work is that we investigate the group testing problem from a new perspective. While most prior contributions rely either on elementary calculations and/or information-theoretic arguments [12, 13, 30, 39], here we bring to bear techniques from the theory of random constraint satisfaction problems [5, 32].

Supported by DFG CO 646/3 and Stiftung Polytechnische Gesellschaft. An extended abstract of this work appeared in the 2019 ICALP proceedings. A revised version is to appear in IEEE Transactions on Information Theory (Copyright (c) 2017 IEEE DOI: 10.1109/TIT.2020.3023377).

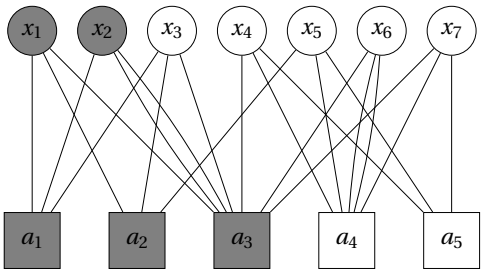

FIGURE 1. The graph illustrates a small example of a group testing instance, with the individuals $x_1,\dots,x_7$ at the top and the tests $a_1,\dots,a_5$ at the bottom. Infected individuals and positive tests are coloured in grey.

Indeed, group testing can be viewed naturally as a constraint satisfaction problem: the tests provide the constraints and the task is to find all possible ways of assigning a status ('infected' or 'not infected') to the n individuals in a way consistent with the given test results. Since the allocation of individuals to tests is random, this question is similar in nature to, e.g., the random k-SAT problem that asks for a Boolean assignment that satisfies a random collection of clauses [4, 6, 20, 23]. It also puts the group testing problem in the same framework as the considerable body of recent work on other inference problems on random graphs such as the stochastic block model (e.g., [1, 18, 22, 35, 37, 43]) or decoding from pooled data [7, 8].

We proceed to state the main results of the paper precisely, followed by a detailed discussion of the prior literature on group testing. The proofs of the information-theoretic and algorithmic bounds follow in 3, Section 4, and 5. The technical details can be found in the appendix.

1.2. **The information-theoretic threshold.** Throughout the paper we labour under the assumptions commonly made in the context of group testing; we will revisit their merit in Section 1.4. Specifically, we assume that the number k of infected individuals satisfies $k \sim n^\theta$ for a fixed $0 < \theta < 1$ [1]. Moreover, let $\boldsymbol{\sigma} \in \{0,1\}^{\{x_1,\dots,x_n\}}$ be a vector of Hamming weight k chosen uniformly at random. The (one-)entries of $\boldsymbol{\sigma}$ indicate which of the n individuals are infected. Moreover, let $\boldsymbol{G} = \boldsymbol{G}(n,m,\Delta)$ signify the aforementioned random bipartite graph with multi-edges. Then $\boldsymbol{\sigma}$ induces a vector $\hat{\boldsymbol{\sigma}} \in \{0,1\}^{\{a_1,\dots,a_m\}}$ that indicates which of the m tests come out positive. To be precise, $\hat{\boldsymbol{\sigma}}_i = 1$ iff test a_i is adjacent to an individual x_j with $\boldsymbol{\sigma}_{x_j} = 1$. For what m is it possible to recover $\boldsymbol{\sigma}$ from $\boldsymbol{G}, \hat{\boldsymbol{\sigma}}$? (Throughout the paper all logarithms are base e.)

Theorem 1.1. *Suppose that* $0 < \theta < 1$, $k \sim n^\theta$ *and* $\varepsilon > 0$ *and let*

$$m_{\inf} = m_{\inf}(n,\theta) = \frac{k\log(n/k)}{\min\left\{1, \frac{1-\theta}{\theta}\log 2\right\}\log 2}.$$

(i) If $m > (1+\varepsilon)m_{\inf}(n,\theta)$, *then there exists an algorithm that given* $\boldsymbol{G}, \hat{\boldsymbol{\sigma}}$ *outputs* $\boldsymbol{\sigma}$ *with high probability.*
(ii) If $m < (1-\varepsilon)m_{\inf}(n,\theta)$, *then there does not exist any algorithm that given* $\boldsymbol{G}, \hat{\boldsymbol{\sigma}}, k$ *outputs* $\boldsymbol{\sigma}$ *with a non-vanishing probability.*

Since for $\theta \le \log(2)/(1+\log(2))$ the first part of Theorem 1.1 readily follows from a folklore argument [25], the interesting regime is $\theta > \log(2)/(1+\log(2)) \approx 0.41$. The negative part of Theorem 1.1 strengthens a result from [13], who showed that for $m < (1-\varepsilon)m_{\inf}$ any inference algorithm has a strictly positive error probability. By comparison, Theorem 1.1 shows that any algorithm fails with *high* probability.

But the main contribution of Theorem 1.1 is the first, positive statement. While the problem was solved for $\theta < 1/3$ for a different test design [39, 40] and the case $\theta > 1/2$ is easy because a plain greedy algorithm succeeds [30], the case $1/3 < \theta < 1/2$ proved more challenging. Only heuristic arguments predicting the result of Theorem 1.1 have been put forward for this regime so far [33]. Indeed, Aldridge et al. [12] conjectured that in this case inferring $\boldsymbol{\sigma}$ from $\boldsymbol{G}, \hat{\boldsymbol{\sigma}}$ is equivalent to solving a hypergraph minimum vertex cover problem. The proof of Theorem 1.1 vindicates this conjecture. Specifically, the vertex set of the hypergraph comprises all 'potentially infected' individuals, i.e., those that do not appear in any negative test. The hyperedges are the neighbourhoods

[1] While we write that $k \sim n^\theta$ for the sake of brevity, our results immediately extend to the case $k \sim Cn^\theta$ for some constant C.

∂a_i of the positive tests a_i in $\boldsymbol{G}$. Exhaustive search solves this vertex cover problem in time $\exp(O(n^\theta \log n))$. But how about efficient algorithms for general θ?

1.3. **Efficient algorithms for group testing.** Several polynomial time group testing algorithms have been proposed. A very simple greedy strategy called DD (for 'definitive defectives') first labels all individuals that are members of negative test groups as uninfected. Subsequently it checks for positive tests in which all individuals but one have been identified as uninfected in the first step. Clearly, the single as yet unlabelled individual in such a test group must be infected. Up to this point all decisions made by DD are correct. But in the final step DD marks all as yet unclassified individuals as uninfected, possibly causing false negatives. In fact, the output of DD may be inconsistent with the test results as possibly some positive tests may fail to include an individual classified as 'infected'. While an achievability result is known for the DD algorithm, a corollary of the work in this paper is a matching converse.

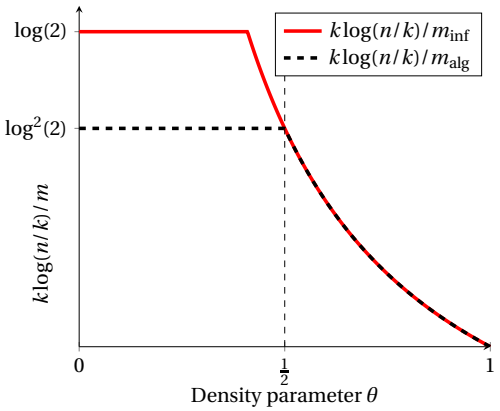

FIGURE 2. The red line shows the information theoretic threshold $m_{\inf}$, the dashed black line signifies the bound m_{alg} which is achieved by the both the SCOMP and the DD algorithm.

The more sophisticated SCOMP algorithm is roughly equivalent to the well-known greedy algorithm for the hypergraph vertex cover problem applied to the hypergraph from the previous paragraph. Specifically, in its first step SCOMP proceeds just like DD, classifying all individuals that occur in negative tests as uninfected. Then SCOMP identifies as infected all unmarked individuals that appear in at least one test whose other participants are already known to be uninfected. Subsequently the algorithm keeps picking an individual that appears in the largest number of as yet 'unexplained' (viz. uncovered) positive tests and marks that individual as infected, with ties broken randomly, until every positive test contains an individual classified as infected. Clearly, SCOMP may produce false positives as well as false negatives. But at least the output is consistent with the test results. Algorithm 1 summarises the procedure of SCOMP.

Input: $\boldsymbol{G}$, $\hat{\boldsymbol{\sigma}}$, k
Output: estimate of $\boldsymbol{\sigma}$
1 Classify all individuals in negative tests as healthy & remove such individuals and tests from $\boldsymbol{G}$;
2 Classify all individuals that appear in at least one positive test as the only yet unclassified individuals as infected & remove such individuals and tests from $\boldsymbol{G}$;
3 **while** *there exists at least one test in* $\boldsymbol{G}$ **do**
4 Classify the individual appearing in the largest number of remaining tests as infected & remove this individual and all adjacent tests from $\boldsymbol{G}$
5 Classify all remaining individuals as healthy;

Algorithm 1: Description of the SCOMP algorithm

Analysing SCOMP has been prominently posed as an open problem in the group testing literature [9, 12, 30]. Indeed, Aldridge et al. [12] opined that "the complicated sequential nature of SCOMP makes it difficult to analyse mathematically". On the positive side, [12] proved that SCOMP succeeds in recovering $\boldsymbol{\sigma}$ correctly given $(\boldsymbol{G}, \hat{\boldsymbol{\sigma}})$ if $m > (1+\varepsilon) m_{\text{alg}}(n,\theta)$ w.h.p.[2], where

$$m_{\text{alg}} = m_{\text{alg}}(n,\theta) = \frac{k \log(n/k)}{\min\left\{1, \frac{1-\theta}{\theta}\right\} \log^2 2}. \tag{1}$$

[2]W.h.p.refers to a probability of $1 - o(1)$ as $n \to \infty$.

However, the algorithm succeeds for a trivial reason; namely, for $m > (1+\varepsilon)m_{\mathrm{alg}}$ even DD suffices to recover $\boldsymbol{\sigma}$ w.h.p. Yet based on experimental evidence [12, 30] conjectured that SCOMP strictly outperforms DD. The following theorem refutes this conjecture.

Theorem 1.2. *Suppose that* $0 < \theta < 1$ *and* $\varepsilon > 0$. *If* $m < (1-\varepsilon)m_{\mathrm{alg}}(n,\theta)$, *then given* $\boldsymbol{G}, \hat{\boldsymbol{\sigma}}$ *w.h.p. both* SCOMP *and* DD *fail to output* $\boldsymbol{\sigma}$.

For $\theta < 1/2$ the information-theoretic bound provided by Theorem 1.1 and the algorithmic bound m_{alg} supplied by Theorem 1.2 remain a modest constant factor apart; see Figure 2. Whether there exists an efficient algorithm for group testing that can close the gap to the information-theoretic bound has long been an open research question. A recent result by Coja-Oghlan et al. [19] shows that such a polynomial-time algorithm indeed exists. The proposed algorithm which is inspired by the notion of spatial coupling from coding theory is able to recover $\boldsymbol{\sigma}$ whenever $m > (1+\varepsilon)m_{\mathrm{inf}}$. Moreover, the authors prove that below the information-theoretic threshold from Theorem 1.1 no non-adaptive algorithm can succeed under any test design (not only the random regular test design considered here) thereby establishing the presence of an adaptivity gap in the group testing problem. An exciting avenue for future research is to investigate the merits of the results and techniques of this paper and [19, 28] for the noisy variant of group testing.

1.4. **Discussion and related work.** Dorfman's original group testing scheme, intended to test the American army for syphilis, was *adaptive*. In a first round of tests each soldier would be allocated to precisely one test group. If the test result came out negative, none of the soldiers in the group were infected. In a second round the soldiers whose group was tested positively would be tested individually. Of course, Dorfman's scheme was not information-theoretically optimal. A first-order optimal adaptive scheme that involves several test stages, with the tests conducted in the present stage governed by the results from the previous stages, is known [15, 25]. In the adaptive scenario the information-theoretic threshold works out to be

$$m_{\mathrm{inf}}^{\mathrm{adapt}}(n,\theta) = \frac{k\log(n/k)}{\log 2}.$$

The lower bound, i.e., that no adaptive design gets by with $(1-\varepsilon)m_{\mathrm{inf}}^{\mathrm{adapt}}(n,\theta)$ tests, follows from a very simple information-theoretic consideration. Namely, with a total of m tests at our disposal there are merely 2^m possible test outcomes, and we need this number to exceed the count $\binom{n}{k}$ of possible vectors $\boldsymbol{\sigma}$, i.e., [14].

More recently there has been a great deal of interest in non-adaptive group testing, where the infection status of each individual is to be determined after just one round of tests [14, 17, 27, 33]. This is the version of the problem that we deal with in the present paper. An important advantage of the non-adaptive scenario is that tests, which may be time-consuming, can be conducted in parallel. Indeed, some of today's most popular applications of group testing are non-adaptive such as DNA screening [17, 31, 38] or protein interaction experiments [36, 42] in computational molecular biology. The randomised test design that we deal with here is the best currently known non-adaptive design (in terms of the number of tests required).

The most interesting regime for the group testing problem is when the number k of infected individuals scales as a power n^θ of the entire population. Mathematically this is because in the linear regime $k = \Omega(n)$ the optimal strategy is to perform n individual tests [11] in order to achieve a vanishing error probability. Similarly, the case of constant k has been solved for some time [41]. Thus, for k linear in n and k constant the theory is already well established. But the sublinear case is also of practical relevance, as witnessed by Heap's law in epidemiology [16] or biological applications [27].

Apart from the randomised test design $\boldsymbol{G}$ where each individual chooses precisely Δ tests (with replacement), the so-called Bernoulli design assigns each individual to every test with a certain probability independently. A considerable amount of attention has been devoted to this model, and its information-theoretic threshold as well as the thresholds for various algorithms have been determined [9, 10, 12, 39]. However, the Bernoulli test design, while easier to analyse, for $\theta > 1/3$ is provably inferior to the test design $\boldsymbol{G}$ that we study here. This is because in the Bernoulli design there are likely quite a few individuals that participate in far fewer tests than expected due to degree fluctuations. We note that our proofs can easily be adapted to reprove the known results for the Bernoulli design. In fact, many technical parts of the proofs become significantly easier and shorter, since we can assume independence between tests, whereas for the constant-column design under consideration here gives rise to subtle dependencies between the tests. A significant portion of the tests is devoted to getting a handle o these dependencies.

1.5. **Notation.** Throughout the paper $\boldsymbol{G} = \boldsymbol{G}(n, m, \Delta)$ denotes the random bipartite graph that describes which individuals take part in which test groups, the vector $\boldsymbol{\sigma} \in \{0,1\}^{\{x_1,\dots,x_n\}}$ encodes which individuals are infected, and $\hat{\boldsymbol{\sigma}} \in \{0,1\}^{\{a_1,\dots,a_m\}}$ indicates the test results. Clearly, $\boldsymbol{G}$ is independent of $\boldsymbol{\sigma}$. Moreover, $k \sim n^\theta$ signifies the number of infected individuals. Additionally, we write

$$V = V_n = \{x_1, \dots, x_n\}, \qquad V_0 = \{x_i \in V : \boldsymbol{\sigma}_{x_i} = 0\} \qquad \text{and} \qquad V_1 = V \setminus V_0$$

for the set of all individuals, the set of uninfected and infected individuals, respectively. For an individual $x \in V$ we write ∂x for the multi-set of tests a_i adjacent to x with $|\partial x| = \Delta$. Analogously, for a test a_i we denote by ∂a_i the multi-set of individuals that take part in the test and $\Gamma_i = |\partial a_i|$. These are multi-sets since individuals are assigned to tests uniformly at random with replacement and therefore $\boldsymbol{G}$ features multi-edges w.h.p.. Let Γ be the vector $(\Gamma_i)_{i\in[m]}$. Furthermore, all asymptotic notation refers to the limit $n \to \infty$. Thus, $o(1)$ denotes a term that vanishes in the limit of large n, while $\omega(1)$ stands for a function that diverges to ∞ as $n \to \infty$. We also let $c, d > 0$ denote reals such that

$$m = ck\log(n/k) \qquad\qquad \Delta = d\log(n/k).$$

Later, we will prove that $c, d = \Theta(1)$ as $n \to \infty$ is optimal for inference. Finally, let $\Gamma_{\min} = \min_{i\in[m]} \Gamma_i$, $\Gamma_{\max} = \max_{i\in[m]} \Gamma_i$. The following sections will outline the proofs of the information-theoretic bounds and the analysis of the SCOMP algorithm and feature the important proofs. The technical details are left to the appendix

2. Getting started

The very first item on the agenda is to get a handle on the posterior distribution of $\boldsymbol{\sigma}$ given $\boldsymbol{G}$ and $\hat{\boldsymbol{\sigma}}$. To this end, let $S_k(\boldsymbol{G}, \hat{\boldsymbol{\sigma}})$ be the set of all vectors $\sigma \in \{0,1\}^V$ of Hamming weight k such that

$$\hat{\boldsymbol{\sigma}}_{a_i} = \mathbf{1}\{\exists x \in \partial a_i : \sigma_x = 1\} \qquad\qquad \text{for all } i \in [m].$$

In words, $S_k(\boldsymbol{G}, \hat{\boldsymbol{\sigma}})$ contains the set of all vectors σ with k ones that label the individuals infected/uninfected in a way consistent with the test results, i.e. that are "satisfying sets" [12, 14]. Let $Z_k(\boldsymbol{G}, \hat{\boldsymbol{\sigma}}) = |S_k(\boldsymbol{G}, \hat{\boldsymbol{\sigma}})|$. The following proposition shows that the posterior of $\boldsymbol{\sigma}$ given $\boldsymbol{G}, \hat{\boldsymbol{\sigma}}$ is uniform on $S_k(\boldsymbol{G}, \hat{\boldsymbol{\sigma}})$.

Proposition 2.1 ([10])**.** *For all $\tau \in \{0,1\}^{\{x_1,\dots,x_n\}}$ we have $\mathbb{P}[\boldsymbol{\sigma} = \tau \mid \boldsymbol{G}, \hat{\boldsymbol{\sigma}}] = \dfrac{\mathbf{1}\{\tau \in S_k(\boldsymbol{G}, \hat{\boldsymbol{\sigma}})\}}{Z_k(\boldsymbol{G}, \hat{\boldsymbol{\sigma}})}$.*

Adopting the jargon of the recent literature on inference problems on random graphs, we refer to Proposition 2.1 as the *Nishimori identity* [18, 43]. The proposition shows that apart from the actual test results, there is no further 'hidden information' about $\boldsymbol{\sigma}$ encoded in $\boldsymbol{G}, \hat{\boldsymbol{\sigma}}$. In particular, the information-theoretically optimal inference algorithm just outputs a uniform sample from $S_k(\boldsymbol{G}, \hat{\boldsymbol{\sigma}})$. In effect, we obtain the following.

Corollary 2.2. *(1) If $Z_k(\boldsymbol{G}, \hat{\boldsymbol{\sigma}}) = \omega(1)$ w.h.p., then for any algorithm $\mathscr{A}$ we have*

$$\mathbb{P}[\mathscr{A}(\boldsymbol{G}, \hat{\boldsymbol{\sigma}}, k) = \boldsymbol{\sigma}] = o(1).$$

(2) If $Z_k(\boldsymbol{G}, \hat{\boldsymbol{\sigma}}) = 1$ w.h.p., then there is an algorithm $\mathscr{A}$ such that

$$\mathbb{P}[\mathscr{A}(\boldsymbol{G}, \hat{\boldsymbol{\sigma}}, k) = \boldsymbol{\sigma}] = 1 - o(1).$$

Both the positive and the negative part of Corollary 2.2 assume that the precise number k of infected individuals is known to the algorithm. This assumption makes the negative part stronger, but weakens the positive part. Yet we will see in due course how in the positive scenario the assumption that k be known can be removed.

For the information-theoretic bound, the proof hinges on analysing the number of individuals that can be flipped without affecting the test results. We encounter two kinds of such individuals. The first kind consists of healthy individuals that only appear in positive tests and which we will denote by V_0^+. In symbols,

$$V_0^+ = \{x_i \in V_0 : \forall a \in \partial x_i \exists y \in \partial a : \boldsymbol{\sigma}_y = 1\}. \tag{2}$$

Similarly, let V_1^+ be the set of all infected individuals x_i such that every test in which x_i occurs features another infected individual; in symbols,

$$V_1^+ = \{x_i \in V_1 : \forall a \in \partial x_i \exists y \in \partial a \setminus \{x_i\} : \boldsymbol{\sigma}_y = 1\}.$$

We think of the individuals in V_0^+ as the 'potential false positives'. Indeed, if for any $x_i \in V_0^+$ we obtain $\boldsymbol{\sigma}'$ from $\boldsymbol{\sigma}$ by setting x_i to one, then $\boldsymbol{\sigma}'$ will render the same test results as $\boldsymbol{\sigma}$. Similarly, the individuals in V_1^+ are potential false negatives. For completeness, we also define V_0^- and V_1^- as

$$V_0^- = V_0 \setminus V_0^+ \qquad \text{and} \qquad V_1^- = V_1 \setminus V_1^+ \tag{3}$$

In the following, let us get a handle on the size of sets V_0^+ and V_1^+. Specifically, we prove the following five statements.

Proposition 2.3. *Let $c, d = \Theta(1)$. Then, the following statements hold w.h.p.*

(1) $\left|V_0^+\right| = (1+n^{-\Omega(1)})n\left(1-\exp(-d/c)\right)^{\Delta}$.
(2) If $k(1-\exp(-d/c))^{\Delta} \geq n^{\Omega(1)}$, *then* $\left|V_1^+\right| = n^{\Omega(1)}$.
(3) If $k(1-\exp(-d/c))^{\Delta} = o(1)$, *then* $\left|V_1^+\right| = o(1)$.
(4) If $c < \frac{\theta}{1-\theta}\frac{1}{\log^2 2}$, *then* $\left|V_1^+\right|, \left|V_0^+\right| = n^{\Omega(1)}$.
(5) If $c > \frac{\theta}{1-\theta}\frac{1}{\log^2 2}$, *then* $\left|V_1^+\right| = o(1)$.

The proof of Proposition 2.3, while not fundamentally difficult, requires a bit of care because we are dealing with a random bipartite multi-graph whose (test-)degrees scale as a power of n. In effect, the diameter of the bipartite graph is quite small and the neighbourhoods of different tests may have a sizeable intersection. The technical workout follows in Section B.6. In the next step, let us get a handle on the size of the test degrees.

Lemma 2.4. *With probability at least $1-o(n^{-2})$ we have*

$$\Delta n/m - \sqrt{\Delta n/m}\log n \leq \Gamma_{\min} \leq \Gamma_{\max} \leq \Delta n/m + \sqrt{\Delta n/m}\log n.$$

The proof of this and the subsequent elementary lemmas are included in Section B. Next, we calculate the number of positive and negative tests. Let $\boldsymbol{m}_1$ be the number of positive tests and let $\boldsymbol{m}_0$ be the number of negative tests. Clearly $\boldsymbol{m}_0 + \boldsymbol{m}_1 = m$.

Lemma 2.5. *With probability at least $1-o(n^{-2})$ we have*

$$\boldsymbol{m}_0 = \exp(-d/c)m + O(\sqrt{m}\log^2 n).$$

Finally, we justify that setting $c, d = \Theta(1)$ as $n \to \infty$ is optimal for inference. The fact that $c = \Theta(1)$ immediately follows from the information-theoretic counting bound, i.e., [14].

Lemma 2.6. *(1) If $\Delta = o(\log(n/k))$ and $m = \Theta(k\log(n/k))$, then $Z_k(\boldsymbol{G}, \hat{\boldsymbol{\sigma}}) = \omega(1)$ w.h.p.*
(2) If $\Delta = \omega(\log(n/k))$ and $m = \Theta(k\log(n/k))$, then $Z_k(\boldsymbol{G}, \hat{\boldsymbol{\sigma}}) = \omega(1)$ w.h.p.

3. The information-theoretic upper bound

We proceed to discuss the proof of Theorem 1.1. The proof of the first, positive statement and of the second, negative statement hinge on two separate arguments. We begin with the proof of the information-theoretic upper bound which is the principal achievement of the present work. The proof rests upon techniques that have come to play an important role in the theory of random constraint satisfaction problems. Specifically, we need to show that $Z_k(\boldsymbol{G}, \hat{\boldsymbol{\sigma}}) = 1$ w.h.p., i.e., that $\boldsymbol{\sigma}$ is the only assignment compatible with the test results w.h.p. We establish this result by combining two separate arguments. First, we use a moment calculation to show that w.h.p. there are no other solutions that have a small 'overlap' with $\boldsymbol{\sigma}$. Then we use an expansion argument to show that w.h.p. there are no alternative solutions with a big overlap. Both these arguments are variants of the arguments that have been used to study the solution space geometry of random constraint satisfaction problems such as random k-SAT or random k-XORSAT [3, 4, 26], as well as the freezing thresholds of random constraint satisfaction problems [2, 34]. Yet to our knowledge these methods have thus far not been applied to the group testing problem. In this section we choose $\Delta = \lceil \frac{m}{k}\log 2 \rceil$ which maximises the entropy of the test results. Formally, we define

$$Z_{k,\ell}(\boldsymbol{G}, \hat{\boldsymbol{\sigma}}) = |\{\sigma \in S_k(\boldsymbol{G}, \hat{\boldsymbol{\sigma}}) : \langle \boldsymbol{\sigma}, \sigma \rangle = \ell\}|$$

as the number of assignments $\sigma \in S_k(\boldsymbol{G}, \hat{\boldsymbol{\sigma}})$ different from the true configuration $\boldsymbol{\sigma}$ whose *overlap*

$$\langle \boldsymbol{\sigma}, \sigma \rangle = \sum_{i=1}^{n} \mathbf{1}\{\sigma_{x_i} = \boldsymbol{\sigma}_{x_i} = 1\}$$

with $\boldsymbol{\sigma}$ is equal to ℓ. The following two propositions rule out assignments with a small and a big overlap, respectively. In either case we choose $\Delta = \lceil \frac{m}{k} \log 2 \rceil$ to take its optimal value.

Proposition 3.1. *Let $\varepsilon > 0$ and $0 < \theta < 1$ and assume that $m > (1+\varepsilon) m_{\inf}(k,\theta)$. W.h.p. we have $Z_{k,\ell}(\boldsymbol{G}, \hat{\boldsymbol{\sigma}}) = 0$ for all $\ell < (1 - 1/\log n)k$.*

Proof. For $i \in [m]$ let Γ_i be the degree of a_i in $\boldsymbol{G}$, i.e., the number of edges incident with a_i; this number may exceed the number of different individuals that participate in test a_i as $\boldsymbol{G}$ may feature multi-edges. Let Γ be the σ-algebra generated by the random variables $(\Gamma_i)_{i \in [m]}$. Whenever we condition on Γ, we assume that the bounds from Lemma 2.4 and 2.5 hold. Given Γ we can generate $\boldsymbol{G}$ from the well-known *pairing model* [29]. Specifically, we create a set $\{x_i\} \times [\Delta]$ of Δ clones of each individual as well as sets $\{a_i\} \times [\Gamma_i]$ of clones of the tests. Then we draw a perfect matching of the complete bipartite graph on the vertex sets $\bigcup_{i=1}^{n} \{x_i\} \times [\Delta], \bigcup_{i=1}^{m} \{a_i\} \times [\Gamma_i]$ uniformly at random. For each matching edge linking a clone of x_i with a clone of a_j we insert an i-j-edge. The resulting bipartite random multi-graph has the same distribution as $\boldsymbol{G}$ given Γ. As an application of this observation we obtain for every integer $0 \le \ell < k$

$$\mathbb{E}[Z_{k,\ell}(\boldsymbol{G}, \hat{\boldsymbol{\sigma}}) \mid \Gamma] \le O\big((\Delta k)^{3/2}\big) \cdot \binom{k}{\ell}\binom{n-k}{k-\ell} \prod_{i=1}^{m} \big(1 - 2(1-k/n)^{\Gamma_i} + 2\,(1 - 2k/n + \ell/n)^{\Gamma_i}\big) \tag{4}$$

To see why (4) holds we use the linearity of expectation. The product of the two binomial coefficients simply accounts for the number of assignments σ that have overlap ℓ with $\boldsymbol{\sigma}$. Hence, with $\mathscr{S}$ the event that one specific $\sigma \in \{0,1\}^V$ that has overlap ℓ with $\boldsymbol{\sigma}$ belongs to $S_{k,\ell}(\boldsymbol{G}, \hat{\boldsymbol{\sigma}})$, we need to show that

$$\mathbb{P}[\mathscr{S} \mid \Gamma] \le O\big((\Delta k)^{3/2}\big) \prod_{i=1}^{m} 1 - 2(1-k/n)^{\Gamma_i} + 2\,(1 - 2k/n + \ell/n)^{\Gamma_i}. \tag{5}$$

By symmetry we may assume that $\boldsymbol{\sigma}_{x_i} = \mathbf{1}\{i \le k\}$ and that $\sigma_{x_i} = \mathbf{1}\{i \le \ell\} + \mathbf{1}\{k < i \le 2k - \ell\}$.

To establish (5) we harness the pairing model. Namely, given Γ we can think of each test a_i as a bin of capacity Γ_i. Moreover, we think of each clone (x_i, h), $h \in [\Delta]$, of an individual as a ball. The ball is labelled $(\boldsymbol{\sigma}_{x_i}, \sigma_{x_i}) \in \{0,1\}^2$. The random matching that creates $\boldsymbol{G}$ effectively tosses the Δn balls randomly into the bins. Hence, for $i \in [m]$ and for $j \in [\Gamma_i]$ let us write $\boldsymbol{A}_{i,j} = (\boldsymbol{A}_{i,j,1}, \boldsymbol{A}_{i,j,2}) \in \{0,1\}^2$ for the label of the jth ball that ends up in bin number i. Then we are left to calculate the probability that for every test a_i either $\boldsymbol{A}_{i,j,1} = \boldsymbol{A}_{i,j,2} = 0$ for every $j \in [\Gamma_i]$ or there is at least one pair $(j,k) \in [\Gamma_i]^2$ such that $\boldsymbol{A}_{i,j,1} = \boldsymbol{A}_{i,k,2} = 1$

$$\mathbb{P}[\mathscr{S} \mid \Gamma] = \mathbb{P}\left[\forall i \in [m] : \max_{j \in [\Gamma_i]} \boldsymbol{A}_{i,j,1} = \max_{j \in [\Gamma_i]} \boldsymbol{A}_{i,j,2} \,\middle|\, \Gamma\right], \tag{6}$$

To calculate this probability we borrow a trick from the analysis of the random k-SAT model [20]. Namely, we consider a new set of $\{0,1\}^2$-valued random variables $\boldsymbol{A}'_{i,j} = (\boldsymbol{A}'_{i,j,1}, \boldsymbol{A}'_{i,j,2})$ such that $(\boldsymbol{A}'_{i,j})_{i \in [m], j \in [\Gamma_i]}$ are mutually independent and such that

$$\mathbb{P}\left[\boldsymbol{A}'_{i,j} = (1,1)\right] = \ell/n, \qquad \mathbb{P}\left[\boldsymbol{A}'_{i,j} = (0,1)\right] = \mathbb{P}\left[\boldsymbol{A}'_{i,j} = (1,0)\right] = (k-\ell)/n,$$
$$\mathbb{P}\left[\boldsymbol{A}'_{i,j} = (0,0)\right] = (n - 2k + \ell)/n$$

for all i, j. Due to their independence, these multinomially distributed random variables are much easier to handle than $\boldsymbol{A}_{i,j}$. It will turn out, that given a (not too unlikely) event, it suffices to analyse these independent variables instead of $\boldsymbol{A}_{i,j}$. Now, let $\mathscr{T}$ be the event that

$$\sum_{i=1}^{m}\sum_{j=1}^{\Gamma_i} \mathbf{1}\left\{\boldsymbol{A}'_{i,j} = (1,1)\right\} = \ell\Delta, \qquad \sum_{i=1}^{m}\sum_{j=1}^{\Gamma_i} \mathbf{1}\left\{\boldsymbol{A}'_{i,j} = (0,0)\right\} = (n - 2k + \ell)\Delta, \tag{7}$$

$$\sum_{i=1}^{m}\sum_{j=1}^{\Gamma_i} \mathbf{1}\left\{\boldsymbol{A}'_{i,j} = (1,0)\right\} = \sum_{i=1}^{m}\sum_{j=1}^{\Gamma_i} \mathbf{1}\left\{\boldsymbol{A}'_{i,j} = (0,1)\right\} = (k-\ell)\Delta, \tag{8}$$

i..e, that all of the sums on the l.h.s. are *precisely* equal to their expected values. Then $\boldsymbol{A}' = (\boldsymbol{A}'_{i,j})_{i,j}$ given $\mathscr{T}$ is distributed precisely as $\boldsymbol{A} = (\boldsymbol{A}_{i,j})_{i,j}$. Hence, (6) yields

$$\mathbb{P}[\mathscr{S} \mid \Gamma] = \mathbb{P}\left[\forall i \in [m] : \max_{j \in [\Gamma_i]} \boldsymbol{A}'_{i,j,1} = \max_{j \in [\Gamma_i]} \boldsymbol{A}'_{i,j,2} \mid \Gamma, \mathscr{T}\right]. \tag{9}$$

Thus, let

$$\mathscr{A} = \left\{ \forall i \in [m] : \max_{j\in[\Gamma_i]} \boldsymbol{A}'_{i,j,1} = \max_{j\in[\Gamma_i]} \boldsymbol{A}'_{i,j,2} \right\}.$$

The grand idea is now to calculate the probability $\mathbb{P}[\mathscr{A} \mid \Gamma]$. Subsequently, we employ Bayes' Theorem to derive a bound for the conditional probability $\mathbb{P}[\mathscr{A} \mid \mathscr{T}, \Gamma]$ for which we know by the above application of the balls-into-bins principle

$$\mathbb{P}[\mathscr{S} \mid \Gamma] = \mathbb{P}[\mathscr{A} \mid \mathscr{T}, \Gamma].$$

Because the $(\boldsymbol{A}'_{i,j})_{i,j}$ are mutually independent, we can easily compute the unconditional probability $\mathbb{P}[\mathscr{A} \mid \Gamma]$: by inclusion/exclusion,

$$\mathbb{P}[\mathscr{A} \mid \Gamma] = \prod_{i=1}^{m} \left(1 - 2(1-k/n)^{\Gamma_i} + 2(1-2k/n+\ell/n)^{\Gamma_i}\right) \tag{10}$$

(the probability that $\max_{j\in[\Gamma_i]} \boldsymbol{A}'_{i,j,1} = \max_{j\in[\Gamma_i]} \boldsymbol{A}'_{i,j,2} = 1$, i.e., both tests positive, equals one minus the probability that $\max_{j\in[\Gamma_i]} \boldsymbol{A}'_{i,j,1} = 0$ minus the probability that $\max_{j\in[\Gamma_i]} \boldsymbol{A}'_{i,j,2} = 0$ plus the probability that $\max_{j\in[\Gamma_i]} \boldsymbol{A}'_{i,j,1} = \max_{j\in[\Gamma_i]} \boldsymbol{A}'_{i,j,2} = 0$; then add the probability that $\max_{j\in[\Gamma_i]} \boldsymbol{A}'_{i,j,1} = \max_{j\in[\Gamma_i]} \boldsymbol{A}'_{i,j,2} = 0$, i.e., both tests negative).

Finally, to deal with the conditioning we use Bayes' rule:

$$\mathbb{P}[\mathscr{A} \mid \mathscr{T}, \Gamma] = \frac{\mathbb{P}[\mathscr{A} \mid \Gamma]\,\mathbb{P}[\mathscr{T} \mid \mathscr{A}, \Gamma]}{\mathbb{P}[\mathscr{T} \mid \Gamma]}. \tag{11}$$

Since the $(\boldsymbol{A}'_{i,j})_{i,j}$ are independent, Stirling's formula yields

$$\mathbb{P}[\mathscr{T} \mid \Gamma] = \Omega\left((\Delta k)^{-3/2}\right).$$

A short justification can be found in Section B.1. Moreover, by definition we have $\mathbb{P}[\mathscr{T} \mid \mathscr{A}, \Gamma] \le 1$. Hence, (5) follows from (9)–(11). To complete the proof of the proposition, we claim that

$$\sum_{0\le\ell\le\lceil(1-1/\log n)k\rceil} O\left((\Delta k)^{3/2}\right)\binom{k}{\ell}\binom{n-k}{k-\ell}\prod_{i=1}^{m}\left(1-2(1-k/n)^{\Gamma_i}+2(1-2k/n+\ell/n)^{\Gamma_i}\right) = o(1). \tag{12}$$

To prove Equation (12), let $\alpha = \ell/k$. Using Lemma 2.4 and recalling $m = ck\log(n/k)$ and $\Delta = d\log(n/k)$, we find

$$\begin{aligned}
\mathbb{E}[Z_{k,l}(\boldsymbol{G},\boldsymbol{\sigma})] &\le O\left((\Delta k)^{3/2}\right)\binom{k}{(1-\alpha)k}\binom{n-k}{(1-\alpha)k}\prod_{i=1}^{m}\left(1-2\left(1-\frac{k}{n}\right)^{\Gamma_i}+2\left(1-\frac{2k}{n}+\frac{\alpha k}{n}\right)^{\Gamma_i}\right)\\
&\le O\left((\Delta k)^{3/2}\right)\left(\frac{e}{(1-\alpha)}\frac{en}{(1-\alpha)k}\right)^{(1-\alpha)k}\left(1-2\left(1-\frac{k}{n}\right)^{\Gamma_{\max}}+2\left(1-\frac{2k}{n}+\frac{\alpha k}{n}\right)^{\Gamma_{\min}}\right)^{m}\\
&\le O\left((\Delta k)^{3/2}\right)\left(\frac{e}{(1-\alpha)}\frac{en}{(1-\alpha)k}\right)^{(1-\alpha)k}\Bigg(1-2\left(1-\frac{k}{n}\right)^{\frac{n\log 2}{k}\left(1+n^{-\Omega(1)}\right)}\\
&\qquad\qquad +2\left(1-\frac{2k}{n}+\frac{\alpha k}{n}\right)^{\frac{n\log 2}{k}\left(1+n^{-\Omega(1)}\right)}\Bigg)^{m} \qquad (13)\\
&\le O\left((\Delta k)^{3/2}\right)\left(\frac{e}{(1-\alpha)}\frac{en}{(1-\alpha)k}\right)^{(1-\alpha)k}\left(1-\left(1-2^{-(1-\alpha)}\right)\exp\left(n^{-\Omega(1)}\right)\right)^{m}\\
&= O\left((\Delta k)^{3/2}\right)\left(\frac{e}{(1-\alpha)}\frac{en}{(1-\alpha)k}(k/n)^{c\log(2)+n^{-\Omega(1)}}(1+o(1))\right)^{(1-\alpha)k}\\
&= O\left((\Delta k)^{3/2}\right)\left(\frac{e^2(k/n)^{c\log(2)-1+n^{-\Omega(1)}}}{(1-\alpha)^2}\right)^{(1-\alpha)k}. \qquad (14)
\end{aligned}$$

By the definition of $m > (1+\varepsilon)m_{\inf}$ and $\ell < \lceil k(1-\log^{-1} n)\rceil$, we have

$$c\log 2 = 1+\varepsilon \qquad \text{and} \qquad (1-\alpha)^2 \ge 1/\log^2 n \tag{15}$$

Moreover, as $\ell < \lceil k(1-\log^{-1} n)\rceil$ we have $(1-\alpha)k = \omega(1)$. Thus (15) implies that (14) tends to zero with $n\to\infty$. Therefore, the proposition follows from Equations (14), (15) and Markov's inequality.

□

The argument from Proposition 3.1 does not extend to large overlaps (close to k) because the expression on the r.h.s. of (4) gets too large. In other words, merely computing the expected number of solutions with a given overlap does not do the trick. This 'lottery phenomenon' is ubiquitous in random constraint satisfaction problems: for big overlap values rare solution-rich instances drive up the expected number of solutions [4, 5]. Fortunately, we can find a remedy.

Proposition 3.2. *Let $\varepsilon > 0$ and $0 < \theta < 1$ and assume that $m > (1+\varepsilon)m_{\inf}(k,\theta)$. W.h.p. we have $Z_{k,\ell}(\boldsymbol{G},\hat{\boldsymbol{\sigma}}) = 0$ for all $(1-1/\log n)k \le \ell < k$.*

In order to cope with this issue we take another leaf out of the random CSP literature [2, 34]. Namely, we show that the solution $\boldsymbol{\sigma}$ is locally rigid. That is, the expansion properties of the random bipartite graph $\boldsymbol{G}$ preclude the existence of other solutions that have a big overlap with $\boldsymbol{\sigma}$. The following lemma holds the key to this effect.

Lemma 3.3. *For any $\varepsilon > 0$ there exists $\delta = \delta(\varepsilon) > 0$ such that for all $m > (1+\varepsilon)m_{\inf}$ the following is true. Let $\mathscr{R}$ be the event that for every x_i with $\boldsymbol{\sigma}_{x_i} = 1$ there are at least $\delta\Delta$ tests $a \in \partial x_i$ such that $\partial a \setminus \{x_i\} \subseteq V_0$. Then $\mathbb{P}[\mathscr{R}] = 1 - o(1)$.*

Proof. Let $(\boldsymbol{X}_i)_{i\in[m]}$ be a sequence of independent $\mathrm{Bin}(\Gamma_i, k/n)$-variables as in Section 2. Also let $W = \sum_{i=1}^{m} \mathbf{1}\{\boldsymbol{Y}_i = 1\}$ as in Section 2. Proceeding along the lines of the proof of Lemma 2.3 (see (35) in Section B.6), we obtain

$$\mathbb{P}\left[W = \left(1 + n^{-\Omega(1)}\right) k\Delta/2 \mid \Gamma\right] = 1 - o(n^{-7}). \tag{16}$$

Let T be the number of infected individuals which only show up less than $\delta\Delta$ of their tests as the only infected individual, i.e.

$$T = \left| x \in V_1 : \sum_{a \in \partial x} \mathbf{1}\{\partial a \setminus \{x\} \subseteq V_0\} < \delta\Delta \right|.$$

Moreover, let $\boldsymbol{H} = \boldsymbol{H}(N, K, n')$ be a hypergeometric random variable with parameters $N = k\Delta$ (total eligible assignments for infected individuals), $K = W$ (tests with only one infected individual) and $n' = \Delta$ (number of tests per individuals). Then the union bound over k infected individuals yields

$$\mathbb{E}[T \mid \Gamma, W] \le k\mathbb{P}[\boldsymbol{H} < \delta\Delta]. \tag{17}$$

Further, the Chernoff bound for the hypergeometric distribution implies

$$\mathbb{P}[\boldsymbol{H} < \delta\Delta] \le \exp(-\Delta D_{\mathrm{KL}}(\delta \| W/(k\Delta))) \tag{18}$$

Recall $\Delta = d\log(n/k)$. Since $D_{\mathrm{KL}}(\delta\|1/2 + o(1)) = \delta\log\delta + (1-\delta)\log(1-\delta) + \log 2 + o(1)$ and $\delta\log\delta + (1-\delta)\log(1-\delta) \nearrow 0$ as $\delta \to 0$ and $c > \frac{\theta}{(1-\theta)\log^2 2}$, we can choose $\delta > 0$ small enough so that

$$\Delta(\delta\log\delta + (1-\delta)\log(1-\delta) + \log 2 + o(1)) > \log k \tag{19}$$

Finally, the assertion follows from (16)–(19). □

Hence, w.h.p. any infected individual appears in plenty of tests where all the other individuals are uninfected. This property causes $\boldsymbol{\sigma}$ to be locally rigid. To see why, consider the repercussions of just changing the status of a single individual x_i from infected to uninfected. Because given $\mathscr{R}$ the individual x_i appears as the only infected individual in at least $\delta\Delta$ tests, in order to maintain the same tests results we will also need to flip at least one individual in each of these tests from 'uninfected' to 'infected'. Since tests typically have relatively few individuals in common, the necessary number of flips from 0 to 1 will be $\Omega(\Delta) = \Omega(\log n)$. But then in order to keep the total number of infected individuals constant k, we will need to perform another $\Omega(\Delta)$ flips from 1 to 0. Yet given $\mathscr{R}$ each of these 'second generation' individuals that we flip from infected to uninfected is itself the only infected individual in many tests. Thus, the single flip that we started from triggers a veritable avalanche of flips, which will stop only after the overlap has dropped significantly. The next lemma formalises this intuition. The lemma shows that while the unconditional expectation of $Z_{k,\ell}(\boldsymbol{G},\hat{\boldsymbol{\sigma}})$ is 'too big', the conditional expectation of $Z_{k,\ell}(\boldsymbol{G},\hat{\boldsymbol{\sigma}})$ given $\mathscr{R}$ (as defined in Lemma 3.3) is much smaller. Let $\boldsymbol{m}_0 = \boldsymbol{m}_0(\boldsymbol{G},\hat{\boldsymbol{\sigma}})$ be the total number of negative tests.

Lemma 3.4. *Suppose that $(1-1/\log n)k \le \ell < k$ and let $\Gamma_{\min} = \min_{i\in[m]}\Gamma_i$, $\Gamma_{\max} = \max_{i\in[m]}\Gamma_i$. Then*

$$\mathbb{E}[Z_{k,\ell}(\boldsymbol{G},\hat{\boldsymbol{\sigma}}) \mid \Gamma, \mathscr{R}, \boldsymbol{m}_0] \le O\left((\Delta k)^{3/2}\right)\binom{k}{\ell}\binom{n-k}{k-\ell}\left(1 - \left(1 - \frac{k-\ell}{n-k}\right)^{\Gamma_{\max}}\right)^{\delta\Delta(k-\ell)}\left(\frac{n-2k+\ell}{n-k}\right)^{\left(1+n^{-\Omega(1)}\right)\Gamma_{\min}\boldsymbol{m}_0}. \tag{20}$$

The proof of Lemma 3.4 is somehow subtle as we need to get a handle on the dependencies in $\boldsymbol{G}$ and is included in Section C.1. To convey the intuition behind the expression in Lemma 3.4, the term $\binom{k}{\ell}\binom{n-k}{k-\ell}$ accounts for the number of assignments $\tau \in \{0,1\}^V$ of Hamming weight k whose overlap with $\boldsymbol{\sigma}$ is equal to ℓ. The terms thereafter capture the probability that such an assignment τ exhibits the same test results as the true configuration $\boldsymbol{\sigma}$. The first term provides a necessary condition for a positive test under $\boldsymbol{\sigma}$ to stay positive under τ. By Lemma 3.3, we know that every infected individual shows up in at least $\delta\Delta$ tests as the only infected individual. Now, there are $k-\ell$ infected under $\boldsymbol{\sigma}$, but healthy under τ. For any of these $\delta\Delta(k-\ell)$ tests, we need to have at least one individual that is healthy under $\boldsymbol{\sigma}$, but infected under τ included in this test. Next, we need to ensure that any negative test under $\boldsymbol{\sigma}$ stay negative under τ. To this end, every individual included in a negative test under $\boldsymbol{\sigma}$ of which we have at least $\Gamma_{\min}\boldsymbol{m}_0$ must be healthy under τ. The second term captures this probability.

Proof of Proposition 3.2. In order to establish the proposition it suffices to show that there is $\varepsilon' \le (1-1/\log(n))k$ such that

$$\sum_{\varepsilon' \le \ell \le k} \mathbb{E}[Z_{k,\ell}(\boldsymbol{G},\hat{\boldsymbol{\sigma}})|\Gamma,\mathscr{R},\boldsymbol{m}_0] = o(1). \tag{21}$$

Starting from the expression in Lemma 3.4, setting $\alpha = \ell/k$ and recalling $m = ck\log(n/k)$ and $\Delta = d\log(n/k)$, we obtain

$$\begin{aligned}
&\mathbb{E}[Z_{k,\ell}(\boldsymbol{G},\hat{\boldsymbol{\sigma}})|\Gamma,\mathscr{R},\boldsymbol{m}_0] \\
&\le O\big((\Delta k)^{3/2}\big)\binom{k}{k-\ell}\binom{n-k}{k-\ell}\left(\frac{n-2k+\ell}{n-k}\right)^{\left(1+n^{-\Omega(1)}\right)\Gamma_{\min}\boldsymbol{m}_0}\left(1-\left(1-\frac{k-\ell}{n-k}\right)^{\Gamma_{\max}}\right)^{\delta\Delta(k-\ell)} \\
&\le O\big((\Delta k)^{3/2}\big)\left(\frac{e}{1-\alpha}\right)^{(1-\alpha)k}\left(\frac{e(n-k)}{(1-\alpha)k}\right)^{(1-\alpha)k}\left(1-\frac{(1-\alpha)k}{n-k}\right)^{\frac{mn\log 2}{2k}\left(1+n^{-\Omega(1)}\right)}\left(1-2^{-(1-\alpha)\left(1+n^{-\Omega(1)}\right)}\right)^{\delta\Delta(1-\alpha)k} \qquad (22)\\
&\le O\big((\Delta k)^{3/2}\big)\left(\frac{e^2 n}{(1-\alpha)^2 k}\right)^{(1-\alpha)k}\exp\left((1-\alpha)k\frac{c\log 2}{2}\left(1+n^{-\Omega(1)}\right)\log(k/n)\right) \\
&\qquad\qquad \cdot\exp\left(-c\delta\log(2)\log\left(1-2^{-(1-\alpha)\left(1+n^{-\Omega(1)}\right)}\right)\log(k/n)(1-\alpha)k\right) \\
&\le O\big((\Delta k)^{3/2}\big)\left(\frac{e^2 n}{(1-\alpha)^2 k}\exp\left(\log(k/n)\left(1+n^{-\Omega(1)}\right)\left(\frac{c\log 2}{2}-c\delta\log(2)\log\left(1-2^{-(1-\alpha)\left(1+n^{-\Omega(1)}\right)}\right)\right)\right)\right)^{(1-\alpha)k}. \qquad (23)
\end{aligned}$$

As long as $1-\alpha = o(1)$, we find

$$(k/n)^{-\log\left(1-2^{-(1-\alpha)}\right)}(1-\alpha)^{-2} \to 0 \qquad \text{as } n\to\infty.$$

Moreover, $(1-\alpha)k \ge 1$. Thus, the expression (23) is of order

$$O\big((\Delta k)^{3/2}\big)(k/n)^{\omega(1)} = n^{-\omega(1)}. \tag{24}$$

Since (24) holds for any constant $c>0$ and any value of α s.t. $1-\alpha = o(1)$, it also holds for $\alpha \ge 1-1/\log n$. Consequently (21) is established w.h.p. □

Propositions 3.1 and 3.2 readily imply that $Z_k(\boldsymbol{G},\hat{\boldsymbol{\sigma}}) = 1$ w.h.p. if $m > (1+\varepsilon)m_{\inf}(k,\theta)$. Hence, Corollary 2.2 shows that there exists an inference algorithm that given $\boldsymbol{G},\hat{\boldsymbol{\sigma}}$ *and* k outputs $\boldsymbol{\sigma}$ w.h.p. Up to now, the algorithm relies on exactly knowing the number of infected individuals k, which in practice could be rather difficult to learn. Fortunately, this assumption can be removed. Namely, the following proposition shows that w.h.p. there is no assignment σ that is compatible with the test results and that has Hamming weight less than k.

Proposition 3.5. *Let $\varepsilon>0$ and $0<\theta<1$ and assume that $m > (1+\varepsilon)m_{\inf}(k,\theta)$. W.h.p. we have $\sum_{k'<k} Z_{k'}(\boldsymbol{G},\hat{\boldsymbol{\sigma}}) = 0$.*

Proof. To get started, suppose that $0<\theta<1$ and $c<\log^{-2}2$. We claim that for any value of $d>0$, $\left|V_0^+\right| \ge k\log n$ w.h.p.. Indeed, from Proposition 2.3(1), we know that

$$\left|V_0^+\right| = \left(1+n^{-\Omega(1)}\right)n\left(1-\exp(-d/c)\right)^{\Delta}.$$

Recalling $\Delta = d\log(n/k)$, the expression takes the minimum at $d = c\log 2$. It follows that

$$\left|V_0^+\right| \ge \left(1+n^{-\Omega(1)}\right)n(k/n)^{c\log^2 2}.$$

If $c = (1-\varepsilon)\log^{-2} 2$ for $\epsilon > 0$, then

$$\left|V_0^+\right| \geq \left(1+n^{-\Omega(1)}\right) n(k/n)^{1-\epsilon} = \left(1+n^{-\Omega(1)}\right) kn^{(1-\theta)\epsilon} \geq k\log n \qquad w.h.p. \tag{25}$$

Now, the following two statements establish that if there does not exist a second satisfying set of Hamming weight k, there does also not exist a satisfying set with smaller Hamming weight w.h.p..

First, we claim that if $m > (1+\epsilon)m_{\text{inf}}(k,\theta)$, w.h.p. there does not exist a satisfying configuration with Hamming weight smaller than the correct configuration, where the set of infected individuals is not a subset of the true set of infected individuals. To see why, suppose there existed a satisfying configuration with a smaller Hamming weight, whose infected individuals are not a subset of the true infected individuals. By (25), we know that $\left|V_0^+\right| \gg k$ for $m < (1-\epsilon)m_{\text{alg}}$ w.h.p. Therefore, we could construct a satisfying configuration of identical Hamming weight as the true configuration by flipping individuals in V_0^+ from healthy to infected. Observe that by the definition of V_0^+, flipping individuals in V_0^+ does not change the test result. Therefore, we would be left with a second satisfying configuration of identical Hamming weight as the true configuration, a contradiction to Propositions 3.1 and 3.2.

Second, we argue that if $m > (1+\epsilon)m_{\text{inf}}(k,\theta)$, w.h.p. there does not exist a satisfying configuration with Hamming weight smaller than the correct configuration, where the set of infected individuals is a subset of the true set of infected individuals. Suppose there existed a satisfying configuration with a smaller Hamming weight, whose infected individuals are a subset of the true infected individuals. Then, the true configuration would need to contain individuals in V_1^+, which can be flipped from infected to healthy without affecting the test result. However, Proposition 2.3(5) shows that for $m > (1+\epsilon)m_{\text{inf}}$, $V_1^+ = \emptyset$ w.h.p. □

As an immediate consequence of Proposition 3.5 we conclude that for $m > (1+\varepsilon)m_{\text{inf}}(k,\theta)$ the problem of inferring $\boldsymbol{\sigma}$ boils down to a minimum vertex cover problem, as previously conjectured by Aldridge, Baldassini and Johnson [12]. Namely, let $\mathscr{P}$ be the set of all positive tests, i.e., all tests a_i, $i \in [m]$, with $\hat{\boldsymbol{\sigma}}_{a_i} = 1$. Moreover, let V^+ be the set of all variables $x_i \in V$ such that $\partial x_i \subseteq \mathscr{P}$; in words, x_i takes part in positive tests only. We set up a hypergraph $\boldsymbol{H}$ with vertex set V^+ and hyperedges $\partial a_i \cap V^+$, $a_i \in \mathscr{P}$. Clearly, the set of all individuals x_i with $\boldsymbol{\sigma}_{x_i} = 1$ provides a valid vertex cover of $\boldsymbol{H}$ (as any positive test must feature an infected individual). Conversely, Propositions 3.1 and 3.2 show that w.h.p. this is the unique vertex cover of size k, and Proposition 3.5 shows that there is no strictly smaller vertex cover w.h.p. Therefore, w.h.p. we can infer $\boldsymbol{\sigma}$ even without prior knowledge of k by way of solving this minimum vertex cover instance.

4. The information-theoretic lower bound

We proceed with the negative statement that w.h.p. $\boldsymbol{\sigma}$ cannot be inferred if $m < (1-\varepsilon)m_{\text{inf}}$. In light of Corollary 2.2 in order to prove the first part of Theorem 1.1 we need to show that the number $Z_k(\boldsymbol{G},\hat{\boldsymbol{\sigma}})$ of assignments consistent with the test results $\hat{\boldsymbol{\sigma}}$ is unbounded w.h.p. The proof of this fact is based on a very simple idea: we just identify a moderately large number of individuals whose infection status could be flipped without affecting the test results. The following lemma yields a bound on m below which the number of such potential false positives ($\left|V_0^+\right|$) and negatives ($\left|V_1^+\right|$) abound.

Proposition 4.1. *Let $\varepsilon > 0$ and $0 < \theta < 1$ and assume that*

$$m < \frac{(1-\varepsilon)\theta}{(1-\theta)\log^2 2} n^\theta (1-\theta)\log n.$$

Then for any choice of Δ we have $|V_0^+|, |V_1^+| = n^{\Omega(1)}$ w.h.p.

Proof. Thanks to Lemma 2.6 we may assume that $\Delta = d(\log(n/k))$, for a constant d as this choice minimizes the number of individuals in V_1^+. Then Proposition 2.3(4) guarantees that for every such constant as long as $c < \frac{\theta}{1-\theta}\frac{1}{\log^2 2}$, there are $n^{\Omega(1)}$ individuals in both V_1^+ and V_0^+, which yields to Proposition 4.1. □

As an immediate application we obtain the following information-theoretic lower bound.

Corollary 4.2. *Let $\varepsilon > 0$ and $0 < \theta < 1$ and assume that*

$$m < \frac{(1-\varepsilon)\theta}{(1-\theta)\log^2 2} n^\theta (1-\theta)\log n. \tag{26}$$

Then $Z_k(\boldsymbol{G},\hat{\boldsymbol{\sigma}}) = \omega(1)$ w.h.p.

Proof. We need to exhibit alternative vectors $\boldsymbol{\sigma}' \in \{0,1\}^V$ with Hamming weight k that render the same test results as $\boldsymbol{\sigma}$. Thus, pick any $x_i \in V_0^+$ and any $x_j \in V_1^+$ and obtain $\boldsymbol{\sigma}'$ from $\boldsymbol{\sigma}$ by setting $\boldsymbol{\sigma}'_{x_i} = 1$ and $\boldsymbol{\sigma}'_{x_j} = 0$. By construction, $\boldsymbol{\sigma}'$ has Hamming weight k and renders the same test results. Hence, Proposition 4.1 shows that $Z_k(\boldsymbol{G}, \hat{\boldsymbol{\sigma}}) \geq |V_0^+ \times V_1^+| = \Omega(n^{2\theta}) \gg 1$ w.h.p. □

The bound (26) matches m_{inf} for $\theta \gtrapprox 0.41$. A simpler, purely information-theoretic argument covers the remaining θ.

Proposition 4.3. *Let $\varepsilon > 0$, $0 < \theta < 1$. If $m < \frac{1-\varepsilon}{\log 2} n^\theta (1-\theta) \log n$, then $Z_k(\boldsymbol{G}, \hat{\boldsymbol{\sigma}}) = \omega(1)$ w.h.p.*

Proof. This Lemma follows from the classical information-theoretic lower bound for the group testing problem. Namely, m tests allow for 2^m possible test results. Hence, if

$$m < \frac{(1-\varepsilon)}{\log 2} n^\theta (1-\theta) \log n,$$

then the number of possible test results is far smaller than the number of vectors $\boldsymbol{\sigma} \in \{0,1\}^V$ with Hamming weight k. Therefore, w.h.p. there exists an unbounded number of vectors of Hamming weight k that render the same test results as $\boldsymbol{\sigma}$. □

We thus conclude that for all $0 < \theta < 1$, w.h.p. $Z_k(\boldsymbol{G}, \hat{\boldsymbol{\sigma}}) = \omega(1)$ if $m < (1-\varepsilon) m_{\mathrm{inf}}$. Therefore, the desired information-theoretic lower bound follows from Corollary 2.2.

5. The SCOMP algorithm

For $\theta \geq 1/2$ we have $m_{\mathrm{alg}} = m_{\mathrm{inf}}$ and thus Theorem 1.1 implies that SCOMP as described in Section 1.3 w.h.p. fails to infer $\boldsymbol{\sigma}$ for $m < (1-\varepsilon) m_{\mathrm{alg}}$. Therefore, we are left to establish Theorem 1.2 for $\theta < 1/2$, in which case

$$m_{\mathrm{alg}} = \frac{k \log(n/k)}{\log^2 2}. \tag{27}$$

The proof of Theorem 1.2 for $\theta < 1/2$ hinges on two propositions. First we show that below m_{alg}, the set V_1^{--} of infected individuals that the second step of SCOMP identifies correctly is empty. Formally, with V_0^- from (3), let

$$V_1^{--} = \left\{ x \in V_1 : \exists a \in \partial x : \partial a \setminus \{x\} \subseteq V_0^- \right\}.$$

Proposition 5.1. *Suppose that $0 < \theta < 1/2$ and $\varepsilon > 0$. If $m < (1-\varepsilon) m_{\mathrm{alg}}$, then for all $\Delta > 0$ we have $V_1^{--}(\boldsymbol{G}, \hat{\boldsymbol{\sigma}}^*) = \emptyset$ w.h.p.*

The proofs of Propositions 5.1 and 5.2 are based on moment calculations that turn out to be mildly subtle due to the potentially very large degrees of the underlying graph $\boldsymbol{G}$. The technical workout in included in Section D.1 and D.2.

With the second step of SCOMP failing to 'explain' (viz. cover) any positive tests, the greedy vertex cover algorithm takes over. This algorithm is applied to the hypergraph whose vertices are the as yet unclassified individuals and whose edges are the neighbourhoods of the positive tests. Our second lemma shows that the set $V^{+,\Delta}$ of potentially false positive individuals $x \in V_0^+$ that participate in the maximum number Δ of different tests is far greater than the actual number k of infected individuals. Formally, let

$$V_0^{+,\Delta} = \left\{ x \in V_0^+ : |\partial x| = \Delta \right\}.$$

Proposition 5.2. *Suppose that $0 < \theta < 1/2$ and $\varepsilon > 0$. If $m < (1-\varepsilon) m_{\mathrm{alg}}$, then for $\Delta = d \log(n/k)$ for all constant d we have $\left| V_0^{+,\Delta} \right| \geq k \log n$ w.h.p.*

We complete the proof of Theorem 1.2 as follows.

Proof of Theorem 1.2. The first step of SCOMP (correctly) marks all individuals that appear in negative tests as healthy. Moreover, Proposition 5.1 implies that the second step of SCOMP is void w.h.p., because there is no single infected individual that appears in a test whose other individuals have already been identified as healthy by the first step. Consequently, SCOMP simply applies the greedy vertex cover algorithm. Now, thanks to Proposition 5.2 it suffices to prove that SCOMP will fail w.h.p. if $\left| V_0^{+,\Delta} \right| = \omega(k)$. Because they belong to positive tests only, all the individuals of $V_0^{+,\Delta}$ are present in the vertex cover instance that SCOMP attempts to solve. Moreover, in the hypergraph no vertex

has degree greater than Δ, because the degrees of $x_1, \ldots, x_n$ in $\boldsymbol{G}$ are equal to Δ. (Some of the hypergraph degrees may be strictly smaller than Δ because $\boldsymbol{G}$ is a multi-graph.) Therefore, since $|V_0^{+,\Delta}| \geq k \log n$ while the actual set of infected individuals only has size k, w.h.p. the individual classified as infected by the very first step of the greedy set cover algorithm belongs to V_0^+. Hence, this individual is not actually infected, i.e., SCOMP errs w.h.p. □

Since the success probability of the SCOMP algorithm is at least as high as of the DD algorithm, we can prove the conjecture of [30] regarding the upper bound of the DD algorithm.

Corollary 5.3. *If $m < (1-\varepsilon) m_{\mathrm{alg}}$, the* DD *algorithm will fail to retrieve the correct set of infected individuals w.h.p..*

Acknowledgment. We thank Arya Mazumdar for bringing the group testing problem to our attention.

A. NOTATION

Notation	Definition & Properties	Description
n		population size
k	$k \sim n^\theta$ for $\theta \in (0,1)$	number of infected individuals
m	$m = ck\log(n/k)$	number of tests
$x_1, \dots, x_n$		variable nodes
$V = V_n$	$\{x_i, \dots, x_n\}$	set of all individuals
$a_1, \dots, a_n$		factor nodes
$F = F_m$	$\{a_i, \dots, a_m\}$	set of all tests
Δ	$\Delta = d\log(n/k)$	tests per individual, variable node degree
$\Gamma_1, \dots, \Gamma_m$	$\left(\sum_{i=1}^m \Gamma_i\right)/m = dn/(ck)$	individuals per test, factor node degree
Γ	$(\Gamma_i)_{i\in[m]}$	σ-algebra generated by the random variables $(\Gamma_i)_{i\in[m]}$
$\boldsymbol{\sigma} \in \{0,1\}^V$	$\sum_{i=1}^n \boldsymbol{\sigma}_i = k$	n-dimensional vector of Hamming weight k indicating the individuals' infection status
$\boldsymbol{G} = \boldsymbol{G}(n,m,\Delta)$		random bipartite graph on n variable nodes, m factor nodes and variable degree Δ
$\partial x_i = \partial_{\boldsymbol{G}} x_i$ for $i \in [n]$	$\partial x_i \subseteq F, \lvert\partial x_i\rvert = \Delta$	set of tests that individual x_i participates in under $\boldsymbol{G}$
$\partial a_i = \partial_{\boldsymbol{G}} a_i$ for $i \in [m]$	$\partial a_i \subseteq V, \lvert\partial a_i\rvert = \Gamma_i$	set of individuals in test a_i under $\boldsymbol{G}$
$\hat{\boldsymbol{\sigma}} \in \{0,1\}^F$	$\hat{\boldsymbol{\sigma}}_i = \mathbf{1}\{\exists x \in \partial a_i : \boldsymbol{\sigma}_x = 1\}$	m-dimensional vector indicating the test outcomes
$\boldsymbol{m}_1, \boldsymbol{m}_0$	$\boldsymbol{m}_1 = \lvert\{a \in F : \hat{\boldsymbol{\sigma}}_a = 1\}\rvert, \boldsymbol{m}_0 = m - \boldsymbol{m}_1$	number of positive and negative tests
V_0	$V_0 = \{x \in V : \boldsymbol{\sigma}_x = 0\}$	set of healthy individuals
V_1	$V_1 = V \setminus V_0, \lvert V\rvert = k$	set of infected individuals
V_0^+	$\{x \in V_0 : \forall a \in \partial x : \hat{\boldsymbol{\sigma}}_a = 1\}$	set of healthy individuals only included in positive tests
V_0^-	$V_0^- = V_0 \setminus V_0^+$	set of healthy individuals included in at least one negative test
V_1^+	$\left\{x \in V_1 : \forall a \in \partial x : \exists y \in \partial a \setminus \{x\} : \boldsymbol{\sigma}_y = 1\right\}$	set of infected individuals that have another infected individual in all their tests
V_1^{--}	$\left\{x \in V_1 : \exists a \in \partial x : \partial a \setminus \{x\} \subseteq V_0^-\right\}$	Set of infected individuals that occur in at least one test with only healthy individuals
$\Gamma_{\min}, \Gamma_{\max}$	$\Gamma_{\min} = \min_{i\in[m]} \Gamma_i, \Gamma_{\max} = \max_{i\in[m]} \Gamma_i$	minimum and maximum test degree
$S_k(\boldsymbol{G}, \hat{\boldsymbol{\sigma}})$	$S_k(\boldsymbol{G}, \hat{\boldsymbol{\sigma}}) = \left\{\sigma \in \{0,1\}^V : \forall a_i \in [m] : \hat{\boldsymbol{\sigma}}_{a_i} = \mathbf{1}\{\exists x \in \partial a_i : \sigma_x = 1\}\right\}$	set of configurations consistent with the test results under $\boldsymbol{G}$
$Z_k(\boldsymbol{G}, \hat{\boldsymbol{\sigma}})$	$Z_k(\boldsymbol{G}, \hat{\boldsymbol{\sigma}}) = \lvert S_k(\boldsymbol{G}, \hat{\boldsymbol{\sigma}})\rvert$	number of configurations consistent with the test results
$Z_{k,\ell}(\boldsymbol{G}, \hat{\boldsymbol{\sigma}})$	$Z_{k,\ell}(\boldsymbol{G}, \hat{\boldsymbol{\sigma}}) = \lvert\{\sigma \in S_k(\boldsymbol{G}, \hat{\boldsymbol{\sigma}}) : \langle \boldsymbol{\sigma}, \sigma\rangle = \ell\}\rvert$	number of configuration consistent with the test results and with overlap ℓ with $\boldsymbol{\sigma}$
$\boldsymbol{Y}_i$ for $i \in [m]$	$\boldsymbol{Y}_i = \lvert\{x \in \partial a_i : \boldsymbol{\sigma}_x = 1\}\rvert$	number of edges that connect test a_i with an infected individual
$\boldsymbol{X}_i$ for $i \in [m]$	$\boldsymbol{X}_i \sim \mathrm{Bin}(\Gamma_i, k/n)$	binomially-distributed random variable with parameters Γ_i and k/n

W, W'	$W = \sum_{i=1}^{m} \mathbf{1}\{\boldsymbol{Y}_i = 1\}, W' = \sum_{i=1}^{m} \mathbf{1}\{\boldsymbol{X}_i = 1\}$	W is the number of tests containing a single infected individual, W' is a random variable depending on $(\boldsymbol{X}_i)_{i\in[m]}$
U	$U = \lvert\{x \in V_1 : \forall a_i \in \partial x : \boldsymbol{Y}_i > 1\}\rvert$	number of infected individuals not adjacent not any test with precisely one infected individual
T	$\left\lvert\left\{x \in V_1 : \sum_{a\in\partial x} \mathbf{1}\{\partial a \setminus \{x\} \subseteq V_0\} < \delta\Delta\right\}\right\rvert$	number of infected individuals who appear in less than $\delta\Delta$ tests as the only infected individual for some constant $\delta > 0$
R	$R = \lvert\{x \in V_1 : \exists a_i \in \partial x :$ $\boldsymbol{Y}_i > 1$ and $\partial a \setminus \{x\} \subseteq V_0\}\rvert$	number of infected individual adjacent to some test multiple times with no other infected individual besides themselves
$\boldsymbol{A}'_{i,j}, \boldsymbol{A}'_{i,j,k}$		auxiliary random variables, defined in proof of Proposition 3.1
$\mathscr{A}$	$\mathscr{A} = \{\forall i \in [m] :$ $\max_{j\in[\Gamma_i]} \boldsymbol{A}'_{i,j,1} = \max_{j\in[\Gamma_i]} \boldsymbol{A}'_{i,j,2}\}$	event that every test under the balls-and-bins experiment features the same test result
$\mathscr{E}$	$\mathscr{E} = \{\sum_{i\in[m]} \boldsymbol{X}_i = k\Delta\}$	event that the sum of $\boldsymbol{X}_i$ is exactly $k\Delta$
$\mathscr{M}$		set of all indices $i \in [m]$ for which there exists precisely one $g_i \in [\Gamma_i]$ such that $\boldsymbol{A}'_{i,g_i,1} = 1$
$\mathscr{N}$		set of indices $i \in [m]$ such that $\max_{j\in[\Gamma_i]} \boldsymbol{A}'_{i,j,1} = 0$
$\mathscr{R}$	$\mathscr{R} = \{\forall x \in V_1 : \lvert\{a \in \partial x : \partial a \setminus \{x\} \subseteq V_0\}\rvert \geq \delta\Delta\}$	event that for every $x \in V_1$ there are at least $\delta\Delta$ tests $a \in \partial x$ for some $\delta > 0$ such that $\partial a \setminus \{x\} \subseteq V_0$.
$\mathscr{S}$		event that one specific σ that has overlap ℓ with $\boldsymbol{\sigma}$ belongs to $S_k(\boldsymbol{G}, \hat{\boldsymbol{\sigma}})$
$\mathscr{T}$		event that sum of independent random variable is equal to specific value, defined in (7)
$\mathscr{V}$	$\mathscr{V} = \{\boldsymbol{m}_1 = \frac{m}{2}(1 + o(1))\}$	event that around half of the tests are positive
$\mathscr{W}$	$\mathscr{W} =$ $\{\lvert V_0^+\rvert = (1 + o(1))(n-k)(1 - \exp(-d/c))^{\Delta}\}$	event that the size of V_0^+ is concentrated around its mean
$o(1), \omega(1)$		$o(1)$ [$\omega(1)$] denotes a term that vanishes [diverges] in the limit of large n
w.h.p.		probability of $1 - o(1)$ as $n \to \infty$

The following sections contain the proofs of the lemmas omitted so far.

B. Preliminaries

B.1. **Preliminaries.** We start by stating the Chernoff bound as applied in this paper.

Lemma B.1 (Chernoff bound, [29] (Section 2.1))**.** *Let $\boldsymbol{X} \sim \mathrm{Bin}(n, p)$ be a binomially-distributed random variable with $\lambda = \mathbb{E}[\boldsymbol{X}]$. Further, let*

$$\varphi : (-1, \infty) \to \mathbb{R}_{\geq 0}, x \mapsto (1 + x)\log(1 + x) - x$$

Then for some $t \geq 0$,

$$\mathbb{P}\left(\lvert\boldsymbol{X} - \lambda\rvert \geq t\right) \leq \exp\left(-\lambda\varphi\left(t/\lambda\right) - (n - \lambda)\varphi\left(-t/(n - \lambda)\right)\right)$$

As an application, we readily find

$$\mathbb{P}\left(|\boldsymbol{X}-\lambda| \geq \sqrt{n}\log n\right) \leq n^{-\omega(1)}$$

Next, we justify that the Stirling approximation of Section 3 is accurate. Namely, let $\boldsymbol{A}'_{i,j} = (\boldsymbol{A}'_{i,j,1}, \boldsymbol{A}'_{i,j,2})$ be $\{0,1\}^2$-valued random variables such that $(\boldsymbol{A}'_{i,j})_{i\in[m],j\in[\Gamma_i]}$ are mutually independent and such that

$$\mathbb{P}\left[\boldsymbol{A}'_{i,j}=(1,1)\right] = \ell/n, \qquad \mathbb{P}\left[\boldsymbol{A}'_{i,j}=(0,1)\right] = \mathbb{P}\left[\boldsymbol{A}'_{i,j}=(1,0)\right] = (k-\ell)/n,$$
$$\mathbb{P}\left[\boldsymbol{A}'_{i,j}=(0,0)\right] = (n-2k+\ell)/n$$

for all i,j. As before, we denote by $\mathcal{T}$ the event that

$$\sum_{i=1}^{m}\sum_{j=1}^{\Gamma_i}\mathbf{1}\left\{\boldsymbol{A}'_{i,j}=(1,1)\right\} = \ell\Delta, \qquad \sum_{i=1}^{m}\sum_{j=1}^{\Gamma_i}\mathbf{1}\left\{\boldsymbol{A}'_{i,j}=(0,0)\right\} = (n-2k+\ell)\Delta,$$
$$\sum_{i=1}^{m}\sum_{j=1}^{\Gamma_i}\mathbf{1}\left\{\boldsymbol{A}'_{i,j}=(1,0)\right\} = \sum_{i=1}^{m}\sum_{j=1}^{\Gamma_i}\mathbf{1}\left\{\boldsymbol{A}'_{i,j}=(0,1)\right\} = (k-\ell)\Delta,$$

i..e, that all of the sums on the l.h.s. are *precisely* equal to their expected values. Since the $(\boldsymbol{A}'_{i,j})_{i,j}$ are independent, Stirling's formula yields

$$\mathbb{P}\left[\mathcal{T}\right] = \Omega\left((\Delta k)\right)^{-3/2}. \tag{28}$$

This can be seen as follows. For the sake of brevity, define

$$p_{00} = (n-2k+\ell)/n, \quad p_{11} = \ell/n, \quad \text{and} \quad p_{10} = p_{01} = (k-\ell)/n.$$

As $\boldsymbol{A}'_{i,j}$ is a family of independent multinomial variables

$$\boldsymbol{A}'_{i,j} \sim \mathrm{Mult}\left(1,(p_{11},p_{00},p_{10},p_{01})\right),$$

we find

$$\boldsymbol{X} \sim \sum_{i=1}^{m}\sum_{j=1}^{\Gamma_i}\boldsymbol{A}'_{i,j} \sim \mathrm{Mult}\left(n\Delta,(p_{11},p_{00},p_{10},p_{01})\right).$$

Hence, the probability of event $\mathcal{T}$ occurring is the probability, that $\boldsymbol{X}$ hits its expectation. Thus, using the very basic approximation $n! = \Theta\left(\sqrt{n}\right)(n/\mathrm{e})^n$ we find

$$\begin{aligned}
\mathbb{P}\left(\mathcal{T}\right) &= \frac{(n\Delta)!(\ell/n)^{\ell\Delta}((n-2k+\ell)/n)^{(n-2k+\ell)\Delta}((k-\ell)/n)^{2(k-\ell)\Delta}}{(\ell\Delta)!((n-2k+\ell)\Delta)!((k-\ell)\Delta)!((k-\ell)\Delta)!} \\
&= \Theta\left(\frac{\sqrt{n\Delta}}{\sqrt{\ell(n-2k\ell)(k-\ell)^2\Delta^4}}\right)\left(\frac{(n\Delta)^n(\ell/n)^{\ell}((n-2k+\ell)/n)^{n-2k+\ell}((k-\ell)/n)^{2(k-\ell)}}{\ell^{\ell}(n-2k+\ell)^{n-2k+\ell}(k-\ell)^{2(k-\ell)}}\right)^{\Delta} \\
&= (1+O(1/n))\Theta\left(\frac{\sqrt{n}}{\sqrt{n}\sqrt{\ell k^2-2\ell^2k+\ell^3-k\left(2\ell k^2/n-2k\ell^2/n\right)+\ell^4/n}}\right) \\
&= \Omega\left(\sqrt{\Delta^{-3}(\ell k^2+\ell^2k+\ell^3)^{-1}}\right) = \Omega\left((\Delta k)^{-3/2}\right),
\end{aligned} \tag{29}$$

where (29) follows immediately from $\ell \leq k = o(n)$ and directly implies (28). In due course we apply similar calculations often, some calculations involve conditional probabilities. These conditions are only restricting Γ_i to take specific (common) values and clearly the above argument is totally invariant under different values of Γ_i, as long as $\sum_i^m \Gamma_i = n\Delta$.

B.2. **Getting started.** In the next step, recall that neighbourhoods of different tests in the random multi-graph seizably intersect. To cope with the ensuing correlations, we introduce a new family of random variables that, as we will see, are closely related to the statistics of the appearances of infected/uninfected individuals in the various tests. Specifically, recalling that Γ_i signifies the degree of test a_i and that $\sum_{i=1}^m \Gamma_i = n\Delta$, let $(\boldsymbol{X}_i)_{i\in[m]}$ be a sequence of independent $\mathrm{Bin}(\Gamma_i, k/n)$-variables. Moreover, let

$$\mathscr{E} = \left\{ \sum_{i\in[m]} \boldsymbol{X}_i = k\Delta \right\}.$$

Because the $\boldsymbol{X}_i$ are mutually independent, Stirling's formula shows that

$$\mathbb{P}[\mathscr{E}] = \Omega(1/\sqrt{\Delta k}), \tag{30}$$

which follows along the lines of Section B.1. Additionally, let $\boldsymbol{Y}_i$ be the number of edges that connect test a_i with an infected individual. (Since $\boldsymbol{G}$ is a multi-graph, it is possible that an infected individual contributes more than one to $\boldsymbol{Y}_i$.) Further, let Γ be the σ-algebra generated by the random variables $(\Gamma_i)_{i\in[m]}$. Whenever we condition on Γ, we assume that the bounds from Lemma 2.4 and 2.5 hold.

Lemma B.2. *Given Γ, the vectors $(\boldsymbol{Y}_1,\ldots,\boldsymbol{Y}_m)$ and $(\boldsymbol{X}_1,\ldots,\boldsymbol{X}_m)$ given $\mathscr{E}$ are identically distributed.*

Proof. For any integer sequence $(y_i)_{i\in[m]}$ with $y_i \geq 0$ and $\sum_{i\in[m]} y_i = k\Delta$ we have

$$\mathbb{P}\left[\forall i \in [m] : \boldsymbol{Y}_i = y_i \mid \Gamma\right] = \frac{\binom{k\Delta}{y_1,\ldots,y_m}\binom{(n-k)\Delta}{\Gamma_1-y_1,\ldots,\Gamma_m-y_m}}{\binom{n\Delta}{\Gamma_1,\ldots,\Gamma_m}} = \frac{\prod_{i=1}^m \frac{\Gamma_i!}{y_i!(\Gamma_i-y_i)!}}{\frac{(n\Delta)!}{(k\Delta)!((n-k)\Delta)!}} = \binom{n\Delta}{k\Delta}^{-1}\prod_{i=1}^m \binom{\Gamma_i}{y_i}.$$

Hence, for any sequences $(y_i), (y_i')$ we obtain

$$\frac{\mathbb{P}\left[\forall i \in [m] : \boldsymbol{Y}_i = y_i \mid \Gamma\right]}{\mathbb{P}\left[\forall i \in [m] : \boldsymbol{Y}_i = y_i' \mid \Gamma\right]} = \prod_{i=1}^m \frac{\binom{\Gamma_i}{y_i}}{\binom{\Gamma_i}{y_i'}} = \frac{\mathbb{P}\left[\forall i \in [m] : \boldsymbol{X}_i = y_i \mid \Gamma, \mathscr{E}\right]}{\mathbb{P}\left[\forall i \in [m] : \boldsymbol{X}_i = y_i' \mid \Gamma, \mathscr{E}\right]},$$

as claimed. □

B.3. **Proof of Lemma 2.4.** Since each variable draws a sequence of Δ tests uniformly at random, for every $i \in [m]$ the degree Γ_i has distribution $\mathrm{Bin}(n\Delta, 1/m)$. Therefore, the assertion follows from the Chernoff bound.

B.4. **Proof of Lemma 2.5.** Let $\boldsymbol{m}_0' = \sum_{i=1}^m \mathbf{1}\{\boldsymbol{X}_i = 0\}$. Then $\mathbb{E}[\boldsymbol{m}_0'] = \sum_{i=1}^m \mathbb{P}[\mathrm{Bin}(\Gamma_i, k/n)) = 0] = \sum_{i=1}^m (1-k/n)^{\Gamma_i}$. Hence, Lemma 2.4 shows that with probability $1-o(n^{-2})$,

$$\mathbb{E}[\boldsymbol{m}_0' \mid \Gamma] \geq m(1-k/n)^{\Gamma_{\max}} = m\exp\left((\Delta n/m + O(\sqrt{\Delta n/m}\log n))\log(1-k/n)\right) \tag{31}$$

$$= m\left(\exp(-d/c) + O(\sqrt{k/n}\log n)\right), \tag{32}$$

$$\mathbb{E}[\boldsymbol{m}_0' \mid \Gamma] \leq m(1-k/n)^{\Gamma_{\min}} = m\left(\exp(-d/c) + O(\sqrt{k/n}\log n)\right). \tag{33}$$

Because the $\boldsymbol{X}_i$ are mutually independent, $\boldsymbol{m}_0'$ is a binomial variable. Therefore, the Chernoff bound (e.g. Lemma B.1) shows that

$$\mathbb{P}\left[\left|\boldsymbol{m}_0' - \mathbb{E}[\boldsymbol{m}_0' \mid \Gamma]\right| > \sqrt{m}\log n \mid \Gamma\right] = o(n^{-10}). \tag{34}$$

Finally, the assertion follows from (30), (31)–(34) and Lemma B.2.

B.5. **Proof of Lemma 2.6.** The expected degree of a test a_i equals $\Delta n/m$. Therefore, if $\Delta = o(\log(n/k))$, then by Lemma 2.5, $\boldsymbol{m}_1 = o(m)$ w.h.p. To exploit this fact, call $\sigma \in \{0,1\}^V$ of Hamming weight k *bad* for $\boldsymbol{G}$ if given $\boldsymbol{\sigma} = \sigma$ we indeed have $\boldsymbol{m}_1 = o(m)$. Let $B(\boldsymbol{G})$ be the set of all such bad σ. Then w.h.p. $\boldsymbol{G}$ has the property that $|B(\boldsymbol{G})| \sim \binom{n}{k}$, i.e. asymptotically most configurations will have few positive tests. Now, condition on the event that $|B(\boldsymbol{G})| \sim \binom{n}{k}$ and let $\mathscr{B}$ be the set of all subsets of $[m]$ of size $o(m)$. Further, let $f_{\boldsymbol{G}} : B(\boldsymbol{G}) \to \mathscr{B}$ map $\sigma \in \{0,1\}^V$ to the corresponding set of positive tests. Finally, let $B'(\boldsymbol{G})$ be the set of all $\sigma \in B(\boldsymbol{G})$ such that $|f_{\boldsymbol{G}}^{-1}(f_{\boldsymbol{G}}(\sigma))| < n$, i.e. the set of all configurations for which there are less than n other configurations rendering the same test results. Then

$$|B'(\boldsymbol{G})| \leq n|\mathscr{B}| \leq n\binom{m}{o(m)} = \exp(o(m)) = o\left(\binom{n}{k}\right).$$

Consequently, w.h.p. over the choice of $\boldsymbol{G}$ and $\boldsymbol{\sigma}$ we have $Z_k(\boldsymbol{G},\hat{\boldsymbol{\sigma}}) \geq n$. The same argument applies for $\log(n/k) = o(\Delta)$ with the term 'positive test' replaced by 'negative test'.

B.6. **Proof of Proposition 2.3.** We start by proving part (1) using a straightforward second-moment calculation. Recall $\Delta = d\log(n/k)$ and $m = ck\log(n/k)$. Lemma 2.4 and Lemma 2.5 show that with probability at least $1-o(n^{-2})$ the total degree of the negative tests comes to

$$\begin{aligned}\sum_{i=1}^{m}\mathbf{1}\{\partial a_i \subseteq V_0\}\Gamma_i &= \Delta n\exp(-d/c) + O\left(\sqrt{m}\log^2(n)\Delta n/m + m\sqrt{\Delta n/m}\log n\right)\\ &= \Delta n\exp(-d/c) + O\left(\left(\sqrt{nk}+n/\sqrt{k}\right)\log^3 n\right) = \Delta n\left(\exp(-d/c)+n^{-\Omega(1)}\right).\end{aligned}$$

Consequently, with probability at least $1-o(n^{-2})$ the total number of edges between V_0 and the set of positive tests is $\Delta n\left(1-\exp(-d/c)+n^{-\Omega(1)}\right)$. Moreover, the total number of edges between V_0 and all tests comes down to $\Delta(n-k)$. Given these events and since each individual is assigned to tests uniformly at random with replacement, the probability that a given $x \in V_0$ belongs to V_0^+ comes out as

$$\binom{\Delta n\left(1-\exp(-d/c)+n^{-\Omega(1)}\right)}{\Delta}\binom{\Delta(n-k)}{\Delta}^{-1} = \left(1+n^{-\Omega(1)}\right)\left(1-\exp(-d/c)\right)^{\Delta}.$$

Next, we estimate the probability that $x, x' \in V_0$ both belong to V_0^+:

$$\binom{\Delta n\left(1-\exp(-d/c)+n^{-\Omega(1)}\right)}{2\Delta}\binom{\Delta(n-k)}{2\Delta}^{-1} = \left(1+n^{-\Omega(1)}\right)\left(1-\exp(-d/c)\right)^{2\Delta},$$

Hence, $\mathbb{E}[|V_0^+|^2 \mid \Gamma] - \mathbb{E}[|V_0^+| \mid \Gamma]^2 = O(n^{2-\Omega(1)})$. Therefore, the assertion follows from Chebyshev's inequality.

Proceeding with part (2), let the number of tests containing a single infected individual be

$$W = \sum_{i=1}^{m}\mathbf{1}\{\boldsymbol{Y}_i = 1\}, \qquad W' = \sum_{i=1}^{m}\mathbf{1}\{\boldsymbol{X}_i = 1\}.$$

Then Lemma 2.4 shows that w.h.p.

$$\begin{aligned}\mathbb{E}[W'] &= \sum_{i=1}^{m}\frac{\Gamma_i k}{n}(1-k/n)^{\Gamma_i-1} \leq \frac{\Gamma_{\max}km}{n}(1-k/n)^{\Gamma_{\min}-1}\\ &= \left(1+n^{-\Omega(1)}\right)k\Delta(1-k/n)^{\Delta n/m} = \left(1+n^{-\Omega(1)}\right)k\Delta\exp(-d/c)\end{aligned}$$

Analogously,

$$\mathbb{E}[W'] \geq \frac{\Gamma_{\min}km}{n}(1-k/n)^{\Gamma_{\max}} = \left(1+n^{-\Omega(1)}\right)k\Delta\exp(-d/c).$$

Hence, because W' is a binomial random variable, the Chernoff bound (e.g. Lemma B.1) shows that

$$\mathbb{P}\left[W' = \left(1+n^{-\Omega(1)}\right)k\Delta\exp(-d/c) \mid \Gamma\right] = 1-o(n^{-9}).$$

Therefore, (30) yields

$$\mathbb{P}\left[W = \left(1+n^{-\Omega(1)}\right)k\Delta\exp(-d/c) \mid \Gamma\right] = 1-o(n^{-7}). \tag{35}$$

Now, let U be the number of $x \in V_1$ that are not adjacent to any test with precisely one positive individual. An individual $x \in V_1$ counts towards U, if out of all possible assignment $k\Delta$, it is only assigned to those tests where it is not the only infected individual (there are a total of $k\Delta - W$ such assignments). Using the notation $n^{\underline{k}} = n(n-1)\dots(n-k+1)$ and recalling $\Delta = \Theta(\log n)$, the bound on W yields

$$\begin{aligned}\mathbb{E}[U \mid \Gamma, W] &= k\binom{k\Delta-W}{\Delta}\binom{k\Delta}{\Delta}^{-1} = k\frac{(k\Delta-W)^{\underline{\Delta}}}{(k\Delta)^{\underline{\Delta}}} = \left(1+n^{-\Omega(1)}\right)k\left(\frac{k\Delta-W}{k\Delta}\right)^{\Delta}\\ &= \left(1+n^{-\Omega(1)}\right)k(1-W/k\Delta)^{\Delta} = \left(1+n^{-\Omega(1)}\right)k(1-\exp(-d/c))^{\Delta}.\end{aligned}$$

By a similar token we obtain

$$\mathbb{E}[U^2 \mid \Gamma, W] = k^2\binom{k\Delta-W}{2\Delta}\binom{k\Delta}{2\Delta}^{-1} = \left(1+n^{-\Omega(1)}\right)\mathbb{E}[U \mid \Gamma, W]^2.$$

Therefore, Chebyshev's inequality shows that w.h.p.

$$U = \left(1 + n^{-\Omega(1)}\right) k(1 - \exp(-d/c))^{\Delta}. \tag{36}$$

To complete the proof we need to compare U and $\left|V_1^+\right|$. Clearly, $U \geq \left|V_1^+\right|$. But the inequality may be strict because U includes positive individuals that appear twice in the same test. To be precise, an individual might be assigned to one test twice as the only infected individual. Such an individual should not be in V_1^+, but it shows up in U. Indeed, letting R be the number of such individuals, we obtain $\left|V_1^+\right| \geq U - R$. Hence, we are left to estimate R. To this end, we observe that the probability that an individual appears in a specific test twice is upper-bounded by $(\Delta/m)^2$. Recall $m = ck\log(n/k)$ and $\Delta = d\log(n/k)$. Consequently, taking the union bound over all tests and infected individuals we yield

$$\mathbb{E}[R \mid \Gamma] \leq km\left(\frac{\Delta}{m}\right)^2 = O(\log n).$$

Since by assumption the r.h.s. of (36) is $n^{\Omega(1)}$, we conclude that $\left|V_1^+\right| \geq U - R = n^{\Omega(1)}$ w.h.p., as claimed.

Next, we consider (3). Define U as in the proof of Proposition 2.3(2). Then we know that $U \geq \left|V_1^+\right|$. Hence, if $k(1-\exp(-d/c))^{\Delta} = o(1)$ then $\left|V_1^+\right| = o(1)$ due to (36).

For part (4), we observe for a given c that $\min_d (1-\exp(-d/c))^{\Delta}$ is attained at $d = c\log 2$. To see this, consider the function $f(d) = (1-\exp(-d/c))^{\Delta} = n^{(1-\theta)d\log(1-\exp(-d/c))}$ and observe that the minimum of $f(d)$ coincides with the minimum of $g(d) = d\log(1-\exp(-d/c))$. Letting $x = d/c$, the derivatives read as

$$\begin{aligned} g(x) &= cx\log(1-\exp(-x)) \\ g'(x) &= c\left(\log(1-\exp(-x)) + \frac{x\exp(-x)}{1-\exp(-x)}\right) \\ g''(x) &= c\left(-\frac{(x-2)\exp(x)+2}{(\exp(x)-1)^2}\right) \end{aligned}$$

For $d > 0$, the unique maximum is attained at $x = \log 2$ and accordingly, $d = c\log 2$. Furthermore, it is the case that $k(1-\exp(-\log 2))^{c\log 2\log(n/k)} \geq n^{\Omega(1)}$ and therefore by Proposition 2.3(2), $\left|V_1^+\right| = n^{\Omega(1)}$. By a similar token by Proposition 2.3(1), $\left|V_0^+\right| = n^{\Omega(1)}$.

Finally, for part (5), setting $d = c\log 2$, we see that $k(1-\exp(-\log 2))^{c\log 2\log(n/k)} = o(1)$ and therefore by Proposition 2.3(3), $\left|V_1^+\right| = o(1)$.

C. The information-theoretic upper bound

C.1. **Proof of Lemma 3.4.** The term $\binom{k}{\ell}\binom{n-k}{k-\ell}$ accounts for the number of assignments $\sigma \in \{0,1\}^V$ of Hamming weight k whose overlap with $\boldsymbol{\sigma}$ is equal to ℓ. Hence, with $\mathscr{S}$ being the event that one specific $\sigma \in \{0,1\}^V$ that has overlap ℓ with $\boldsymbol{\sigma}$ belongs to $S_{k,\ell}(\boldsymbol{G}, \hat{\boldsymbol{\sigma}})$, we need to show that

$$\mathbb{P}\left[\mathscr{S} \mid \Gamma, \mathscr{R}, \boldsymbol{m}_0\right] \leq O\left((\Delta k)^{3/2}\right) \cdot \left(1 - \left(1 - \frac{k-\ell}{n-k}\right)^{\Gamma_{\max}}\right)^{\delta\Delta(k-\ell)} \left(\frac{n-2k+\ell}{n-k}\right)^{\Gamma_{\min}\boldsymbol{m}_0} \tag{37}$$

Due to symmetry we may assume that $\boldsymbol{\sigma}_{x_i} = \mathbf{1}\{i \leq k\}$ and that $\sigma_{x_i} = \mathbf{1}\{i \leq \ell\} + \mathbf{1}\{k < i \leq 2k-\ell\}$.

Proceeding as in the proof of Proposition 3.1, we think of each test a_i as a bin of capacity Γ_i and of each clone (x_i, h), $h \in [\Delta]$, of an individual as a ball labelled $(\boldsymbol{\sigma}_{x_i}, \sigma_{x_i}) \in \{0,1\}^2$. We toss the Δn balls randomly into the bins. For $i \in [m]$ and for $j \in [\Gamma_i]$ we let $\boldsymbol{A}_{i,j} = (\boldsymbol{A}_{i,j,1}, \boldsymbol{A}_{i,j,2}) \in \{0,1\}^2$ be the label of the jth ball that ends up in bin number i. To cope with this experiment we introduce a new set $\{0,1\}^2$-valued random variables $\boldsymbol{A}'_{i,j} = (\boldsymbol{A}'_{i,j,1}, \boldsymbol{A}'_{i,j,2})$ such that $(\boldsymbol{A}'_{i,j})_{i\in[m], j\in[\Gamma_i]}$ are mutually independent and

$$\begin{aligned} &\mathbb{P}\left[\boldsymbol{A}'_{i,j} = (1,1)\right] = \ell/n, && \mathbb{P}\left[\boldsymbol{A}'_{i,j} = (0,1)\right] = \mathbb{P}\left[\boldsymbol{A}'_{i,j} = (1,0)\right] = (k-\ell)/n, \\ &\mathbb{P}\left[\boldsymbol{A}'_{i,j} = (0,0)\right] = (n-2k+\ell)/n \end{aligned}$$

for all i, j. With $\mathscr{T}$ being the event that

$$\sum_{i=1}^{m}\sum_{j=1}^{\Gamma_i}\mathbf{1}\left\{\boldsymbol{A}'_{i,j}=(1,1)\right\}=\ell\Delta,\qquad \sum_{i=1}^{m}\sum_{j=1}^{\Gamma_i}\mathbf{1}\left\{\boldsymbol{A}'_{i,j}=(0,0)\right\}=(n-2k+\ell)\Delta, \tag{38}$$

$$\sum_{i=1}^{m}\sum_{j=1}^{\Gamma_i}\mathbf{1}\left\{\boldsymbol{A}'_{i,j}=(1,0)\right\}=\sum_{i=1}^{m}\sum_{j=1}^{\Gamma_i}\mathbf{1}\left\{\boldsymbol{A}'_{i,j}=(0,1)\right\}=(k-\ell)\Delta, \tag{39}$$

the vector $\boldsymbol{A}'=(\boldsymbol{A}'_{i,j})_{i,j}$ given $\mathscr{T}$ is distributed as $\boldsymbol{A}=(\boldsymbol{A}_{i,j})_{i,j}$ given Γ. Moreover, with similar arguments as in Section B.1, Stirling's formula yields

$$\mathbb{P}[\mathscr{T}]=\Omega((\Delta k)^{-3/2}). \tag{40}$$

Let $\mathscr{N}$ be the set of indices $i\in[m]$ such that $\max_{j\in[\Gamma_i]}\boldsymbol{A}'_{i,j,1}=0$. Moreover, let $\mathscr{M}$ be the set of all indices $i\in[m]$ for which there exists precisely one $g_i\in[\Gamma_i]$ such that $\boldsymbol{A}'_{i,g_i,1}=1$ and such that for this index we have $\boldsymbol{A}'_{i,g_i,2}=0$. Further, let

$$\mathscr{S}'=\left\{\forall i\in\mathscr{N}:\max_{j\in[\Gamma_i]}\boldsymbol{A}'_{i,j,2}=0\right\},\qquad\qquad \mathscr{S}''=\left\{\forall i\in\mathscr{M}:\max_{j\in[\Gamma_i]}\boldsymbol{A}'_{i,j,2}=1\right\}.$$

Then

$$\mathscr{A}=\left\{\forall i\in[m]:\max_{j\in[k]}\boldsymbol{A}'_{i,j,1}=\max_{j\in[k]}\boldsymbol{A}'_{i,j,2}\right\}\subseteq\mathscr{S}'\cap\mathscr{S}''.$$

Furthermore, given $\mathscr{N},\mathscr{M}$ the events $\mathscr{S}',\mathscr{S}''$ are independent and

$$\mathbb{P}\left[\mathscr{S}'\mid\mathscr{N}\right]=\prod_{i\in\mathscr{N}}\left(\frac{n-2k+\ell}{n-k}\right)^{\Gamma_i}\le\left(\frac{n-2k+\ell}{n-k}\right)^{\Gamma_{\min}|\mathscr{N}|},$$

$$\mathbb{P}\left[\mathscr{S}''\mid\mathscr{M}\right]=\prod_{i\in\mathscr{M}}\left(1-\left(1-\frac{k-\ell}{n-k}\right)^{\Gamma_i-1}\right)\le\left(1-\left(1-\frac{k-\ell}{n-k}\right)^{\Gamma_{\max}}\right)^{|\mathscr{M}|}.$$

For an intuitive explanation of the above expressions, please refer to the section immediately following the statement of the Lemma 3.4. Given $|\mathscr{N}|\ge\left(1-n^{-\Omega(1)}\right)\boldsymbol{m}_0$ and $|\mathscr{M}|\ge\delta\Delta(k-\ell)$, we obtain

$$\mathbb{P}\left[\mathscr{A}\mid|\mathscr{N}|\ge\left(1-n^{-\Omega(1)}\right)\boldsymbol{m}_0,|\mathscr{M}|\ge\delta\Delta(k-\ell)\right]\le\left(\frac{n-2k+\ell}{n-k}\right)^{\Gamma_{\min}\boldsymbol{m}_0}\left(1-\left(1-\frac{k-\ell}{n-k}\right)^{\Gamma_{\max}}\right)^{\delta\Delta(k-\ell)}. \tag{41}$$

Moreover, we find by 3.3, the concentration of $|\mathscr{N}|$ and the fact that $\mathbb{E}[|\mathscr{N}|]=\mathbb{E}[\boldsymbol{m}_0]=m/2$

$$\mathbb{P}\left(|\mathscr{N}|\ge\left(1-n^{-\Omega(1)}\right)\boldsymbol{m}_0,|\mathscr{M}|\ge\delta\Delta(k-\ell)\right)=1-o(1)$$

and thus

$$\mathbb{P}\left[\mathscr{T}\,||\mathscr{N}|\ge\left(1-n^{-\Omega(1)}\right)\boldsymbol{m}_0,|\mathscr{M}|\ge\delta\Delta(k-\ell)\right]=\Omega((\Delta k)^{-3/2}).$$

Combining (40)–(41) and using the trivial bound

$$\mathbb{P}\left[\mathscr{T}\mid\mathscr{S},\mathscr{S}',|\mathscr{N}|\ge\left(1-n^{-\Omega(1)}\right)\boldsymbol{m}_0,|\mathscr{M}|\ge\delta\Delta(k-\ell)\right]\le 1, \tag{42}$$

we obtain by Bayes Theorem

$$\mathbb{P}\left[\mathscr{A}\mid\mathscr{T},|\mathscr{N}|\ge\left(1-n^{-\Omega(1)}\right)\boldsymbol{m}_0,|\mathscr{M}|\ge\delta\Delta(k-\ell)\right]\le O\left((\Delta k)^{3/2}\right)\left(\frac{n-2k+\ell}{n-k}\right)^{\left(1-n^{-\Omega(1)}\right)\Gamma_{\min}\boldsymbol{m}_0}\left(1-\left(1-\frac{k-\ell}{n-k}\right)^{\Gamma_{\max}}\right)^{\delta\Delta(k-\ell)}. \tag{43}$$

Because $\boldsymbol{A}'=(\boldsymbol{A}'_{i,j})_{i,j}$ given $\mathscr{T}$ is distributed as $\boldsymbol{A}=(\boldsymbol{A}_{i,j})_{i,j}$ given Γ, (37) follows from (43).

D. THE SCOMP ALGORITHM

D.1. **Proof of Proposition 5.1.** The proof of Proposition 5.1 proceeds in three steps. First, we show that $\left|V_0^+\right|$ is concentrated around its expectation. $\mathscr{W}$ denotes the corresponding event. Second, we need to get a handle on the subtle dependencies in $\boldsymbol{G}$. To this end, we introduce a set of independent multinomial random variables indexed over the tests. Whereas $\boldsymbol{Y}_1^i, \boldsymbol{Y}_{0+}^i, \boldsymbol{Y}_{0-}^i$ denotes the number of infected, potentially false positive and definitively healthy individuals in test a_i, respectively, the triple $(\boldsymbol{X}_1^i, \boldsymbol{X}_{0+}^i, \boldsymbol{X}_{0-}^i)$ denote the corresponding multinomial random variable. We will show that conditioned on the sum of $\boldsymbol{X}_1^i, \boldsymbol{X}_{0+}^i, \boldsymbol{X}_{0-}^i$ hitting the total number of individuals of the three types, $(\boldsymbol{X}_1^i, \boldsymbol{X}_{0+}^i, \boldsymbol{X}_{0-}^i)$ is distributed like $\boldsymbol{Y}_1^i, \boldsymbol{Y}_{0+}^i, \boldsymbol{Y}_{0-}^i$. The technical workout is delicate, but is based on standard results from balls-into-bins experiments. Third, we show that for $m < (1-\varepsilon)m_{\mathrm{alg}}$, the number of tests W for which $\boldsymbol{X}_1^i = 1$ and $\boldsymbol{X}_{0+}^i = 0$ decays exponentially in n, which implies that $V_1^{--} = \emptyset$ w.h.p.

Proof. Lemma 2.6 implies that the optimal choice for the variable degree is $\Delta = d\log(n/k)$ for a constant d. Let $\boldsymbol{m}_1$ be the amount of positive tests and, w.l.o.g. assume that $a_1 \ldots a_{\boldsymbol{m}_1}$ are the positive tests and define

$$\mathscr{W} = \left\{|V_0^+| = (1+o(1))(n-k)(1-\exp(-d/c))^{\Delta}\right\}.$$

as the event that the number of 'potential false positives' $\left|V_0^+\right|$ is highly concentrated around its mean. Then by Proposition 2.3(1), we find

$$\mathbb{P}[\mathscr{W}] \geq 1 - o(1) \tag{44}$$

Similarly as before, we introduce a family of independent random variables corresponding to the tests.

Let $\boldsymbol{Y}_1^1, \ldots, \boldsymbol{Y}_1^{\boldsymbol{m}_1}$ be the number of ones in the tests corresponding to $a_1, \ldots, a_{\boldsymbol{m}_1}$ respectively. Let $\boldsymbol{Y}_{0+}^1, \ldots, \boldsymbol{Y}_{0+}^{\boldsymbol{m}_1}$ count the V_0^+ occurrences in $a_1, \ldots, a_{\boldsymbol{m}_1}$. Let $\boldsymbol{Y}_{0-}^1, \ldots, \boldsymbol{Y}_{0-}^{\boldsymbol{m}_1}$ count the V_0^- occurrences in $a_1, \ldots, a_{\boldsymbol{m}_1}$. By definition we find $\boldsymbol{Y}_{0-}^i = \Gamma_i - \boldsymbol{Y}_{0+}^i - \boldsymbol{Y}_1^i$. We introduce auxiliary variables $\boldsymbol{X}_1^1, \ldots, \boldsymbol{X}_1^{\boldsymbol{m}_1}$, $\boldsymbol{X}_{0+}^1, \ldots, \boldsymbol{X}_{0+}^{\boldsymbol{m}_1}, \boldsymbol{X}_{0-}^1, \ldots, \boldsymbol{X}_{0-}^{\boldsymbol{m}_1}$ such that $(\boldsymbol{X}_1^i, \boldsymbol{X}_{0+}^i, \boldsymbol{X}_{0-}^i)$ have distribution

$$\mathrm{Mult}_{\geq(1,0,0)}(\Gamma_i, p, q, 1-p-q),$$

a multinomial distribution conditioned on the first variable being at least one. The triples $\left((X_1^i, X_{0^+}^i, X_{0-}^i)\right)_{i\in\boldsymbol{m}_1}$ are mutually independent. We seek a choice of p satisfying the equation

$$p := \frac{k\Delta}{\sum_{i=1}^{\boldsymbol{m}_1} \frac{\Gamma_i}{1-(1-p)^{\Gamma_i}}} \qquad \text{and} \qquad q := \frac{\left|V_0^+\right|\Delta}{\sum_{i=1}^{\boldsymbol{m}_1} \frac{\Gamma_i}{1-(1-p)^{\Gamma_i}}}.$$

and will show following equation (48) that such a choice exists. Define

$$\mathscr{E} = \left\{\sum_{i=1}^{\boldsymbol{m}_1} \boldsymbol{X}_1^i = k\Delta, \sum_{i=1}^{\boldsymbol{m}_1} \boldsymbol{X}_{0+}^i = |V_0^+|\Delta\right\}.$$

Along the lines of Section B.1 , Stirling's formula implies

$$\mathbb{P}[\mathscr{E}] = \Omega(1/n). \tag{45}$$

Moreover, $(\boldsymbol{Y}_1^1, \boldsymbol{Y}_{0+}^1, \boldsymbol{Y}_{0-}^1, \ldots, \boldsymbol{Y}_1^{\boldsymbol{m}_1}, \boldsymbol{Y}_{0+}^{\boldsymbol{m}_1}, \boldsymbol{Y}_{0-}^{\boldsymbol{m}_1})$ and $(\boldsymbol{X}_1^1, \boldsymbol{X}_{0+}^1, \boldsymbol{X}_{0-}^1, \ldots, \boldsymbol{X}_1^{\boldsymbol{m}_1}, \boldsymbol{X}_{0+}^{\boldsymbol{m}_1}, \boldsymbol{X}_{0-}^{\boldsymbol{m}_1})$ given $\mathscr{E}$ are identically distributed. This can be seen as follows:

$$\begin{aligned}
&\mathbb{P}\left[\forall i \in [\boldsymbol{m}_1] : (\boldsymbol{Y}_1^i, \boldsymbol{Y}_{0+}^i, \boldsymbol{Y}_{0-}^i) = (y_i, y_i', y_i'') \mid \Gamma, |V_0^+|, \boldsymbol{m}_1\right] \\
&= \frac{\binom{k\Delta}{y_1 \ldots y_{\boldsymbol{m}_1}}\binom{|V_0^+|\Delta}{y_1' \ldots y_{\boldsymbol{m}_1}'}\binom{\sum_{i=1}^{\boldsymbol{m}_1}\Gamma_i - (k+|V_0^+|)\Delta}{\Gamma_1 - y_1 - y_1', \ldots, \Gamma_{\boldsymbol{m}_1} - y_{\boldsymbol{m}_1} - y_{\boldsymbol{m}_1}'}}{\binom{\sum_{i=1}^{\boldsymbol{m}_1}\Gamma_i}{\Gamma_1, \ldots, \Gamma_{\boldsymbol{m}_1}}} \mathbf{1}\{\forall i \in [\boldsymbol{m}_1] : y_i'' = \Gamma_i - y_i - y_i'\} \\
&= \binom{\sum_{i=1}^{\boldsymbol{m}_1}\Gamma_i}{k\Delta, |V_0^+|\Delta, \sum_{i=1}^{\boldsymbol{m}_1}\Gamma_i - \left(k+|V_0^+|\right)\Delta} \prod_{i=1}^{\boldsymbol{m}_1} \binom{\Gamma_i}{y_i, y_i', \Gamma - y_i - y_i'} \mathbf{1}\{\forall i \in [\boldsymbol{m}_1] : y_i'' = \Gamma_i - y_i - y_i'\}.
\end{aligned}$$

Thus, given $y_i'' = \Gamma_i - y_i - y_i'$ and $\tilde{y}_i'' = \Gamma_i - \tilde{y}_i - \tilde{y}_i'$ for all $i \in [\boldsymbol{m}_1]$, we find

$$\frac{\mathbb{P}\left[\forall i \in [\boldsymbol{m}_1] : (\boldsymbol{Y}_1^1, \boldsymbol{Y}_{0+}^1, \boldsymbol{Y}_{0-}^1) = (y_i, y_i', y'') \mid \Gamma, |V_0|^+, \boldsymbol{m}_1\right]}{\mathbb{P}\left[\forall i \in [\boldsymbol{m}_1] : (\boldsymbol{Y}_1^1, \boldsymbol{Y}_{0+}^1, \boldsymbol{Y}_{0-}^1) = (\tilde{y}_i, \tilde{y}_i', \tilde{y}_i'') \mid \Gamma, |V_0|^+, \boldsymbol{m}_1\right]} = \prod_{i=1}^{\boldsymbol{m}_1} \frac{\binom{\Gamma_i}{y_i, y_i', \Gamma - y_i - y_i'}}{\binom{\Gamma_i}{\tilde{y}_i, \tilde{y}_i', \Gamma - \tilde{y}_i - \tilde{y}_i'}}. \tag{46}$$

Given $x_i'' = \Gamma_i - x_i - x_i'$, we find:

$$\begin{aligned}
&\mathbb{P}\left[\forall i \in [\boldsymbol{m}_1] : (\boldsymbol{X}_1^1, \boldsymbol{X}_{0+}^1, \boldsymbol{X}_{0-}^1) = (x_i, x_i', x_i'') \mid \mathscr{E}, \Gamma, |V_0|^+, \boldsymbol{m}_1\right] \\
&= \prod_{i=1}^{\boldsymbol{m}_1} \binom{\Gamma_i}{x_i, x_i', x_i''} p^{x_i} q^{x_i'} (1-p-q)^{x_i''} \frac{1}{1-(1-p)^{\Gamma_i}} \\
&= p^{k\Delta} q^{|V_0^+|\Delta} (1-p-q)^{\sum_{i=1}^{\boldsymbol{m}_1} \Gamma_i - \Delta(k+|V_0^+|)} \prod_{i=1}^{\boldsymbol{m}_1} \frac{1}{1-(1-p)^{\Gamma_i}} \binom{\Gamma_i}{x_i, x_i', x_i''}
\end{aligned}$$

where the last equality follows from the fact that we conditioned on $\mathscr{E}$. Since the first terms are independent of x_i, x_i', x_i'', we find

$$\frac{\mathbb{P}\left[\forall i \in [\boldsymbol{m}_1] : (\boldsymbol{X}_1^i, \boldsymbol{X}_{0+}^i, \boldsymbol{X}_{0-}^i) = (x_i, x_i', x_i'') \mid \mathscr{E}, \Gamma, |V_0|^+, \boldsymbol{m}_1\right]}{\mathbb{P}\left[\forall i \in [\boldsymbol{m}_1] : (\boldsymbol{X}_1^i, \boldsymbol{X}_{0+}^i, \boldsymbol{X}_{0-}^i) = (\tilde{x}_i, \tilde{x}_i', \tilde{x}_i'') \mid \mathscr{E}, \Gamma, |V_0|^+, \boldsymbol{m}_1\right]} = \prod_{i=1}^{\boldsymbol{m}_1} \frac{\binom{\Gamma_i}{x_i, x_i', \Gamma - x_i - x_i'}}{\binom{\Gamma_i}{\tilde{x}_i, \tilde{x}_i', \Gamma - \tilde{x}_i - \tilde{x}_i'}}.$$

Therefore, given $\Gamma_i = x_i + x_i' + x_i'' = \tilde{x}_i + \tilde{x}_i' + \tilde{x}_i''$, we have by comparison with (46),

$$\begin{aligned}
&\frac{\mathbb{P}\left[\forall i \in [\boldsymbol{m}_1] : (\boldsymbol{X}_1^i, \boldsymbol{X}_{0+}^i, \boldsymbol{X}_{0-}^i) = (x_i, x_i', x_i'') \mid \mathscr{E}, \Gamma, |V_0|^+, \boldsymbol{m}_1\right]}{\mathbb{P}\left[\forall i \in [\boldsymbol{m}_1] : (\boldsymbol{X}_1^i, \boldsymbol{X}_{0+}^i, \boldsymbol{X}_{0-}^i) = (\tilde{x}_i, \tilde{x}_i', \tilde{x}_i'') \mid \mathscr{E}, \Gamma, |V_0|^+, \boldsymbol{m}_1\right]} \\
=&\frac{\mathbb{P}\left[\forall i \in [\boldsymbol{m}_1] : (\boldsymbol{Y}_1^i, \boldsymbol{Y}_{0+}^i, \boldsymbol{Y}_{0-}^i) = (x_i, x_i', x_i'') \mid \Gamma, \boldsymbol{m}_1\right]}{\mathbb{P}\left[\forall i \in [\boldsymbol{m}_1] : (\boldsymbol{Y}_1^i, \boldsymbol{Y}_{0+}^i, \boldsymbol{Y}_{0-}^i) = (\tilde{x}_i, \tilde{x}_i', \tilde{x}_i'') \mid \Gamma, \boldsymbol{m}_1\right]},
\end{aligned}$$

which yields the claim. Let

$$W = \sum_{i=1}^{\boldsymbol{m}_1} \mathbf{1}\left\{\boldsymbol{X}_1^i + \boldsymbol{X}_{0+}^i = 1\right\}.$$

be the number of positive tests that contain exactly one infected individual and no healthy individuals in V_0^+. Note that this split is the only possibility for the test to be positive. Then

$$\mathbb{E}[W|\Gamma, \mathscr{E}, \left|V_0^+\right|, \boldsymbol{m}_1] = \sum_{i=1}^{\boldsymbol{m}_1} \mathbb{P}[\boldsymbol{X}_1^i = 1, \boldsymbol{X}_{0+}^i = 0, \boldsymbol{X}_{0-}^i = \Gamma_i - 1] = \sum_{i=1}^{\boldsymbol{m}_1} \frac{\Gamma_i p(1-p-q)^{\Gamma_i - 1}}{1-(1-p)^{\Gamma_i}}.$$

By Lemma 2.5 we readily find for any choice of $c, d = \Theta(1)$ that

$$\sum_{i=1}^{\boldsymbol{m}_1} \frac{\Gamma_i p(1-p-q)^{\Gamma_i - 1}}{1-(1-p)^{\Gamma_i}} = (1+o(1)) \sum_{i=1}^{m} \Gamma_i p(1-p-q)^{\Gamma_i - 1} \tag{47}$$

Hence,

$$m\Gamma_{\min} p(1-p-q)^{\Gamma_{\max}} \le \mathbb{E}[W \mid \Gamma, \mathscr{E}, |V_0^+|, \boldsymbol{m}_1] \le m\Gamma_{\max} p(1-p-q)^{\Gamma_{\min}-1}.$$

Moreover, since W is a binomial random variable, the Chernoff bound (e.g. Lemma B.1) shows that

$$\mathbb{P}\left[\left|W - \mathbb{E}[W \mid \Gamma, \mathscr{E}, |V_0^+|]\right| > \sqrt{m}\log n\right] \le O(n^{-2}).$$

Further, Lemma 2.4 yields approximations for $\Gamma_{\min}$ and $\Gamma_{\max}$. Now assume that $c < \log^{-2} 2$. Using a similar reformulation as in (47), we find that $p = (1+o(1))k/n$. Thus, we have

$$\begin{aligned}
&\mathbb{E}[W|\Gamma, \mathscr{E}, \mathscr{W}] \\
&= (1+o(1)) m \frac{dn}{ck} \frac{k}{n} \exp\left((1+o(1)) \frac{dn}{ck} \log\left((1-k/n)\left(1+n^{-\Omega(1)}\right)(1-(1-\exp(-d/c)))^{\Delta}\right)\right) \\
&= (1+o(1)) m \exp(-d/c) \frac{d}{c} \left(1-(k/n)^{-d\log(1-\exp(-d/c))}\right)^{dn/(ck)}
\end{aligned} \tag{48}$$

As Lemma 2.6 shows, the optimal value of d is a constant. For a fixed c the same d that maximizes $-d/c\log(1-\exp(-d/c))$ in (48), also maximizes $\mathbb{E}[W|\Gamma, \mathscr{E}, \left|V_0^+\right|]$. This maximum is attained at $d = c\log 2$. Consequently $p = o(q)$ and

$$q \sim \left(\frac{k}{n}\right)^{c\log^2 2}.$$

Hence,

$$\mathbb{E}[W \mid \Gamma, \mathscr{E}, \mathscr{W}] \sim \frac{k\Delta}{2} \exp\left(-(\log 2)\left(\frac{n}{k}\right)^{1-c\log^2 2}\right) = \exp(-n^{\Omega(1)}).$$

As before, we find $\mathbb{E}[W] \to 0$ w.h.p. since $\mathbb{P}(\mathscr{W}) = 1 - o(1)$ and $\mathbb{P}(\mathscr{E}) = \Omega(1/\sqrt{\Delta k})$ and Markov's inequality leads to $V_1^{--} = \emptyset$. Proposition 5.1 follows. □

D.2. **Proof of Proposition 5.2.** By Lemma 2.3, we have $\left|V_0^+\right| \geq k\log n$ for $m < (1-\varepsilon)m_{\text{alg}}$. To prove Proposition 5.2, we need to show that for such m, we also have $\left|V_0^{+,\Delta}\right| \geq k\log n$. We proceed in two steps. First, we show that every individual $x \in V$ is assigned to at least $\Delta - O(1)$ distinct tests. Second, we show that a constant fraction of individuals $x \in V_0^+$ are assigned to exactly Δ tests establishing Proposition 5.2.

Proof. Let $d^\star(x)$ be the number of distinct neighbors of a vertex x. We claim that w.h.p. the following statements are true.

$$\min_{x \in V} d^\star(x) \geq \Delta - 2/\theta^2.$$

The probability that a given $x \in V$ appears $\ell \geq 2$ times in the same test is upper-bounded by

$$\binom{\Delta}{\ell} m^{1-\ell} \leq \frac{m}{\ell!}\left(\frac{d}{ck}\right)^\ell = \frac{ck\log(n/k)}{\ell!}\left(\frac{d}{ck}\right)^\ell \leq \frac{c(d/c)^\ell}{\ell!} n^{(1-\ell)\theta + o(1)} = o(1/n),$$

provided that $\ell > 1 + 1/\theta$. Moreover, the probability that x appears in one test twice is upper-bounded by $\Delta\dot{\Delta}/m$. Thus, the probability that x appears in at least ℓ tests at least twice is upper-bounded by

$$\sum_{i=\ell}^{\lfloor \Delta/2 \rfloor} \left(\frac{\Delta^2}{m}\right)^i = (1+o(1))\left(\frac{\Delta^2}{m}\right)^\ell \leq (1+o(1))\left(\frac{O(\log^2 n)}{ck\log(n/k)}\right)^\ell = n^{-\theta\ell + o(1)} = o(1/n),$$

provided that $\ell > 1/\theta$ and since $m = ck\log(n/k)$ and $\Delta = d\log(n/k)$. The bound follows.

By Lemma 2.3, we know that for $m < (1-\varepsilon)m_{\text{alg}}$, $\left|V_0^+\right| \geq k\log n$ w.h.p.. Since the SCOMP algorithm in its third stage selects the individual with the highest number of adjacent unexplained tests, we are left to show that also $\left|V_0^{+,\Delta}\right| \geq k\log n$, which implies that w.h.p. we erroneously classify a healthy individual as infected. The prior bounds ensure that each individual is in at least $\Delta - O(1)$ tests. The question remains which fraction of individuals in V_0^+ are in $V_0^{+,\Delta}$. In principle, it could be the case that most potentially false positive individuals of V_0^+ appear in less than Δ different tests. Indeed, it is more likely for such an individual in V_0^+ to be in fewer than Δ different tests since each additional test increases the probability for such an individual to be assigned to a negative test. However, we claim that a constant fraction of all potentially false positive individuals in V_0^+ will have degree Δ, thus be in $V_0^{+,\Delta}$. To see this, let p be the maximum proportion of $\left|V_0^{+,\Delta-i}\right|$ and $\left|V_0^{+,\Delta-i+1}\right|$ for $i \in [2/\theta^2]$, i.e.

$$p = \max_{i \in [2/\theta^2]} \frac{\left|V_0^{+,\Delta-i}\right|}{\left|V_0^{+,\Delta-i+1}\right|}$$

By conditioning on a test degree sequence $\Gamma_1, \ldots, \Gamma_m$, we find

$$p \geq (1 - (1-(k/n))^{\Gamma_{\min}}) = \Theta(1),$$

as long as $c, d = \Theta(1)$, which by Lemma 2.6 we can safely assume. Since each individual in V_0^+ is in at least $\Delta - O(1)$ different tests and the probability of being in any number of different tests $\Delta, \Delta-1\ldots$ is constant, a constant fraction of individuals in V_0^+ will be in exactly Δ tests. Since $\left|V_0^+\right| = \Omega(k\log n)$, the claim follows. □

REFERENCES

[1] E. Abbe: Community detection and stochastic block models: recent developments. The Journal of Machine Learning Research **18** (2017) 6446–6531.
[2] D. Achlioptas, A. Coja-Oghlan: Algorithmic barriers from phase transitions. Proc. 49th FOCS (2008) 793–802.
[3] D. Achlioptas, A. Coja-Oghlan, F. Ricci-Tersenghi: On the solution space geometry of random formulas. Random Structures and Algorithms **38** (2011) 251–268.
[4] D. Achlioptas, C. Moore: Random k-SAT: two moments suffice to cross a sharp threshold. SIAM Journal on Computing **36** (2006) 740–762.
[5] D. Achlioptas, A. Naor, and Y. Peres: Rigorous location of phase transitions in hard optimization problems. Nature **435** (2005) 759–764.
[6] D. Achlioptas, Y. Peres: The threshold for random k-SAT is $2^k \log 2 - O(k)$. Journal of the AMS **17** (2004) 947–973.
[7] A. Alaoui, A. Ramdas, F. Krzakala, L. Zdeborová, M. Jordan: Decoding from pooled data: Sharp information-theoretic bounds. SIAM Journal on Mathematics of Data Science **1** (2019) 161–188.
[8] A. Alaoui, A. Ramdas, F. Krzakala, L. Zdeborová, M. Jordan: Decoding from pooled data: Phase transitions of message passing. IEEE Transactions on Information Theory **65** (2019) 572–585.
[9] M. Aldridge: On the optimality of some group testing algorithms. IEEE International Symposium on Information Theory (2017).
[10] M. Aldridge: The capacity of Bernoulli nonadaptive group testing. IEEE Transactions on Information Theory **63** (2017) 7142–7148.
[11] M. Aldridge: Individual testing is optimal for nonadaptive group testing in the linear regime. IEEE Transactions on Information Theory **65** (2019) 2058–2061.
[12] M. Aldridge, L. Baldassini, O. Johnson: Group testing algorithms: bounds and simulations. IEEE Transactions on Information Theory **60** (2014) 3671–3687.
[13] M. Aldridge, O. Johnson, J. Scarlett: Improved group testing rates with constant column weight designs. IEEE International Symposium on Information Theory (2016).
[14] M. Aldridge, O. Johnson, J. Scarlett: Group testing: an information theory perspective. arXiv preprint arXiv:1902.06002 (2019).
[15] A. Alleman: An efficient algorithm for combinatorial group testing. H. Aydinian, F. Cicalese, C. Deppe (eds) Information Theory, Combinatorics, and Search Theory. Lecture Notes in Computer Science. Springer, Berlin, Heidelberg **7777** (2013), 569–596.
[16] R. Benz, S. Swamidass, P. Baldi: Discovery of power-laws in chemical space. Journal of Chemical Information and Modeling **48** (2008) 1138–1151.
[17] H. Chen, F. Hwang: A survey on nonadaptive group testing algorithms through the angle of decoding. Journal of Combinatorial Optimization **15** (2008) 49–59.
[18] A. Coja-Oghlan, F. Krzakala, W. Perkins , L. Zdeborová: Information-theoretic thresholds from the cavity method. Advances in Mathematics **333** (2018) 694–795.
[19] A. Coja-Oghlan, O. Gebhard, M. Hahn-Klimroth, P. Loick: Optimal non-adaptive group testing. (2019) arXiv: 1911.02287.
[20] A. Coja-Oghlan, K. Panagiotou: The asymptotic k-SAT threshold. Advances in Mathematics **288** (2016) 985–1068.
[21] B. Davis, D. McDonald: An elementary proof of the local central limit theorem. Journal of Theoretical Probability **8** (1995) 693–702.
[22] A. Decelle, F. Krzakala, C. Moore, L. Zdeborová: Asymptotic analysis of the stochastic block model for modular networks and its algorithmic applications. Phys. Rev. E **84** (2011) 066106.
[23] J. Ding, A. Sly, N. Sun: Proof of the satisfiability conjecture for large k. Proc. 47th STOC (2015) 59–68.
[24] R. Dorfman: The detection of defective members of large populations. Annals of Mathematical Statistics **14** (1943) 436–440.
[25] D. Du, F. Hwang: Combinatorial group testing and its applications. World Scientific (1993).
[26] O. Dubois, J. Mandler: The 3-XORSAT Threshold. Proc. 43rd FOCS (2002) 769–778.
[27] A. Emad, O. Milenkovic: Poisson group testing: a probabilistic model for nonadaptive streaming Boolean compressed sensing. Proc. ICASSP (2014) 3335–3339.
[28] M. Hahn-Klimroth, P. Loick: Optimal adaptive group testing. (2019) arXiv:1911.06647.
[29] S. Janson, T. Luczak, A. Ruciński: Random Graphs, Wiley 2000.
[30] O. Johnson, M. Aldridge, J. Scarlett: Performance of group testing algorithms with near-constant tests per item. IEEE Transactions on Information Theory **65** (2019) 707–723.
[31] H. Kwang-Ming, D. Ding-Zhu: Pooling designs and nonadaptive group testing: important tools for DNA sequencing. World Scientific (2006)
[32] M. Mézard, A. Montanari: Information, physics and computation. Oxford University Press 2009.
[33] M. Mézard, M. Tarzia, C. Toninelli: Group Testing with Random Pools: Phase Transitions and Optimal Strategy. Journal of Statistical Physics **131** (2008) 783–801.
[34] M. Molloy: The freezing threshold for k-colourings of a random graph. Proc. 43rd STOC (2012) 921–930.
[35] C. Moore: The computer science and physics of community detection: landscapes, phase transitions, and hardness. Bulletin of the EATCS **121** (2017).
[36] R. Mourad, Z. Dawy, F. Morcos: Designing pooling systems for noisy high-throughput protein-protein interaction experiments using Boolean compressed sensing. IEEE/ACM Transactions on Computational Biology and Bioinformatics **10** (2013) 1478–1490.
[37] E. Mossel, J. Neeman, A. Sly: Reconstruction and estimation in the planted partition model. Probability Theory and Related Fields (2014) 1–31.
[38] H. Ngo, D. Du: A survey on combinatorial group testing algorithms with applications to DNA library screening. Discrete Mathematical Problems with Medical Applications **7** (2000) 171–182.
[39] J. Scarlett, V. Cevher: Phase transitions in group testing. Proc. 27th SODA (2016) 40–53.
[40] J. Scarlett, V. Cevher: Limits on support recovery with probabilistic models: an information-theoretic framework. IEEE Transactions on Information Theory **63** (2017) 593-620.
[41] A. Sebo: On two random search problems. Journal of Statistical Planning and Inference **11** (1985) 23–31.

[42] N. Thierry-Mieg: A new pooling strategy for high-throughput screening: the shifted transversal design. BMC Bioinformatics **7** (2006) 28
[43] L. Zdeborová, F. Krzakala: Statistical physics of inference: thresholds and algorithms. Advances in Physics **65** (2016) 453–552.

AMIN COJA-OGHLAN, acoghlan@math.uni-frankfurt.de, GOETHE UNIVERSITY, MATHEMATICS INSTITUTE, 10 ROBERT MAYER ST, FRANKFURT 60325, GERMANY.

OLIVER GEBHARD, gebhard@math.uni-frankfurt.de, GOETHE UNIVERSITY, MATHEMATICS INSTITUTE, 10 ROBERT MAYER ST, FRANKFURT 60325, GERMANY.

MAX HAHN-KLIMROTH, hahnklim@math.uni-frankfurt.de, GOETHE UNIVERSITY, MATHEMATICS INSTITUTE, 10 ROBERT MAYER ST, FRANKFURT 60325, GERMANY.

PHILIPP LOICK, loick@math.uni-frankfurt.de, GOETHE UNIVERSITY, MATHEMATICS INSTITUTE, 10 ROBERT MAYER ST, FRANKFURT 60325, GERMANY.

C. Optimal Group Testing

OPTIMAL GROUP TESTING

AMIN COJA-OGHLAN, OLIVER GEBHARD, MAX HAHN-KLIMROTH, PHILIPP LOICK

ABSTRACT. In the group testing problem the aim is to identify a small set of $k \sim n^\theta$ infected individuals out of a population size n, $0 < \theta < 1$. We avail ourselves of a test procedure capable of testing groups of individuals, with the test returning a positive result iff at least one individual in the group is infected. The aim is to devise a test design with as few tests as possible so that the set of infected individuals can be identified correctly with high probability. We establish an explicit sharp information-theoretic/algorithmic phase transition m_{inf} for non-adaptive group testing, where all tests are conducted in parallel. Thus, with more than m_{inf} tests the infected individuals can be identified in polynomial time w.h.p., while learning the set of infected individuals is information-theoretically impossible with fewer tests. In addition, we develop an optimal adaptive scheme where the tests are conducted in two stages. *MSc: 05C80, 60B20, 68P30*

1. INTRODUCTION

1.1. **Background and motivation.** Various intriguing combinatorial problems come as inference tasks where we are to learn a hidden ground truth by means of indirect queries. The goal is to get by with as small a number of queries as possible. The ultimate solution to such a problem should consist of a positive algorithmic result showing that a certain number of queries suffice to learn the ground truth efficiently, complemented by a matching information-theoretic lower bound showing that with fewer queries the problem is insoluble, regardless of computational resources.

Group testing is a prime example of such an inference problem [6]. The objective is to identify within a large population of size n a subset of k individuals infected with a rare disease. We presume that the number of infected individuals scales as a power $k = \lceil n^\theta \rceil$ of the population size with an exponent $\theta \in (0,1)$, a parametrisation suited to modelling the pivotal early stages of an epidemic [36]. Indeed, since early on in an epidemic test kits might be in short supply, it is vital to get the most diagnostic power out the least number of tests. To this end we assume that the test gear is capable of not merely testing a single individual but an entire group. The test comes back positive if any one individual in the group is infected and negative otherwise. While in *non-adaptive* group testing all tests are conducted in parallel, in *adaptive* group testing test are conducted in several stages. In either case we are free to allocate individuals to test groups as we please. Randomisation is allowed. What is the least number of tests required so that the set of infected individuals can be inferred from the test results with high probability? Furthermore, in adaptive group testing, what is the smallest depth of test stages required?

Closing the considerable gaps that the best prior bounds left, the main results of this paper furnish matching algorithmic and information-theoretic bounds for both adaptive and non-adaptive group testing. Specifically, the best prior information-theoretic lower bound derives from the following folklore observation. Suppose that we conduct m tests that each return either 'positive' or 'negative'. Then to correctly identify the set of infected individuals we need the total number 2^m of conceivable test results to asymptotically exceed the number $\binom{n}{k}$ of possible sets of infected individuals. Hence, $2^m \geq (1+o(1))\binom{n}{k}$. Thus, Stirling's formula yields the lower bound

$$m_{\mathrm{ad}} = \frac{1-\theta}{\ln 2} n^\theta \ln n, \tag{1.1}$$

which applies to both adaptive and non-adaptive testing. On the positive side, a randomised non-adaptive test design with

$$m_{\mathrm{DD}} \sim \frac{\max\{\theta, 1-\theta\}}{\ln^2 2} n^\theta \ln n \tag{1.2}$$

Supported by DFG CO 646/3 and Stiftung Polytechnische Gesellschaft. An extended abstract version of this work has been submitted to the COLT 2020 conference.

tests exists from which a greedy algorithm called DD correctly infers the set of infected individuals w.h.p. [22]. Clearly, $m_{\mathrm{ad}} < m_{\mathtt{DD}}$ for all infection densities θ and $m_{\mathtt{DD}}/m_{\mathrm{ad}} \to \infty$ as $\theta \to 1$. In addition, there is an efficient adaptive three-stage group testing scheme that asymptotically matches the lower bound m_{ad} [33].

We proceed to state the main results of the paper. First, improving both the information-theoretic and the algorithmic bounds, we present optimal results for non-adaptive group testing. Subsequently we show how the non-adaptive result can be harnessed to perform adaptive group testing with the least possible number $(1+o(1))m_{\mathrm{ad}}$ of tests in only two stages.

1.2. **Non-adaptive group testing.** A *non-adaptive test design* is a bipartite graph $G = (V \cup F, E)$ with one vertex class $V = V_n = \{x_1, \ldots, x_n\}$ representing individuals and the other class $F = F_m = \{a_1, \ldots, a_m\}$ representing tests. For a vertex v of G denote by $\partial v = \partial_G v$ the set of neighbours of v. Thus, an individual x_j takes part in a test a_i iff $x_j \in \partial a_i$. Since we can shuffle the individuals randomly, we may safely assume that the vector $\boldsymbol{\sigma} \in \{0,1\}^V$ whose 1-entries mark the infected individuals is a uniformly random vector of Hamming weight k. Furthermore, the test results induced by $\boldsymbol{\sigma}$ read

$$\hat{\boldsymbol{\sigma}}_{a_i} = \hat{\boldsymbol{\sigma}}_{G,a_i} = \max_{x \in \partial a_i} \boldsymbol{\sigma}_x.$$

Hence, given $\hat{\boldsymbol{\sigma}} = \hat{\boldsymbol{\sigma}}_G = (\hat{\boldsymbol{\sigma}}_{G,a})_{a \in F}$ and G we aim to infer $\boldsymbol{\sigma}$. Thus, we can represent an inference procedure by a function $\mathscr{A}_G : \{0,1\}^m \to \{0,1\}^n$. The following theorem improves the lower bound on the number of tests required for successful inference. Let

$$m_{\mathrm{inf}} = m_{\mathrm{inf}}(n,\theta) = \max\left\{\frac{\theta}{\ln^2 2}, \frac{1-\theta}{\ln 2}\right\} n^\theta \ln n. \tag{1.3}$$

Theorem 1.1. *For any $0 < \theta < 1$, $\varepsilon > 0$ there exists $n_0 = n_0(\theta,\varepsilon)$ such that for all $n > n_0$, all test designs G with $m \le (1-\varepsilon)m_{\mathrm{inf}}$ tests and for every function $\mathscr{A}_G : \{0,1\}^m \to \{0,1\}^n$ we have*

$$\mathbb{P}\left[\mathscr{A}_G(\hat{\boldsymbol{\sigma}}_G) = \boldsymbol{\sigma}\right] < \varepsilon. \tag{1.4}$$

Theorem 1.1 rules out both deterministic and randomised test designs and inference procedures because (1.4) holds uniformly for all G and all $\mathscr{A}_G$. Thus, no test design, randomised or not, with fewer than m_{inf} tests allows to infer the set of infected individuals with a non-vanishing probability. Since m_{inf} matches $m_{\mathtt{DD}}$ from (1.2) for $\theta \ge 1/2$, Theorem 1.1 shows that the positive result from [22] is optimal in this regime. The following theorem closes the remaining gap by furnishing an optimal positive result for all θ.

Theorem 1.2. *For any $0 < \theta < 1$, $\varepsilon > 0$ there is $n_0 = n_0(\theta,\varepsilon)$ such that for every $n > n_0$ there exist a randomised test design $\boldsymbol{G}$ comprising $m \le (1+\varepsilon)m_{\mathrm{inf}}$ tests and a polynomial time algorithm* SPIV *that given $\boldsymbol{G}$ and the test results $\hat{\boldsymbol{\sigma}}_{\boldsymbol{G}}$ outputs $\boldsymbol{\sigma}$ w.h.p.*

An obvious candidate for an optimal test design appears be a plain random bipartite graph. In fact, prior to the present work the best known test design consisted of a uniformly random bipartite graph where all vertices in V_n have the same degree Δ. In other words, every individual independently joins Δ random test groups. Applied to this random Δ-out test design the DD algorithm correctly recovers the set of infected individuals in polynomial time provided that the number of tests exceeds $m_{\mathtt{DD}}$ from (1.2). However, $m_{\mathtt{DD}}$ strictly exceeds m_{inf} for $\theta < 1/2$. While the random Δ-out test design with $(1+o(1))m_{\mathrm{inf}}$ tests is known to admit an exponential time algorithm that successfully infers the set of infected individuals w.h.p. [11], we do not know of a polynomial time that solves this inference problem. Instead, to facilitate the new efficient inference algorithm SPIV the test design for Theorem 1.2 relies on a blend of a geometric and a random construction that is inspired by recent advances in coding theory known as spatially coupled low-density parity check codes [18, 26].

Finally, for

$$\theta \le \frac{\ln 2}{1+\ln 2} \approx 0.41 \tag{1.5}$$

the number m_{inf} of tests required by Theorem 1.2 matches the folklore lower bound m_{ad} from (1.2) that applies to both adaptive and non-adaptive group testing. Hence, in this regime adaptivity confers no advantage. By contrast, for $\theta > \ln(2)/(1+\ln 2)$ the adaptive bound m_{ad} is strictly smaller than m_{inf}. Consequently, in this regime at least two test stages are necessary to match the lower bound. Indeed, the next theorem shows that two stages suffice.

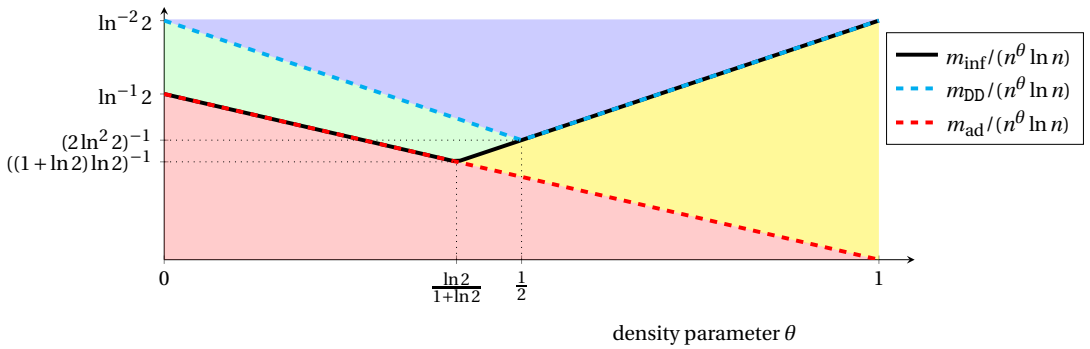

FIGURE 1. The phase transitions in group testing. The best previously known algorithm DD succeeds in the blue but not in the green region. The new algorithm SPIV succeeds in both the blue and the green region. The black line indicates the non-adaptive information-theoretic threshold m_{inf}, below which non-adaptive group testing is impossible. In the red area even (multi-stage) adaptive inference is impossible. Finally, the two-stage adaptive group testing algorithm from Theorem 1.3 succeeds in the yellow region.

1.3. **Adaptive group testing.** A *two-stage test design* consists of a bipartite graph $G = (V, F)$ along with a second bipartite graph $G' = G'(G, \hat{\boldsymbol{\sigma}}_G) = (V', F')$ with $V' \subset V$ that may depend on the tests results $\hat{\boldsymbol{\sigma}}_G$ of the first test design G. Hence, the task is to learn $\boldsymbol{\sigma}$ correctly w.h.p. from $G, \hat{\boldsymbol{\sigma}}_G, G'$ and the test results $\hat{\boldsymbol{\sigma}}_{G'}$ from the second stage while minimising the total number $|F| + |F'|$ of tests. The following theorem shows that a two-stage test design and an efficient inference algorithm exist that meet the multi-stage adaptive lower bound (1.1).

Theorem 1.3. *For any $0 < \theta < 1$, $\varepsilon > 0$ there is $n_0 = n_0(\theta, \varepsilon)$ such that for every $n > n_0$ there exist a two-stage test design with no more than $(1+\varepsilon)m_{\text{ad}}$ tests in total and a polynomial time inference algorithm that outputs $\boldsymbol{\sigma}$ with high probability.*

Theorem 1.3 improves over [33] by reducing the number of stages from three to two, thus potentially significantly reducing the overall time required to complete the test procedure [10, 28]. The proof of Theorem 1.3 combines the test design and efficient algorithm from Theorem 1.2 with ideas from [32].

The question of whether an 'adaptivity gap' exists for group testing, i.e., if the number of tests can be reduced by allowing multiple stages, has been raised prominently [6]. Theorems 1.1–1.3 answer this question comprehensively. While for $\theta \leq \ln(2)/(1+\ln(2)) \approx 0.41$ adaptivity confers no advantage, Theorem 1.1 shows that for $\theta > \ln(2)/(1+\ln(2))$ there is a widening gap between m_{ad} and the number m_{inf} of tests required by the optimal non-adaptive test design. Further, Theorem 1.3 demonstrates that this gap can be closed by allowing merely two stages. Figure 1 illustrates the thresholds from Theorems 1.1–1.3.

1.4. **Discussion.** The group testing problem was first raised in 1943, when Dorfman [15] proposed a two-stage adaptive test design to test the US Army for syphilis: in a first stage disjoint groups of equal size are tested. All members of negative test groups are definitely uninfected. Then, in the second stage the members of positive test groups get tested individually. Of course, this test design is far from optimal, but Dorfman's contribution triggered attempts at devising improved test schemes.

At first combinatorial group testing, where the aim is to construct a test design that is guaranteed to succeed on *all* vectors $\boldsymbol{\sigma}$, attracted significant attention. This version of the problem was studied, among others, by Erdős and Rényi [17], D'yachkov and Rykov [16] and Kautz and Singleton [23]. Hwang [20] was the first to propose an adaptive test design that asymptotically meets the information-theoretic lower bound m_{ad} from (1.1) for all $\theta \in [0,1]$. However, this test design requires an unbounded number of stages. Conversely, D'yachkov and Rykov [16] showed that m_{ad} tests do not suffice for non-adaptive group testing. Indeed, $m \geq \min\{\Omega(k^2), n\}$ tests are required non-adaptively, making individual testing optimal for $\theta > 1/2$. For an excellent survey of combinatorial group testing see [6].

Since the early 2000s attention has shifted to probabilistic group testing, which we study here as well. Thus, instead of asking for test designs and algorithms that are guaranteed to work for *all* $\boldsymbol{\sigma}$, we are content with recovering $\boldsymbol{\sigma}$ with high probability. Berger and Levenshtein [8] presented a two-stage probabilistic group testing design and

algorithm requiring

$$m_{\mathrm{BL,ad}} \sim 4n^{\theta}\ln n$$

tests in expectation. Their test design, known as the Bernoulli design, is based on a random bipartite graph where each individual joins every test independently with a carefully chosen edge probability. For a fixed θ the number $m_{\mathrm{BL,ad}}$ of tests is within a bounded factor of the information-theoretic lower bound m_{ad} from (1.1), although the gap $m_{\mathrm{ad}}/m_{\mathrm{BL,ad}}$ diverges as $\theta \to 1$. Unsurprisingly, the work of Berger and Levenshtein spurred efforts at closing the gap. Mézard, Tarzia and Toninelli proposed a different two-stage test design whose first stage consists of a random bipartite graph called the constant weight design [29]. Here each individual independently joins an equal number of random tests. For their two-stage design they obtained an inference algorithm that gets by with about

$$m_{\mathrm{MTT,ad}} \sim \frac{1-\theta}{\ln^2 2}n^{\theta}\ln n. \tag{1.6}$$

tests, a factor of $1/\ln 2$ above the elementary bound m_{ad}. Conversely, Mézard, Tarzia and Toninelli showed by means of the FKG inequality and positive correlation arguments that two-stage test algorithms from a certain restricted class cannot beat the bound (1.6). Furthermore, Aldridge, Johnson and Scarlett analysed non-adaptive test designs and inference algorithms [4, 22]. For the Bernoulli test design their best efficient algorithm DD requires

$$m_{\mathrm{DD,Be}} \sim \mathrm{e}\cdot\max\{\theta, 1-\theta\}\, n^{\theta}\ln n.$$

tests. For the constant weight design they obtained the bound m_{DD} from (1.2). In addition, in a previous article [11] we showed that on the constant weight design an exponential time algorithm correctly identifies the set of infected individuals w.h.p. if the number of tests exceeds m_{inf} from (1.3). Furthermore, Scarlett [33] discovered the aforementioned three-stage test design and polynomial time algorithm that matches the universal lower bound m_{ad} from (1.1). Finally, concerning lower bounds, in the case of a linear number $k = \Theta(n)$ infected individuals Aldridge [5] showed via arguments similar to [29] that individual testing is optimal in the non-adaptive case, while Ungar [35] proved that individual testing is optimal even adaptively once $k \geq (3-\sqrt{5})n/2$.

A further variant of group testing is known as the quantitative group testing or the coin weighing problem. In this problem tests are assumed to not merely indicate the presence of at least one infected individual but to return the number of infected individuals. Thus, the tests are significantly more powerful. For quantitative group testing with k infected individuals Alaoui, Ramdas, Krzakala, Zdeborová and Jordan [3] presented a test design with

$$m_{\mathrm{QGT}} \sim 2\left(1+\frac{(n-k)\ln(1-k/n)}{k\ln(k/n)}\right)\frac{k\ln(n/k)}{\ln(k)} \qquad \text{for} \qquad k=\Theta(n)$$

tests from which the set of infected individuals can be inferred in exponential time; the paper actually deals with the slightly more general pooled data problem. However, no efficient algorithm is known to come within a constant factor of m_{QGT}. Indeed, the best efficient algorithm, due to the same authors [2], requires $\Omega(k\ln(n/k))$ tests.

More broadly, the idea of harnessing random graphs to tackle inference problems has been gaining momentum. One important success has been the development of capacity achieving linear codes called spatially coupled low-density parity check ('LDPC') codes [26, 27]. The Tanner graphs of these codes, which represent their check matrices, consist of a linear sequence of sparse random bipartite graphs with one class of vertices corresponding to the bits of the codeword and the other class corresponding to the parity checks. The bits and the checks are divided equitably into a number of compartments, which are arranged along a line. Each bit of the codeword takes part in random checks in a small number of preceding and subsequent compartments of checks along the line. This combination of a spatial arrangement and randomness facilitates efficient decoding by means of the Belief Propagation message passing algorithm. Furthermore, the general design idea of combining a linear spatial structure with a random graph has been extended to other inference problems. Perhaps the most prominent example is compressed sensing, i.e., solving an underdetermined linear system subject to a sparsity constraint [13, 14, 24, 25], where a variant of Belief Propagation called Approximate Message Passing matches an information-theoretic lower bound from [37].

While in some inference problems such as LDPC decoding or compressed sensing the number of queries required to enable an efficient inference algorithm matches the information-theoretic lower bound, in many other problems gaps remain. A prominent example is the stochastic block model [1, 12, 30], an extreme case of which is the notorious planted clique problem [7]. For both these models the existence of a genuine computationally

intractable phase where the problem can be solved in exponential but not in polynomial time appears to be an intriguing possibility. Further examples include code division multiple access [34, 38], quantitative group testing [2], sparse principal component analysis [9] and sparse high-dimensional regression [31]. The problem of solving the group testing inference problem on the test design from [22] could be added to the list. Indeed, while an exponential time algorithm (that reduces the problem to minimum hypergraph vertex cover) infers the set of infected individuals w.h.p. with only $(1+\varepsilon)m_{\inf}$ tests, the best known polynomial algorithm requires $(1+\varepsilon)m_{\mathtt{DD}}$ tests.

Instead of developing a better algorithm for the test design from [22], here we exercise the discretion of constructing a different test design that the group testing problem affords. The new design is tailored to enable an efficient algorithm SPIV for Theorem 1.2 that gets by with $(1+\varepsilon)m_{\inf}$ tests. While prior applications of the idea of spatial coupling such as coding and compressed sensing required sophisticated message passing algorithms [18, 26, 27], the SPIV algorithm is purely combinatorial and extremely transparent. The main step of the algorithm merely computes a weighted sum to discriminate between infected individuals and 'disguised' healthy individuals. Furthermore, the analysis of the algorithm is based on a technically subtle but conceptually clean large deviations analysis. This technique of blending combinatorial ideas and large deviations methods with spatial coupling promises to be an exciting route for future research. Applications might include noisy versions of group testing, quantitative group testing or the coin weighing problem [2]. Beyond these immediate extensions, it would be most interesting to see if the SPIV strategy extends to other inference problems for sparse data.

1.5. **Organisation.** After collecting some preliminaries and introducing notation in Section 2, we prove Theorem 1.1 in Section 3. Section 4 then deals with the test design and the inference algorithm for Theorem 1.2. Finally, in Section 5 we prove Theorem 1.3.

2. Preliminaries

As we saw in Section 1.2 a non-adaptive test design can be represented by a bipartite graph $G = (V \cup F, E)$ with one vertex class V representing the individuals and the other class F representing the tests. We refer to the number $|V|$ of individuals as the *order* of the test design and to the number $|F|$ of tests as its *size*. For a vertex v of G we denote by $\partial_G v$ the set of neighbours. Where G is apparent from the notation we just write ∂v. Furthermore, for an integer $k \le |V|$ we denote by $\boldsymbol{\sigma}_{G,k} = (\boldsymbol{\sigma}_{G,k,x})_{x\in V} \in \{0,1\}^V$ a random vector of Hamming weight k. Additionally, we let

$$\hat{\boldsymbol{\sigma}}_{G,k} = (\hat{\boldsymbol{\sigma}}_{G,k,a})_{a\in F} \in \{0,1\}^F \qquad \text{with} \qquad \hat{\boldsymbol{\sigma}}_{G,k,a} = \max_{x\in\partial_G a} \boldsymbol{\sigma}_{G,k,x} \tag{2.1}$$

be the associated vector of test results. Where G and/or k are apparent from the context, we drop them from the notation. More generally, for a given vector $\tau \in \{0,1\}^V$ we introduce a vector $\hat{\tau}_G = (\hat{\tau}_{G,a})_{a\in F}$ by letting $\hat{\tau}_{G,a} = \max_{x\in\partial_G a}\tau_x$, just as in (2.1). Furthermore, for a given $\tau \in \{0,1\}^V$ we let

$$V_0(G,\tau) = \{x \in V : \tau_x = 0\}, \quad V_1(G,\tau) = \{x \in V : \tau_x = 1\}, \quad F_0(G,\tau) = \{a \in F : \hat{\tau}_{G,a} = 0\}, \quad F_1(G,\tau) = \{a \in F : \hat{\tau}_{G,a} = 1\}.$$

The *Kullback-Leibler divergence* of $p, q \in (0,1)$ is denoted by

$$D_{\mathrm{KL}}(q\|p) = q\ln\left(\frac{q}{p}\right) + (1-q)\ln\left(\frac{1-q}{1-p}\right).$$

We will occasionally apply the following Chernoff bound.

Lemma 2.1 ([21])**.** *Let $\boldsymbol{X}$ be a binomial random variable with parameters N, p. Then*

$$\mathbb{P}[X \ge qN] \le \exp\left(-ND_{\mathrm{KL}}(q\|p)\right) \quad \text{for } p < q < 1, \tag{2.2}$$

$$\mathbb{P}[X \le qN] \le \exp\left(-ND_{\mathrm{KL}}(q\|p)\right) \quad \text{for } 0 < q < p. \tag{2.3}$$

In addition, we recall that the *hypergeometric distribution* $\mathrm{Hyp}(L,M,N)$ is defined by

$$\mathbb{P}[\mathrm{Hyp}(L,M,N) = k] = \binom{M}{k}\binom{L-M}{N-k}\binom{L}{N}^{-1}. \qquad (k \in \{0,1,\dots,M\wedge N\}).$$

Hence, out of a total of L items of which M are special we draw N items without replacement and count the number of special items in the draw. The mean of the hypergeometric distribution equals MN/L. It is well known that the Chernoff bound extends to the hypergeometric distribution.

Lemma 2.2 ([19])**.** *For a hypergeometric variable $\boldsymbol{X} \sim \mathrm{Hyp}(L,M,N)$ the bounds* (2.2)–(2.3) *hold with $p = M/L$.*

Throughout the paper we use asymptotic notation $o(\cdot), \omega(\cdot), O(\cdot), \Omega(\cdot), \Theta(\cdot)$ to refer to limit $n \to \infty$. It is understood that the constants hidden in, e.g., a $O(\cdot)$-term may depend on the density parameter θ or other parameters.

3. The information theoretic lower bound

In this section we prove Theorem 1.1. The proof combines techniques based on the FKG inequality and positive correlation that were developed in [6, 29] with new combinatorial ideas. Throughout this section we fix a number $\theta \in (0,1)$ and we let $k = \lceil n^\theta \rceil$.

3.1. Outline.

The starting point is a simple and well known observation. Namely, for a test design $G = G_{n,m} = (V_n, F_m)$ and a vector $\tau \in \{0,1\}^{F_m}$ of test results let

$$\mathscr{S}_k(G,\tau) = \left\{\sigma \in \{0,1\}^{V_n} : \sum_{x \in V_n} \sigma_x = k,\ \hat{\sigma}_G = \tau\right\}$$

be the set of all possible vectors σ of Hamming weight k that give rise to the test results τ. Further, let $Z_k(G,\tau) = |\mathscr{S}_k(G,\tau)|$ be the number of such vectors σ. Also recall that $\boldsymbol{\sigma} = \boldsymbol{\sigma}_{G,k} \in \{0,1\}^{V_n}$ is a random vector of Hamming weight k and that $\hat{\boldsymbol{\sigma}} = \hat{\boldsymbol{\sigma}}_{G,k}$ comprises the test results that $\boldsymbol{\sigma}$ renders under the test design G. We observe that the posterior of $\boldsymbol{\sigma}$ given $\hat{\boldsymbol{\sigma}}$ is the uniform distribution on $\mathscr{S}_k(G,\hat{\boldsymbol{\sigma}})$.

Fact 3.1. *For any $G, \sigma \in \{0,1\}^{V_n}$ we have $\mathbb{P}[\boldsymbol{\sigma} = \sigma \mid \hat{\boldsymbol{\sigma}}] = \mathbf{1}\{\sigma \in \mathscr{S}_k(G,\hat{\boldsymbol{\sigma}})\}/Z_k(G,\hat{\boldsymbol{\sigma}})$.*

As an immediate consequence of Fact 3.1, the success probability of any inference scheme $\mathscr{A}_G : \{0,1\}^{F_m} \to \{0,1\}^{V_n}$ is bounded by $1/Z_k(G,\hat{\boldsymbol{\sigma}})$. Indeed, an optimal inference algorithm is to simply return a uniform sample from $\mathscr{S}_k(G,\hat{\boldsymbol{\sigma}})$.

Fact 3.2. *For any test design G and for any $\mathscr{A}_G : \{0,1\}^{F_m} \to \{0,1\}^{V_n}$ we have $\mathbb{P}[\mathscr{A}_G(\hat{\boldsymbol{\sigma}}) = \boldsymbol{\sigma} \mid \hat{\boldsymbol{\sigma}}] \le 1/Z_k(G,\hat{\boldsymbol{\sigma}})$.*

Hence, in order to prove Theorem 1.1 we just need to show that $Z_k(G,\hat{\boldsymbol{\sigma}})$ is large for any test design G with $m < (1-\varepsilon)m_{\inf}$ tests. In other words, we need to show that w.h.p. there are many vectors $\sigma \in \mathscr{S}_k(G,\hat{\boldsymbol{\sigma}})$ that give rise to the test results $\hat{\boldsymbol{\sigma}}$.

We obtain these σ by making diligent local changes to $\boldsymbol{\sigma}$. More precisely, we identify two sets $V_{0+} = V_{0+}(G,\boldsymbol{\sigma})$, $V_{1+} = V_{1+}(G,\boldsymbol{\sigma})$ of individuals whose infection status can be flipped without altering the test results. Specifically, following [5] we call an individual $x \in V_n$ *disguised* if every test $a \in \partial_G x$ contains another individual $y \in \partial_G a \setminus \{x\}$ with $\boldsymbol{\sigma}_y = 1$. Let $V_+ = V_+(G,\boldsymbol{\sigma})$ be the set of all disguised individuals. Moreover, let

$$V_{0+} = V_{0+}(G,\boldsymbol{\sigma}) = \{x \in V_+ : \boldsymbol{\sigma}_x = 0\}, \qquad V_{1+} = V_{1+}(G,\boldsymbol{\sigma}) = \{x \in V_+ : \boldsymbol{\sigma}_x = 1\}. \tag{3.1}$$

Hence, V_{0+} is the set of all healthy disguised individuals while V_{1+} contains all infected disguised individuals.

Fact 3.3. *We have $Z_k(G,\hat{\boldsymbol{\sigma}}) \ge |V_{0+}(G,\boldsymbol{\sigma})| \cdot |V_{1+}(G,\boldsymbol{\sigma})|$.*

Proof. For a pair $(x,y) \in V_{0+}(G,\boldsymbol{\sigma}) \times V_{1+}(G,\boldsymbol{\sigma})$ obtain $\boldsymbol{\tau}$ from $\boldsymbol{\sigma}$ by letting $\boldsymbol{\tau}_x = 1, \boldsymbol{\tau}_y = 0$ and $\boldsymbol{\tau}_z = \boldsymbol{\sigma}_z$ for all $z \neq x, y$. Then $\boldsymbol{\tau}$ has Hamming weight k and $\hat{\boldsymbol{\tau}}_G = \hat{\boldsymbol{\sigma}}$. Thus, $\boldsymbol{\tau} \in \mathscr{S}_k(G,\hat{\boldsymbol{\sigma}})$. □

Hence, an obvious proof strategy for Theorem 1.1 is to exhibit a large number of disguised individuals. A similar strategy has been pursued in the proof of the conditional lower bound of Mézard, Tarzia and Toninelli [29] and the proof of Aldridge's lower bound for the linear case $k = \Theta(n)$ [5]. Both [5, 29] exhibit disguised individuals via positive correlation and the FKG inequality. However, we do not see how to stretch such arguments to obtain the desired lower bound for all $\theta \in (0,1)$. Yet for θ *extremely* close to one it is possible to combine the positive correlation argument with new combinatorial ideas to obtain the following.

Proposition 3.4. *For any $\varepsilon > 0$ there exists $\theta_0 = \theta_0(\varepsilon) < 1$ such that for every $\theta \in (\theta_0, 1)$ there exists $n_0 = n_0(\theta,\varepsilon)$ such that for all $n > n_0$ and all test designs $G = G_{n,m}$ with $m \le (1-\varepsilon)m_{\inf}$ we have*

$$\mathbb{P}[|V_{0+}(G,\boldsymbol{\sigma})| \wedge |V_{1+}(G,\boldsymbol{\sigma})| \ge \ln n] > 1 - \varepsilon.$$

The proof of Proposition 3.4 can be found in Section 3.2.

The second step towards Theorem 1.1 is a reduction from larger to smaller values of θ. Suppose we wish to apply a test scheme designed for an infection density $\theta \in (0,1)$ to a larger infection density $\theta' \in (\theta, 1)$. Then we could dilute the larger infection density by adding a large number of healthy 'dummy' individuals. A careful analysis of this dilution process yields the following result. Due to the elementary lower bound (1.1) we need not worry about $\theta \le \ln(2)/(1+\ln 2)$.

Proposition 3.5. *For any* $\ln(2)/(1+\ln(2)) < \theta < \theta' < 1$, $t > 0$ *there exists* $n_0 = n_0(\theta,\theta',t) > 0$ *such that for every* $n > n_0$ *and for every test design* G *of order* n *there exist an integer* n' *such that*

$$k = \lceil n^{\theta} \rceil = \lceil n'^{\theta'} \rceil$$

and a test design G' *of order* n' *with the same number of tests as* G *such that the following is true. Let* $\boldsymbol{\tau} \in \{0,1\}^{V_{n'}}$ *be a random vector of Hamming weight* k *and let* $\hat{\boldsymbol{\tau}}_a = \max_{x \in \partial_{G'} a} \boldsymbol{\tau}_x$ *comprise the tests results of* G'. *Then*

$$\mathbb{P}\left[Z_k(G,\hat{\boldsymbol{\sigma}}) \le t\right] \le \mathbb{P}\left[Z_k(G',\hat{\boldsymbol{\tau}}) \le t\right].$$

Hence, if a test design exists for $\theta < \theta'$ that beats $m_{\inf}(n,\theta)$, then there is a test design for infection density θ' that beats $m_{\inf}(n',\theta')$. We prove Proposition 3.4 in Section 3.2. Theorem 1.1 is an easy consequence of Propositions 3.4 and 3.5.

Proof of Theorem 1.1. For $\theta \le \ln(2)/(1+\ln(2))$ the assertion follows from the elementary lower bound (1.1). Hence, fix $\varepsilon > 0$ and assume for contradiction that some $\theta \in (\ln(2)/(1+\ln(2)),1)$ for infinitely many n admits a test design G of order n and size $m \le (1-\varepsilon)m_{\inf}(n,\theta)$ such that $\mathbb{P}[Z_k(G,\hat{\boldsymbol{\sigma}}_G) \le t] \ge \varepsilon$. Then Proposition 3.5 shows that for $\theta' > \theta$ arbitrarily close to one for an integer n' with $k = \lceil n'^{\theta'} \rceil$ a test design $G' = G_{n',m}$ exists such that

$$\mathbb{P}\left[Z_k(G',\hat{\boldsymbol{\tau}}) \le 1/\varepsilon\right] \ge \varepsilon. \tag{3.2}$$

Furthermore, (1.3) shows that for large n,

$$m_{\inf}(n',\theta') = \frac{\theta'}{\ln^2 2} n'^{\theta'} \ln n' = \frac{\theta + o(1)}{\ln^2 2} n^{\theta} \ln n = (1+o(1)) m_{\inf}(n,\theta).$$

Hence, the number m of tests of G' satisfies $m \le (1-\varepsilon+o(1))m_{\inf}(n',\theta')$. Thus, (3.2) contradicts Fact 3.3 and Proposition 3.4. □

3.2. **Proof of Proposition 3.4.** Given a small $\varepsilon > 0$ we choose $\theta_0 = \theta_0(\varepsilon) \in (0,1)$ sufficiently close to one and fix $\theta \in (\theta_0,1)$. Additionally, pick $\xi = \xi(\varepsilon,\theta) \in (0,1)$ such that

$$2(1-\theta) < \xi < \theta\varepsilon. \tag{3.3}$$

We fix ε,θ,ξ throughout this section.

To avoid the (mild) stochastic dependencies that result from the total number of infected individuals being fixed, instead of $\boldsymbol{\sigma}$ we will consider a vector $\boldsymbol{\chi} \in \{0,1\}^{V_n}$ whose entries are stochastically independent. Specifically, every entry of $\boldsymbol{\chi}$ equals one with probability

$$p = \frac{k - \sqrt{k}\ln n}{n}$$

independently. Let $\hat{\boldsymbol{\chi}}_G \in \{0,1\}^{F_m}$ be the corresponding vector of test results. The following lemma shows that it suffices to estimate $|V_{0+}(G,\boldsymbol{\chi})|, |V_{1+}(G,\boldsymbol{\chi})|$. Let G denote an arbitrary test design with individuals $V_n = \{x_1,\ldots,x_n\}$ and tests $F_m = \{a_1,\ldots,a_m\}$.

Lemma 3.6. *There is* $n_0 = n_0(\theta,\varepsilon)$ *such that for all* $n > n_0$ *and for all* G *with* $m \le m_{\inf}$ *the following is true:*

$$\textit{if } \mathbb{P}\left[|V_{0+}(G,\boldsymbol{\chi})| \wedge |V_{1+}(G,\boldsymbol{\chi})| \ge 2\ln n\right] > 1-\varepsilon/4, \textit{ then } \mathbb{P}\left[|V_{0+}(G,\boldsymbol{\sigma})| \wedge |V_{1+}(G,\boldsymbol{\sigma})| \ge \ln n\right] > 1-\varepsilon.$$

Proof. Let $\mathscr{X} = \{k - 2\sqrt{k}\ln n \le \sum_{x \in V_n} \boldsymbol{\chi}_x \le k\}$. The Chernoff bound shows for large enough n,

$$\mathbb{P}[\mathscr{X}] > 1-\eta/4. \tag{3.4}$$

Further, given $\mathscr{X}$ we can couple $\boldsymbol{\chi},\boldsymbol{\sigma}$ such that the latter is obtained by turning $k - \sum_{x \in V_n} \boldsymbol{\chi}_x$ random zero entries of the former into ones. Since turning zero entries into ones can only increase the number of disguised individuals, on $\mathscr{X}$ we have

$$V_{1+}(G,\boldsymbol{\sigma}) \ge V_{1+}(G,\boldsymbol{\chi}). \tag{3.5}$$

Of course, it is possible that $|V_{0+}(G,\boldsymbol{\sigma})| < |V_{0+}(G,\boldsymbol{\chi})|$. But since on $\mathscr{X}$ the two vectors $\boldsymbol{\sigma},\boldsymbol{\chi}$ differ in no more than $2\sqrt{k}\ln n$ entries, we obtain the bound

$$\mathbb{E}\left[|V_{0+}(G,\boldsymbol{\chi})| - |V_{0+}(G,\boldsymbol{\sigma})| \,|\, \mathscr{X}\right] \le \frac{2\sqrt{k}\ln n}{n-k} |V_{0+}(G,\boldsymbol{\chi})| < n^{-1/3} |V_{0+}(G,\boldsymbol{\chi})|,$$

provided n is sufficiently large. Hence, Markov's inequality shows that for large enough n,

$$\mathbb{P}\left[|V_{0+}(G,\boldsymbol{\chi})| - |V_{0+}(G,\boldsymbol{\sigma})| > |V_{0+}(G,\boldsymbol{\chi})|/2 \,|\, \mathscr{X}\right] < \varepsilon/4. \tag{3.6}$$

Combining (3.4), (3.5) and (3.6) completes the proof. □

As a next step we show that there is no point in having very big tests a that contain more than, say, $\Gamma = \Gamma(n,\theta) = n^{1-\theta}\ln n$ individuals. This is because anyway all such tests are positive w.h.p., so there is little point in actually conducting them. Indeed, the following lemma shows that w.h.p. all tests of very high degree contain at least two infected individuals.

Lemma 3.7. *There exists $n_0 = n_0(\theta,\varepsilon) > 0$ such that for all $n > n_0$ and all test designs G with $m \le m_{\inf}$ tests,*

$$\mathbb{P}\left[\exists a \in F_m : |\partial_G a| > \Gamma \wedge |\partial_G a \cap V_1(G,\boldsymbol{\chi})| \le 1\right] < \varepsilon/8.$$

Proof. Consider a test a of degree $\gamma = |\partial_G a| \ge \Gamma$. Because in $\boldsymbol{\chi}$ each of the γ individuals that take part in a is infected with probability p independently, we have

$$\mathbb{P}\left[|\partial_G a \cap V_1(G,\boldsymbol{\sigma})| \le 1\right] = \mathbb{P}\left[\mathrm{Bin}(\gamma,p) \le 1\right] = (1-p)^\gamma + \gamma p(1-p)^{\gamma-1} \le (1+\gamma p/(1-p))\exp(-\gamma p) = n^{o(1)-1}. \tag{3.7}$$

Since $m \le m_{\inf} = O(n^\theta)$ for a fixed $\theta < 1$, the assertion follows from (3.7) and the union bound. □

Let G^* be test design obtained from $G = G_{n,m}$ by deleting all tests of degree larger than Γ. If indeed every test of degree at least Γ contains at least two infected individuals, then $V_{0+}(G^*,\boldsymbol{\chi}) = V_{0+}(G,\boldsymbol{\chi})$ and $V_{1+}(G^*,\boldsymbol{\chi}) = V_{1+}(G,\boldsymbol{\chi})$. Hence, Lemma 3.7 shows that it suffices to bound $|V_{0+}(G^*,\boldsymbol{\chi})|, |V_{1+}(G^*,\boldsymbol{\chi})|$. To this end we observe that G^* contains few individuals of very high degree.

Lemma 3.8. *There is $n_0 = n_0(\theta,\varepsilon) > 0$ such that for all $n > n_0$ and all test designs G with $m \le m_{\inf}$ we have*

$$\left|\left\{x \in V_n : |\partial_{G^*} x| > \ln^3 n\right\}\right| \le \frac{n\ln\ln n}{\ln n}.$$

Proof. Since $\max_{a\in F_m} |\partial_{G^*} a| \le \Gamma = n^{1-\theta}\ln n$, double counting yields

$$\sum_{x\in V_n} |\partial_{G^*} x| = \sum_{a\in F_m} |\partial_{G^*} a| \le m_{\inf}\Gamma = O(n\ln^2 n).$$

Consequently, there are no more than $O(n/\ln n)$ individuals $x \in V_n$ with $|\partial_{G^*} x| > \ln^3 n$. □

Further, obtain $G^{(0)}$ from G^* by deleting all individuals of degree greater than $\ln^3 n$ (but keeping all tests). Then the degrees of $G^{(0)}$ satisfy

$$\max_{a\in F(G^{(0)})} |\partial_{G^{(0)}} a| \le \Gamma, \qquad \max_{x\in V(G^{(0)})} |\partial_{G^{(0)}} x| \le \ln^3 n. \tag{3.8}$$

Let $\boldsymbol{\chi}^{(0)} = (\boldsymbol{\chi}_x)_{x\in V(G^{(0)})}$ signify the restriction of $\boldsymbol{\chi}$ to the individuals that remain in $G^{(0)}$.

With these preparations in place we are ready to commence the main step of the proof of Proposition 3.4. Given a test design G with $m \le (1-\varepsilon)m_{\inf}$ we are going to construct a sequence $y_1, y_2, \ldots, y_N$, $N = \lceil n^{1-\xi}\rceil$, of individuals of $G^{(0)}$ such that each y_i individually has a moderately high probability of being disguised. Of course, to conclude that in the end a large number of disguised y_i actually materialise, we need to cope with stochastic dependencies. To this end we will pick individuals y_i that have pairwise distance at least five in $G^{(0)}$. The degree bounds (3.8) guarantee a sufficient supply of such far apart individuals.

To be precise, starting from $G^{(0)}$ we construct a sequence of test designs $G^{(1)}, G^{(2)}, \ldots, G^{(N)}$ inductively as follows. For each $i \ge 1$ select a variable $y_{i-1} \in V(G^{(i-1)})$ whose probability of being disguised is maximum; ties are broken arbitrarily. In formulas,

$$\mathbb{P}\left[y_{i-1} \in V_+(G^{(i-1)},\boldsymbol{\chi}^{(i-1)})\right] = \max_{y\in V(G^{(i-1)})} \mathbb{P}\left[y \in V_+(G^{(i-1)},\boldsymbol{\chi}^{(i-1)})\right],$$

where, of course, $\boldsymbol{\chi}^{(i-1)}$ is the only random object. Then obtain $G^{(i)}$ from $G^{(i-1)}$ by removing y_{i-1} along with all vertices (i.e., tests or individuals) at distance at most four from y_{i-1}. Moreover, let $\boldsymbol{\chi}^{(i)}$ denote the restriction $(\boldsymbol{\chi}_x)_{x\in V(G^{(i)})}$ of $\boldsymbol{\chi}$ to $G^{(i)}$. The following lemma estimates the probability of y_i being disguised. Let $m^* = |F(G^*)|$ be the total number of tests of G of degree at most Γ.

Lemma 3.9. *There exists $n_0 = n_0(\varepsilon,\theta,\xi)$ such that for all $n > n_0$ and all G with $m \le (1-\varepsilon)m_{\inf}$ we have*

$$\min_{1\le i\le N} \mathbb{P}\left[y_i \in V_+(G^{(i)})\right] \ge \exp\left(-\frac{m\ln^2 2}{n^\theta} - 1\right).$$

The proof of Lemma 3.9 requires three intermediate steps. First, we need a lower bound on number of individuals in $G^{(i)}$. Recall that $N = \lceil n^{1-\xi}\rceil$.

Claim 3.10. *We have* $\min_{0\le i\le N} |V(G^{(i)})| \ge n - N\Gamma^2\ln^6 n$.

Proof. Since throughout the construction of the $G^{(i)}$ we only delete vertices, the degree bound (3.8) implies

$$\max_{a\in F(G^{(i)})} |\partial_{G^{(i)}} a| \le \Gamma = n^{1-\theta}\ln n, \qquad \max_{x\in V(G^{(i)})} |\partial_{G^{(i)}} x| \le \ln^3 n \qquad \text{for all } i \le N. \tag{3.9}$$

We now proceed by induction on i. For $i = 0$ there is nothing to show. Going from i to $i+1 \le N$, we notice that because all individuals $x \in V(G^{(i)}) \setminus V(G^{(i+1)})$ have distance at most four from y_{i+1}, (3.9) ensures that

$$|V(G^{(i)}) \setminus V(G^{(i+1)})| \le \Gamma^2\ln^6 n. \tag{3.10}$$

Iterating (3.10), we obtain $|V(G^{(0)}) \setminus V(G^{(i+1)})| \le (i+1)\Gamma^2\ln^6 n$, whence $|V(G^{(i+1)})| \ge n - (i+1)\Gamma^2\ln^6 n$. □

The following claim resembles the proof of [5, Theorem 1] (where the case $k = \Omega(n)$ is considered).

Claim 3.11. *Let $\mathscr{D}^{(i)}(x) = \{x \in V_+(G^{(i)})\}$ and let*

$$L^{(i)} = \frac{1}{|V(G^{(i)})|}\sum_{x\in V(G^{(i)})} \ln\mathbb{P}\left[\mathscr{D}^{(i)}(x)\right]. \tag{3.11}$$

Then

$$L^{(i)} \ge \frac{|F(G^{(i)})|}{|V(G^{(i)})|}\min_{a\in F(G^{(i)})} \left|\partial_{G^{(i)}} a\right| \ln\left(1-(1-p)^{|\partial_{G^{(i)}} a|-1}\right). \tag{3.12}$$

Proof. For an individual $x \in V(G^{(i)})$ and a test $a \in \partial_{G^{(i)}} x$ let $\mathscr{D}^{(i)}(x,a)$ be the event that there is another individual $z \in \partial_{G^{(i)}} a \setminus \{x\}$ such that $\boldsymbol{\chi}_z = 1$. Then for every $x \in V(G^{(i)})$ we have

$$\mathbb{P}\left[\mathscr{D}^{(i)}(x)\right] = \mathbb{P}\left[\bigcap_{a\in\partial_{G^{(i)}} x} \mathscr{D}^{(i)}(x,a)\right]. \tag{3.13}$$

Furthermore, the events $\mathscr{D}^{(i)}(x,a)$ are increasing with respect to $\boldsymbol{\chi}$. Therefore, (3.13) and the FKG inequality imply

$$\mathbb{P}\left[\mathscr{D}^{(i)}(x)\right] \ge \prod_{a\in\partial_{G^{(i)}} x} \mathbb{P}\left[\mathscr{D}^{(i)}(x,a)\right]. \tag{3.14}$$

Moreover, because each entry of $\boldsymbol{\chi}$ is one with probability p independently, we obtain

$$\mathbb{P}\left[\mathscr{D}^{(i)}(x,a)\right] = 1-(1-p)^{|\partial_{G^{(i)}} a|-1} \tag{3.15}$$

Finally, combining (3.13)–(3.15), we obtain

$$\begin{aligned}|V(G^{(i)})|L^{(i)} &\ge \sum_{x\in V(G^{(i)})}\sum_{a\in F(G^{(i)})} \mathbf{1}\{a\in\partial_{G^{(i)}} x\}\ln\left(1-(1-p)^{|\partial_{G^{(i)}} a|-1}\right)\\ &= \sum_{a\in F(G^{(i)})} \left|\partial_{G^{(i)}} a\right|\ln\left(1-(1-p)^{|\partial_{G^{(i)}} a|-1}\right) \ge |F(G^{(i)})|\min_{a\in F(G^{(i)})}\left|\partial_{G^{(i)}} a\right|\ln\left(1-(1-p)^{|\partial_{G^{(i)}} a|-1}\right),\end{aligned}$$

as claimed. □

As a final preparation for the proof of Lemma 3.9 we need the following estimate.

Claim 3.12. *The function $z \in (0,\infty) \mapsto z\ln(1-(1-p)^{z-1})$ attains its minimum at $z = \left(1+O(n^{-\Omega(1)})\right)\ln(2)/p$.*

Proof. We consider three separate cases.

Case 1: $z = o(1/p)$**:** we obtain

$$\begin{aligned}z\ln\left(1-(1-p)^{z-1}\right) &= z\ln\left(1-\exp\left(-pz+O(p^2z)\right)\right) = z\ln\left(1-\left(1-pz+O(p^2z^2)\right)\right)\\ &= \frac{z}{\ln}(zp+O(zp)^2) = o(1/p).\end{aligned} \tag{3.16}$$

Case 2: $z = \omega(1/p)$**:** we find

$$z\ln\bigl(1-(1-p)^{z-1}\bigr) = z\ln\bigl(1-\exp\bigl(-pz+O(p^2z)\bigr)\bigr) = -z\bigl(\exp(-pz)+O\bigl(\exp(-2pz)\bigr)\bigr)$$
$$= -\frac{1}{p}pz\bigl(\exp\bigl(-pz\bigr)+\exp\bigl(-2pz\bigr)\bigr) = o(1/p). \tag{3.17}$$

Case 3: $z = \Theta(1/p)$**:** letting $d = zp$, we obtain

$$z\ln\bigl(1-(1-p)^{z-1}\bigr) = \frac{d}{p}\ln\bigl(1-\exp\bigl(-d+O(p)\bigr)\bigr) = \frac{d}{p}\ln\bigl(1-\exp\bigl(-d\bigr)\bigr)+O(1). \tag{3.18}$$

Since the strictly convex function $d \in (0,\infty) \mapsto d\ln(1-\exp(-d))$ attains its minimum at $d = \ln 2$, (3.18) dominates (3.16) and (3.17). Thus, the minimiser reads $z = \ln(2)/p + O(p^{-1/2})$. □

Proof of Lemma 3.9. Combining Claims 3.11 and 3.12, we see that for all test designs G with $m \le (1-\varepsilon)m_{\inf}$ and for all $i \le N$,

$$L^{(i)} \ge -\bigl(1+O(n^{-\Omega(1)})\bigr)\frac{|F(G^{(i)})|\ln^2 2}{|V(G^{(i)})|p} \ge -\bigl(1+O(n^{-\Omega(1)})\bigr)\frac{m\ln^2 2}{|V(G^{(i)})|p}.$$

Hence, Claim 3.10, (3.3) and the choice $p = (k+\sqrt{k}\ln n)/n$ imply that for all $i \le N$,

$$L^{(i)} \ge -\bigl(1+O(n^{-\Omega(1)})\bigr)\frac{m\ln^2 2}{(n-N\Delta^2\ln^6 n)p} \ge -\bigl(1+O(n^{-\Omega(1)})\bigr)\frac{m\ln^2 2}{n^\theta}. \tag{3.19}$$

Further, combining the definition (3.11) of $L^{(i)}$ with (3.19), we conclude that for every $i \le N$ there exists an individual $y_i \in V(G^{(i)})$ such that

$$\mathbb{P}\Bigl[y_i \in V_+(G^{(i)})\Bigr] = \mathbb{P}\Bigl[\mathscr{D}^{(i)}(y_i)\Bigr] \ge \exp\Bigl(L^{(i)}\Bigr) \ge \exp\biggl(-\bigl(1+O(n^{-\Omega(1)})\bigr)\frac{m\ln^2(2)}{n^\theta}\biggr),$$

which implies the assertion. □

Lemma 3.9 implies the following bound on $|V_{0+}(G^*,\boldsymbol{\chi})|$, $|V_{1+}(G^*,\boldsymbol{\chi})|$.

Corollary 3.13. *There exists $n_0 = n_0(\varepsilon,\theta,\xi)$ such that for all $n > n_0$ and all $G = G_{n,m}$ with $m \le (1-\varepsilon)m_{\inf}$ we have*

$$\mathbb{P}\bigl[\bigl|V_{0+}(G^*,\boldsymbol{\chi})\bigr| \wedge \bigl|V_{1+}(G^*,\boldsymbol{\chi})\bigr| < \ln^4 n\bigr] < \varepsilon/8.$$

Proof. We observe that $V_+(G^{(i)},\boldsymbol{\chi}) \subset V_+(G^*,\boldsymbol{\chi})$ for all $i \le N$ because by construction for any individual $x \in V(G^{(i)})$ every test $a \in \partial_{G^*}x$ of G^* that x belongs to is still present in $G^{(i)}$. Consequently, we obtain the bound

$$\mathbb{P}\bigl[x \in V_+(G^*)\bigr] \ge \mathbb{P}\Bigl[x \in V(G^{(i)})\Bigr] \qquad \text{for all } i \in [N],\ x \in V(G^*). \tag{3.20}$$

Combining (3.20) with Lemma 3.9 we obtain

$$\mathbb{P}\Bigl[y^{(i)} \in V_+(G^*)\Bigr] \ge \exp\Bigl(-\ln^2(2)n^{-\theta}m-1\Bigr) \ge \exp\Bigl(-(1-\varepsilon)\ln^2(2)n^{-\theta}m_{\inf}-1\Bigr) \qquad \text{for all } i \in [N].$$

Hence, recalling the definition of $m_{\inf}$ from (1.3), we obtain

$$\mathbb{P}\Bigl[y^{(i)} \in V_+(G^*)\Bigr] \ge \exp\bigl(-(1-\varepsilon)\theta\ln(n)-1\bigr) = n^{(\varepsilon-1)\theta}/\mathrm{e}. \qquad \text{for all } i \in [N]. \tag{3.21}$$

Since the entry $\boldsymbol{\chi}_{y^{(i)}}$ is independent of the event $\{y^{(i)} \in V_+(G^*)\}$, the definitions (3.1) of $V_{0+}(G^*,\boldsymbol{\chi})$ and $V_{1+}(G^*,\boldsymbol{\chi})$ and (3.21) yield

$$\mathbb{P}\Bigl[y^{(i)} \in V_{0+}(G^*,\boldsymbol{\chi})\Bigr] \ge (1-p)\cdot\frac{n^{(\varepsilon-1)\theta}}{\mathrm{e}} \ge \frac{n^{\varepsilon\theta-1}}{3}, \quad \mathbb{P}\Bigl[y^{(i)} \in V_{1+}(G^*,\boldsymbol{\chi})\Bigr] \ge p\cdot\frac{n^{(\varepsilon-1)\theta}}{\mathrm{e}} \ge \frac{n^{\varepsilon\theta-1}}{3} \quad \text{for all } i \in [N],$$

provided n is sufficiently large. Therefore, recalling $N = \lceil n^{1-\xi}\rceil$ we obtain for large enough n,

$$\mathbb{E}\bigl|\{y^{(1)},\ldots,y^{(N)}\} \cap V_{0+}(G^*,\boldsymbol{\chi})\bigr| \ge n^{\varepsilon\theta-\xi}/3, \qquad \mathbb{E}\bigl|\{y^{(1)},\ldots,y^{(N)}\} \cap V_{1+}(G^*,\boldsymbol{\chi})\bigr| \ge n^{\varepsilon\theta-\xi}/3. \tag{3.22}$$

Further, because the pairwise distances of $y^{(1)},\dots,y^{(N)}$ in G^* exceed four, the events $\{y^{(i)} \in V_{0+}(G^*,\boldsymbol{\chi})\}_{i\le N}$ are mutually independent. So are the events $\{y^{(i)} \in V_{1+}(G^*,\boldsymbol{\chi})\}_{i\le N}$. Finally, since (3.3) ensures that $\varepsilon\theta-\xi>0$, (3.22) and the Chernoff bound yield

$$\mathbb{P}\left[\left|\{y^{(1)},\dots,y^{(N)}\}\cap V_{0+}(G^*,\boldsymbol{\chi})\right|\le\ln^2 n\right]\le\mathbb{P}\left[\mathrm{Bin}(N,n^{\varepsilon\theta-1}/3)\le\ln^2 n\right]\le\exp(-n^{\Omega(1)}),$$
$$\mathbb{P}\left[\left|\{y^{(1)},\dots,y^{(N)}\}\cap V_{1+}(G^*,\boldsymbol{\chi})\right|\le\ln^2 n\right]\le\mathbb{P}\left[\mathrm{Bin}(N,n^{\varepsilon\theta-1}/3)\le\ln^2 n\right]\le\exp(-n^{\Omega(1)}),$$

whence the assertion is immediate. □

Proof of Proposition 3.4. Suppose that $n>n_0(\varepsilon,\theta,\xi)$ is large enough and let $G=G_{n,m}$ be a test design with $m\le(1-\varepsilon)m_{\inf}$ tests. If for every test $a\in F_m$ of degree $|\partial_G a|>\Gamma$ we have $|\partial_G a\cap V_1(G,\boldsymbol{\chi})|\ge 2$, then $V_{0+}(G,\boldsymbol{\chi})=V_{0+}(G^*,\boldsymbol{\chi})$ and $V_{1+}(G,\boldsymbol{\chi})=V_{1+}(G^*,\boldsymbol{\chi})$. Therefore, the assertion is an immediate consequence of Lemma 3.6, Lemma 3.7 and Corollary 3.13. □

3.3. **Proof of Proposition 3.5.** Given $\varepsilon>0$ and $\ln(2)/(1+\ln(2))\le\theta<\theta'<1$ we choose a large enough $n_0=n_0(\varepsilon,\theta,\theta')$ and assume that $n>n_0$. Furthermore, let G be a test design with $m\le(1-\varepsilon)m_{\inf}(n,\theta)$ for the purpose of identifying $k=\lceil n^\theta\rceil$ infected individuals. Starting from the test design G infection for density θ we are going to construct a random test design $\boldsymbol{G}'$ for infection density θ' with the same number m of tests as G. The following lemma fixes the order of $\boldsymbol{G}'$.

Lemma 3.14. *There exists an integer $n^{\theta/\theta'}/2\le n'\le 2n^{\theta/\theta'}\wedge n$ such that $k'=\lceil n'^{\theta'}\rceil=k$.*

Proof. Let $n''=\lceil n^{\theta/\theta'}/2\rceil$. Then $(4n'')^{\theta'}>k$ but $n''^{\theta'}<k$ because the function $z\in(1,\infty)\mapsto z^{\theta'}$ has derivative less than one. For the same reason for any integer $n''<N<4n''$ we have $(N+1)^{\theta'}-N^{\theta'}\le 1$ and thus

$$\lceil(N+1)^{\theta'}\rceil-\lceil N^{\theta'}\rceil\le 1.$$

Consequently, there exists an integer $n'\in(n'',4n'')$ such that $\lceil n'^{\theta'}\rceil=k$. □

Given the test design G with individuals $V_n=\{x_1,\dots,x_n\}$ and tests $F_m=\{a_1,\dots,a_m\}$ we now construct the test design $\boldsymbol{G}'$ as follows. Choose a subset $V(\boldsymbol{G}')\subset V_n$ of n' individuals uniformly at random. Then $\boldsymbol{G}'$ is the subgraph that G induces on $V(\boldsymbol{G}')\cup F_m$. Thus, $\boldsymbol{G}'$ has the same tests as G but we simply leave out from every test the individuals that do not belong to the random subset $V(\boldsymbol{G}')$. Let $\boldsymbol{\tau}\in\{0,1\}^{V(\boldsymbol{G}')}$ be a random vector of Hamming weight k and let $\hat{\boldsymbol{\tau}}\in\{0,1\}^{F_m}$ be the induced vector of tests results

$$\hat{\boldsymbol{\tau}}_a=\max_{x\in\partial_{\boldsymbol{G}'}a}\boldsymbol{\tau}_x\qquad\qquad(a\in F_m).$$

Lemma 3.15. *For any integer $t>0$ we have $\mathbb{P}[Z_k(G,\hat{\boldsymbol{\sigma}})\ge t]\ge\mathbb{P}\left[Z_k(\boldsymbol{G}',\hat{\boldsymbol{\tau}})\ge t\right]$.*

Proof. The choice of n' ensures that $k'=\lceil n'^{\theta'}\rceil=k$. Therefore, the random sets $\{x\in V:\boldsymbol{\sigma}_x=1\}$ and $\{x\in V(\boldsymbol{G}'):\boldsymbol{\tau}_x=1\}$ are identically distributed. Indeed, we obtain the latter by first choosing the random subset $V(\boldsymbol{G}')$ of V_n and then choosing a random subset of $V(\boldsymbol{G}')$ size k. Clearly, this two-step procedure is equivalent to just choosing a random subset of size k out of V_n. Hence, we can couple $\boldsymbol{\sigma},\boldsymbol{\tau}$ such that the sets $\{x\in V:\boldsymbol{\sigma}_x=1\}$, $\{x\in V:\boldsymbol{\tau}_x=1\}$ are identical. Then the construction of $\boldsymbol{G}'$ ensures that the vectors $\hat{\boldsymbol{\sigma}},\hat{\boldsymbol{\tau}}$ coincide as well.

Now consider a vector $\sigma'\in\mathscr{S}_k(\boldsymbol{G}',\hat{\boldsymbol{\tau}})$ that explains the test results. Extend σ' to a vector $\sigma\in\{0,1\}^{V_n}$ by setting $\sigma_x=0$ for all $x\in V_n\setminus V(\boldsymbol{G}')$. Then $\sigma\in\mathscr{S}_k(\boldsymbol{G},\hat{\boldsymbol{\sigma}})$. Hence, $Z_k(G,\hat{\boldsymbol{\sigma}})\ge Z_k(\boldsymbol{G}',\hat{\boldsymbol{\tau}})$. □

Proof of Proposition 3.5. Lemma 3.15 shows that for any $t>0$,

$$\mathbb{P}[Z_k(G,\hat{\boldsymbol{\sigma}})\ge t]\ge\mathbb{P}\left[Z_k(\boldsymbol{G}',\hat{\boldsymbol{\tau}})\ge t\right]=\mathbb{E}\left[\mathbb{P}\left[Z_k(\boldsymbol{G}',\hat{\boldsymbol{\tau}})\ge t\mid\boldsymbol{G}'\right]\right].$$

Consequently, there exists an outcome G' of $\boldsymbol{G}'$ such that $\mathbb{P}[Z_k(G,\hat{\boldsymbol{\sigma}})\ge t]\ge\mathbb{P}\left[Z_k(G',\hat{\boldsymbol{\tau}})\ge t\right]$. □

4. The non-adaptive group testing algorithm SPIV

In this section we describe the new test design and the associated inference algorithm SPIV for Theorem 1.2. Throughout we fix $\theta\in(0,1)$ and $\varepsilon>0$ and we tacitly assume that $n>n_0(\varepsilon,\theta)$ is large enough for the various estimates to hold.

4.1. **The random bipartite graph and the DD algorithm.** To motivate the new test design we begin with a brief discussion of the plain random design used in prior work and the best previously known inference algorithm DD [11, 22]. At first glance a promising candidate test design appears to be a random bipartite graph with one vertex class $V_n = \{x_1, \ldots, x_n\}$ representing individuals and the other class $F_m = \{a_1, \ldots, a_m\}$ representing tests. Indeed, two slightly different random graph models have been proposed [6]. First, in the *Bernoulli model* each V_n–F_m-edge is present with a certain probability (the same for every pair) independently of all others. However, due to the relatively heavy lower tail of the degrees of the individuals, this test design turns out to be inferior to a second model where the degrees of the individuals are fixed. Specifically, in the Δ-*out model* every individual independently joins an equal number of Δ tests drawn uniformly at random without replacement [29].

Clearly, in order to extract the maximum amount of information Δ should be chosen so as to maximise the entropy of the vector of test results. Specifically, since the average test degree equals $\Delta n/m$ and a total of k individuals are infected, the average number of infected individuals per test comes to $\Delta k/m$. Indeed, since $k \sim n^\theta$ for a fixed $\theta < 1$, the number of infected individuals in test a_i can be well approximated by a Poisson variable. Therefore, setting

$$\Delta \sim \frac{m}{k} \ln 2 \tag{4.1}$$

ensures that about half the tests are positive w.h.p.

With respect to the performance of the Δ-out model, [11, Theorem 1.1] implies together with Theorem 1.1 that this simple construction is information-theoretically optimal. Indeed, $m = (1+\varepsilon+o(1))m_{\mathrm{inf}}$ test suffice so that an exponential time algorithm correctly infers the set of infected individuals. Specifically, the algorithm solves a minimum hypergraph vertex cover problem with the individuals as the vertex set and the positive test groups as the hyperedges. For $m = (1+\varepsilon+o(1))m_{\mathrm{inf}}$ the unique optimal solution is precisely the correct set of infected individuals w.h.p. While the worst case NP-hardness of hypergraph vertex cover does not, of course, preclude the existence of an algorithm that is efficient on random hypergraphs, despite considerable efforts no such algorithm has been found. In fact, as we saw in Section 1.4 for a good number of broadly similar inference and optimisation problems on random graphs no efficient information-theoretically optimal algorithms are known.

But for m exceeding the threshold $m_{\mathtt{DD}}$ from (1.2) an efficient greedy algorithm DD correctly recovers $\boldsymbol{\sigma}$ w.h.p. The algorithm proceeds in three steps.

DD1: declare every individual that appears in a negative test uninfected and subsequently remove all negative tests and all individuals that they contain.
DD2: for every remaining (positive) test of degree one declare the individual that appears in the test infected.
DD3: declare all other individuals as uninfected.

The decisions made by the first two steps **DD1**–**DD2** are clearly correct but **DD3** might produce false negatives. Prior to the present work DD was the best known polynomial time group testing algorithm. While DD correctly identifies the set of infected individuals w.h.p. if $m > (1+\varepsilon)m_{\mathtt{DD}}$ [22], the algorithm fails if $m < (1-\varepsilon)m_{\mathtt{DD}}$ w.h.p. [11].

4.2. **Spatial coupling.** The new efficient algorithm SPIV for Theorem 1.2 that gets by with the optimal number $(1+\varepsilon+o(1))m_{\mathrm{inf}}$ of tests comes with a tailor-made test design that, inspired by spatially coupled codes [18, 26, 27], combines randomisation with a superimposed geometric structure. Specifically, we divide both the individuals and the tests into

$$\ell = \lceil \ln^{1/2} n \rceil \tag{4.2}$$

compartments of equal size. The compartments are arranged along a ring and each individual joins an equal number of random tests in the

$$s = \lceil \ln \ln n \rceil = o(\ell) \tag{4.3}$$

topologically subsequent compartments. Additionally, to get the algorithm started we equip the first s compartments with extra tests so that they can be easily diagnosed via the DD algorithm. Then, having diagnosed the initial compartments correctly, SPIV will work its way along the ring, diagnosing one compartment after the other.

To implement this idea precisely we partition the set $V = V_n = \{x_1, \ldots, x_n\}$ of individuals into pairwise disjoint subsets $V[1], \ldots, V[\ell]$ of sizes $|V[j]| \in \{\lfloor n/\ell \rfloor, \lceil n/\ell \rceil\}$. With each compartment $V[i]$ of individuals we associate a compartment $F[i]$ of tests of size $|F[i]| = m/\ell$ for an integer m that is divisible by ℓ. Additionally, we introduce

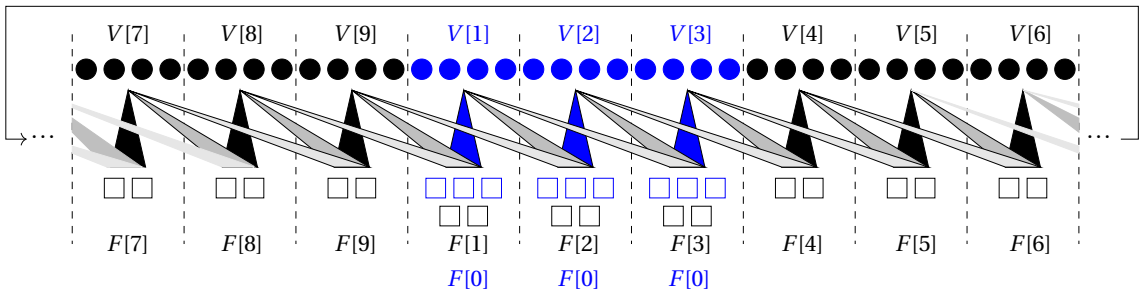

FIGURE 2. The spatially coupled test design with $n = 36, \ell = 9, s = 3$. The individuals in the seed groups $V[1] \cup \cdots \cup V[s]$ (blue) are equipped with additional test $F[0]$ (blue rectangles). The black rectangles represent the tests $F[1] \cup \cdots \cup F[\ell]$.

a set $F[0]$ of $10\lceil (ks/\ell)\ln n\rceil$ extra tests to facilitate the greedy algorithm for diagnosing the first s compartments. Thus, the total number of tests comes to

$$|F[0]| + \sum_{i=1}^{\ell} |F[i]| = (1 + O(s/\ell))m = (1 + o(1))m. \tag{4.4}$$

Finally, for notational convenience we define $V[\ell + i] = V[i]$ and $F[\ell + i] = F[i]$ for $i = 1, \ldots, s$.

The test groups are composed as follows: let

$$k = \lceil n^{\theta} \rceil \qquad \text{and let} \qquad \Delta = \frac{m \ln 2}{k} + O(s) \tag{4.5}$$

be an integer divisible by s; cf. (4.1). Then we construct a random bipartite graph as follows.

SC1: for $i = 1, \ldots, \ell$ and $j = 1, \ldots, s$ every individual $x \in V[i]$ joins Δ/s tests from $F[i + j - 1]$ chosen uniformly at random without replacement. The choices are mutually independent for all individuals x and all j.

SC2: additionally, each individual from $V[1] \cup \cdots \cup V[s]$ independently joins $\lceil 10 \ln(2) \ln n \rceil$ random tests from $F[0]$, drawn uniformly without replacement.

Thus, **SC1** provides that the individuals in compartment $V[i]$ take part in the next s compartments $F[i], \ldots, F[i + s - 1]$ of tests along the ring. Furthermore, **SC2** supplies the tests required by the `DD` algorithm to diagnose the first s compartments. Figure 2 provides an illustration of the resulting random test design,

From here on the test design produced by **SC1–SC2** is denoted by $\boldsymbol{G}$. Furthermore $\boldsymbol{\sigma} \in \{0,1\}^V$ denotes a uniformly random vector of Hamming weight k, drawn independently of $\boldsymbol{G}$, and $\hat{\boldsymbol{\sigma}} = (\hat{\boldsymbol{\sigma}}_a)_{a \in F[0] \cup \cdots \cup F[\ell]}$ signifies the vector of test results

$$\hat{\boldsymbol{\sigma}}_a = \max_{x \in \partial a} \boldsymbol{\sigma}_x.$$

In addition, let $V_1 = \{x \in V : \boldsymbol{\sigma}_x = 1\}$ be the set of infected individuals and let $V_0 = V \setminus V_1$ be the set of healthy individuals. Moreover, let $F = F[0] \cup F[1] \cup \cdots \cup F[\ell]$ be the set of all tests, let $F_1 = \{a \in F : \hat{\boldsymbol{\sigma}}_a = 1\}$ be the set of all positive tests and let $F_0 = F \setminus F_1$ be the set of all negative tests. Finally, let

$$V_0[i] = V[i] \cap V_0, \qquad V_1[i] = V[i] \cap V_1, \qquad F_0[i] = F[i] \cap F_0, \qquad F_1[i] = F[i] \cap F_1.$$

The following proposition summarises a few basic properties of the test design $\boldsymbol{G}$.

Proposition 4.1. *If $m = \Theta(n^{\theta} \ln n)$ then $\boldsymbol{G}$ enjoys the following properties with probability $1 - o(n^{-2})$.*

(i) The infected individual counts in the various compartments satisfy

$$\frac{k}{\ell} - \sqrt{\frac{k}{\ell}} \ln n \le \min_{i \in [\ell]} |V_1[i]| \le \max_{i \in [\ell]} |V_1[i]| \le \frac{k}{\ell} + \sqrt{\frac{k}{\ell}} \ln n.$$

(ii) For all $i \in [\ell]$ and all $j \in [s]$ the test degrees satisfy

$$\frac{\Delta n}{ms} - \sqrt{\frac{\Delta n}{ms}} \ln n \le \min_{a \in F[i+j-1]} |V[i] \cap \partial a| \le \max_{a \in F[i+j-1]} |V[i] \cap \partial a| \le \frac{\Delta n}{ms} + \sqrt{\frac{\Delta n}{ms}} \ln n.$$

(iii) For all $i \in [\ell]$ the number of negative tests in compartment $F[i]$ satisfies

$$\frac{m}{2\ell} - \sqrt{m}\ln^3 n \le |F_0[i]| \le \frac{m}{2\ell} + \sqrt{m}\ln^3 n.$$

We prove Proposition 4.1 in Section 4.4. Finally, as a preparation for things to come we point out that for any specific individual $x \in V[i]$ and any particular test $a \in F[i+j]$, $j = 0,\dots,s-1$, we have

$$\mathbb{P}[x \in \partial a] = 1 - \mathbb{P}[x \notin \partial a] = 1 - \binom{|F[i+j]|-1}{\Delta/s}\binom{|F[i+j]|}{\Delta/s}^{-1} = \frac{\Delta\ell}{ms} + O\left(\left(\frac{\Delta\ell}{ms}\right)^2\right). \tag{4.6}$$

4.3. **The Spatial Inference Vertex Cover ('SPIV') algorithm.** The SPIV algorithm for Theorem 1.2 proceeds in three phases. The plan of attack is for the algorithm to work its way along the ring, diagnosing one compartment after the other aided by what has been learned about the preceding compartments. Of course, we need to start somewhere. Hence, in its first phase SPIV diagnoses the seed compartments $V[1],\dots,V[s]$.

4.3.1. *Phase 1: the seed.* Specifically, the first phase of SPIV applies the DD greedy algorithm from Section 4.1 to the subgraph of $\boldsymbol{G}$ induced on the individuals $V[1]\cup\ldots\cup V[s]$ and the tests $F[0]$. Throughout the vector $\tau \in \{0,1\}^V$ signifies the algorithm's current estimate of the ground truth $\boldsymbol{\sigma}$.

Input: $\boldsymbol{G}$, $\hat{\boldsymbol{\sigma}}$
Output: an estimate of $\boldsymbol{\sigma}$
1 Let $(\tau_x)_{x\in V[1]\cup\cdots\cup V[s]} \in \{0,1\}^{V[1]\cup\cdots\cup V[s]}$ be the result of applying DD to the tests $F[0]$;
2 Set $\tau_x = 0$ for all individuals $x \in V \setminus (V[1]\cup\cdots\cup V[s])$;

Algorithm 1: SPIV, phase 1

The following proposition, whose proof can be found in Section 4.5, summarises the analysis of phase 1.

Proposition 4.2. *W.h.p. the output of* DD *satisfies $\tau_x = \boldsymbol{\sigma}_x$ for all $x \in V[1]\cup\cdots\cup V[s]$.*

4.3.2. *Phase 2: enter the ring.* This is the main phase of the algorithm. Thanks to Proposition 4.2 we may assume that the seed has been diagnosed correctly. Now, the programme is to diagnose one compartment after the other, based on what the algorithm learned previously. Hence, assume that we managed to diagnose compartments $V[1],\dots,V[i]$ correctly. How do we proceed to compartment $V[i+1]$?

For a start, we can safely mark as uninfected all individuals in $V[i+1]$ that appear in a negative test. But a simple calculation reveals that this will still leave us with many more than k undiagnosed individuals w.h.p. To be precise, consider the set of uninfected disguised individuals

$$V_{0+}[i+1] = \{x \in V_0[i+1] : \hat{\boldsymbol{\sigma}}_a = 1 \text{ for all } a \in \partial x\},$$

i.e., uninfected individuals that fail to appear in a negative test. In Section 4.6 we prove the following.

Lemma 4.3. *Suppose that $(1+\varepsilon)m_{\mathrm{ad}} \le m = O(n^\theta \ln n)$. Then w.h.p. for all $s \le i < \ell$ we have*

$$|V_{0+}[i+1]| = \left(1 + O\left(n^{-\Omega(1)}\right)\right)\frac{n}{\ell 2^\Delta}.$$

Hence, by the definition (4.5) of Δ for m close to $m_{\inf}$ the set $V_{0+}[i+1]$ has size $k^{1+\Omega(1)} \gg k$ w.h.p.

Thus, the challenge is to discriminate between $V_{0+}[i+1]$ and the set $V_1[i+1]$ of actual infected individuals in compartment $i+1$. The key observation is that we can tell these sets apart by counting currently 'unexplained' positive tests. To be precise, for an individual $x \in V[i+1]$ and $1 \le j \le s$ let $\boldsymbol{W}_{x,j}$ be the number of tests in compartment $F[i+j]$ that contain x but that do not contain an infected individual from the preceding compartments $V[1]\cup\cdots\cup V[i]$. In formulas,

$$\boldsymbol{W}_{x,j} = \left|\left\{a \in \partial x \cap F[i+j] : \partial a \cap (V_1[1]\cup\cdots\cup V_1[i]) = \emptyset\right\}\right|. \tag{4.7}$$

Crucially, the following back-of-the-envelope calculation shows that the mean of this random variable depends on whether x is infected or healthy but disguised.

Infected individuals ($x \in V_1[i+1]$): consider a test $a \in \partial x \cap F[i+j]$, $j = 1,\dots,s$. Because the individuals join tests independently, conditioning on x being infected does not skew the distribution of the individuals from the $s-j$ prior compartments $V[i+j-s+1],\dots,V[i]$ that appear in a. Furthermore, we chose Δ so that for each of these compartments $V[h]$ the expected number of infected individuals that join a has mean $(\ln 2)/s$. Indeed, due to independence it is not difficult to see that $|V_1[h] \cap \partial a|$ is approximately a Poisson variable. Consequently,

$$\mathbb{P}\left[(V_1[i+j-s+1] \cup \dots \cup V_1[i]) \cap \partial a = \emptyset\right] \sim 2^{-(s-j)/s}. \tag{4.8}$$

Hence, because x appears in Δ/s tests $a \in F[i+j]$, the linearity of expectation yields

$$\mathbb{E}\left[\boldsymbol{W}_{x,j} \mid x \in V_1[i+1]\right] \sim 2^{j/s-1}\frac{\Delta}{s}. \tag{4.9}$$

Disguised healthy individuals ($x \in V_{0+}[i+1]$): similarly as above, for any individual $x \in V[i+1]$ and any $a \in \partial x \cap F[i+j]$ the *unconditional* number of infected individuals in a is asymptotically $\mathrm{Po}(\ln 2)$. But given $x \in V_{0+}[i+1]$ we know that a is positive. Thus, $\partial a \setminus \{x\}$ contains at least one infected individual. In effect, the number of positives in a approximately turns into a conditional Poisson $\mathrm{Po}_{\geq 1}(\ln 2)$. Consequently, for test a not to include any infected individual from one of the known compartments $V[h]$, $h = i+j-s+1,\dots,i$, every infected individual in test a must stem from the j yet undiagnosed compartments. Summing up the conditional Poisson and recalling that x appears in Δ/s tests $a \in F[j]$, we thus obtain

$$\mathbb{E}\left[\boldsymbol{W}_{x,j} \mid x \in V_{0+}[i+1]\right] \sim \frac{\Delta}{s}\sum_{t\geq 1}\mathbb{P}\left[\mathrm{Po}_{\geq 1}(\ln 2) = t\right](j/s)^t = (2^{j/s}-1)\frac{\Delta}{s}. \tag{4.10}$$

A first idea to tell $V_{0+}[i+1]$ and $V_1[i+1]$ apart might thus be to simply calculate

$$\boldsymbol{W}_x = \sum_{j=1}^{s-1}\boldsymbol{W}_{x,j} \qquad (x \in V[i+1]). \tag{4.11}$$

Indeed, (4.9) and (4.10) yield

$$\mathbb{E}\left[\boldsymbol{W}_x \mid x \in V_1[i+1]\right] \sim \frac{\Delta}{2\ln 2} = 0.721\dots\Delta \qquad \text{whereas} \qquad \mathbb{E}\left[\boldsymbol{W}_x \mid x \in V_{0+}[i+1]\right] \sim \frac{\Delta(1-\ln 2)}{\ln 2} = 0.442\dots\Delta.$$

But unfortunately a careful large deviations analysis reveals that $\boldsymbol{W}_x$ is not sufficiently concentrated. More precisely, even for $m = (1+\varepsilon+o(1))m_{\inf}$ there are as many as $k^{1+\Omega(1)}$ 'outliers' $x \in V_{0+}[i+1]$ whose $\boldsymbol{W}_x$ grows as large as the mean $\Delta/(2\ln 2)$ of actual infected individuals w.h.p.

At second thought the plain sum (4.11) does seem to leave something on the table. While $\boldsymbol{W}_x$ counts all as yet unexplained positive tests equally, not all of these tests reveal the same amount of information. In fact, we should really be paying more attention to 'early' unexplained tests $a \in F[i+1]$ than to 'late' ones $b \in F[i+s]$. For we already diagnosed $s-1$ out of the s compartments of individuals that a draws on, whereas only one of the s compartments that contribute to b has already been diagnosed. Thus, the unexplained test a is a much stronger indication that x might be infected. Consequently, it seems promising to replace $\boldsymbol{W}_x$ by a weighted sum

$$\boldsymbol{W}^\star_x = \sum_{j=1}^{s-1} w_j \boldsymbol{W}_{x,j} \tag{4.12}$$

with $w_1,\dots,w_{s-1} \geq 0$ chosen so as to gauge the amount of information carried by the different compartments.

To find the optimal weights $w_1,\dots,w_{s-1}$ we need to investigate the rate function of $\boldsymbol{W}^\star_x$ given $x \in V_{0+}[i+1]$. More specifically, we should minimise the probability that $\boldsymbol{W}^\star_x$ given $x \in V_{0+}[i+1]$ grows as large as the mean of $\boldsymbol{W}^\star_x$ given $x \in V_1[i+1]$, which we read off (4.9) easily:

$$\mathbb{E}\left[\boldsymbol{W}^\star_x \mid x \in V_1[i+1]\right] \sim \frac{\Delta}{s}\sum_{j=1}^{s-1} 2^{j/s-1} w_j. \tag{4.13}$$

A careful large deviations analysis followed by a Lagrangian optimisation leads to the optimal choice

$$w_j = \ln\frac{(1-2\zeta)2^{j/s-1}(2-2^{j/s})}{(1-(1-2\zeta)2^{j/s-1})(2^{j/s}-1)} \qquad \text{where} \qquad \zeta = 1/s^2. \tag{4.14}$$

The following two lemmas show that with these weights the scores $\boldsymbol{W}_x^\star$ discriminate well between the potential false positives and the infected individuals. More precisely, thresholding $\boldsymbol{W}_x^\star$ we end up misclassifying no more than $o(k)$ individuals x w.h.p.

Lemma 4.4. *Suppose that* $(1+\varepsilon)m_{\mathrm{ad}} \le m = O(n^\theta \ln n)$. *W.h.p. we have*

$$\sum_{s\le i<\ell}\sum_{x\in V_1[i]} \mathbf{1}\left\{\boldsymbol{W}_x^\star < (1-\zeta/2)\frac{\Delta}{s}\sum_{j=1}^{s-1} 2^{j/s-1}w_j\right\} \le k\exp\left(-\frac{\Omega(\ln n)}{(\ln\ln n)^4}\right). \tag{4.15}$$

Lemma 4.5. *Suppose that* $(1+\varepsilon)m_{\mathrm{ad}} \le m = O(n^\theta \ln n)$. *W.h.p. we have*

$$\sum_{s\le i<\ell}\sum_{x\in V_{0+}[i]} \mathbf{1}\left\{\boldsymbol{W}_x^\star > (1-2\zeta)\frac{\Delta}{s}\sum_{j=1}^{s-1} 2^{j/s-1}w_j\right\} \le k^{1-\Omega(1)}. \tag{4.16}$$

We prove these two lemmas in Sections 4.7 and 4.8.

Lemmas 4.4–4.5 leave us with only one loose end. Namely, calculating the scores $\boldsymbol{W}_x^\star$ requires knowledge of the correct infection status $\boldsymbol{\sigma}_x$ of *all* the individuals $x \in V[1]\cup\cdots\cup V[i]$ from the previous compartments. But since the r.h.s. expressions in (4.15) and (4.16) are non-zero, it is unrealistic to assume that the algorithm's estimates τ_x will consistently match the ground truth $\boldsymbol{\sigma}_x$ beyond the seed compartments. Hoping that the algorithm's estimate will not stray too far, we thus have to make do with the approximate scores

$$W_x^\star(\tau) = \sum_{j=1}^{s-1} w_j W_{x,j}(\tau), \qquad \text{where} \qquad W_{x,j}(\tau) = \left|\left\{a\in\partial x\cap F[i+j-1] : \max_{y\in\partial a\cap(V[1]\cup\cdots V[i])}\tau_y = 0\right\}\right|. \tag{4.17}$$

Hence, phase 2 of SPIV reads as follows.

```
for i = s, ..., ℓ − 1 do
    for x ∈ V[i + 1] do
        if ∃a ∈ ∂x : σ̂_a = 0 then
            τ_x = 0 // classify as uninfected
        else if W*_x(τ) < (1 − ζ)(Δ/s) Σ_{j=1}^{s−1} 2^{j/s−1} w_j then
            τ_x = 0 // tentatively classify as uninfected
        else
            τ_x = 1 // tentatively classify as infected
```

Algorithm 2: SPIV, phase 2.

Since phase 2 of SPIV uses the approximations from (4.17), there seems to be a risk of errors amplifying as we move along. Fortunately, it turns out that errors proliferate only moderately and the second phase of SPIV will misclassify only $o(k)$ individuals. The following proposition summarises the analysis of phase 2.

Proposition 4.6. *Suppose that* $(1+\varepsilon)m_{\mathrm{ad}} \le m = O(k\ln n)$. *W.h.p. the assignment* τ *obtained after steps 1–10 satisfies*

$$\sum_{x\in V}\mathbf{1}\{\tau_x \ne \boldsymbol{\sigma}_x\} \le k\exp\left(-\frac{\ln n}{(\ln\ln n)^6}\right).$$

The proof of Proposition 4.6 can be found in Section 4.9.

4.3.3. *Phase 3: cleaning up.* The final phase of the algorithm rectifies the errors incurred during phase 2. The combinatorial insight that makes this possible is that for $m \ge (1+\varepsilon)m_{\mathrm{inf}}$ every infected individual has at least $\Omega(\Delta)$ positive tests to itself w.h.p. Thus, these tests do not feature a second infected individual. Phase 3 of the algorithm exploits this observation by simply thresholding the number S_x of tests where there is no other infected individual besides potentially x. Thanks to the expansion properties of the graph $\boldsymbol{G}$, each iteration of the thresholding procedure reduces the number of misclassified individuals by at least a factor of three. In effect, after $\ln n$ iterations all individuals will be classified correctly w.h.p. Of course, due to Proposition 4.2 we do not need to reconsider the seed $V[1]\cup\cdots\cup V[s]$.

Let $\tau^{(1)} = \tau$;
for $i = 1, \dots, \lceil \ln n \rceil$ **do**
  For all $x \in V[s+1] \cup \dots \cup V[\ell]$ calculate
    $S_x(\tau^{(i)}) = \sum_{a \in \partial x : \hat{\boldsymbol{\sigma}}_a = 1} \mathbf{1}\left\{\forall y \in \partial a \setminus \{x\} : \tau_y^{(i)} = 0\right\}$;
  Let $\tau_x^{(i+1)} = \begin{cases} \tau_x^{(i)} & \text{if } x \in V[1] \cup \dots \cup V[s], \\ \mathbf{1}\left\{S_x\left(\tau^{(i)}\right) > \ln^{1/4} n\right\} & \text{otherwise} \end{cases}$;
return $\tau^{(\lceil \ln n \rceil)}$

Algorithm 3: `SPIV`, phase 3.

Proposition 4.7. *Suppose that* $(1+\varepsilon)m_{\inf} \le m = O(n^\theta \ln n)$. *W.h.p. for all* $1 \le i \le \lceil \ln n \rceil$ *we have*

$$\sum_{x \in V} \mathbf{1}\{\tau_x^{(i+1)} \neq \boldsymbol{\sigma}_x\} \le \frac{1}{3} \sum_{x \in V} \mathbf{1}\{\tau_x^{(i)} \neq \boldsymbol{\sigma}_x\}.$$

We prove Proposition 4.7 in Section 4.10.

Proof of Theorem 1.2. The theorem is an immediate consequence of Propositions 4.2, 4.6 and 4.7. □

4.4. **Proof of Proposition 4.1.** The number $|V_1[i]|$ of infected individuals in compartment $V[i]$ has distribution $\mathrm{Hyp}(n, k, |V[i]|)$. Since $||V[i]| - n/\ell| \le 1$, (i) is an immediate consequence of the Chernoff bound from Lemma 2.2.

With respect to (ii), we recall from (4.6) that $\mathbb{P}[x \in \partial a] = \frac{\Delta\ell}{ms}(1 + O(\frac{\Delta\ell}{ms}))$. Hence, because the various individuals $x \in V[i]$ join tests independently, the number $|V[i] \cap \partial a|$ of test participants from $V[i]$ has distribution

$$|V[i] \cap \partial a| \sim \mathrm{Bin}(|V[i]|, \Delta\ell/(ms) + O((\Delta\ell/ms)^2)).$$

Since $|V[i]| = n/\ell + O(1)$, assertion (ii) follows from (4.5) and the Chernoff bound from Lemma 2.1.

Coming to (iii), due to part (i) we may condition on $\mathscr{E} = \{\forall i \in [\ell] : |V_1[i]| = k/\ell + O(\sqrt{k/\ell}\ln n)\}$. Hence, with h ranging over the s compartments whose individuals join tests in $F[i]$, (4.6) implies that for every test $a \in F[i]$ the number of infected individuals $|V_1 \cap \partial a|$ is distributed as a sum of independent binomial variables

$$|V_1 \cap \partial a| \sim \sum_h \boldsymbol{X}_h \qquad \text{with} \qquad \boldsymbol{X}_h \sim \mathrm{Bin}\left(V_1[h], \frac{\Delta\ell}{ms} + O\left(\left(\frac{\Delta\ell}{ms}\right)^2\right)\right).$$

Consequently, (4.5) ensures that the event $V_1 \cap \partial a = \emptyset$ has conditional probability

$$\begin{aligned}\mathbb{P}[V_1 \cap \partial a = \emptyset \mid \mathscr{E}] &= \prod_h \mathbb{P}[\boldsymbol{X}_h = 0 \mid \mathscr{E}] = \exp\left[s\left(\frac{k}{\ell} + O\left(\sqrt{\frac{k}{\ell}}\ln n\right)\right)\ln\left(1 - \frac{\Delta\ell}{ms} + O\left(\left(\frac{\Delta\ell}{ms}\right)^2\right)\right)\right] \\ &= \exp\left[-\frac{sk}{\ell}\cdot\frac{\Delta\ell}{ms} + O\left(\sqrt{\frac{k}{\ell}}\cdot\frac{\Delta\ell}{m}\right) + O\left(\frac{sk}{\ell}\cdot\left(\frac{\Delta\ell}{ms}\right)^2\right)\right] = \frac{1}{2} + O(\sqrt{\ell/k}).\end{aligned}$$

Therefore, we obtain the estimate

$$\mathbb{E}[|F_0[i]| \mid \mathscr{E}] = \frac{m}{2\ell} + O(\sqrt{m}\ln n). \tag{4.18}$$

Finally, changing the set of tests that a specific infected individual $x \in V_1[h]$ joins shifts $|F_0[i]|$ by at most Δ (while tinkering with uninfected ones does not change $|F_0[i]|$ at all). Therefore, the Azuma–Hoeffding inequality yields

$$\mathbb{P}[||F_0[i]| - \mathbb{E}[|F_0[i]| \mid \mathscr{E}]| \ge t \mid \mathscr{E}] \le 2\exp\left(-\frac{t^2}{2k\Delta^2}\right) \qquad \text{for any } t > 0. \tag{4.19}$$

Thus, (iii) follows from (4.5), (4.18) and (4.19) with $t = \sqrt{m}\ln^3 n$.

4.5. **Proof of Proposition 4.2.** Let $D = \lceil 10\ln(2)\ln n\rceil$ and recall that $|F[0]| = \lceil 10ks\ln(n)/\ell\rceil$. Since by **SC2** every individual from $\in V[1]\cup\cdots\cup V[s]$ joins D random tests from $F[0]$, in analogy to (4.6) for every $x \in V[1]\cup\cdots\cup V[s]$ and every test $a \in F[0]$ we obtain

$$\mathbb{P}[x \in \partial a] = 1 - \mathbb{P}[x \notin \partial a] = 1 - \binom{|F[0]|-1}{D}\binom{|F[0]|}{D}^{-1} = \frac{D}{|F[0]|}\left(1 + O\left(\frac{D}{|F[0]|}\right)\right) = \frac{\ell\ln 2}{ks}\left(1 + O(n^{-\Omega(1)})\right). \tag{4.20}$$

Let $F_1[0]$ be the set of tests $a \in F[0]$ with $\hat{\boldsymbol{\sigma}}_a = 1$.

Lemma 4.8. *W.h.p. the number of positive tests $a \in F[0]$ satisfies $|F_1[0]| = |F[0]|(\frac{1}{2} + O(n^{-\Omega(1)}))$.*

Proof. By Proposition 4.1 we may condition on the event $\mathscr{E}$ that $|V_1[1]\cup\cdots\cup V_1[s]| = \frac{ks}{\ell}(1+O(n^{-\Omega(1)}))$. Hence, (4.20) implies that given $\mathscr{E}$ the expected number of infected individuals in a test $a \in F[0]$ comes to

$$\mathbb{E}[|\partial a \cap V_1| \mid \mathscr{E}] = \ln 2 + O(n^{-\Omega(1)}). \tag{4.21}$$

Moreover, since individuals join tests independently, $|\partial a \cap V_1|$ is a binomial random variable. Hence, (4.21) implies $\mathbb{P}[\partial a \cap V_1 = \emptyset \mid \mathscr{E}] = \frac{1}{2} + O(n^{-\Omega(1)})$. Consequently, since $\mathbb{P}[\mathscr{E}] = 1 - o(n^{-2})$ by Proposition 4.1,

$$\mathbb{E}|F_1 \cap F[0]| = \mathbb{E}|F_1[0]| = \frac{|F[0]|}{2}(1 + O(n^{-\Omega(1)})). \tag{4.22}$$

Finally, changing the set ∂x of neighbours of an infected individual can shift $|F_1[0]|$ by at most D. Therefore, the Azuma–Hoeffding inequality implies that

$$\mathbb{P}[||F_1[0]| - \mathbb{E}|F_1[0]|| > t] \le 2\exp\left(-\frac{t^2}{2D^2k}\right) \qquad \text{for any } t > 0. \tag{4.23}$$

Since $D = O(\ln n)$, combining (4.22) and (4.23) and setting, say, $t = k^{2/3}$ completes the proof. □

As an application of Lemma 4.8 we show that w.h.p. every seed individual x appears in a test $a \in F[0]$ whose other individuals are all healthy.

Corollary 4.9. *W.h.p. every individual $x \in V[1]\cup\cdots\cup V[s]$ appears in a test $a \in F[0]\cap\partial x$ such that $\partial a \setminus \{x\} \subset V_0$.*

Proof. We expose the random bipartite graph induced on $V[1]\cup\cdots\cup V[s]$ and $F[0]$ in two rounds. In the first round we expose $\boldsymbol{\sigma}$ and all neighbourhoods $(\partial y)_{y\in(V[1]\cup\cdots\cup V[s])\setminus\{x\}}$. In the second round we expose ∂x. Let $\boldsymbol{X}$ be the number of negative tests $a \in F[0]$ after the first round. Since x has degree $D = O(\ln n)$, Lemma 4.8 implies that $\boldsymbol{X} = |F[0]|(\frac{1}{2} + O(n^{-\Omega(1)}))$ w.h.p. Furthermore, given $\boldsymbol{X}$ the number of tests $a \in \partial x$ all of whose other individuals are uninfected has distribution $\mathrm{Hyp}(|F[0]|, \boldsymbol{X}, D)$. Hence,

$$\mathbb{P}[\forall a \in \partial x : V_1 \cap \partial a \setminus \{x\} \neq \emptyset \mid \boldsymbol{X}] = \binom{|F[0]|-\boldsymbol{X}}{D}\binom{|F[0]|}{D}^{-1} \le \exp(-D\boldsymbol{X}/|F[0]|). \tag{4.24}$$

Assuming $\boldsymbol{X}/|F[0]| = \frac{1}{2} + O(n^{-\Omega(1)})$ and recalling that $D = \lceil 10\ln(2)\ln n\rceil$, we obtain $\exp(-D\boldsymbol{X}/|F[0]|) = o(1/n)$. Thus, the assertion follows from (4.24) and the union bound. □

Proof of Proposition 4.2. Due to Corollary 4.9 we may assume that for every $x \in V[1]\cup\cdots\cup V[s]$ there is a test $a_x \in F[0]$ such that $\partial a_x \setminus \{x\} \subset V_0$. Hence, recalling the DD algorithm from Section 4.1, we see that the first step **DD1** will correctly identify all healthy individuals $x \in V_0[1]\cup\cdots\cup V_0[s]$. Moreover, the second step **DD2** will correctly classify all remaining individuals $V_1[1]\cup\cdots\cup V_1[s]$ as infected, and the last step **DD3** will be void. □

4.6. **Proof of Lemma 4.3.** Let $\mathscr{E}$ be the event that properties (i) and (iii) from Proposition 4.1 hold; then $\mathbb{P}[\mathscr{E}] = 1 - o(n^{-2})$. Moreover, let $\mathfrak{E}$ be the σ-algebra generated by $\boldsymbol{\sigma}$ and the neighbourhoods $(\partial x)_{x\in V_1}$. Then the event $\mathscr{E}$ is $\mathfrak{E}$-measurable while the neighbourhoods $(\partial x)_{x\in V_0}$ of the healthy individuals are independent of $\mathfrak{E}$. Recalling from **SC1** that the individuals $x \in V_0[i]$ choose Δ/s random tests in each of the compartments $F[i+j]$, $0 \le j \le s-1$ independently and remembering that $x \in V_{0+}[i]$ iff none of these tests is negative, on $\mathscr{E}$ we obtain

$$\begin{aligned}\mathbb{P}[x \in V_{0+}[i] \mid \mathfrak{E}] &= \binom{m/(2\ell) + O(\sqrt{m}\ln^3 n)}{\Delta/s}^s \binom{m/\ell}{\Delta/s}^{-s} = \left(\frac{1 + O(m^{-1/2}\ell\ln^3 n)}{2}\right)^\Delta \\ &= 2^{-\Delta} + O(m^{-1/2}\Delta\ell\ln^3 n) = 2^{-\Delta}(1 + O(n^{-\theta/2}\ln^4 n)) \qquad \text{[due to (4.2) and (4.5)]}.\end{aligned} \tag{4.25}$$

Because all $x \in V_0[i]$ choose their neighbourhoods independently, (4.25) implies that the conditional random variable $|V_{0+}[i]|$ given $\mathfrak{E}$ has distribution $\mathrm{Bin}(|V_0[i]|, 2^{-\Delta}(1+O(n^{-\Omega(1)})))$. Therefore, since on $\mathscr{E}$ we have $|V_0[i]| = |V[i]| + O(n^\theta) = n/\ell + O(n^\theta)$, the assertion follows from the Chernoff bound from Lemma 2.1.

4.7. **Proof of Lemma 4.4.** The aim is to estimate the weighted sum $\boldsymbol{W}^\star_x$ for infected individuals $x \in V[i+1]$ with $s \le i < \ell$. These individuals join tests in the s compartments $F[i+j]$, $j \in [s]$. Conversely, for each such j the tests $a \in F[i+j]$ recruit their individuals from the compartments $V[i+j-s+1], \ldots, V[i+j]$. Thus, the compartments preceding $V[i+1]$ that the tests in $F[i+j]$ draw upon are $V[h]$ with $i+j-s < h \le i$. We begin by investigating the set $\mathscr{W}_{i,j}$ of tests $a \in F[i+j]$ without an infected individual from these compartments, i.e.,

$$\mathscr{W}_{i,j} = \big\{a \in F[i+j] : (V_1[1] \cup \cdots \cup V_1[i]) \cap \partial a = \emptyset\big\} = \left\{a \in F[i+j] : \bigcup_{i+j-s+1<h\le i} V_1[h] \cap \partial a = \emptyset\right\}.$$

Claim 4.10. *With probability* $1-o(n^{-2})$ *for all* $s \le i < \ell$, $j \in [s]$ *we have* $|\mathscr{W}_{i,j}| = 2^{-(s-j)/s}\frac{m}{\ell}(1+O(n^{-\Omega(1)}))$.

Proof. We may condition on the event $\mathscr{E}$ that (i) from Proposition 4.1 occurs. To compute the mean of $|\mathscr{W}_{i,j}|$ fix a test $a \in F[i+j]$ and an index $i+j-s < h \le i$. Then (4.6) shows that the probability that a fixed individual $x \in V[h]$ joins a equals $\mathbb{P}[x \in \partial a] = \frac{\Delta\ell}{ms}(1+O(\frac{\Delta\ell}{ms}))$. Hence, the choices (4.2) and (4.5) of Δ and ℓ and the assumption $m = \Theta(k\ln n)$ ensure that

$$\begin{aligned}\mathbb{E}\big[\big|(V_1[i+j-s+1] \cup \cdots \cup V_1[i]) \cap \partial a\big| \mid \mathscr{E}\big] &= (s-j)\left(\frac{\Delta\ell}{ms}\cdot\frac{k}{\ell} + O\left(\frac{\Delta^2 k}{m^2 s^2}\right) + O\left(\frac{\Delta\ell\sqrt{k}\ln n}{ms}\right)\right)\\ &= \frac{s-j}{s}\ln 2 + O(n^{-\Omega(1)}).\end{aligned} \tag{4.26}$$

Since by **SC1** the events $\{x \in \partial a\}_x$ are independent, $|V_1[h] \cap \partial a|$ is a binomial random variable for every h and all these random variables $(|V_1[h] \cap \partial a|)_h$ are mutually independent. Therefore, (4.26) implies that

$$\mathbb{P}\big[(V_1[i+j-s+1] \cup \cdots V_1[i]) \cap \partial a = \emptyset \mid \mathscr{E}\big] = 2^{-(s-j)/s} + O(n^{-\Omega(1)}). \tag{4.27}$$

Hence,

$$\mathbb{E}\big[|\mathscr{W}_{i,j}| \mid \mathscr{E}\big] = \sum_{a \in F[i+j]} \mathbb{P}\big[(V_1[i+j-s+1] \cup \cdots \cup V_1[i]) \cap \partial a = \emptyset \mid \mathscr{E}\big] = \frac{m}{\ell}2^{-(s-j)/s}(1+O(n^{-\Omega(1)})). \tag{4.28}$$

Finally, changing the neighbourhood ∂x of one infected individual $x \in V_1$ can alter $|\mathscr{W}_{i,j}|$ by at most Δ. Therefore, the Azuma–Hoeffing inequality shows that for any $t > 0$,

$$\mathbb{P}\big[\big||\mathscr{W}_{i,j}| - \mathbb{E}[|\mathscr{W}_{i,j}| \mid \mathscr{E}]\big| > t \mid \mathscr{E}\big] \le 2\exp\left(-\frac{t^2}{2k\Delta^2}\right). \tag{4.29}$$

Combining (4.28) and (4.29), applied with $t = \sqrt{m}\ln^2 n$, and taking a union bound on i, j completes the proof. □

As a next step we use Claim 4.10 to estimate the as yet unexplained tests counts $\boldsymbol{W}_{x,j}$ from (4.7).

Claim 4.11. *For all* $s \le i < \ell$, $x \in V_1[i+1]$ *and* $j \in [s]$ *we have*

$$\mathbb{P}\left[\boldsymbol{W}_{x,j} < (1-\varepsilon/2)2^{j/s-1}\Delta/s\right] \le \exp\left(-\frac{\Omega(\ln n)}{(\ln\ln n)^4}\right).$$

Proof. Fix a pair of indices i, j and an individual $x \in V_1[i+1]$. We also condition on the event $\mathscr{E}$ that (i) from Proposition 4.1 occurs. Additionally, thanks to Claim 4.10 we may condition on the event

$$\mathscr{E}' = \left\{|\mathscr{W}_{i,j}| = 2^{-(s-j)/s}\frac{m}{\ell}(1+O(n^{-\Omega(1)}))\right\}.$$

Further, let $\mathfrak{E}$ be the σ-algebra generated by $\boldsymbol{\sigma}$ and by the neighbourhoods $(\partial y)_{y \in V[1]\cup\cdots\cup V[i]}$. Recall from **SC1** that x simply joins Δ/s random tests in compartment $F[i+j]$, independently of all other individuals, and remember from (4.7) that $\boldsymbol{W}_{x,j}$ counts tests $a \in \mathscr{W}_{i+j} \cap \partial x$. Therefore, since the events $\mathscr{E}, \mathscr{E}'$ and the random variable $|\mathscr{W}_{i,j}|$ are $\mathfrak{E}$-measurable while ∂x is independent of $\mathfrak{E}$, given $\mathfrak{E}$ the random variable $\boldsymbol{W}_{x,j}$ has a hypergeometric distribution $\mathrm{Hyp}(m/\ell, |\mathscr{W}_{i,j}|, \Delta/s)$. Thus, the assertion follows from the hypergeometric Chernoff bound from Lemma 2.2 and the choice (4.14) of ζ. □

Proof of Lemma 4.4. Since $\boldsymbol{W}^\star_x = \sum_{j=1}^s w_j \boldsymbol{W}_{x,j}$, the lemma is an immediate consequence of Markov's inequality and Claim 4.11. □

4.8. **Proof of Lemma 4.5.** We need to derive the rate functions of the random variable $\boldsymbol{W}_{x,j}$ that count as yet unexplained tests for $x \in V_{0+}[i+1]$. To this end we first investigate the set of positive tests in compartment $i+j$ that do not contain any infected individuals from the first i compartments. In symbols,

$$\mathscr{P}_{i+1,j} = \{a \in F_1[i+j] : \partial a \cap (V_1[1] \cup \cdots \cup V_1[i]) = \emptyset\} \qquad (s \le i < \ell,\ j \in [s]).$$

Claim 4.12. *W.h.p. for all $s \le i < \ell$, $j \in [s]$ we have $|\mathscr{P}_{i+1,j}| = \big(1+O(n^{-\Omega(1)})\big)\big(2^{j/s}-1\big)\frac{m}{2\ell}$.*

Proof. We may condition on the event $\mathscr{E}$ that (i) from Proposition 4.1 occurs. As a first step we calculate the probability that $(V_1[i+1] \cup \cdots \cup V_1[i+j]) \cap \partial a \neq \emptyset$ for a specific test $a \in F[i+j]$. To this end we follow the steps of the proof of Claim 4.10. Since by (4.6) a specific individual $x \in V[h]$, $i < h \le i+j$, joins a with probability $\mathbb{P}[x \in \partial a] = (\Delta\ell/(ms))(1+O(\Delta\ell/(ms)))$ and since given $\mathscr{E}$ each compartment $V[h]$ contains $k/\ell + O(\sqrt{k/\ell}\ln n)$ infected individuals, we obtain, in perfect analogy to (4.26),

$$\mathbb{E}\big[\big|(V_1[i+1] \cup \cdots \cup V_1[i+j]) \cap \partial a\big| \mid \mathscr{E}\big] = \frac{j}{s}\ln 2 + O(n^{-\Omega(1)}). \tag{4.30}$$

Since the individuals $x \in V[i+1] \cup \cdots \cup V[i+j]$ join tests independently, (4.30) implies that

$$\mathbb{P}\big[(V_1[i+1] \cup \cdots \cup V_1[i+j]) \cap \partial a \neq \emptyset \mid \mathscr{E}\big] = 1 - 2^{-j/s} + O(n^{-\Omega(1)}). \tag{4.31}$$

Furthermore, we already verified in (4.27) that

$$\mathbb{P}\big[(V_1[i+j-s+1] \cup \cdots V_1[i]) \cap \partial a = \emptyset \mid \mathscr{E}\big] = 2^{-(s-j)/s} + O(n^{-\Omega(1)}). \tag{4.32}$$

Because the choices for the compartments $V[i+j-s+1] \cup \cdots \cup V[i+j]$ from which a draws its individuals are mutually independent, we can combine (4.31) with (4.32) to obtain

$$\mathbb{P}\left[\bigcup_{i+j-s<h\le i} V_1[h] \cap \partial a = \emptyset \neq \bigcup_{i<h\le i+j} V_1[h] \cap \partial a \mid \mathscr{E}\right] = \frac{2^{j/s}-1}{2} + O(n^{-\Omega(1)}). \tag{4.33}$$

Further, (4.33) implies

$$\mathbb{E}\big[|\mathscr{P}_{i+1,j}| \mid \mathscr{E}\big] = \mathbb{E}\left[\left|\left\{a \in F_1[i+j] : \bigcup_{h\le i} V_1[h] \cap \partial a = \emptyset \neq \bigcup_{i<h} V_1[h] \cap \partial a\right\}\right| \mid \mathscr{E}\right] = (2^{j/s}-1)\frac{m}{2\ell}\big(1+O(n^{-\Omega(1)})\big). \tag{4.34}$$

Finally, altering the neighbourhood ∂x of any infected individual can shift $|\mathscr{P}_{i+1,j}|$ by at most Δ. Therefore, the Azuma–Hoeffding inequality implies that

$$\mathbb{P}\big[\big||\mathscr{P}_{i+1,j}| - \mathbb{E}[|\mathscr{P}_{i+1,j}| \mid \mathscr{E}]\big| > t \mid \mathscr{E}\big] \le 2\exp\left(-\frac{t^2}{2k\Delta^2}\right). \tag{4.35}$$

Thus, the assertion follows from (4.5), (4.34) and (4.35) by setting $t = \sqrt{m}\ln^2 n$. □

Thanks to Proposition 4.1 (iii) and Lemma 4.12 in the following we may condition on the event

$$\mathscr{U} = \left\{\forall s < i \le \ell, j \in [s] : |F_1[i+j]| = \big(1+O(n^{-\Omega(1)})\big)\frac{m}{2\ell} \wedge |\mathscr{P}_{i+1,j}| = \big(1+O(n^{-\Omega(1)})\big)\Big(2^{j/s}-1\Big)\frac{m}{2\ell}\right\}. \tag{4.36}$$

As a next step we will determine the conditional distribution of $\boldsymbol{W}_{x,j}$ for $x \in V_{0+}[i+1]$ given $\mathscr{U}$.

Claim 4.13. *Let $s < i \le \ell$ and $j \in [s]$. Given $\mathscr{U}$ for every $x \in V_{0+}[i+1]$ we have*

$$\boldsymbol{W}_{x,j} \sim \mathrm{Hyp}\left(\big(1+O(n^{-\Omega(1)})\big)\frac{m}{2\ell}, \big(1+O(n^{-\Omega(1)})\big)\Big(2^{j/s}-1\Big)\frac{m}{2\ell}, \frac{\Delta}{s}\right). \tag{4.37}$$

Proof. By **SC1** each individual $x \in V_{0+}[i+1]$ joins Δ/s positive test from $F[i+j]$, drawn uniformly without replacement. Moreover, by (4.7) given $x \in V_{0+}[i+1]$ the random variable $\boldsymbol{W}_{x,j}$ counts the number of tests $a \in \mathscr{P}_{i+1,j} \cap \partial x$. Therefore, $\boldsymbol{W}_{x,j} \sim \mathrm{Hyp}(|F_1[i+j]|, |\mathscr{P}_{i+1,j}|, \Delta/s)$. Hence, given $\mathscr{U}$ we obtain (4.37). □

The estimate (4.37) enables us to bound the probability that $\boldsymbol{W}^\star_x$ gets 'too large'.

Claim 4.14. *Let*

$$\mathscr{M} = \min \frac{1}{s}\sum_{j=1}^{s-1} \mathbf{1}\left\{z_j \geq 2^{j/s}-1\right\} D_{\mathrm{KL}}\left(z_j \| 2^{j/s}-1\right)$$
$$s.t. \quad \sum_{j=1}^{s-1}\left(z_j - (1-2\zeta)2^{j/s-1}\right) w_j = 0, \qquad z_1,\ldots,z_{s-1} \in [0,1].$$

Then for all $s \leq i < \ell$ and all $x \in V[i+1]$ we have

$$\mathbb{P}\left[\boldsymbol{W}_x^\star > (1-2\zeta)\frac{\Delta}{s}\sum_{j=1}^{s-1} 2^{j/s-1} w_j \mid \mathscr{U}, x \in V_{0+}[i+1]\right] \leq \exp(-(1+o(1))\mathscr{M}\Delta).$$

Proof. Let $s \leq i < \ell$ and $x \in V_{0+}[i+1]$. Step **SC1** of the construction of $\boldsymbol{G}$ ensures that the random variables $(\boldsymbol{W}_{x,j})_{j\in[s]}$ are independent because the tests in the various compartments $F[i+j]$, $j \in [s]$, that x joins are drawn independently. Therefore, the definition (4.12) of $\boldsymbol{W}_x^\star$ and Lemma 4.13 yield

$$\mathbb{P}\left[\boldsymbol{W}_x^\star > (1-2\zeta)\frac{\Delta}{s}\sum_{j=1}^{s-1} 2^{j/s-1} w_j \mid \mathscr{U}, x \in V_{0+}[i+1]\right] = \mathbb{P}\left[\sum_{j=1}^{s-1} w_j \boldsymbol{W}_{x,j} \geq \frac{1-2\zeta}{s}\sum_{j=1}^{s-1} 2^{j/s-1} w_j \mid \mathscr{U}, x \in V_{0+}[i+1]\right]$$
$$\leq \sum_{y_1,\ldots,y_s=0}^{\Delta} \mathbf{1}\left\{\sum_{j=1}^{s-1} w_j y_j \geq \frac{1-2\zeta}{s}\sum_{j=1}^{s-1} 2^{j/s-1} w_j\right\} \prod_{j=1}^{s-1} \mathbb{P}[\boldsymbol{W}_{x,j} \geq y_j \mid \mathscr{U}, x \in V_{0+}[i+1]]. \tag{4.38}$$

Further, let

$$\mathscr{Z} = \left\{(z_1,\ldots,z_{s-1}) \in [0,1]^{s-1} : \sum_{j=1}^{s-1}\left(z_j - (1-2\zeta)2^{j/s-1}\right) w_j = 0\right\}.$$

Substituting $y_j = \Delta z_j / s$ in (4.38) and bounding the total number of summands by $(\Delta+1)^s$, we obtain

$$\mathbb{P}\left[\boldsymbol{W}_x^\star > (1-2\zeta)\frac{\Delta}{s}\sum_{j=1}^{s-1} 2^{j/s-1} w_j \mid \mathscr{U}, x \in V_{0+}[i+1]\right] \leq (\Delta+1)^s \max_{(z_1,\ldots,z_s)\in\mathscr{Z}} \prod_{j=1}^{s-1} \mathbb{P}[\boldsymbol{W}_{x,j} \geq \Delta z_j/s \mid \mathscr{U}, x \in V_{0+}[i+1]]. \tag{4.39}$$

Moreover, Claim 4.13 and the Chernoff bound from Lemma 2.2 yield

$$\mathbb{P}[\boldsymbol{W}_{x,j} \geq \Delta z_j/s \mid \mathscr{U}, x \in V_{0+}[i+1]] \leq \exp\left(-\mathbf{1}\left\{z_j \geq p_j\right\}\frac{\Delta}{s} D_{\mathrm{KL}}\left(z_j \| p_j\right)\right) \qquad \text{where } p_j = 2^{j/s}-1+O(n^{-\Omega(1)}).$$

Consequently, since (4.5) and the assumption $m = \Theta(k\ln n)$ ensure that $\Delta = \Theta(\ln n)$, we obtain

$$\mathbb{P}[\boldsymbol{W}_{x,j} \geq \Delta z_j/s \mid \mathscr{U}, x \in V_{0+}[i+1]] \leq \exp\left(-\mathbf{1}\left\{z_j \geq 2^{j/s}-1\right\}\frac{\Delta}{s} D_{\mathrm{KL}}\left(z_j \| 2^{j/s}-1\right) + O(n^{-\Omega(1)})\right). \tag{4.40}$$

Finally, the assertion follows from (4.39) and (4.40). □

As a next step we solve the optimisation problem $\mathscr{M}$ from Claim 4.14.

Claim 4.15. *We have $\mathscr{M} = 1 - \ln 2 + O(\ln(s)/s)$.*

Proof. Fixing an auxiliary parameter $\delta \geq 0$ we set up the Lagrangian

$$\mathscr{L}_\delta(z_1,\ldots,z_s,\lambda) = \sum_{j=1}^{s-1}\left(\mathbf{1}\left\{z_j \geq 2^{j/s}-1\right\} + \delta\mathbf{1}\left\{z_j < 2^{j/s}-1\right\}\right) D_{\mathrm{KL}}\left(z_j \| 2^{j/s}-1\right) + \frac{\lambda}{s}\sum_{j=1}^{s-1} w_j\left(z_j - (1-2\zeta)2^{j/s-1}\right).$$

The partial derivatives come out as

$$\frac{\partial\mathscr{L}_\delta}{\partial\lambda} = -\frac{1}{s}\sum_{j=1}^{s-1}((1-2\zeta)2^{j/s-1} - z_j) w_j, \quad \frac{\partial\mathscr{L}_\delta}{\partial z_j} = -\lambda w_j + \left(\mathbf{1}\left\{z_j \geq 2^{j/s}-1\right\} + \delta\mathbf{1}\left\{z_j < 2^{j/s}-1\right\}\right)\ln\frac{z_j(2-2^{j/s})}{(1-z_j)(2^{j/s}-1)}.$$

Set $z_j^* = (1-2\zeta)2^{j/s-1}$ and $\lambda^* = 1$. Then clearly

$$\frac{\partial\mathscr{L}_\delta}{\partial\lambda}\Big|_{\lambda^*, z_1^*,\ldots,z_{s-1}^*} = 0. \tag{4.41}$$

Moreover, the choice (4.14) of ζ guarantees that $z_j^* \geq 2^{j/s}-1$. Hence, by the choice (4.14) of the weights w_j,

$$\frac{\partial \mathscr{L}_\delta}{\partial z_j}\Big|_{\lambda^*, z_1^*, \dots, z_{s-1}^*} = 0. \tag{4.42}$$

Since $\mathscr{L}_\delta(y_1,\dots,y_s,\lambda)$ is strictly convex in $z_1,\dots,z_s$ for every $\delta>0$, (4.41)–(4.42) imply that $\lambda^*, z_1^*, \dots, z_{s-1}^*$ is a global minimiser. Furthermore, since this is true for any $\delta>0$ and since $z_j^* \geq 2^{j/s}-1$, we conclude that $(z_1^*,\dots,z_{s-1}^*)$ is an optimal solution to the minimisation problem $\mathscr{M}$. Hence,

$$\mathscr{M} = \frac{1}{s}\sum_{j=1}^{s-1} D_{\mathrm{KL}}\left(z_j^* \| 2^{j/s}-1\right) = \frac{1}{s}\sum_{j=1}^{s-1} D_{\mathrm{KL}}\left((1-2\zeta)2^{j/s-1}\|2^{j/s}-1\right). \tag{4.43}$$

Since

$$\frac{\partial}{\partial\alpha} D_{\mathrm{KL}}\left((1-2\alpha)2^{z-1}\|2^z-1\right) = 2^z\left[-z\ln(2)+\ln(1-2^{z-1}+\alpha 2^z)-\ln(1-2^{z-1})-\ln(1-2\alpha)+\ln(2^z-1)\right],$$

we obtain $\frac{\partial}{\partial\alpha} D_{\mathrm{KL}}\left((1-2\alpha)2^{z-1}\|2^z-1\right) = O(\ln s)$ for all $z=1/s,\dots,(s-1)/s$ and $\alpha\in[0,2\zeta]$. Combining this bound with (4.43), we arrive at the estimate

$$\mathscr{M} = O(\zeta\ln s) + \frac{1}{s}\sum_{j=1}^{s-1} D_{\mathrm{KL}}\left(2^{j/s-1}\|2^{j/s}-1\right). \tag{4.44}$$

Additionally, the function $f: z\in[0,1]\mapsto D_{\mathrm{KL}}\left(2^{z-1}\|2^z-1\right)$ is strictly decreasing and convex. Indeed,

$$f'(z) = \frac{2^{z-1}\ln 2}{2^z-1}\left((2^z-1)\ln\left(\frac{2^z}{2^z-1}\right)-1\right), \qquad f''(z) = \left(2^{z-1}\ln^2 2\right)\left(\ln\left(\frac{2^z}{2^z-1}\right)+\frac{2-2^z}{(2^z-1)^2}\right).$$

The first derivative is negative because $2^{z-1}/(2^z-1)>0$ while $(2^z-1)\ln\left(2^z/(2^z-1)\right)<1$ for all $z\in(0,1)$. Moreover, since evidently $f''(z)>0$ for all $z\in(0,1)$, we obtain convexity. Further, l'Hôpital's rule yields

$$D_{\mathrm{KL}}\left(2^{1/s-1}\|2^{1/s}-1\right) = O(\ln s).$$

As a consequence, we can approximate the sum (4.44) by an integral and obtain

$$\begin{aligned}\mathscr{M} &= O(\ln(s)/s) + \int_0^1 D_{\mathrm{KL}}\left(2^{z-1}\|2^z-1\right)\mathrm{d}z \\ &= O(\ln(s)/s) + \frac{2(1-z)\ln^2(2)+2^z\ln 2^z+(1-2^z)\ln(2^z-1)}{2\ln 2}\Big|_{z=0}^{z=1} = 1-\ln(2)+O(\ln(s)/s),\end{aligned}$$

as claimed. □

Proof of Lemma 4.5. Fix $s\leq i<\ell$ and let $\boldsymbol{X}_i$ be the number of $x\in V_{0+}[i]$ such that $\boldsymbol{W}_x^\star > (1-2\zeta)\frac{\Delta}{s}\sum_{j=1}^{s-1}2^{j/s-1}w_j$. Also recall that Proposition 4.1 (iii) and Claim 4.12 imply that $\mathbb{P}[\mathscr{U}]=1-o(1)$. Combining Lemma 4.3 with Claims 4.14 and 4.15, we conclude that

$$\mathbb{E}[\boldsymbol{X}_i\,|\,\mathscr{U}] \leq \left(1+O\left(n^{-\Omega(1)}\right)\right)2^{-\Delta}n\exp(-(1-\ln(2)+o(1))\Delta) = \exp\left(\ln n-(1+o(1))\Delta\right). \tag{4.45}$$

Recalling the definition (4.5) of Δ and using the assumption that $m\geq(1+\varepsilon)m_{\mathrm{ad}}$ for a fixed $\varepsilon>0$, we obtain $\Delta\geq(1-\theta+\Omega(1))\ln n$. Combining this estimate with (4.45), we find

$$\mathbb{E}[\boldsymbol{X}_i\,|\,\mathscr{U}] \leq n^{\theta-\Omega(1)}. \tag{4.46}$$

Finally, the assertion follows from (4.46) and Markov's inequality. □

4.9. **Proof of Proposition 4.6.** The following lemma establishes an expansion property of $\boldsymbol{G}$. Specifically, if T is a small set of individuals, then there are few individuals x that share many tests with another individual from T.

Lemma 4.16. *Suppose that $m=\Theta(n^\theta\ln n)$. W.h.p. for any set $T\subset V$ of size at most $\exp(-\ln^{7/8}n)k$ we have*

$$\left|\left\{x\in V:\sum_{a\in\partial x\setminus F[0]}\mathbf{1}\{T\cap\partial a\setminus\{x\}\neq\emptyset\}\geq\ln^{1/4}n\right\}\right| \leq \frac{|T|}{3}.$$

Proof. Fix a set $T \subset V$ of size $t = |T| \le \exp(-\ln^{7/8} n)k$, a set $R \subset V$ of size $r = \lceil t/3 \rceil$ and let $\gamma = \lceil \ln^{1/4} n \rceil$. Furthermore, let $U \subset F[1] \cup \cdots \cup F[\ell]$ be a set of tests of size $\gamma r \le u \le \Delta t$. Additionally, let $\mathscr{E}(R,T,U)$ be the event that every test $a \in U$ contains two individuals from $R \cup T$. Then

$$\mathbb{P}\left[R \subset \left\{x \in V : \sum_{a \in \partial x \setminus F[0]} \mathbf{1}\left\{T \cap \partial a \setminus \{x\} \neq \emptyset\right\} \ge \gamma\right\}\right] \le \mathbb{P}\left[\mathscr{E}(R,T,U)\right]. \tag{4.47}$$

Hence, it suffices to estimate $\mathbb{P}[\mathscr{E}(R,T,U)]$.

Given a test $a \in U$ there are at most $\binom{r+t}{2}$ way to choose two individuals $x_a, x_a' \in R \cup T$. Moreover, (4.6) shows that the probability of the event $\{x_a, x_a' \in \partial a\}$ is bounded by $(1+o(1))(\Delta\ell/(ms))^2$. Therefore,

$$\mathbb{P}\left[\mathscr{E}(R,T,U)\right] \le \left[\binom{r+t}{2}\left(\frac{(1+o(1))\Delta\ell}{ms}\right)^2\right]^u.$$

Consequently, the event $\mathscr{E}(t,u)$ that there exist sets R,T,U of sizes $|R| = r = \lceil t/3 \rceil, |T| = t, |U| = u$ such that $\mathscr{E}(R,T,U)$ occurs has probability

$$\mathbb{P}\left[\mathscr{E}(t,u)\right] \le \binom{n}{r}\binom{n}{t}\binom{m}{u}\left[\binom{r+t}{2}\left(\frac{(1+o(1))\Delta\ell}{ms}\right)^2\right]^u.$$

Hence, the bounds $\gamma t/3 \le \gamma r \le u \le \Delta t$ yield

$$\begin{aligned}\mathbb{P}\left[\mathscr{E}(t,u)\right] &\le \binom{n}{t}^2\binom{m}{u}\left[\binom{2t}{2}\left(\frac{(1+o(1))\Delta\ell}{ms}\right)^2\right]^u \le \left(\frac{\mathrm{e}n}{t}\right)^{2t}\left(\frac{2\mathrm{e}\Delta^2\ell^2t^2}{ms^2u}\right)^u \\ &\le \left[\left(\frac{\mathrm{e}n}{t}\right)^{3/\gamma}\frac{6\mathrm{e}\Delta^2\ell^2 t}{\gamma ms^2}\right]^u \le \left[\left(\frac{\mathrm{e}n}{t}\right)^{3/\gamma}\cdot\frac{t\ln^4 n}{m}\right]^u \qquad \text{[due to (4.2), (4.5)].}\end{aligned}$$

Further, since $\gamma = \Omega(\ln^{1/4} n)$ and $m = \Omega(k\ln n)$ while $t \le \exp(-\ln^{7/8} n)k$, we obtain $\mathbb{P}[\mathscr{E}(t,u)] \le \exp(-u\sqrt{\ln n})$. Thus,

$$\sum_{\substack{1 \le t \le k^{1-\alpha} \\ \gamma t/3 \le u \le \Delta t}} \mathbb{P}\left[\mathscr{E}(t,u)\right] \le \sum_{1 \le u \le \Delta t} u\exp(-u\sqrt{\ln n}) = o(1). \tag{4.48}$$

Finally, the assertion follows from (4.47) and (4.48). □

Proof of Proposition 4.6. With τ the result of steps 1–10 of SPIV let $\mathscr{M}[i] = \{x \in V[i] : \tau_x \neq \boldsymbol{\sigma}_x\}$ be the set of misclassified individuals in compartment $V[i]$. Proposition 4.2 shows that w.h.p. $\mathscr{M}[i] = \emptyset$ for all $i \le s$. Further, we claim that for every $s \le i < \ell$ and any individual $x \in \mathscr{M}[i+1]$ one of the following three statements is true.

- **M1:** $x \in V_1[i+1]$ and $\boldsymbol{W}_x^\star < (1-\zeta/2)\frac{\Delta}{s}\sum_{j=1}^{s-1} 2^{j/s-1} w_j$,
- **M2:** $x \in V_{0+}[i+1]$ and $\boldsymbol{W}_x^\star > (1-2\zeta)\frac{\Delta}{s}\sum_{j=1}^{s-1} 2^{j/s-1} w_j$, or
- **M3:** $x \in V[i+1]$ and $\sum_{a \in \partial x} \mathbf{1}\{\partial a \cap (\mathscr{M}[1] \cup \cdots \cup \mathscr{M}[i]) \neq \emptyset\} \ge \ln^{1/4} n$.

To see this, assume that $x \in \mathscr{M}[i+1]$ while **M3** does not hold. Then comparing (4.7) and (4.17) we obtain

$$\left|W_{x,j}(\tau) - \boldsymbol{W}_{x,j}\right| \le \ln^{1/4} n \qquad \text{for all } 1 \le j < s. \tag{4.49}$$

Moreover, the definition (4.14) of the weights, the choice (4.3) of s, and the choices (4.14) of ζ and the weights w_j ensure that $0 \le w_j \le O(s) = O(\ln\ln n)$. This bound implies together with the definition (4.12) of the scores $\boldsymbol{W}_x^\star$ and (4.49) that

$$|\boldsymbol{W}_x^\star - W_x^\star(\tau)| = o(\zeta\Delta). \tag{4.50}$$

Thus, combining (4.50) with the definition of τ_x in Steps 5–10 of SPIV, we conclude that either **M1** or **M2** occurs.

Finally, to bound $\mathscr{M}[i+1]$ let $\mathscr{M}_1[i+1]$, $\mathscr{M}_2[i+1]$, $\mathscr{M}_3[i+1]$ be the sets of individuals $x \in V[i+1]$ for which **M1**, **M2** or **M3** occurs, respectively. Then Lemmas 4.4 and 4.5 imply that w.h.p.

$$|\mathscr{M}_1[i+1]|, |\mathscr{M}_2[i+1]| \le k\exp\left(-\frac{\ln n}{(\ln\ln n)^5}\right).$$

Furthermore, Lemma 4.16 shows that $|\mathscr{M}_3[i+1]| \le \sum_{h=1}^{i} |\mathscr{M}[h]|$ w.h.p. Hence, we obtain the relation

$$|\mathscr{M}[i+1]| \le k\exp\left(-\frac{\ln n}{(\ln\ln n)^5}\right) + \sum_{h=1}^{i} |\mathscr{M}[h]|. \tag{4.51}$$

Because (4.2) ensures that the total number of compartments is $\ell = O(\ln^{1/2} n)$, the bound (4.51) implies that $|\mathscr{M}[i+1]| \le O(\ell^2 k\exp(-(\ln n)/(\ln\ln n)^5)$ for all $i \in [\ell]$ w.h.p. Summing on i completes the proof. □

4.10. **Proof of Proposition 4.7.** For an infected individual $x \in V$ let

$$\boldsymbol{S}_x[j] = \left|\left\{a \in F[j] \cap \partial x : V_1 \cap \partial a = \{x\}\right\}\right| \qquad \text{and} \qquad \boldsymbol{S}_x = \sum_{j=1}^{\ell} \boldsymbol{S}_x[j].$$

Thus, $\boldsymbol{S}_x[j]$ is the number of positive sets $a \in F[j]$ that x has to itself, i.e., tests that do not contain a second infected individual, and $\boldsymbol{S}_x$ is the total number of such tests.

Lemma 4.17. *Assume that $m \ge (1+\varepsilon)m_{\inf}$. W.h.p. we have $\min_{x\in V_1} \boldsymbol{S}_x \ge \sqrt{\Delta}$.*

Proof. Due to Proposition 4.1 we may condition on the event

$$\mathscr{N} = \left\{\forall i \in [\ell] : \frac{m}{2\ell} - \sqrt{m}\ln n \le |F_0[i]| \le \frac{m}{2\ell} + \sqrt{m}\ln n\right\}.$$

We claim that given $\mathscr{N}$ for each $x \in V_1[i]$, $i \in [\ell]$, the random variable $\boldsymbol{S}_x$ has distribution

$$\boldsymbol{S}_x[i+j-1] \sim \mathrm{Hyp}\left(\frac{m}{\ell}, \frac{m}{2\ell} + O(\sqrt{m}\ln n), \frac{\Delta}{s}\right). \tag{4.52}$$

To see this, consider the set $F_x[i+j-1] = \left\{a \in F[i+j-1] : \partial a \cap V_1 \setminus \{x\} = \emptyset\right\}$ of all tests in compartment $F[i+j-1]$ without an infected individual besides possibly x. Since x joins $\Delta/s = O(\ln n)$ tests in $F[i+j-1]$, given $\mathscr{N}$ we have

$$\left|F_{0,x}[i+j]\right| = |F_0[i+j]| + O(\ln n) = \frac{m}{2\ell} + O(\sqrt{m}\ln n). \tag{4.53}$$

Furthermore, consider the experiment of first constructing the test design $\boldsymbol{G}$ and then re-sampling the set ∂x of neighbours of x; i.e., independently of $\boldsymbol{G}$ we have x join Δ/s random tests in each compartment $F[i+j]$. Then the resulting test design $\boldsymbol{G}'$ has the same distribution as $\boldsymbol{G}$ and hence the random variable $\boldsymbol{S}'_x[i+j-1]$ that counts tests $a \in F[i+j-1] \cap \partial x$ that do not contain another infected individual has the same distribution as $\boldsymbol{S}_x[i+j-1]$. Moreover, the conditional distribution of $\boldsymbol{S}'_x[i+j-1]$ given $\boldsymbol{G}$ reads

$$\boldsymbol{S}'_x[i+j-1] \sim \mathrm{Hyp}\left(\frac{m}{\ell}, |F_{0,x}[i+j-1]|, \frac{\Delta}{s}\right). \tag{4.54}$$

Combining (4.53) and (4.54), we obtain (4.52).

To complete the proof we combine (4.52) with Lemma 2.2, which implies that

$$\mathbb{P}\left[\boldsymbol{S}_x[i+j-1] \le \sqrt{\Delta} \mid x \in V_1\right] \le \exp\left(-\frac{\Delta}{s} D_{\mathrm{KL}}\left((1+o(1))s/\sqrt{\Delta}\|1/2+o(1)\right)\right) = \exp\left(-(1+o(1))\frac{\Delta\ln 2}{s}\right). \tag{4.55}$$

Since **SC1** ensures that the random variables $(\boldsymbol{S}_x[i+j-1])_{j\in[s]}$ are mutually independent, (4.55) yields

$$\mathbb{P}\left[\boldsymbol{S}_x \le \sqrt{\Delta} \mid x \in V_1\right] \le 2^{-(1+o(1))\Delta}. \tag{4.56}$$

Finally, the assumption $m \ge (1+\varepsilon)m_{\inf}$ for a fixed $\varepsilon > 0$ and the choice (4.5) of Δ ensure that $2^{-(1+o(1))\Delta} = o(1/k)$. Thus, the assertion follows from (4.56) by taking a union bound on $x \in V_1$. □

Proof of Proposition 4.7. For $j = 1 \dots \lceil \ln n \rceil$, let

$$\mathscr{M}_j = \left\{x \in V : \tau_x^{(j)} \neq \boldsymbol{\sigma}_x\right\}$$

contain all individuals that remain misclassified at the j-th iteration of the clean-up step. Proposition 4.6 shows that w.h.p.

$$|\mathscr{M}_1| \le k\exp\left(-\frac{\ln n}{(\ln\ln n)^6}\right). \tag{4.57}$$

Furthermore, in light of Lemma 4.17 we may condition on the event $\mathscr{A} = \{\min_{x\in V_1} \boldsymbol{S}_x \ge \sqrt{\Delta}\}$.

We now claim that given $\mathscr{A}$ for every $j \geq 1$

$$\mathscr{M}_{j+1} \subset \left\{ x \in V : \sum_{a \in \partial x \setminus F[0]} \left| \partial a \cap \mathscr{M}_j \setminus \{x\} \right| \geq \lceil \ln^{1/4} n \rceil \right\}. \tag{4.58}$$

To see this, suppose that $x \in \mathscr{M}_{j+1}$ and recall that the assumption $m \geq m_{\inf}$ and (4.5) ensure that $\Delta = \Omega(\ln n)$. Also recall that SPIV's Step 15 thresholds the number

$$S_x(\tau^{(j)}) = \sum_{a \in \partial x : \hat{\boldsymbol{\sigma}}_a = 1} \mathbf{1}\left\{ \forall y \in \partial a \setminus \{x\} : \tau_y^{(j)} = 0 \right\}$$

of positive tests containing x whose other individuals are deemed uninfected. There are two cases to consider.

- **Case 1:** $x \in V_0$**:** in this case every positive tests $a \in \partial x$ contains an individual that is actually infected. Hence, if $\tau_y^{(j)} = 0$ for all $y \in \partial a \setminus \{x\}$, then $\partial a \cap \mathscr{M}_j \setminus \{x\} \neq \emptyset$. Consequently, since Step 15 of SPIV applies the threshold of $S_x(\tau^{(j)}) \geq \ln^{1/4} n$, there are at least $\ln^{1/4} n$ tests $a \in \partial x$ such that $\partial a \cap \mathscr{M}_j \setminus \{x\} \neq \emptyset$.
- **Case 2:** $x \in V_1$**:** given $\mathscr{A}$ every infected x participates in at least $\boldsymbol{S}_x \geq \sqrt{\Delta} = \Omega(\ln^{1/2} n)$ tests that do not actually contain another infected individual. Hence, if $S_x(\tau^{(j)}) \leq \ln^{1/4} n$, then at least $\sqrt{\Delta} - \ln^{1/4} n \geq \ln^{1/4} n$ tests $a \in \partial x$ contain an individual from $\mathscr{M}_j \setminus \{x\}$.

Thus, we obtain (4.58). Finally, (4.57), (4.58) and Lemma 4.16 show that w.h.p. $|\mathscr{M}_{j+1}| \leq |\mathscr{M}_j|/3$ for all $j \geq 1$. Consequently, $\mathscr{M}_{\lceil \ln n \rceil} = \emptyset$ w.h.p. □

5. Optimal adaptive group testing

In this final section we show how the test design $\boldsymbol{G}$ from Section 4 can be extended into an optimal two-stage adaptive design. The key observation is that Proposition 4.6, which summarises the analysis of the first two phases of SPIV (i.e., steps 1–10) only requires $m \geq (1+\varepsilon)m_{\mathrm{ad}}$ tests. In other words, the excess number $(1+\varepsilon)(m_{\inf} - m_{\mathrm{ad}})$ of tests required for non-adaptive group testing is necessary only to facilitate the clean-up step, namely phase 3 of SPIV.

Replacing phase 3 of SPIV by a second test stage, we obtain an optimal adaptive test design. To this end we follow Scarlett [32], who observed that a single-stage group testing scheme that correctly diagnoses all but $o(k)$ individuals with $(1+o(1))m_{\mathrm{ad}}$ tests could be turned into a two-stage design that diagnoses all individuals correctly w.h.p. with $(1+o(1))m_{\mathrm{ad}}$ tests in total. (Of course, at the time no such optimal single-stage test design and algorithm were known.) The second test stage works as follows. Let τ denote the outcome of phases 1 and 2 of SPIV applied to $\boldsymbol{G}$ with $m = (1+\varepsilon)m_{\mathrm{ad}}$.

- **T1:** Test every individual from the set $V_1(\tau) = \{x \in V : \tau_x = 1\}$ of individuals that SPIV diagnosed as infected separately.
- **T2:** To the individuals $V_0(\tau) = \{x \in V : \tau_x = 0\}$ apply the random d-out design and the DD-algorithm from Section 4.1 with a total of $m = k$ tests and $d = \lceil 10 \ln n \rceil$.

Let $\tau' \in \{0,1\}^V$ be the result of **T1**–**T2**.

Proposition 5.1. *W.h.p. we have $\tau'_x = \boldsymbol{\sigma}_x$ for all $x \in V$.*

As a matter of course **T1** renders correct results, i.e., for all individuals $x \in V_1(\tau)$ we have $\tau'_x = \boldsymbol{\sigma}_x$. Further, to analyse **T2** we use a similar argument as in the analysis of the first phase of SPIV in Section 4.5; we include the analysis for the sake of completeness. We begin by investigating the number of negative tests. Let $\boldsymbol{G}'$ denote the test design set up by **T2**, let $F' = \{b_1, \ldots, b_k\}$ denote its set of tests and let $\hat{\boldsymbol{\sigma}}_{b_1}, \ldots, \hat{\boldsymbol{\sigma}}_{b_k}$ signify the corresponding test results. Further, let $F'_0 = \{b \in F' : \hat{\boldsymbol{\sigma}}_b = 0\}$ and $F'_1 = \{b \in F' : \hat{\boldsymbol{\sigma}}_b = 1\}$ be the set of negative and positive tests, respectively.

Lemma 5.2. *W.h.p. we have $|F'_1| \leq \frac{k}{2}$.*

Proof. Proposition 4.6 implies that w.h.p.

$$|V_0(\tau) \cap V_1| \leq \sum_{x \in V} \mathbf{1}\{\tau_x \neq \boldsymbol{\sigma}_x\} \leq k \exp\left(-\frac{\ln n}{(\ln \ln n)^6} \right). \tag{5.1}$$

Moreover, since every individual $x \in V_0(\tau)$ joins d random tests, for any specific test $b \in F'$ we have

$$\mathbb{P}\left[x \in \partial_{\boldsymbol{G}'} b\right] = 1 - \mathbb{P}\left[x \notin \partial_{\boldsymbol{G}'} b\right] = 1 - \binom{k-1}{d}\binom{k}{d}^{-1} = \frac{d}{k}(1 + O(n^{-\Omega(1)})).$$

Hence, for every test $b \in F'$,

$$\mathbb{E}\left[|\partial b \cap V_1| \,\middle|\, |V_0(\tau) \cap V_1| \le k \exp\left(-\frac{\ln n}{(\ln\ln n)^6}\right)\right] = O(1/\ln n).$$

Consequently,

$$\mathbb{E}\left[|F_1'| \,|\, |V_0(\tau) \cap V_1| \le k/\ln n\right] = O(k/\ln n). \tag{5.2}$$

Finally, combining (5.1) and (5.2) and applying Markov's inequality, we conclude that $|F_1'| \le \frac{k}{2}$ w.h.p. □

Corollary 5.3. *W.h.p. for every $x \in V_0(\tau)$ there is a test $b \in F'$ such that $\partial b \setminus \{x\} \subset V_0$.*

Proof. We construct the random graph $\boldsymbol{G}'$ in two rounds. In the first round we first expose the neighbourhoods $(\partial_{\boldsymbol{G}'} y)_{y \in V_0(\tau)\setminus\{x\}}$. Lemma 5.2 implies that after the first round the number $\boldsymbol{X}$ of tests that do not contain an infected individual $y \in V_0(\tau) \cap V_1$ exceeds $k/2$ w.h.p. In the second round we expose $\partial_{\boldsymbol{G}'} x$. Because $\partial_{\boldsymbol{G}'} x$ is chosen independently of the neighbourhoods $(\partial_{\boldsymbol{G}'} y)_{y \in V_0(\tau)\setminus\{x\}}$, the number of tests $b \in \partial_{\boldsymbol{G}'} x$ that do not contain an infected individual $y \in V_0(\tau) \cap V_1$ has distribution $\mathrm{Hyp}(k, \boldsymbol{X}, d)$. Therefore, since $d \ge 10 \ln n$ we obtain

$$\mathbb{P}\left[\forall b \in \partial x : V_1 \cap \partial b \setminus \{x\} \ne \emptyset \mid \boldsymbol{X} \le k/2\right] \le \mathbb{P}\left[\mathrm{Hyp}(k, k/2, d) = 0\right] \le 2^{-d} = o(1/n). \tag{5.3}$$

Finally, the assertion follows (5.3) and the union bound. □

Proof of Proposition 5.1. Corollary 5.3 shows that we may assume that for every $x \in V_0(\tau)$ there is a test $b_x \in F'$ with $\partial b_x \setminus \{x\} \subset V_0$. As a consequence, upon executing the first step **DD1** of the DD algorithm, **T2** will correctly diagnose all individuals $x \in V_0(\tau) \cap V_0$. Therefore, if $x \in V_0(\tau) \cap V_1$, then **DD2** will correctly identify x as infected because all other individuals $y \in \partial b_x$ were already identified as healthy by **DD1**. Thus, $\tau'_x = \boldsymbol{\sigma}_x$ for all $x \in V$. □

Proof of Theorem 1.3. Proposition 5.1 already establishes that the output of the two-stage adaptive test is correct w.h.p. Hence, to complete the proof we just observe that the total number of tests comes to $(1+\varepsilon)m_{\mathrm{ad}}$ for the first stage plus $|V_1(\tau)| + k$ for the second stage. Furthermore, Proposition 4.6 implies that w.h.p.

$$|V_1(\tau)| \le |V_1| + \sum_{x \in V} \mathbf{1}\{\tau_x \ne \boldsymbol{\sigma}_x\} \le k\left(1 + \exp\left(-\frac{\ln n}{(\ln\ln n)^6}\right)\right) = (1 + o(1))k.$$

Thus, the second stage conducts $O(k) = o(m_{\mathrm{ad}})$ tests. □

Acknowledgment. We thank Arya Mazumdar for bringing the group testing problem to our attention.

References

[1] E. Abbe: Community detection and stochastic block models: recent developments. The Journal of Machine Learning Research **18** (2017) 6446–6531.

[2] A. Alaoui, A. Ramdas, F. Krzakala, L. Zdeborová, M. Jordan: Decoding from pooled data: phase transitions of message passing. IEEE Transactions on Information Theory **65** (2019) 572–585.

[3] A. Alaoui, A. Ramdas, F. Krzakala, L. Zdeborová, M. Jordan: Decoding from pooled data: Sharp information-theoretic bounds. SIAM Journal on Mathematics of Data Science **1** (2019) 161–188.

[4] M. Aldridge, L. Baldassini, O. Johnson: Group testing algorithms: bounds and simulations. IEEE Transactions on Information Theory **60** (2014) 3671–3687.

[5] M. Aldridge: Individual testing is optimal for nonadaptive group testing in the linear regime. IEEE Transactions on Information Theory **65** (2019) 2058–2061.

[6] M. Aldridge, O. Johnson, J. Scarlett: Group testing: an information theory perspective. Foundations and Trends in Communications and Information Theory (2019).

[7] N. Alon, M. Krivelevich, B. Sudakov: Finding a large hidden clique in a random graph. Proc. 9th SODA (1998) 594–598.

[8] T. Berger, V. Levenshtein: Asymptotic efficiency of two-stage disjunctive testing. IEEE Transactions on Information Theory, **48** (2002) 1741–1749.

[9] M. Brennan, G. Bresler: Optimal average-case reductions to sparse pca: From weak assumptions to strong hardness. arXiv:1902.07380.

[10] H. Chen, F. Hwang: A survey on nonadaptive group testing algorithms through the angle of decoding. Journal of Combinatorial Optimization **15** (2008) 49–59.

[11] A. Coja-Oghlan, O. Gebhard, M. Hahn-Klimroth, P. Loick: Information-theoretic and algorithmic thresholds for group testing. Proc. 46th ICALP (2019) #43.

[12] A. Decelle, F. Krzakala, C. Moore, L. Zdeborová: Asymptotic analysis of the stochastic block model for modular networks and its algorithmic applications. Phys. Rev. E **84** (2011) 066106.
[13] D. Donoho: Compressed sensing. IEEE Transactions on Information Theory **52** (2006) 1289–1306.
[14] D. Donoho, A. Javanmard , A. Montanari: Information-theoretically optimal compressed sensing via spatial coupling and approximate message passing. IEEE Transactions on Information Theory **59** (2013) 7434–7464.
[15] R. Dorfman: The detection of defective members of large populations. Annals of Mathematical Statistics **14** (1943) 436–440.
[16] A. D'yachkov, V. Rykov: Bounds on the length of disjunctive codes. Problemy Peredachi Informatsii **18** (1982) 166–171.
[17] P. Erdős, A. Rényi: On Two Problems of Information Theory. Magyar Tud. Akad. Mat. Kutató Int. Közl **8** (1963) 229–243.
[18] A. Felstrom, K. Zigangirov: Time-varying periodic convolutional codes with low-density parity-check matrix. IEEE Transactions on Information Theory **45** (1999) 2181–2191.
[19] W. Hoeffding: Probability inequalities for sums of bounded random variables. In N. Fisher, P. Sen (eds.): The collected works of Wassily Hoeffding. Springer Series in Statistics (Perspectives in Statistics). Springer, New York, NY (1994) 409–426.
[20] F. Hwang: A method for detecting all defective members in a population by group testing. Journal of the American Statistical Association **67** (1972) 605–608.
[21] S. Janson, T. Luczak, A. Rucinski: Random Graphs. John Wiley & Sons (2011).
[22] O. Johnson, M. Aldridge, J. Scarlett: Performance of group testing algorithms with near-constant tests per item. IEEE Transactions on Information Theory **65** (2018) 707–723.
[23] W. Kautz, R. Singleton: Nonrandom binary superimposed codes. IEEE Transactions on Information Theory **10** (1964), 363–377.
[24] F. Krzakala, M. Mézard, F. Sausset, Y. Sun, L. Zdeborová: Statistical-physics-based reconstruction in compressed sensing. Physical Review X **2** (2012) 021005.
[25] S. Kudekar, H. Pfister: The effect of spatial coupling on compressive sensing. Proc. 48th Allerton (2010) 347–353.
[26] S. Kudekar, T. Richardson, R. Urbanke: Threshold saturation via spatial coupling: why convolutional LDPC ensembles perform so well over the BEC. IEEE Transaction on Information Theory **57** (2011) 803–834.
[27] S. Kudekar, T. Richardson, R. Urbanke: Spatially coupled ensembles universally achieve capacity under belief propagation. IEEE Transaction on Information Theory **59** (2013) 7761–7813.
[28] H. Kwang-Ming, D. Ding-Zhu: Pooling designs and nonadaptive group testing: important tools for DNA sequencing. World Scientific (2006).
[29] M. Mézard, M. Tarzia, C. Toninelli: Group testing with random pools: phase transitions and optimal strategy. Journal of Statistical Physics **131** (2008) 783–801.
[30] C. Moore: The computer science and physics of community detection: landscapes, phase transitions, and hardness. Bulletin of the EATCS **121** (2017).
[31] G. Reeves, H. Pfister (2019). Understanding phase transitions via mutual information and MMSE. arXiv:1907.02095.
[32] J. Scarlett: Noisy adaptive group testing: Bounds and algorithms. IEEE Transactions on Information Theory **65** (2018) 3646–3661.
[33] J. Scarlett: An efficient algorithm for capacity-approaching noisy adaptive group testing. Proc. IEEE International Symposium on Information Theory (2019) 2679–2683.
[34] K. Takeuchi, T. Tanaka, T. Kawabata: Improvement of BP-based CDMA multiuser detection by spatial coupling. Proc. IEEE International Symposium on Information Theory Proceedings (2011) 1489–1493.
[35] P. Ungar: The cutoff point for group testing. Communications on Pure and Applied Mathematics **13** (1960) 49–54.
[36] L. Wang, X. Li, Y. Zhang, K. Zhang: Evolution of scaling emergence in large-scale spatial epidemic spreading. PLoS ONE **6** (2011).
[37] Y. Wu, S. Verdu, Renyi information dimension: fundamental limits of almost lossless analog compression. IEEE Transactions on Information Theory **56** (2010) 3721–3748.
[38] L. Zdeborová, F. Krzakala: Statistical physics of inference: thresholds and algorithms. Advances in Physics **65** (2016) 453–552.

AMIN COJA-OGHLAN, acoghlan@math.uni-frankfurt.de, GOETHE UNIVERSITY, MATHEMATICS INSTITUTE, 10 ROBERT MAYER ST, FRANKFURT 60325, GERMANY.

OLIVER GEBHARD, gebhard@math.uni-frankfurt.de, GOETHE UNIVERSITY, MATHEMATICS INSTITUTE, 10 ROBERT MAYER ST, FRANKFURT 60325, GERMANY.

MAX HAHN-KLIMROTH, hahnklim@math.uni-frankfurt.de, GOETHE UNIVERSITY, MATHEMATICS INSTITUTE, 10 ROBERT MAYER ST, FRANKFURT 60325, GERMANY.

PHILIPP LOICK, loick@math.uni-frankfurt.de, GOETHE UNIVERSITY, MATHEMATICS INSTITUTE, 10 ROBERT MAYER ST, FRANKFURT 60325, GERMANY.

D. Near-Optimal Sparsity-Constrained Group Testing: Improved Bounds and Algorithms

NEAR-OPTIMAL SPARSITY-CONSTRAINED GROUP TESTING: IMPROVED BOUNDS AND ALGORITHMS

OLIVER GEBHARD, MAX HAHN-KLIMROTH, OLAF PARCZYK, MANUEL PENSCHUCK,
MAURICE ROLVIEN, JONATHAN SCARLETT, AND NELVIN TAN

Abstract

Recent advances in noiseless non-adaptive group testing have led to a precise asymptotic characterization of the number of tests required for high-probability recovery in the sublinear regime $k = n^\theta$ (with $\theta \in (0,1)$), with n individuals among which k are infected. However, the required number of tests may increase substantially under real-world practical constraints, notably including bounds on the maximum number Δ of tests an individual can be placed in, or the maximum number Γ of individuals in a given test. While previous works have given recovery guarantees for these settings, significant gaps remain between the achievability and converse bounds. In this paper, we substantially or completely close several of the most prominent gaps. In the case of Δ-divisible items, we show that the definite defectives (DD) algorithm coupled with a random regular design is asymptotically optimal in dense scaling regimes, and optimal to within a factor of e more generally; we establish this by strengthening both the best known achievability and converse bounds. In the case of Γ-sized tests, we provide a comprehensive analysis of the regime $\Gamma = \Theta(1)$, and again establish a precise threshold proving the asymptotic optimality of DD equipped with a tailored pooling scheme. Finally, for each of these two settings, we provide near-optimal adaptive algorithms based on sequential splitting, and provably demonstrate gaps between the performance of optimal adaptive and non-adaptive algorithms.

1. INTRODUCTION

The group testing problem, originally introduced by Dorfman [14], is a prominent example of a classical inference problem that has recently regained considerable attention [4, 12, 15]. Briefly, the problem is posed as follows: Among a population of n individuals, a small subset of k individuals is infected with a rare disease. We are able to test groups of individuals at once, and each test result returns positive if (and only if) there is at least one infected individual in the test group. The challenge is to develop strategies for pooling individuals into tests such that the status of every individual can be recovered reliably from the outcomes, and to do so using as few tests as possible.

While the preceding terminology corresponds to medical applications, group testing also has many other key applications [4, Sec. 1.7], ranging from DNA sequencing [27, 33] to protein interaction experiments [31, 39]. Particular attention has been paid to group testing as a tool for the containment of an epidemic crisis. On the one hand, mass testing appears to be an essential tool to face pandemic spread [10], while on the other hand, the capability of efficiently identifying infected individuals fast and at a low cost is indispensable [29]. For the sake of pandemic control, risk surveillance plans aim at an early, fast and efficient identification of infected individuals to prevent diseases from spreading [18, 34, 35].

The group testing problem includes many variants, depending on the presence/absence of noise, possible adaptivity of the tests, recovery requirements, and so on. Our focus in this paper is on the following setup, which has been the focus of numerous recent works (see [4] for a survey):

- The tests are *non-adaptive*, meaning they must all be designed in advance before observing any outcomes. This is highly desirable in applications, as it permits the tests to be implemented in parallel.
- The tests are *noiseless*; this assumption is more realistic in some applications than others, but serves as a crucial starting point for understanding the problem.
- The goal is *high-probability* identification of each individual's defectivity status (i.e., probability approaching one as $n \to \infty$). While a deterministic (probability-one) recovery guarantee is also feasible in the noiseless setting [15], it requires considerably more tests, incurring a k^2 dependence on the number of infected individuals instead of k.

The authors are listed alphabetically. This work was presented in part at the IEEE International Symposium on Information Theory (ISIT), 2020 [38].

Funding: OG was funded by DFG CO 646/3. MHK was partially funded by Stiftung Polytechnische Gesellschaft and DFG FOR 2975. OP was supported by the DFG (Grant PA 3513/1-1). MP was funded by ME 2088/4-2, and ME 2088/5-1 (DFG FOR 2975). JS was funded by an NUS Early Career Research Award.

- The number of infected individuals k is taken to equal n^θ for some $\theta \in (0,1)$,[1] i.e., the *sublinear regime*. Heaps' law of epidemics [7, 40] indicates that this regime is of major interest. In addition, recent hardness results preclude non-trivial recovery guarantees in the linear regime $k = \Theta(n)$ [1], at least under the most widely-adopted recovery criterion.

Under this setup, Coja-Oghlan et al. [11, 12] recently established the exact information-theoretic threshold on the number of tests, in an asymptotic sense including the implied constant. This threshold was originally attained using a *random regular testing* design [11] (see also [26]), improving on earlier results for *Bernoulli testing* [2, 37]. While the recovery algorithm used in [11] is not computationally efficient, the subsequent work [12] attained the same threshold using a *spatially coupled random regular design* and a computationally efficient recovery algorithm.

All of the preceding test designs have in common that every individual takes part in $O(\ln n)$ tests, and each test contains $O(n/k)$ individuals. As a result, these designs face limitations in real-world applications. Firstly, one may face dilution effects: If an infected individual gets tested within a group of many uninfected individuals, the signal of the infection (e.g., concentration of the relevant molecules) might be too low. For instance, a testing scheme for HIV typically should not contain more than 80 individual samples per test [41]. More recently, evidence was found that certain laboratory tests allow pooling of up to 5 individuals [16] or 64 individuals [21] per test for reliably detecting COVID-19 infections. Secondly, it is often the case that each individual can only be tested a certain number of times, due to the limited volume of the sample taken. More generally, test designs with few tests-per-individual and/or individuals-per-test may be favorable due to resource limitations, difficulties in manually placing samples into tests, and so on.

In light of these practical issues, there is substantial motivation to study the group testing problem under the following constraints on the test design:

- Under the Δ-*divisible items constraint* (or *bounded resource model*), any given individual can only be tested at most Δ times;
- Under the Γ-*sized tests constraint* (or *bounded test-size model*), any given test can only contain at most Γ individuals.

Previous studies of group testing under these constraints [20, 23, 28, 38] are surveyed in Section 1.1. We note that some of the above practical motivations may warrant more sophisticated models (e.g., random noise models for dilution effects), but nevertheless, noiseless group testing under the preceding constraints serves as an important starting point towards a full understanding. In addition, as with previous works, we only consider the above two constraints separately, though the case that both are present simultaneously may be of interest for future studies.

1.1. **Related Work.** As outlined above, the asymptotically optimal performance limits are well-understood in the case of unconstrained test designs, with optimal designs placing each item in $\Delta = \Theta(\ln n)$ tests, and each test containing $\Gamma = \Theta\big(\frac{n}{k}\big)$ items. We refer the reader to [4] for a more detailed survey, and subsequently focus our attention on the (much more limited) prior work considering the constrained variants with $\Delta = o(\ln n)$ and $\Gamma = o\big(\frac{n}{k}\big)$.

The most relevant prior work is that of Gandikota et al. [20], who gave information-theoretic lower bounds on the number of tests under both kinds of constraint, as well as upper bounds via the simple COMP algorithm [9].[2] The main results therein are summarized as follows, assuming the sublinear regime $k = n^\theta$ with $\theta \in (0,1)$ throughout:

- Δ-divisible items setting:
  - **(Converse)** For $\Delta = o(\ln n)$, any non-adaptive design with error probability at most ξ requires $m \geq \Delta k\big(\frac{n}{k}\big)^{\frac{1-5\xi}{\Delta}}$, for sufficiently small ξ and sufficiently large n.
  - **(Achievability)** Under a suitably-chosen random test design and COMP recovery, the error probability is at most ξ provided that $m \geq \lceil e\Delta k\big(\frac{n}{\xi}\big)^{\frac{1}{\Delta}}\rceil$.
- Γ-sized tests setting:
  - **(Converse)** For $\Gamma = \Theta\big(\big(\frac{n}{k}\big)^\beta\big)$ with $\beta \in [0,1)$, any non-adaptive design with error probability at most ξ requires $m \geq \frac{1-6\xi}{1-\beta} \cdot \frac{n}{\Gamma}$, for sufficiently large n.

[1] To simplify notation, we assume that $k = n^\theta$ exactly, but all of our analysis and results extend easily to the more general case that $k = cn^\theta$ for any $c = O(1)$.

[2] The COMP algorithm declares any individual in a negative test as uninfected, and all other individuals as infected. It is called Column Matching Algorithm in [20].

– **(Achievability)** Under a suitably-chosen random test design and COMP recovery, for $\Delta = \Theta\big(\big(\frac{n}{k}\big)^{\beta}\big)$ with $\beta \in [0,1)$ and $\xi = n^{-\zeta}$ with $\zeta > 0$, the error probability is at most ξ when $m \geq \lceil \frac{1+\zeta}{(1-\alpha)(1-\beta)} \rceil \cdot \lceil \frac{n}{\Gamma} \rceil$.

A sizable gap remains between the achievability and converse bounds in the case of Δ-divisible items, since typically $\big(\frac{1}{\xi}\big)^{\frac{1}{\Delta}} \gg \big(\frac{1}{k}\big)^{\frac{1}{\Delta}}$. For Γ-sized tests, the bounds match to within a constant factor, but the optimal constant remains unknown. In particular, the two differ by at least a multiplicative $\frac{1}{1-\alpha}$ factor, and even for α close to zero, the two can differ by a factor of 2 due to the rounding in the achievability part.

As we outline further below, we nearly completely close these gaps for Δ-divisible items, and we close them completely for Γ-sized tests in the special case $\beta = 0$ (i.e., $\Gamma = \Theta(1)$) for all $\theta \in (0,1)$. While the regime $\beta \in (0,1)$ is also of interest, it appears to require different techniques, and is deferred to future work.

Gandikota et al. [20] additionally gave explicit designs (i.e., test matrices that can be deterministically constructed in polynomial time), but these give worse scaling laws, and are therefore of less relevance to our results based on random designs. In a distinct but related line of works, Macula [28] and Inan et al. [23, 24] developed designs for the much stronger guarantee of *uniform recovery*, i.e., a single test matrix that uniquely recovers any infected set of size at most k, without allowing any error probability. This stronger guarantee comes at the price of requiring considerably more tests, and we thus omit a direct comparison and refer the interested reader to [23, 24, 28] for details.

1.2. **Contributions.** Our main contributions are informally outlined as follows (with $k = n^{\theta}$ for $\theta \in (0,1)$, and ε being an arbitrarily small constant throughout), with formal statements given in the theorems referenced:

- Δ-divisible items setting. Assuming that $\Delta = (\ln n)^{1-\Omega(1)}$ (and in some cases, any $\Delta = o(\ln n)$ is allowed), we have the following:
  - **(General converse – Theorem 3.1)** If $m \leq (1-\varepsilon)\mathrm{e}^{-1}\Delta k^{1+\frac{1-\theta}{\Delta\theta}}$,[3] then w.h.p. any (possibly adaptive) group testing strategy fails.
  - **(Non-adaptive converse – Theorem 3.2)** Under any non-adaptive test design, if $m \leq (1-\varepsilon)\Delta k^{1+\frac{1}{\Delta}}$, then w.h.p. any inference algorithm fails. Combining with the general lower bound, the same holds for $m \leq (1-\varepsilon)\max\Big\{\mathrm{e}^{-1}\Delta k^{1+\frac{1-\theta}{\Delta\theta}}, \Delta k^{1+\frac{1}{\Delta}}\Big\}$.
  - **(Non-adaptive achievability via DD – Theorem 3.3)** Under a random regular test design, DD succeeds when $m \geq (1+\varepsilon)\max\Big\{\Delta k^{1+\frac{1-\theta}{\Delta\theta}}, \Delta k^{1+\frac{1}{\Delta}}\Big\}$ (w.h.p.when $\Delta = \omega(1)$, and with probability $(1-o(1))\big(1-(1+\varepsilon)^{-\Delta}\big)$ when $\Delta = \Theta(1)$).
  - **(DD-specific converse – Theorem 3.4)** Under random regular testing, DD fails when m is slightly below the achievability bound (w.h.p. when $\Delta = \omega(1)$, and with $\Omega(1)$ probability when $\Delta = \Theta(1)$).
  - **(Adaptive achievability – Theorem 5.1)** There exists an efficient adaptive algorithm succeeding with probability one when $m \geq (1+\varepsilon)\Delta k^{1+\frac{1-\theta}{\Delta\theta}}$.
- Γ-sized tests setting: Assuming that $\Gamma = \Theta(1)$ in the non-adaptive setting (whereas the adaptive results allow general $\Gamma = o\big(\frac{n}{k}\big)$), we have the following:
  - **(Non-adaptive converse– Theorem 4.1)** If $m \leq (1-\varepsilon)\max\Big\{\big(1+\Big\lfloor\frac{\theta}{1-\theta}\Big\rfloor\big)\frac{n}{\Gamma}, 2\frac{n}{\Gamma+1}\Big\}$, then any non-adaptive group testing strategy fails (w.h.p. if $\frac{\theta}{1-\theta}$ is non-integer, and with $\Omega(1)$ probability if $\frac{\theta}{1-\theta}$ is an integer).
  - **(Non-adaptive achievability via DD in the dense regime – Theorem 4.10)** If $\theta \geq \frac{1}{2}$, then under a suitably-chosen random test design, DD succeeds w.h.p. when $m \geq \frac{n}{\Gamma}\big(1+\Big\lfloor\frac{\theta}{1-\theta}\Big\rfloor\big)$.
  - **(Non-adaptive achievability via DD in the sparse regime – Theorem 4.18)** If $\theta < \frac{1}{2}$, then under a suitably-chosen random test design, DD succeeds w.h.p. with $m \geq \frac{2n}{\Gamma+1}$.
  - **(Adaptive achievability – Theorem 6.1)** There exists an efficient adaptive algorithm succeeding with probability one when $m \geq (1+\varepsilon)\frac{n}{\Gamma} + k\log_2\Gamma$. In particular, when $\Gamma = o\big(\frac{n}{k\ln n}\big)$, it suffices that $m \geq (1+\varepsilon)\frac{n}{\Gamma}$.
  - **(General converse – Theorem 6.2)** If $m \leq (1-\varepsilon)\frac{n}{\Gamma}$, then the error probability is bounded away from zero for any (possibly adaptive) group testing strategy.

These results have several interesting implications, which we discuss as follows. In the Δ-divisible setting, our first converse bound strengthens that of [20] (removing the -5ξ term in the exponent) and extends it to the adaptive setting, and our second converse provides a further improvement for non-adaptive designs.

[3] These expressions are obtained after substituting $k = n^{\theta}$. In the more general case that k equals a positive constant times n^{θ}, the results remain unchanged upon replacing $k^{1+\frac{1-\theta}{\Delta\theta}}$ by $k\big(\frac{n}{k}\big)^{\frac{1}{\Delta}}$ everywhere.

Our DD achievability result scales as $O\big(\Delta k\big(\max\{k, \frac{n}{k}\}\big)^{\frac{1}{\Delta}}\big)$, which is strictly better than the $O\big(\Delta k n^{\frac{1}{\Delta}}\big)$ scaling of COMP [20] for all $\theta \in (0,1)$. In fact, for $\theta > \frac{1}{2}$ and $\Delta = \omega(1)$, our results demonstrate that DD is asymptotically optimal among non-adaptive strategies, with a precise phase transition between success and failure at $m \approx \Delta k^{1+\frac{1}{\Delta}}$. For $\theta < \frac{1}{2}$, while establishing a precise phase transition remains an open problem, our results establish DD's optimality up to a multiplicative factor of e, and demonstrate that one cannot reduce the number of tests further under DD and the random regular design. Finally, our results prove a strict adaptivity gap for $\theta > \frac{1}{2}$, and demonstrate that our adaptive algorithm is optimal to within a factor of e for all $\theta \in (0,1)$.

In the Γ-sized tests setting, our results provide an exact asymptotic threshold on the number of tests in the $\Gamma = \Theta(1)$ regime, and we again establish the asymptotic optimality of DD in all such cases. To achieve this, we analyze via novel techniques specific to this scaling, including a novel test design in the case $\theta > \frac{1}{2}$, as described in the next section. We note that the distinction between integer and non-integer valued $\frac{\theta}{1-\theta}$ arises due to rounding issues in the analysis, e.g., counting the number of individuals appearing in at most $\lfloor \frac{\theta}{1-\theta} \rfloor$ tests. Our results again demonstrate a strict adaptivity gap (this time for all $\theta \in (0,1)$), and we provide a precise phase transition at $\frac{n}{\Gamma}$ for adaptive algorithms under most scalings of Γ.

2. Fundamentals of Non-Adaptive Group Testing

2.1. **General Notation.** Given the number of individuals n, the number of infected individuals $k \sim n^\theta (\theta \in (0,1))$, and the number of tests m, we let $\mathscr{G} = (V \cup F, E)$ be a random bipartite (multi-)graph with $|F| = m$ *factor* nodes $(a_1, ..., a_m)$ and $|V| = n$ variable nodes $(x_1, ..., x_n)$. The variable nodes represent individuals, the factor nodes represent tests, and an edge between individual x_i and test a_j indicates, that x_i takes part in test a_j. Furthermore, let $(\partial_{\mathscr{G}} a_1, ..., \partial_{\mathscr{G}} a_m)$ and $(\partial_{\mathscr{G}} x_1, ..., , \partial_{\mathscr{G}} x_n)$ denote the neighborhoods in $\mathscr{G}$. Whenever the context clarifies what $\mathscr{G}$ is, we will drop the subscript. The test-node degrees are given by $\Gamma_i(\mathscr{G}) = |\partial_{\mathscr{G}} a_i|$, and the individual-node degrees by $\Delta_i(\mathscr{G}) = |\partial_{\mathscr{G}} x_i|$. We can visualize any non-adaptive group testing instance by a *pooling scheme* in the form of such a graph $\mathscr{G}$.

We indicate the infection status of each individual of the population by $\boldsymbol{\sigma} \in \{0,1\}^n$, a uniformly chosen vector of Hamming weight k. Formally, $\boldsymbol{\sigma}_x = 1$ iff x is infected. Then, we let $\hat{\boldsymbol{\sigma}} = \hat{\boldsymbol{\sigma}}(\mathscr{G}, \boldsymbol{\sigma}) \in \{0,1\}^m$ denote the sequence of test results, such that $\hat{\boldsymbol{\sigma}}_a = 1$ iff test a contains at least one infected individual, that is

$$\hat{\boldsymbol{\sigma}}_a = \max_{x \in \partial a} \boldsymbol{\sigma}_x.$$

Throughout the paper, we use standard Landau notation, e.g., $o(1)$ is a function converging to 0 while $\omega(1)$ stands for an arbitrarily slowly diverging function. Moreover, we say that a property $\mathscr{P}$ holds *with high probability (w.h.p.)*, if $\mathbb{P}(\mathscr{P}) = 1 - o(1)$ as $n \to \infty$.

2.2. **Pooling Schemes.** The random (almost-)regular bipartite pooling scheme is known to be information-theoretically optimal in the unconstrained variant of group testing [11], and is conceptually simple and easy to implement. In this work, depending on the setup, we sometimes require less standard schemes, as described in the following. It is important to note that in each of these designs, we are constructing a multi-graph rather than a graph, and every multi-edge is counted when referring to a node degree. In the following we will define our choices of the restricted pooling scheme and denote them $\mathscr{G}_\Delta$ and $\tilde{\mathscr{G}}_\Gamma$

2.2.1. Δ-*divisible.* In this setup, we adopt the design of [11, 26], but with fewer tests per individual in accordance with the problem constraint: Each individual chooses Δ tests uniformly at random with replacement; thus, an individual may be placed in the same test more than once. By construction of $\mathscr{G}_\Delta$, any individual has degree *exactly* Δ, whereas the test degrees fluctuate. We denote by $\boldsymbol{\Gamma}(\mathscr{G}_\Delta) = \{\boldsymbol{\Gamma}_1(\mathscr{G}_\Delta), \ldots, \boldsymbol{\Gamma}_m(\mathscr{G}_\Delta)\}$ the (random) sequence of test-degrees.

2.2.2. Γ-*sparse.* In the Γ-sparse case, our choice of pooling scheme requires additional care; we define $\tilde{\mathscr{G}}_\Gamma(\theta)$ separately for two cases:

$$\tilde{\mathscr{G}}_\Gamma(\theta) = \begin{cases} \mathscr{G}_\Gamma & \text{if } \theta \geq 1/2 \\ \mathscr{G}^*_\Gamma & \text{otherwise} \end{cases} \tag{2.1}$$

with $\mathscr{G}_\Gamma$ and $\mathscr{G}^*_\Gamma$ defined in the following. Throughout the paper, we will always clarify which of the cases we assume, and we will therefore refer to $\tilde{\mathscr{G}}_\Gamma(\theta)$ as $\tilde{\mathscr{G}}_\Gamma$. Starting with $\mathscr{G}_\Gamma$, we employ the *configuration model* [25]. Given n, m, Γ, set $\Delta = m\Gamma/n$ and create for each individual $x \in [n]$ exactly Δ clones $\{x\} \times \{1\}, \ldots, \{x\} \times \{\Delta\}$. We assume throughout, that Δ, Γ, n, m are integers, thus all divisibility requirements are fulfilled.[4] Analogously,

[4]It will turn out in due course that $m\Gamma/n$ is an integer under the choice of Γ used in the analysis.

create Γ clones $\{a\}\times\{1\}\dots\{a\}\times\{\Gamma\}$ for each test $a\in[m]$. Then, choose a perfect matching uniformly at random between the individual-clones and the test-clones and construct a random multi-graph by merging the clones to vertices and adding an edge (x,a) whenever there are $i\in[\Delta]$, $j\in[\Gamma]$ such that the edge $(\{x\}\times i,\{a\}\times j)$ is part of the perfect matching. We denote by $\mathscr{G}_\Gamma$ the random regular multi-graph that comes from this procedure.

For $\mathscr{G}^*_\Gamma$, we adopt a different approach. First, we select $\gamma\le\frac{2n}{\Gamma+1}$ individuals randomly and put them apart for the moment (denote by $X=\{x_1\dots x_\gamma\}$ the set of those vertices). The precise γ value is chosen such that we can create a random bipartite regular graph on the remaining vertices with each individual having degree 2 and each test having degree $\Gamma-1$ (thus, an instance of $\mathscr{G}_{\Gamma-1}$). By a simple comparison of degrees, this is only possible if $m\ge 2\frac{n}{\Gamma+1}$. Now, we draw a uniformly random matching between the tests (of degree $\Gamma-1$) and the remaining individuals $x_1\dots x_\gamma$. By definition, each of those individuals takes part in exactly one test.

In both cases above, $\mathscr{G}^*_\Gamma$ is an almost-regular bipartite graph with each test compromising at most Γ individuals.

2.3. **Choice of recovery algorithm.** We make use of the definite defectives (DD) algorithm [2], which is described as follows.

1 Declare every individual x that appears in a negative test as non-infected; remove all such individuals.
2 Declare all individuals that are now the sole individual in a (positive) test as infected.
3 Declare all remaining individuals as uninfected.

Algorithm 1: The DD algorithm as defined by [2].

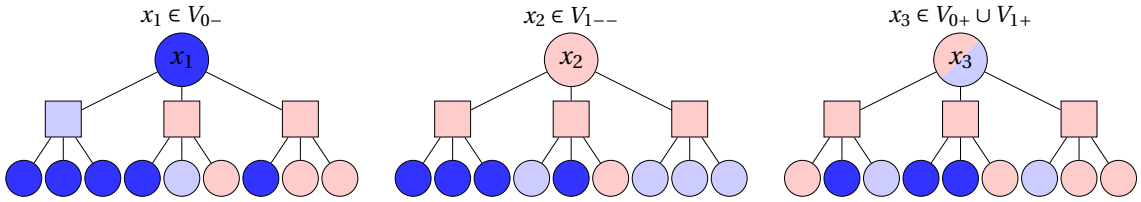

FIGURE 1. Rectangles represent tests and circles individuals. Dark blue individuals are elements of V_{0-} and can be easily identified as uninfected. Light blue individuals are elements of V_{0+}, and even if uninfected themselves, they only appear in positive tests and might be hard to identify. Infected individuals (red) that appear only in such tests are impossible to identify. Finally, infected individuals of V_{1--} appear in at least one test with only elements of V_{0-}. Thus, after identifying all elements of V_{0-}, they can be identified.

2.4. **The combinatorics behind group testing.** In this section, we introduce four types of individuals (see Figure 1) that might appear in any group testing instance and which the student can make use of. It turns out that the sizes of the sets of these individuals are the key to understand group testing combinatorially. Given a pooling scheme $\mathscr{G}$, let

$$V_0(\mathscr{G})=\{x\in V(\mathscr{G}):\boldsymbol{\sigma}_x=0\}\qquad\text{and}\qquad V_1(\mathscr{G})=\{x\in V(\mathscr{G}):\boldsymbol{\sigma}_x=1\}$$

be the uninfected and infected individuals, respectively. Then we can define *easy uninfected* individuals to be the uninfected individuals that appear in a negative test – clearly, they can easily be identified. We will call the set of such individuals V_{0-}; formally,

$$V_{0-}(\mathscr{G})=\{x\in V_0(\mathscr{G}):\exists a\in\partial_{\mathscr{G}}x:\hat{\boldsymbol{\sigma}}_a=0\}\,. \tag{2.2}$$

Then, there the *easy infected* individuals. These are those infected individuals that appear in at least one test with only easy uninfected individuals. Thus, upon removing the easy uninfected individuals, there will be at least one positive test with exactly one undeclared individual, and this individual has to be infected. We call this set

$$V_{1--}(\mathscr{G})=\{x\in V_1(\mathscr{G}):\exists a\in\partial_{\mathscr{G}}x:(\partial_{\mathscr{G}}a\setminus\{x\})\subset V_{0-}(\mathscr{G})\}\,. \tag{2.3}$$

Subsequently, there might be *disguised uninfected* individuals, that are uninfected themselves but only appear in positive tests. It is well known [3, 11, 12] that since the prior probability of being uninfected is very large, a group testing instance can tolerate a certain number of individuals of this type. Formally,

$$V_{0+}(\mathscr{G})=\{x\in V_0(\mathscr{G}):\forall a\in\partial_{\mathscr{G}}x:\hat{\boldsymbol{\sigma}}_a=1\}\,. \tag{2.4}$$

Finally, there might be *disguised infected* individuals, thus infected individuals appearing only in tests that contain at least one more infected individual. Formally,

$$V_{1+}(\mathcal{G}) = \{x \in V_1(\mathcal{G}) : \forall a \in \partial_{\mathcal{G}} x : (\partial_{\mathcal{G}} a \setminus \{x\}) \cap V_1(\mathcal{G}) \neq \emptyset\}. \tag{2.5}$$

2.4.1. *Remarks on information-theoretic and combinatorial bounds.* It turns out that in the sparse group testing problem - as well as in the unrestricted version [2, 12] - the non-adaptive information-theoretic phase transition comes into two installments. First, there are universal information-theoretic bounds, e.g., counting bounds, that account for the fact that a given number of tests can carry only a certain amount of information. Such bounds directly apply to the non-adaptive as well as the adaptive setting. Second, there are combinatorial / graph theoretical restrictions: Given that there exist a large number of totally disguised infected individuals (i.e., individuals such that in each of its tests there is a second infected individual), any non-adaptive algorithm fails with high (conditional) probability [11, 12]. This non-adaptivity gap becomes stronger if we increase the problem density θ, because for larger θ, the chance of finding multiple infected individuals in a small neighborhood increases as well. In this section we deal with the combinatorial part. In our setting, the transition where the combinatorial bound dominates the information-theoretic bound happens at $k \sim \sqrt{n}$, i.e., at the point where we find multiple infected individuals in a bounded neighborhood w.h.p..

2.5. **The Nishimori property.** Given a pooling scheme $\mathcal{G}$, a ground truth infection status vector $\boldsymbol{\sigma}$ (drawn uniformly from the vectors of Hamming weight k) and a sequence of test results $\hat{\boldsymbol{\sigma}}$, we denote by $S_k(\mathcal{G}, \boldsymbol{\sigma})$ the set of all colorings (i.e., infection status assignments) of individuals $\tau \in \{0,1\}^n$ that would have led to the test outcomes $\hat{\boldsymbol{\sigma}}$ (clearly including $\boldsymbol{\sigma}$ itself). Furthermore, we define $Z_k(\mathcal{G}, \boldsymbol{\sigma}) = |S_k(\mathcal{G}, \boldsymbol{\sigma})|$. The following proposition states that all sets in $S_k(\mathcal{G}, \boldsymbol{\sigma})$ are equally likely given the test outcomes.

Proposition 2.1. *[Corollary 2.1 of* [11]*] For all $\tau \in \{0,1\}^{x_1,\dots,x_n}$ we have*

$$\mathbb{P}(\boldsymbol{\sigma} = \tau | \mathcal{G}, \hat{\boldsymbol{\sigma}}) = \frac{\mathbb{1}\{\tau \in S_k(\mathcal{G}, \boldsymbol{\sigma})\}}{Z_k(\mathcal{G}, \boldsymbol{\sigma})}.$$

This immediately implies the following corollary.

Corollary 2.2. *If $Z_k(\mathcal{G}, \boldsymbol{\sigma}) \geq \ell$ w.h.p., then any inference algorithm has success probability at most ℓ^{-1}, when inferring $\boldsymbol{\sigma}$ from $(\mathcal{G}, \hat{\boldsymbol{\sigma}})$.*

The following claim and corollary will also be useful.

Claim 2.3. *For any test design, we have $Z_k(\mathcal{G}, \boldsymbol{\sigma}) \geq |V_{1+}(\mathcal{G})||V_{0+}(\mathcal{G})|$. Hence, conditioned on the sets $V_{1+}(\mathcal{G})$ and $V_{0+}(\mathcal{G})$, any inference algorithm fails with probability at least $1 - \frac{1}{|V_{1+}(\mathcal{G})||V_{0+}(\mathcal{G})|}$.*

Proof. The first statement is straightforward and was already given in [11, Fact 3.3], and the second statement follows directly from Corollary 2.2. □

Finally, we have the following well-known result on the DD algorithm.

Corollary 2.4. *The* DD *algorithm succeeds if and only if $V_1(\mathcal{G}) = V_{1--}(\mathcal{G})$.*

Proof. By definition, DD first classifies all $x \in V_{0-}(\mathcal{G})$ correctly. In the second step, DD classifies those individuals x as infected, which belong to a positive test a such that $\partial a \setminus \{x\} \subset V_{0-}(\mathcal{G})$. Thus, DD finds all $x \in V_1 \cap V_{1--}(\mathcal{G})$. As DD classifies the remaining individuals as uninfected, it fails as soon as there exists an individual $x \in V_1 \setminus V_{1--}(\mathcal{G})$. □

2.6. **The two-round exposure technique.** A key tool to deal with an arbitrary test design is to introduce certain levels of independent randomness. For example, the only randomness in $(\mathcal{G}, \boldsymbol{\sigma}^*)$ is the infection status of each individual, and instead of dealing with k infected individuals, we can assume that each individual is infected independently from all others with probability $p = \frac{k-\sqrt{k}\ln n}{n}$ (see Corollary 3.6). While we want to show that $V_{1+}(\mathcal{G}) \neq \emptyset$, we will establish this in two steps. We denote by $V_+(\mathcal{G})$ the set of *disguised* individuals, i.e., all tests containing this individual x contain at least one other individual (differing from x) that is infected, and hence

$$V_+(\mathcal{G}) = V_{1+}(\mathcal{G}) \cup V_{0+}(\mathcal{G}).$$

Once we find a large enough set $|V_+(\mathcal{G})| \gg n/k$, there will be some infected individuals in $V_+(\mathcal{G})$ w.h.p.. The main challenge is that in order to find the set of disguised individuals, one uses infected individuals, therefore the events $|V_+(\mathcal{G})|$ exceeding a specific size and infected individuals existing in $V_+(\mathcal{G})$ are not independent

in $(\mathcal{G}, \boldsymbol{\sigma}^*)$. This is where the two-round exposure technique, used very prominently in the study of random graphs [25], comes into account.

More specifically, our analysis will take the following steps in which individuals are randomly infected:

(1) We first mark each individual as infected with probability $\alpha k/n$ for some fixed constant $\alpha \in (0,1)$ and find a set $\mathcal{K}_1$ of infected individuals whose neighbourhood (the tests they belong to) has certain properties.
(2) Next, we mark the remaining individuals in the second neighbourhood of $\mathcal{K}_1$ (hence, we look at the individuals that are contained in the tests together with the vertices of $\mathcal{K}_1$) as infected with probability $(1-2\alpha)k/n$ for establishing the property of being disguised.
(3) Since each individual is infected with probability at most $\alpha k/n + (1-\alpha k/n)(1-2\alpha)k/n < p$, in the $\{0,1\}$-vector of infected individuals $\boldsymbol{\sigma}^*$ that we obtain each individual is infected with probability less than p. Adding further random infections we can ensure that each individual in $\boldsymbol{\sigma}^*$ is infected with probability exactly p. In addition, these extra infections can only enlarge (or keep unchanged) the set of individuals that are disguised.

3. Non-Adaptive Group Testing with Δ-Divisible Individuals

In this section, we formally state and prove our main results regarding non-adaptive group testing with Δ-divisible individuals.

3.1. **Model.**

As we highlighted earlier, optimal unconstrained designs are known that place each individual in $\Theta(\ln n)$ tests. Accordingly, we only consider the regime $\Delta = o(\ln n)$, and specifically suppose that $\Delta \le \ln^{1-\delta} n$ for some constant $\delta \in (0,1)$.

3.2. **Results.**

Define

$$m_{\inf}(\Delta) = \min\left\{\max\left\{e^{-1}\Delta k^{1+\frac{(1-\theta)}{\Delta\theta}}, \Delta k^{1+\frac{1}{\Delta}}\right\}, n\right\}, \qquad m_{\mathrm{DD}}(\Delta) = \max\left\{\Delta k^{1+\frac{(1-\theta)}{\Delta\theta}}, \Delta k^{1+\frac{1}{\Delta}}\right\}, \tag{3.1}$$

which will represent the information theoretic converse bound for any non-adaptive group testing scheme and the algorithmic barrier for DD, respectively.

In the following, we tacitly assume that $\theta + \theta\Delta^{-1} < 1$, or equivalently $\Delta > \theta/(1-\theta)$. If this inequality is reversed, then we find that $m_{\mathrm{DD}}(\Delta) = \omega(n)$, in which case one is better off resorting to individual testing.

Our first main result provides a simple counting-based converse bound for any adaptive or non-adaptive test design. This result, and all subsequent results, will be proved throughout the rest of the section.

Theorem 3.1. *Fix $\varepsilon \in (0,1)$, and suppose that $k = n^\theta$ with $\theta = (0,1)$ and $\Delta \in o(\ln(n))$. Then, if $m \le (1-\varepsilon)\mathrm{e}^{-1}\Delta k^{1+\frac{(1-\theta)}{\Delta\theta}}$ for fixed $\varepsilon > 0$, we have w.h.p. that any (possibly adaptive) group testing procedure that tests each individual at most Δ times fails to recover $\boldsymbol{\sigma}$.*

This bound recovers the first term of $\max\{\cdot,\cdot\}$ appearing in the definition of $m_{\inf}(\Delta)$ above, which is dominant for $\theta \le 1/2$. For the second term (which is dominant for $\theta \ge 1/2$), we require a more sophisticated argument that only holds for non-adaptive designs; as we will see in Section 5, adaptive designs can in fact go beyond this threshold.

Theorem 3.2. *Given any non-adaptive pooling scheme $\mathcal{G}$ where any individual gets tested at most Δ times (with $\theta/(1-\theta) < \Delta \le (\ln n)^{1-\delta}$ for some $\delta > 0$), if $m \le (1-\varepsilon)\Delta k^{1+1/\Delta}$ for some $\varepsilon \in (0,1)$, any algorithm (efficient or not) fails at inferring $\boldsymbol{\sigma}$ from $(\mathcal{G}_\Delta, \hat{\boldsymbol{\sigma}})$ with probability at least $\max\{\Omega(\varepsilon^2), 1 - O\big((1-\varepsilon/2)^\Delta\big)\}$.*

Combining these results, we find that any non-adaptive group testing strategy using at most $(1-\varepsilon)m_{\inf}(\Delta)$ tests fails w.h.p. if $\Delta = \omega(1)$, and fails with constant non-zero probability if $\Delta = O(1)$.

Next, we state our main upper bound, corresponding to the random regular design and the DD algorithm.

Theorem 3.3. *Suppose that $m = (1+\varepsilon)m_{\mathrm{DD}}(\Delta)$ for some $\varepsilon > 0$. Then, under the random regular design with parameter Δ, DD recovers $\boldsymbol{\sigma}$ from $(\mathcal{G}_\Delta, \hat{\boldsymbol{\sigma}})$ with probability at least $(1-o(k/n))\big(1-(1+\varepsilon)^{-\Delta}\big)$.*

Note that the success probability tends to one as $\Delta \to \infty$; if $\Delta = O(1)$ then we need to take $\varepsilon \to \infty$ for the probability to approach one (but it can be close to one for finite ε). Comparing this result with Theorem 3.1, we find that DD is asymptotically optimal for $\theta \ge 1/2$. On the other hand, a gap between $m_{\inf}(\Delta)$ and $m_{\mathrm{DD}}(\Delta)$ remains for $\theta < \frac{1}{2}$. In principle, this could be due to a weakness in the converse, a fundamental limitation of DD, or a weakness in our analysis of DD. However, the following theorem rules out the latter of these.

Theorem 3.4. *Let $\theta < 1/2$. Given the random regular pooling scheme $\mathcal{G}_\Delta$ on $m = (1-\varepsilon)m_{\mathrm{DD}}(\Delta)$ tests for fixed $\varepsilon \in (0,1)$, we have the following:*

(1) If $\Delta = \Theta(1)$*, then* DD *fails with positive probability bounded way from zero.*
(2) If $\Delta = (\ln n)^{1-\delta}$ *for* $\delta \in (0,1)$*, then* DD *fails w.h.p..*

Thus, Theorem 3.4 settles a coarse phase transition of DD in the random regular model when there are finitely many tests-per-individual, and a sharp phase transition when the number of tests-per-individual is diverging. We expect that DD is in fact provably suboptimal for $\theta < \frac{1}{2}$, but leave this as an open problem.

3.3. **Universal counting-based converse: Proof of Theorem 3.1.** We first prove a counting-based upper bound on the success probability for any test design and inference algorithm. Afterwards, we will use this bound on the success probability to prove our main converse bound, providing a lower bound on m for attaining a given target error probability.

Let $\mathscr{A}(\mathscr{G}, \hat{\boldsymbol{\sigma}}, k)$ be the output of a group testing inference algorithm with input $\mathscr{G}$ (pooling scheme), $\hat{\boldsymbol{\sigma}}$ (test results), and k (number of infected individuals). The inference algorithm is successful if $\mathscr{A}(\mathscr{G}, \hat{\boldsymbol{\sigma}}, k) = \boldsymbol{\sigma}$, and $\mathbb{P}(\mathscr{A}(\mathscr{G}, \hat{\boldsymbol{\sigma}}, k) = \boldsymbol{\sigma})$ is the success probability. We first prove the following non-asymptotic counting-based bound via a similar approach to [6] with suitable adjustments, and also using the Nishimori property similarly to [11].

Lemma 3.5. *Under the preceding setup, for any pooling scheme* $\mathscr{G}$ *and inference algorithm* $\mathscr{A}(\mathscr{G}, \hat{\boldsymbol{\sigma}}, k)$*, we have*

$$\mathbb{P}(\mathscr{A}(\mathscr{G}, \hat{\boldsymbol{\sigma}}, k) = \boldsymbol{\sigma}) \le \frac{\sum_{i=0}^{\Delta k} \binom{m}{i}}{\binom{n}{k}}. \tag{3.2}$$

Proof. Any given pooling scheme and inference algorithm can be viewed as a deterministic mapping from an infection status vector $\boldsymbol{\sigma} \in \{0,1\}^n$ to an outcome vector $\hat{\boldsymbol{\sigma}} \in \{0,1\}^m$. Recall that in Proposition 2.1, $S_k(\mathscr{G}, \boldsymbol{\sigma})$ is the set of all colorings of individuals that lead to the testing sequence $\hat{\boldsymbol{\sigma}}$, and $Z_k(\mathscr{G}, \boldsymbol{\sigma})$ is its cardinality. In the following, we additionally let $\hat{Z}_k(\mathscr{G}, \hat{\boldsymbol{\sigma}})$ denote $Z_k(\mathscr{G}, \boldsymbol{\sigma})$ when the test outcomes produced by $(\mathscr{G}, \boldsymbol{\sigma})$ are equal to $\hat{\boldsymbol{\sigma}}$, and let $\hat{S}_k(\mathscr{G}, \hat{\boldsymbol{\sigma}})$ be the set of all $\boldsymbol{\sigma}$ sequences that produce test outcomes $\hat{\boldsymbol{\sigma}}$.

Proposition 2.1 shows that the optimal inference algorithm outputs an arbitrary element of $S_k(\mathscr{G}, \boldsymbol{\sigma})$, and is correct with probability (conditioned on $\boldsymbol{\sigma}$) equal to $\frac{1}{Z_k(\mathscr{G}, \boldsymbol{\sigma})}$. Thus, averaging over the $\binom{n}{k}$ possible k-sparse vectors $\boldsymbol{\sigma}$, we have the following:

$$\begin{aligned}
\mathbb{P}(\mathscr{A}(\mathscr{G}, \hat{\boldsymbol{\sigma}}, k) = \boldsymbol{\sigma}) &= \frac{1}{\binom{n}{k}} \sum_{\boldsymbol{\sigma}} \frac{1}{Z_k(\mathscr{G}, \boldsymbol{\sigma})} \\
&= \frac{1}{\binom{n}{k}} \sum_{\hat{\boldsymbol{\sigma}} : \hat{Z}_k(\mathscr{G}, \hat{\boldsymbol{\sigma}}) \ge 1} \sum_{\boldsymbol{\sigma} \in \hat{S}_k(\mathscr{G}, \hat{\boldsymbol{\sigma}})} \frac{1}{\hat{Z}_k(\mathscr{G}, \hat{\boldsymbol{\sigma}})} \\
&\overset{(a)}{\le} \frac{|\{\hat{\boldsymbol{\sigma}} \in \{0,1\}^m : \hat{Z}_k(\mathscr{G}, \hat{\boldsymbol{\sigma}}) \ge 1\}|}{\binom{n}{k}} \\
&\overset{(b)}{\le} \frac{|\{\hat{\boldsymbol{\sigma}} \text{ with at most } \Delta k \text{ ones}\}|}{\binom{n}{k}} \\
&= \frac{\sum_{i=0}^{\Delta k} \binom{m}{i}}{\binom{n}{k}},
\end{aligned}$$

where (a) follows since there are $\hat{Z}_k(\mathscr{G}, \hat{\boldsymbol{\sigma}})$ terms in the second summation, thus canceling the $\frac{1}{\hat{Z}_k(\mathscr{G}, \hat{\boldsymbol{\sigma}})}$ term, and (b) uses the fact that at most Δk test outcomes can be positive, even in the adaptive setting; this is because adding another infected always introduces at most Δ additional positive tests. □

We now use the result in (3.2) to prove Theorem 3.1.

Proof of Theorem 3.1. It suffices to prove the claim for $m = (1-\delta) m_{\text{count}}(\Delta)$, since the inference algorithm could choose to ignore tests. We use the non-asymptotic bound in Lemma 3.5, and upper bound the sum of binomial coefficients via [5, Section 4.7.1] to obtain the following for a fixed target success probability of $1-\xi$ (for some $\xi \in (0,1)$):

$$\mathbb{P}(\mathscr{A}(\mathscr{G}, \hat{\boldsymbol{\sigma}}, k) = \boldsymbol{\sigma}) \le \frac{e^{mh(\frac{\Delta k}{m})}}{\binom{n}{k}} \equiv 1 - \xi, \tag{3.3}$$

where $h(\cdot)$ is the binary entropy function in nats (logs to base e). From (3.3), we have $e^{mh(\frac{\Delta k}{m})}/\binom{n}{k} = 1-\xi$, which implies that

$$\ln\left((1-\xi)\binom{n}{k}\right) = mh\left(\frac{\Delta k}{m}\right) = \Delta k \ln\frac{m}{\Delta k} + (m-\Delta k)\ln\frac{1}{1-\frac{\Delta k}{m}} \overset{(a)}{=} \Delta k \ln\frac{m}{\Delta k} + \Delta k(1+o(1)), \tag{3.4}$$

where (a) uses a Taylor expansion and the fact that $\frac{\Delta k}{m} \in o(1)$ (due to $\Delta = o(\ln n)$ and $m = (1-\delta)m_{\text{count}}(\Delta)$). Hence, we have $(1-\frac{\Delta k}{m})^{-1} = \exp(\frac{\Delta k}{m})(1+o(1))$ which is used to obtain the simplification. Rearranging (3.4), we obtain

$$\ln\frac{m}{\Delta k} = \frac{1}{\Delta k}\ln\left((1-\xi)\binom{n}{k}\right) - (1+o(1)),$$

which gives

$$m = e^{-(1+o(1))}\Delta k\left((1-\xi)\binom{n}{k}\right)^{\frac{1}{\Delta k}} \overset{(a)}{\geq} e^{-(1+o(1))}(1-\xi)^{\frac{1}{\Delta k}}\Delta k^{1+\frac{1-\theta}{\theta\Delta}}, \tag{3.5}$$

where (a) follows from the fact that $\binom{n}{k} \geq \left(\frac{n}{k}\right)^k$ and $k = n^\theta$.

The proof is completed by noting that $(1-\xi)^{1/(\Delta k)} \to 1$ for any fixed $\xi \in (0,1)$, since $k \to \infty$ by assumption. □

3.4. **Universal converse for non-adaptive designs: Proof of Theorem 3.2.** Let $\varepsilon, \theta, \delta \in (0,1)$, and $\theta/(1-\theta) \leq \Delta \leq \ln^{1-\delta} n$. Furthermore, let $\mathscr{G}$ be an arbitrary non-adaptive pooling scheme with $V(\mathscr{G})$ the set of n individuals and $F(\mathscr{G})$ the set of $m \leq (1-\varepsilon)\Delta k^{1+1/\Delta}$ test such that each individual is tested at most Δ times. Let

$$\bar{\ell} = \frac{1}{1-\varepsilon}k^{-\frac{1}{\Delta}} \qquad \text{and} \qquad \bar{\Gamma} = \frac{1}{m}\sum_{a\in F(\mathscr{G})}\Gamma_a = \frac{n\Delta}{m} \geq \bar{\ell}\frac{n}{k}. \tag{3.6}$$

Thus, $\bar{\Gamma}$ represents the average degree of the tests in $F(\mathscr{G})$, where Γ_a is the size of test a. We pick a set of k infected individuals uniformly at random and let $\boldsymbol{\sigma}$ be the $\{0,1\}$-vector representing them. We introduce $p = \frac{k-\sqrt{k}\ln n}{n}$ and $\boldsymbol{\sigma}^*$ as a binomial $\{0,1\}$-vector, such that each entry represents one individual and equals 1 with probability p independently of the others. The following result relates $\boldsymbol{\sigma}$ and $\boldsymbol{\sigma}^*$.

Corollary 3.6. *Under the preceding setup, for fixed $\varepsilon \in (0,1)$ and n large enough, we find that whenever there is a positive integer C such that*

$$\mathbb{P}\left(\left|V_{1+}(\mathscr{G},\boldsymbol{\sigma}^*)\right| > 2C\right) \geq 1-\varepsilon/4 \qquad \textit{and} \qquad \mathbb{P}\left(\left|V_{0+}(\mathscr{G},\boldsymbol{\sigma}^*)\right| > 2C\right) \geq 1-\varepsilon/4,$$

then it also holds that

$$\mathbb{P}\left(|V_{1+}(\mathscr{G},\boldsymbol{\sigma})| > C\right) \geq 1-\varepsilon \qquad \textit{and} \qquad \mathbb{P}\left(|V_{0+}(\mathscr{G},\boldsymbol{\sigma})| > C\right) \geq 1-\varepsilon.$$

Proof. The proof follows the lines of the proof of [12, Lemma 3.6]. Let $\mathscr{B}$ be the event that $|\boldsymbol{\sigma}^*| \in [k - 2\sqrt{k}\ln n, k]$. Then a standard application of the Chernoff bound guarantees that $\mathbb{P}(\mathscr{B}) = 1-o(1)$.

Given $\mathscr{B}$, couple $\boldsymbol{\sigma}^*$ and $\boldsymbol{\sigma}$ by flipping at most $2\sqrt{k}\ln n$ uninfected individuals in $\boldsymbol{\sigma}^*$ to infected, uniformly at random. Clearly, the size of totally disguised infected individuals can only increase, thus

$$\left|V_{1+}(\mathscr{G},\boldsymbol{\sigma}^*)\right| \leq |V_{1+}(\mathscr{G},\boldsymbol{\sigma})|.$$

However, it might happen that previously disguised uninfected individuals do now contribute to $|V_{1+}(\mathscr{G},\boldsymbol{\sigma})|$ instead of $|V_{0+}(\mathscr{G},\boldsymbol{\sigma})|$. Let

$$\boldsymbol{V} := \left||V_{0+}(\mathscr{G},\boldsymbol{\sigma})| - \left|V_{0+}(\mathscr{G},\boldsymbol{\sigma}^*)\right|\right|.$$

By the above coupling argument, we have

$$\mathbb{E}[\boldsymbol{V} \mid \mathscr{B}] \leq \frac{2\sqrt{k}\ln n}{n-k}\left|V_{0+}(\mathscr{G},\boldsymbol{\sigma}^*)\right| < n^{-(1-\theta)}\left|V_{0+}(\mathscr{G},\boldsymbol{\sigma}^*)\right|.$$

Therefore Markov's inequality implies

$$\mathbb{P}\left(|V_{0+}(\mathscr{G},\boldsymbol{\sigma})| < \left|V_{0+}(\mathscr{G},\boldsymbol{\sigma}^*)\right|/2 \mid \mathscr{B}\right) < \varepsilon/4.$$

□

Corollary 3.7. *Under the preceding setup, we have*

$$\mathbb{P}\left(\left|V_{1+}(\mathscr{G},\boldsymbol{\sigma}^*)\right| > 0\right) \geq \mathbb{P}\left(|V_{1+}(\mathscr{G},\boldsymbol{\sigma})| > 0\right) - o(1).$$

Furthermore, if $|V_{1+}(\mathcal{G},\boldsymbol{\sigma}^*)|>0$ *with probability* $\Omega(1)$*, then we have w.h.p. that* $|V_{0+}(\mathcal{G},\boldsymbol{\sigma}^*)|>2\ln n$*, and*

$$\mathbb{P}\left(|V_{0+}(\mathcal{G},\boldsymbol{\sigma})|>\ln n\right)=1-o(1).$$

Proof. The first part follows directly from the first part of the proof of Corollary 3.6. For the second part, we use the fact that the property of being totally disguised is independent of the infection status. Since we have assumed that $|V_{1+}(\mathcal{G},\boldsymbol{\sigma}^*)|>0$ with probability $\Omega(1)$, the expected number of totally disguised individuals is at least n/k, so the actual number is at least $2\ln n$ w.h.p. due to Markov's inequality. The final claim then follows from Corollary 3.6. □

By adopting the two-round exposure technique from Section 2.6, Theorem 3.2 will follow from the next lemma, which utilizes Corollary 3.7.

Lemma 3.8. *For any* $\varepsilon,\theta,\delta\in(0,1)$ *the following holds. Let* $\mathcal{G}$ *be a test design such that any of the* $n=V(\mathcal{G})$ *individuals is tested at most* Δ *times (with* $\theta/(1-\theta)<\Delta\le(\ln n)^{1-\delta}$*) and* $m=|F(\mathcal{G})|\le(1-\varepsilon)\Delta k^{1+1/\Delta}$*, where* $k=n^\theta$*. Then, we have with probability at least*

$$\left(1-\exp\left(-\varepsilon(1-\varepsilon/2)^{-\Delta}/64\right)\right)\frac{\varepsilon(1-\varepsilon/2)^{-\Delta}/16}{1+\varepsilon(1-\varepsilon/2)^{-\Delta}/16}\ge 1-O\left((1-\varepsilon/2)^{\Delta}\right) \tag{3.7}$$

that $|V_{1+}(\mathcal{G})|>0$.

Proof. We first show that $\mathcal{G}$ satisfies certain degree properties, namely, there cannot be any tests that are too small.

Claim 3.9. *For any fixed integer* D*, we can assume without loss of generality (for proving Lemma 3.8) that, for* n *large enough, every test has size at least* D*.*

Proof of Claim 3.9. We obtain an alternative design $\mathcal{G}'$ from $\mathcal{G}$ by iteratively deleting a test of size less than D and all individuals contained in the test, until all tests have size at least D. In each step, we remove one test, between one and D individuals, and at most ΔD edges. Without loss of generality, assume that in $\mathcal{G}$ there are only $o(n)$ individuals that are not contained in a test (otherwise, the error probability would trivially tend to one). Therefore, the test-design $\mathcal{G}'$ contains at least $(1-o(1))n-m\Delta D=(1-o(1))n$ edges, and since the individual degree is still at most Δ, its number of individuals $n'=|V(\mathcal{G}')|$ satisfies $n'\ge(1-o(1))n/\Delta$. This lower bound on n' along with the assumption $\Delta\le(\ln n)^{1-\delta}$ additionally imply that $\Delta\le(\ln n')^{1-\delta/2}$ when n is sufficiently large.

As for the remaining number of tests $m'=|F(\mathcal{G}')|$, we claim that for all large enough n,

$$m'\le(1-\varepsilon)\Delta k^{1+1/\Delta}-(n-n')/D\le(1-\varepsilon/2)\Delta(n')^{\theta+\theta/\Delta}.$$

Indeed, the first inequality follows since $m\le(1-\varepsilon)\Delta k^{1+1/\Delta}$ and the fact that we delete at least one test per D deleted individuals. For the second inequality we distinguish to cases. First, noting that for $0<\zeta<1$ and $x>0$ the mapping $x\mapsto x^\zeta$ is concave, we have for $\zeta:=\theta+\theta/\Delta<1$, $(n-n')\ge\sqrt{n}$, and n large enough that

$$(n-n')/D\ge\Delta(n-n')^\zeta\ge\Delta\left(n^\zeta-(n')^\zeta\right)\ge(1-\varepsilon)\Delta\left(n^\zeta-(n')^\zeta\right),$$

yielding the second inequality in the above claim. If one the other hand $(n-n')<\sqrt{n}$, we get with Bernoulli's inequality and with n large enough that

$$(1-\varepsilon)\Delta n^\zeta<(1-\varepsilon)\Delta(n')^\zeta\cdot(1+\sqrt{n}/n')^\zeta\le(1-\varepsilon)\Delta(n')^\zeta\cdot(1+\zeta\sqrt{n}/n')\le(1-\varepsilon/2)\Delta(n')^\zeta,$$

again yielding the second inequality in the above claim. Since $V_{1+}(\mathcal{G}')\subseteq V_{1+}(\mathcal{G})$, we can continue working with $\mathcal{G}'$ and the desired claim holds. □

Recall that in the multi-step argument in Section 2.6, for some $\alpha>0$, the first step is to infect each individual independently with probability $\alpha k/n$, and denote the resulting set of infected individuals by $\mathcal{K}_1$. We seek to characterize the number of disguised individuals in $\mathcal{K}_1$ following a second step of infections, in which each previously-uninfected individual is infected with probability $(1-2\alpha)k/n$. Given $\mathcal{K}_1$, let $\boldsymbol{X}_v^*$ be the probability that $v\in\mathcal{K}_1$ is totally disguised after this second step, and let $\boldsymbol{X}^*=\sum_{v\in\mathcal{K}_1}\boldsymbol{X}_v^*$. To prove that $\boldsymbol{X}^*$ is large, we need the following two statements.

Fact 3.10. *Let* a *be a test such that* $|\partial a\cap\mathcal{K}_1|\ge 2$*. Then any individual in* $\mathcal{K}_1$ *is totally disguised if and only if it is totally disguised when removing the test* a*.*

This fact is immediate as any infected individual is disguised in a by definition. Furthermore, to get a handle on the subtle dependencies between overlapping tests, we prove that the probability for an individual to be disguised in two tests is minimized when the tests are disjoint. For this, denote by $\partial^{(x)}a=\partial a\setminus\{x\}$ the individuals in test a without x.

Claim 3.11. *Consider marking each individual in $\partial^{(x)}a \cup \partial^{(x)}a'$ as infected with some probability q independent of the others. Then, for any integer $z > 0$, any individual $x \in V(\mathscr{G})$ and any two tests $a, a' \in \partial x$, we have*

$$\begin{aligned}&\mathbb{P}\left(\partial^{(x)}a \cap V_1(\mathscr{G}) \neq \emptyset, \partial^{(x)}a' \cap V_1(\mathscr{G}) \neq \emptyset \mid \partial^{(x)}a \cap \partial^{(x)}a' = \emptyset\right)\\ &\quad\le \mathbb{P}\left(\partial^{(x)}a \cap V_1(\mathscr{G}) \neq \emptyset, \partial^{(x)}a' \cap V_1(\mathscr{G}) \neq \emptyset \mid \left|\partial^{(x)}a \cap \partial^{(x)}a'\right| = z\right).\end{aligned}$$

Proof. We first note that

$$\mathbb{P}\left(\partial^{(x)}a \cap V_1(\mathscr{G}) \neq \emptyset, \partial^{(x)}a' \cap V_1(\mathscr{G}) \neq \emptyset \mid \partial^{(x)}a \cap \partial^{(x)}a' = \emptyset\right) = \left(1-(1-q)^{|\partial^{(x)}a|}\right)\left(1-(1-q)^{|\partial^{(x)}a'|}\right), \tag{3.8}$$

as the infected individuals in the two tests are independent due to the conditioning event.

On the other hand, suppose that $\left|\partial^{(x)}a \cap \partial^{(x)}a'\right| = z \ge 0$. In order to make both tests contain at least one infected individual that is not x, we can either have at least one of the z common individuals which is infected (happening with probability $\left(1-(1-q)^z\right)$, or we need both tests to contain an infected individual outside of the intersection. Hence,

$$\begin{aligned}&\mathbb{P}\left(\partial^{(x)}a \cap V_1(\mathscr{G}) \neq \emptyset, \partial^{(x)}a' \cap V_1(\mathscr{G}) \neq \emptyset \mid \left|\partial^{(x)}a \cap \partial^{(x)}a'\right| = z\right)\\ &\quad= \left(1-\left(1-q\right)^z\right) + \left(1-q\right)^z\left(1-(1-q)^{|\partial^{(x)}a|-z}\right)\left(1-(1-q)^{|\partial^{(x)}a'|-z}\right)\end{aligned} \tag{3.9}$$

Using (3.8) and (3.9), we conclude the proof with a short calculation:

$$\begin{aligned}&\mathbb{P}\left(\partial^{(x)}a \cap V_1(\mathscr{G}) \neq \emptyset, \partial^{(x)}a' \cap V_1(\mathscr{G}) \neq \emptyset \mid \left|\partial^{(x)}a \cap \partial^{(x)}a'\right| = z\right)\\ &\quad-\mathbb{P}\left(\partial^{(x)}a \cap V_1(\mathscr{G}) \neq \emptyset, \partial^{(x)}a' \cap V_1(\mathscr{G}) \neq \emptyset \mid \partial^{(x)}a \cap \partial^{(x)}a' = \emptyset\right)\\ &= \left(1-\left(1-q\right)^z\right) + \left(1-q\right)^z\left(1-(1-q)^{|\partial^{(x)}a|-z}\right)\left(1-(1-q)^{|\partial^{(x)}a'|-z}\right) - \left(1-(1-q)^{|\partial^{(x)}a|}\right)\left(1-(1-q)^{|\partial^{(x)}a'|}\right)\\ &= \left(1-\left(1-q\right)^z\right)\left(1-q\right)^{|\partial^{(x)}a|+|\partial^{(x)}a'|-z} \ge 0.\end{aligned}$$

□

With this in mind, we can consider a simplified model in which the test degrees are unchanged, but the *tests are all disjoint*.[5] More precisely, we define the following: Given an infection rate $q \in (0,1)$, we let $\boldsymbol{Y}_a = \boldsymbol{Y}_a(q) := \left(1-\left(1-q\right)^{\Gamma_a - 1}\right)$ be the probability that in a test a of size Γ_a with one fixed individual x, there is at least one infected individual that is not x. For any individual v, we then denote by $\boldsymbol{X}_v = \boldsymbol{X}_v(q) := \prod_{a\in\partial v}\boldsymbol{Y}_a(q)$ the probability that v is totally disguised in this model, where all tests are mutually disjoint. Observe that, by Claim 3.11, $\boldsymbol{X}^*_v \ge \boldsymbol{X}_v$, and therefore, $\boldsymbol{X}^* \ge \boldsymbol{X}$. The advantage is that in this model, $\boldsymbol{X}_v$ and $\boldsymbol{X}_u$ are independent for $v \neq u$. Recall that

$$\bar{\ell} = \frac{1}{1-\varepsilon}k^{-1/\Delta} = o(1),$$

because $\Delta = O(\ln^{1-\delta} n)$ and $k = n^\theta$, and let $\ell_a = \Gamma_a k/n$.

The following lemma provides a useful lower bound on $n^{-1}\sum_{v\in V(\mathscr{G})}\boldsymbol{X}_v$.

Claim 3.12. *Under the preceding setup with $q = (1-2\alpha)k/n$, we have*

$$n^{-1}\sum_{v\in V(\mathscr{G})}\boldsymbol{X}_v \ge (1-\exp(-(1-3\alpha)\bar{\ell}))^\Delta.$$

Proof. By the inequality of arithmetic and geometric means, we have

$$n^{-1}\sum_{v\in V(\mathscr{G})}\boldsymbol{X}_v \ge \prod_{v\in V(\mathscr{G})}\left(\prod_{a\in\partial v}\boldsymbol{Y}_a\right)^{1/n} = \prod_{a\in F(\mathscr{G})}\boldsymbol{Y}_a^{\Gamma_a/n}. \tag{3.10}$$

Furthermore, by Claim 3.9, we may assume that $\Gamma_a \ge (3\alpha)^{-1}$, and we deduce that

$$\boldsymbol{Y}_a \ge 1-\exp\left(-q\left(\Gamma_a-1\right)\right) \ge 1-\exp\left(-(1-3\alpha)\ell_a\right).$$

Hence, (3.10) yields

$$n^{-1}\sum_{v\in V(\mathscr{G})}\boldsymbol{X}_v \ge \prod_{a\in F(\mathscr{G})}\left(1-\exp(-(1-3\alpha)\ell_a)\right)^{\ell_a/k}. \tag{3.11}$$

[5]This implies an increase in the number of individuals, but the number of individuals does not play a role in this argument.

Using Jensen's inequality applied to the convex function $x \mapsto x\ln(1-\exp(-(1-3\alpha)x))$ on $(0,1)$, and noting that $\bar{\ell} = m^{-1}k\Delta = m^{-1}\sum_{a\in F(\mathscr{G})}\ell_a$, we get

$$\begin{aligned} k^{-1}\sum_{a\in F(\mathscr{G})}\ell_a\ln\big(1-\exp(-(1-3\alpha)\ell_a)\big) &= mk^{-1}\sum_{a\in F(\mathscr{G})}m^{-1}\big(\ell_a\ln\big(1-\exp(-(1-3\alpha)\ell_a)\big)\big) \\ &\geq mk^{-1}\left(\sum_{a\in F(\mathscr{G})}m^{-1}\ell_a\right)\ln\left(1-\exp\left(-(1-3\alpha)m^{-1}\sum_{a\in F(\mathscr{G})}\ell_a\right)\right) = \Delta\ln\big(1-\exp\big(-(1-3\alpha)\bar{\ell}\big)\big). \end{aligned} \tag{3.12}$$

Finally, the assertion of the claim follows from (3.11) and (3.12). □

We note from this claim that if we select an individual v uniformly at random, we have (also using $\bar{\ell} = o(1)$) that

$$\mathbb{E}[\boldsymbol{X}_v] \geq \big(1-\exp\big(-(1-3\alpha)\bar{\ell}\big)\big)^{\Delta} \geq (1-4\alpha)^{\Delta}\bar{\ell}^{\Delta} = \frac{(1-4\alpha)^{\Delta}}{(1-\varepsilon)^{\Delta}k} \geq (1-\varepsilon/2)^{-\Delta}k^{-1},$$

provided that $\alpha \leq \varepsilon/8$.

Now, recall that $\boldsymbol{X} = \sum_{v\in\mathscr{K}_1}\boldsymbol{X}_v$, and that each individual v is in $\mathscr{K}_1$ with probability $\alpha k/n$. Then we deduce from above that

$$\mathbb{E}[\boldsymbol{X}] = \alpha k\mathbb{E}[\boldsymbol{X}_v] \geq \alpha k(1-4\alpha)^{\Delta}\bar{\ell}^{\Delta}.$$

As $\boldsymbol{X}_v$ and $\boldsymbol{X}_u$ are independent for $v \neq u$, we can apply Lemma 7.2 to obtain

$$\mathbb{P}\big(\boldsymbol{X} < \alpha(1-\varepsilon/2)^{-\Delta}/2\big) \leq \exp(-\alpha(1-\varepsilon/2)^{-\Delta}/8). \tag{3.13}$$

Now, as described earlier, consider infecting any uninfected individual with probability $q = (1-2\alpha)k/n$ independent of all the others. Then, as $\sum_{v\in\mathscr{K}_1}\mathbb{P}(v\in V_{1+}(\mathscr{G})) = \boldsymbol{X}^* \geq \boldsymbol{X}$, we find that conditioned on $\mathscr{K}_1$ and $\boldsymbol{X}$, it holds with probability at least

$$1-\prod_{v\in\mathscr{K}_1}(1-\mathbb{P}(v\in V_{1+}(\mathscr{G}))) \geq 1-\left(1-\frac{\boldsymbol{X}}{|\mathscr{K}_1|}\right)^{|\mathscr{K}_1|} \geq \frac{\boldsymbol{X}}{1+\boldsymbol{X}}$$

that at least one individual from $\mathscr{K}_1$ is totally disguised. Here we used the inequality of arithmetic and geometric means to upper bound the product, and used Bernoulli's inequality in the last step. With $\alpha = \varepsilon/8$ and the upper bound (3.13) on the probability that $\boldsymbol{X} < \varepsilon(1-\varepsilon/2)^{-\Delta}/16$, it follows that there exists a totally disguised individual in $\mathscr{K}_1$ with probability at least

$$(1-\exp(-\varepsilon(1-\varepsilon/2)^{-\Delta}/64))\frac{\boldsymbol{X}}{1+\boldsymbol{X}},$$

which matches the statement of Lemma 3.8.

Recall that $p = \frac{k-\sqrt{k}\ln n}{n}$, and note that any individual is infected with probability at most

$$\tilde{p} = \alpha k/n + (1-\alpha k/n)(1-2\alpha)k/n < p,$$

independent of all the others. As discussed in Section 2.6 we can in hindsight raise the infection probability of each individual to p, which can only increase the size of the set $V_{1+}(\mathscr{G})$. Therefore, we proved that in a model where each individual gets infected independently with probability p the assertion of Lemma 3.8 holds. Applying Corollary 3.7, this argument completes the proof of Lemma 3.8. □

Proof of Theorem 3.2. The theorem now follows easily by combining Lemma 3.8 with the last part of Corollary 3.7: With at least one disguised infected individual and at least $\ln n$ disguised uninfected individuals, the conditional error probability is $1-o(1)$ due to Claim 2.3.

As for the overall success probability, the $1-O\big((1-\varepsilon/2)^{\Delta}\big)$ term is immediate from Lemma 3.8, and is dominant when Δ grows large for fixed ε. However, due to the hidden constant, this expression may be negative when Δ is small. For such cases, we simply observe that that the left hand side of (3.7) is $\Omega(\varepsilon^2)$ even when $\Delta = 1$, due to the Taylor expansion $1-\exp(-x) = x+\Theta(x^2)$ as $x\to 0$. □

3.5. **Algorithmic achievability on the random regular model: Proof of Theorem 3.3.**

3.5.1. *Further notation.* Recall the random regular model $\mathscr{G}_{\Delta}$ from Section 2.2.1. We let $(\boldsymbol{\Gamma}_1,\dots,\boldsymbol{\Gamma}_m)$ be the (random) sequence of test-degrees, which satisfies the following by construction:

$$\sum_{i=1}^{m}\boldsymbol{\Gamma}_i = n\Delta. \tag{3.14}$$

Furthermore, given the sequence $(\boldsymbol{\Gamma}_i)_{i\in[m]}$, we define d

$$\Gamma_{\min} = \min_{i\in[m]}\boldsymbol{\Gamma}_i, \qquad \bar{\Gamma} = \frac{1}{m}\sum_{i=1}^{m}\boldsymbol{\Gamma}_i \qquad \text{and} \qquad \Gamma_{\max} = \max_{i\in[m]}\boldsymbol{\Gamma}_i.$$

We stress at this point that the construction of $\mathcal{G}_\Delta$ allows for multi-edges, and hence one individual might take part in a test multiple times and contribute more than one to its degree.

Moreover, we parametrise the average degree as $\bar{\Gamma} = \ell n/k$, such that ℓ denotes the expected number of infected individuals a test would contain in a binomial random bipartite graph. Note that the assumed scaling $\Delta = o(\ln n)$ readily implies that $\omega(k/n) \le \ell \le o(1)$.

We first argue that each test degree is tightly concentrated with high probability, defining the concentration event $\mathcal{C}_\Gamma$ as follows:

$$\mathcal{C}_\Gamma = \left\{\left(1 - O\left(n^{-\Omega(1)}\right)\right)\ell n/k \le \Gamma_{\min} \le \bar{\Gamma} \le \Gamma_{\max} \le \left(1 + O\left(n^{-\Omega(1)}\right)\right)\ell n/k\right\}. \tag{3.15}$$

Lemma 3.13. *If* $\ell = \Omega\left(\min\left\{n^{-(1-\theta)/\Delta}, n^{-\theta/\Delta}\right\}\right)$*, then we have* $\mathbb{P}(\mathcal{C}_\Gamma(\theta)) = 1 - \tilde{O}(n^{-3})$.

Proof. Each individual chooses Δ tests with replacement. Hence, each individual has the chance of picking a given test Δ times independently, yielding

$$\boldsymbol{\Gamma}_i = \sum_{j=1}^{n}\sum_{h=1}^{\Delta} \mathbf{1}\left\{x_j \text{ chooses } a_i \text{ in } h\text{-th selection}\right\} \qquad \text{and} \qquad \boldsymbol{\Gamma}_i \sim \mathrm{Bin}\left(n\Delta, 1/m\right).$$

Thus, we find $\mathbb{E}[\boldsymbol{\Gamma}_i] = \ell n/k$.

As $\Delta \ge \frac{\theta}{1-\theta}$ and therefore $\ell = \omega(n^{-(1-\theta)})$, we find by the Chernoff bound (Lemma 7.1) that

$$\mathbb{P}\left(\boldsymbol{\Gamma}_i < (1-t)\ell n/k\right) \le \exp\left(-t^2 \ell n^{1-\theta}/2\right) \le \exp\left(-\Omega\left(t^2 n^{(1-1/\Delta)(1-\theta)/2}\right)\right).$$

Hence, with $t = n^{-(1-\theta)/4}\ln n$, we find

$$\mathbb{P}\left(\boldsymbol{\Gamma}_i < (1-t)\ell n/k\right) = \tilde{O}(n^{-4}). \tag{3.16}$$

An analogous calculation shows

$$\mathbb{P}\left(\boldsymbol{\Gamma}_i > (1+t)\ell n/k\right) = \tilde{O}(n^{-4}). \tag{3.17}$$

Therefore, the lemma follows from (3.16), (3.17), and a union bound over all $m \le n$ tests. □

3.5.2. *Analysis of the different types of individuals.* Let $\boldsymbol{Y}_i$ denote the number of infected individuals (including all multi-edges) in test a_i (for $i = 1 \dots m$). These variables are not mutually independent, as a single individual takes part in multiple tests. Luckily, it turns out that the family of the $\boldsymbol{Y}_i$ can be approximated by a family of mutually independent random variables quite well. Given $\boldsymbol{\Gamma}_1 \dots \boldsymbol{\Gamma}_m$, let $(\boldsymbol{X}_i)_{i\in[m]}$ be a sequence of mutually independent $\mathrm{Bin}\left(\boldsymbol{\Gamma}_i, k/n\right)$ variables. Furthermore, let

$$\mathcal{E}_\Delta = \left\{\sum_{i=1}^{m} \boldsymbol{X}_i = k\Delta\right\} \tag{3.18}$$

be the event that the sequence $(\boldsymbol{X}_i)$ renders the correct number of infected individuals. Stirling's approximation (Lemma 7.6) guarantees that $\mathcal{E}_\Delta$ is not too unlikely; specifically, $\mathbb{P}(\mathcal{E}_\Delta) = \Omega((n\Delta)^{-1/2})$. Furthermore, the $\boldsymbol{X}_i$ are indeed a good local approximation to the correct distribution, as stated in the following.

Lemma 3.14 (Lemma 3.6 of [11])**.** *The sequences* $(\boldsymbol{Y}_i)_{i\in[m]}$ *and* $(\boldsymbol{X}_i)_{i\in[m]}$ *given* $\mathcal{E}_\Delta$ *are identically distributed.* □

Next, we establish that the number of negative tests $\boldsymbol{m}_0 = \boldsymbol{m}_0(\mathcal{G}_\Delta, \boldsymbol{\sigma})$ and the number of positive tests $\boldsymbol{m}_1 = m - \boldsymbol{m}_0$ are highly concentrated.

Lemma 3.15. *With probability at least* $1 - o(n^{-2})$ *we have*

$$\boldsymbol{m}_0 = \left(1 + O\left(n^{-\Omega(1)}\right)\right) m\exp(-\ell) \qquad \text{and} \qquad \boldsymbol{m}_1 = \left(1 + O\left(n^{-\Omega(1)}\right)\right) m\left(1 - \exp(-\ell)\right).$$

Proof. Let $\boldsymbol{m}_0' = \left|\left\{(\boldsymbol{X}_i)_{i\in[m]} : \boldsymbol{X}_i = 0\right\}\right|$. Combining the definition of $\boldsymbol{X}_i$ with (7.6), we get

$$\mathbb{E}\left[\boldsymbol{m}_0' \mid \mathcal{E}_\Delta, (\boldsymbol{\Gamma}_i)_i\right] = \sum_{i=1}^{m} \mathbb{P}\left(\boldsymbol{X}_i = 0 \mid \boldsymbol{\Gamma}_i\right) = \sum_{i=1}^{m} (1 - k/n)^{\boldsymbol{\Gamma}_i},$$

which represents the expected number of negative tests approximated through $(\boldsymbol{X}_i)_i$. By Lemma 3.13 and a second order Taylor expansion (Lemma 7.5), we find

$$\mathbb{E}\left[\boldsymbol{m}_0' \mid \mathcal{E}_\Delta, \mathcal{C}_\Gamma\right] = \left(1 + O\left(n^{-\Omega(1)}\right)\right) m\exp(-\ell). \tag{3.19}$$

Then, the Chernoff bound implies

$$\mathbb{P}\left(\left|\boldsymbol{m}_0' - \mathbb{E}\left[\boldsymbol{m}_0' \mid \mathcal{E}_\Delta, \mathcal{C}_\Gamma\right]\right| > \sqrt{m}\ln^2(n) \mid \mathcal{E}_\Delta, \mathcal{C}_\Gamma\right) = o(n^{-10}). \tag{3.20}$$

The assertion of the lemma follows from (3.19) , (3.20), Lemma 3.13, Lemma 3.14, and $\mathbb{P}(\mathscr{E}_\Delta) = \Omega((n\Delta)^{-1/2})$. □

Next, we provide a characterization of the size of $V_{0+}(\mathscr{G}_\Delta)$, i.e., the number of disguised uninfected individuals.

Lemma 3.16. *We have with probability at least* $1-\tilde{O}\left(n^{-1}\right)$ *that*

$$|V_{0+}(\mathscr{G}_\Delta)| = \left(1+O\left(n^{-\Omega(1)}\right)\right) n\left(1-\exp(-\ell)\right)^\Delta.$$

Proof. Without loss of generality, given $\boldsymbol{m}_1$ and $\mathscr{C}_\Gamma$, we suppose that tests $a_1 \dots a_{\boldsymbol{m}_1}$ are the positive tests. By Lemma 3.13 and Lemma 3.15, the total number of edges connected to a positive test is w.h.p. given by

$$\sum_{i=1}^{\boldsymbol{m}_1} \boldsymbol{\Gamma}_i = \left(1+O\left(n^{-\Omega(1)}\right)\right) m\bar{\Gamma}\left(1-\exp(-\ell)\right). \tag{3.21}$$

We need to calculate the probability that a given uninfected individual belongs to $V_{0+}(\mathscr{G}_\Delta)$, i.e., each of its Δ edges is connected to a positive test. By a counting argument, we have

$$\mathbb{P}_{\mathscr{G}_\Delta}\left(x \in V_{0+}(\mathscr{G}_\Delta) \mid x \in V_0(\mathscr{G}_\Delta), \boldsymbol{m}_1, \mathscr{C}_\Gamma, (\boldsymbol{\Gamma}_i)_i\right) = \binom{\sum_{i=1}^{\boldsymbol{m}_1}\boldsymbol{\Gamma}_i}{\Delta}\binom{\sum_{i=1}^{m}\boldsymbol{\Gamma}_i}{\Delta}^{-1} = \left(1+O\left(n^{-\Omega(1)}\right)\right)\left(1-\exp(-\ell)\right)^\Delta,$$

where the simplification follows via Claim 7.4 along with (3.21), $\sum_{i=1}^m \boldsymbol{\Gamma}_i = m\bar{\Gamma}$ and $\Delta \le \ln^{1-\delta} n$.

Therefore,

$$\mathbb{E}_{\mathscr{G}_\Delta}\left[|V_{0+}(\mathscr{G}_\Delta)| \mid \mathscr{C}_\Gamma\right] = \left(1+O\left(n^{-\Omega(1)}\right)\right) n\left(1-\exp(-\ell)\right)^\Delta. \tag{3.22}$$

Analogously, again by Claim 7.4, the second moment turns out to be

$$\begin{aligned}\mathbb{E}_{\mathscr{G}_\Delta}\left[|V_{0+}(\mathscr{G}_\Delta)|^2 \mid \mathscr{C}_\Gamma\right] &= \binom{n-k}{2}\binom{\left(1+O\left(n^{-\Omega(1)}\right)\right) m\bar{\Gamma}\left(1-\exp(-\ell)\right)}{2\Delta}\binom{\left(1+O\left(n^{-\Omega(1)}\right)\right) m\bar{\Gamma}}{2\Delta}^{-1} \\ &= \left(1+O\left(n^{-\Omega(1)}\right)\right) n^2\left(1-\exp(-\ell)\right)^{2\Delta}. \end{aligned} \tag{3.23}$$

Thus, (3.22), (3.23) and Chebychev's inequality imply the lemma. □

Let $\boldsymbol{A}$ denote the number of infected individuals that do not belong to the easy uninfected set $V_{1--}(\mathscr{G}_\Delta)$. The following lemma allows us to bound its size.

Lemma 3.17. *Let* $m = (1+\varepsilon)m_{\mathrm{DD}}(\Delta)$*, then* $\mathbb{E}_{\mathscr{G}_\Delta}[\boldsymbol{A}] = (1+\varepsilon)^{-\Delta} + o(n^{-(1-\theta)})$.

Proof. Recall the definition

$$\bar{\Gamma} = n\Delta/m = \ell n/k.$$

Hence, $\ell = \frac{\Delta k}{m}$, and by the definition of $m_{\mathrm{DD}}(\Delta)$ and $k = n^\theta$, we get

$$\ell = \min\left\{(1+\varepsilon)^{-1} n^{-(1-\theta)/\Delta}, (1+\varepsilon)^{-1} k^{-1/\Delta}\right\}.$$

We can then distinguish two cases, depending on the sparsity level θ:

$$\ell = \begin{cases} (1+\varepsilon)^{-1} n^{-(1-\theta)/\Delta}, & \text{if} \quad \theta \le 1/2 \\ (1+\varepsilon)^{-1} k^{-1/\Delta}, & \text{if} \quad \theta > 1/2. \end{cases}$$

Recall that $\boldsymbol{m}_1$ is the number of positive tests, and define

$$\mathscr{F}_\Delta = \left\{\boldsymbol{m}_1 = \left(1+O\left(n^{-\Omega(1)}\right)\right) m\left(1-\exp(-\ell)\right)\right\} \cap \left\{|V_{0+}(\mathscr{G}_\Delta)| = \left(1+O\left(n^{-\Omega(1)}\ln^2 n\right)\right) n\left(1-\exp(-\ell)\right)^\Delta\right\} \tag{3.24}$$

as the event that both the number of positive tests as well as the size of $V_{0+}(\mathscr{G}_\Delta)$ behave as expected. Lemmas 3.15 and 3.16 guarantee that $\mathscr{F}_\Delta$ is a high probability event, namely, $\mathbb{P}\{\mathscr{F}_\Delta\} \ge 1-\tilde{O}(n^{-1})$. Given $\boldsymbol{m}_1$, we suppose without loss of generality that $a_1 \dots a_{\boldsymbol{m}_1}$ are the tests rendering a positive result.

We describe the number of occurrences of different types of individuals by introducing two sequences of random variables. Define $\boldsymbol{R}_i = (\boldsymbol{R}_i^1, \boldsymbol{R}_i^{0+}, \boldsymbol{R}_i^{0-})_{i\in[\boldsymbol{m}_1]}$ as the number of infected individuals, disguised uninfected individuals of $V_{0+}(\mathscr{G}_\Delta)$ and non-disguised uninfected individuals (those of $V_{0-}(\mathscr{G}_\Delta)$) appearing in test i, respectively. By definition, we find $\boldsymbol{R}_i^{0-} = \Gamma_i - \boldsymbol{R}_i^{0+} - \boldsymbol{R}_i^1$.

Given $|V_{0+}(\mathscr{G}_\Delta)|$, we approximate these variables by a sequence of mutually independent multinomials. Specifically, let

$$\boldsymbol{H}_i = (\boldsymbol{H}_i^1, \boldsymbol{H}_i^{0+}, \boldsymbol{H}_i^{0-})_{i\in[\boldsymbol{m}_1]} \sim \mathrm{Mult}_{\ge(1,0,0)}\left(\boldsymbol{\Gamma}_i, (k/n, |V_{0+}(\mathscr{G}_\Delta)|/n, 1-k/n-|V_{0+}(\mathscr{G}_\Delta)|/n)\right)$$

be multinomial conditioned on the first coordinate being at least one. We introduce the event

$$\mathscr{D}_\Delta = \left\{ \sum_{i=1}^{\boldsymbol{m}_1} \boldsymbol{H}_i^1 = k\Delta, \quad \sum_{i=1}^{\boldsymbol{m}_1} \boldsymbol{H}_i^{0+} = |V_{0+}(\mathscr{G}_\Delta)|\,\Delta \right\}.$$

Claim 3.18. *Given $\boldsymbol{m}_1$ and $|V_{0+}(\mathscr{G}_\Delta)|$, the distribution of $\boldsymbol{R}_i$ equals the distribution of $\boldsymbol{H}_i$ given $\mathscr{D}_\Delta$. Furthermore, $\mathbb{P}(\mathscr{D}_\Delta) \geq \Omega(n^{-2})$.*

Proof. Let $(r_i)_{i\in[\boldsymbol{m}_1]}$ be a sequence with $r_i = (r_i^1, r_i^{0+}, r_i^{0-})$ satisfying

$$S_1 := \sum_{i=1}^{\boldsymbol{m}_1} r_i^1 = k\Delta, \qquad S_{0+} := \sum_{i=1}^{\boldsymbol{m}_1} r_i^{0+} = |V_{0+}(\mathscr{G}_\Delta)|\,\Delta \qquad \text{and} \qquad r_i^{0-} = \Gamma_i - r_i^1 - r_i^{0+}.$$

In addition, let

$$S_{0-} := \sum_{i=1}^{\boldsymbol{m}_1} r_i^{0-}$$

denote the number of connections from individuals in $V_{0-}(\mathscr{G}_\Delta)$ to positive tests. Then, a counting argument gives

$$\begin{aligned}\mathbb{P}_{\mathscr{G}_\Delta}(\forall i \in [\boldsymbol{m}_1] : \boldsymbol{R}_i = r_i \mid \boldsymbol{m}_1, |V_{0+}(\mathscr{G}_\Delta)|) &= \frac{\binom{S_1}{r_1^1 \dots r_{\boldsymbol{m}_1}^1}\binom{S_{0+}}{v_1^{0+}\dots r_{\boldsymbol{m}_1}^{0+}}\binom{S_{0-}}{v_1^{0-}\dots r_{\boldsymbol{m}_1}^{0-}}}{\binom{S_1+S_{0+}+S_{0-}}{\Gamma_1,\dots,\Gamma_{\boldsymbol{m}_1}}} \\ &= \binom{S_1+S_{0+}+S_{0-}}{S_1, S_{0+}, S_{0-}}^{-1} \prod_{i=1}^{\boldsymbol{m}_1} \binom{\Gamma_i}{r_i^1, r_i^{0+}, r_i^{0-}}\end{aligned}$$

Let $(r_i^{'})_{i\in[\boldsymbol{m}_1]}$ be a second sequence as above. Then,

$$\frac{\mathbb{P}_{\mathscr{G}_\Delta}(\forall i \in [\boldsymbol{m}_1] : \boldsymbol{R}_i = r_i \mid \boldsymbol{m}_1)}{\mathbb{P}_{\mathscr{G}_\Delta}(\forall i \in [\boldsymbol{m}_1] : \boldsymbol{R}_i = r_i^{'} \mid \boldsymbol{m}_1)} = \prod_{i=1}^{\boldsymbol{m}_1} \frac{\binom{\Gamma_i}{r_i^1 r_i^{0+} r_i^{0-}}}{\binom{\Gamma_i}{r_i^{1'} r_i^{0+'} r_i^{0-'}}}. \tag{3.25}$$

Let

$$R_1 = \sum_{i=1}^{\boldsymbol{m}_1} r_i^1, \qquad R_+ = \sum_{i=1}^{\boldsymbol{m}_1} r_i^{0+}, \qquad \text{and} \qquad R_- = \sum_{i=1}^{\boldsymbol{m}_1} r_i^{0-}$$

and define R_1', R_+', R_-' analogously. By definition, we get

$$R_1 = R_1', \qquad R_+ = R_+' \qquad \text{and} \qquad R_- = R_-'.$$

Then, by the definition of $\boldsymbol{H}$, we have

$$\begin{aligned}&\frac{\mathbb{P}_{\mathscr{G}_\Delta}(\forall i \in [\boldsymbol{m}_1] : \boldsymbol{H}_i = r_i \mid \boldsymbol{m}_1, \mathscr{D}_\Delta)}{\mathbb{P}_{\mathscr{G}_\Delta}(\forall i \in [\boldsymbol{m}_1] : \boldsymbol{H}_i = r_i^{'} \mid \boldsymbol{m}_1, \mathscr{D}_\Delta)} \\ &= \frac{(k/n)^{R_1}(|V_{0+}(\mathscr{G}_\Delta)|/n)^{R_+}(1-k/n-|V_{0+}(\mathscr{G}_\Delta)|/n)^{R_-}}{(k/n)^{R_1'}(|V_{0+}(\mathscr{G}_\Delta)|/n)^{R_+'}(1-k/n-|V_{0+}(\mathscr{G}_\Delta)|/n)^{R_-'}} \cdot \prod_{i=1}^{\boldsymbol{m}_1} \frac{\binom{\Gamma_i}{r_i^1, r_i^{0+}, r_i^{0-}}}{\binom{\Gamma_i}{r_i'^1, r_i'^{0+}, r_i'^{0-}}} = \prod_{i=1}^{\boldsymbol{m}_1} \frac{\binom{\Gamma_i}{r_i^1, r_i^{0+}, r_i^{0-}}}{\binom{\Gamma_i}{r_i'^1, r_i'^{0+}, r_i'^{0-}}}.\end{aligned} \tag{3.26}$$

Thus, Claim 3.18 follows from (3.25) and (3.26). □

We now introduce a random variable that counts (positive) tests featuring only one infected individual and no disguised uninfected individuals. Formally, let

$$\boldsymbol{B} = \sum_{i=1}^{\boldsymbol{m}_1} \mathbf{1}\left\{\boldsymbol{R}_i^1 + \boldsymbol{R}_i^{0+} = 1\right\} \qquad \text{and} \qquad \boldsymbol{B}' = \sum_{i=1}^{\boldsymbol{m}_1} \mathbf{1}\left\{\boldsymbol{H}_i^1 + \boldsymbol{H}_i^{0+} = 1\right\}. \tag{3.27}$$

Therefore, by the definition of $\boldsymbol{H}_i$,

$$\mathbb{E}_{\mathscr{G}_\Delta}\left[\boldsymbol{B}' \mid \mathscr{D}_\Delta, (\boldsymbol{\Gamma}_i)_i, \boldsymbol{m}_1\right] = \sum_{i=1}^{\boldsymbol{m}_1} \binom{\boldsymbol{\Gamma_i}}{1, 0, \Gamma_i - 1} \frac{k/n(1-k/n-|V_{0+}(\mathscr{G}_\Delta)|/n)^{\boldsymbol{\Gamma}_i - 1}}{1-(1-k/n)^{\boldsymbol{\Gamma}_i}}. \tag{3.28}$$

Thus, as $\boldsymbol{\Gamma}_i$ is concentrated around $\bar{\Gamma}$ and $\boldsymbol{m}_1$ is concentrated around $m(1-\exp(-\ell))$, and also using the second-order approximation of $(1-k/n)^{\bar{\Gamma}} \sim \exp(-\ell)$, we find by incorporating the error terms of $\mathscr{E}_\Delta$ that

$$\mathbb{E}_{\mathscr{G}_\Delta}\left[\boldsymbol{B}' \mid \mathscr{D}_\Delta, \mathscr{E}_\Delta, \boldsymbol{m}_1\right] = \left(1 + O\left(n^{-\Omega(1)}\right)\right) \boldsymbol{m}_1 \bar{\Gamma} \frac{n^{-(1-\theta)}(1-n^{-(1-\theta)}-|V_{0+}(\mathscr{G}_\Delta)|/n)^{\bar{\Gamma}}}{1-\exp(-\ell)}$$

Hence, conditioning on the high probability event $\mathscr{F}_\Delta$ of (3.24) leads to

$$\mathbb{E}_{\mathscr{G}_\Delta}\left[\boldsymbol{B}' \mid \mathscr{D}_\Delta, \mathscr{E}_\Delta, \mathscr{F}_\Delta\right] = \left(1+O\left(n^{-\Omega(1)}\right)\right) k\Delta\left(1-\frac{k+|V_{0+}(\mathscr{G}_\Delta)|}{n}\right)^{\bar{\Gamma}}. \tag{3.29}$$

Now, let us distinguish between the cases $\theta \le 1/2$ and $\theta > 1/2$.

Case 1: $\theta > 1/2$: In this case, we have $n/k = o(k)$, and $\ell = (1+\varepsilon)^{-1}k^{-1/\Delta}$. We recall the event $\mathscr{F}_\Delta$ from (3.24) that gives a concentration condition for $|V_{0+}(\mathscr{G}_\Delta)|$ and $\boldsymbol{m}_1$. Plugging in ℓ into (3.24), we find given $\mathscr{F}_\Delta$, there is some $\gamma \in (0,1)$ such that

$$|V_{0+}(\mathscr{G}_\Delta)| \sim (1+\varepsilon)^{-\Delta} n/k = O\left(k^{1-\gamma}\right).$$

Hence, by (3.29)

$$\begin{aligned}\mathbb{E}_{\mathscr{G}_\Delta}\left[\boldsymbol{B}' \mid \mathscr{D}_\Delta, \mathscr{C}_\Gamma, \mathscr{F}_\Delta\right] &= \left(1+O\left(n^{-\Omega(1)}\right)\right) k\Delta\left(1-\frac{\left(1+O\left(n^{-\Omega(1)}\right)\right)k}{n}\right)^{\bar{\Gamma}} \\ &= \left(1+O\left(n^{-\Omega(1)}\right)\right) k\Delta\exp\left(-\ell\right) = \left(1+O\left(n^{-\Omega(1)}\right)\right) k\Delta(1-\ell).\end{aligned} \tag{3.30}$$

Now, $\boldsymbol{B}'$ is (with respect to the graph $\mathscr{G}_\Delta$) a binomial random variable with a random number of trials and a random probability parameter. Clearly, when conditioning on a specific number of trials and a specific probability, the conditioned variable in a binomial random variable. For the sake of brevity, we write such a conditioning as a conditioning on the concentration guaranteeing events $\mathscr{D}_\Delta, \mathscr{C}_\Gamma, \mathscr{F}_\Delta$, as the error terms account for the range of possible fixed parameters in the range of the events. Therefore, the Chernoff bound guarantees that given $\mathscr{D}_\Delta\mathscr{C}_\Gamma, \mathscr{F}_\Delta$ and with respect to $\mathscr{G}_\Delta$,

$$\boldsymbol{B}' = \left(1+O\left(n^{-\Omega(1)}\right)\right)\Delta k\cdot(1-(1+\varepsilon)^{-1}k^{-1/\Delta})$$

with probability at least $o(n^{-10})$. By Claim 3.18, it follows that

$$\boldsymbol{B} = \left(1+O\left(n^{-\Omega(1)}\right)\right)\Delta k\cdot(1-(1+\varepsilon)^{-1}k^{-1/\Delta}). \tag{3.31}$$

with probability at least $1-o(n^{-8})$. Thus, we can calculate the probability of an infected individual not belonging to $V_{1--}(\mathscr{G})$ as follows. Such an individual has to choose all of its Δ edges out of the $k\Delta-\boldsymbol{B}$ edges that would lead to a test in which the individual could be identified by DD. Hence, we have

$$\mathbb{P}\left(x \notin V_{1--}(\mathscr{G}_\Delta) \mid x \in V_1(\mathscr{G})\right) = (1+o(k/n))\binom{k\Delta-\boldsymbol{B}}{\Delta}\binom{k\Delta}{\Delta}^{-1} = (1+o(k/n))\left((1+\varepsilon)^{-1}k^{-1/\Delta}\right)^{\Delta}, \tag{3.32}$$

where the simplification holds w.h.p. using (3.31) and Claim 7.4. Interpreting the average of $\boldsymbol{A}$ as a sum of k probabilities, it follows that

$$\mathbb{E}_{\mathscr{G}_\Delta}[\boldsymbol{A}] \le (1+o(k/n))(1+\varepsilon)^{-\Delta}. \tag{3.33}$$

Case 2: $\theta \le 1/2$: In this case, we have $\ell = (1+\varepsilon)^{-1}n^{-(1-\theta)/\Delta}$. Hence, given $\mathscr{F}_\Delta$,

$$|V_{0+}(\mathscr{G}_\Delta)| = \left(1-O\left(n^{-\Omega(1)}\right)\right)k(1+\varepsilon)^{-\Delta}. \tag{3.34}$$

In contrast to the first case, here we find that the influence of the size of disguised non-infected individuals does not vanish asymptotically in relation to the number of infected individuals in (3.29).

By a similar argument as the first case, (3.34) and (3.29) imply

$$\begin{aligned}\mathbb{E}_{\mathscr{G}_\Delta}\left[\boldsymbol{B}' \mid \mathscr{D}_\Delta, \mathscr{C}_\Gamma, \boldsymbol{m}_1\right] &= \left(1+O\left(n^{-\Omega(1)}\right)\right)k\Delta\left(1-\frac{k}{n}\left(1-(1+\varepsilon)^{-\Delta}-O\left(\frac{n^{-\Omega(1)}}{(1+\varepsilon)^{\Delta}}\right)\right)\right)^{\bar{\Gamma}} \\ &= \left(1+O\left(n^{-\Omega(1)}\right)\right)k\Delta\exp\left(-\left(1-(1+\varepsilon)^{-\Delta}-O\left(\frac{n^{-\Omega(1)}}{(1+\varepsilon)^{\Delta}}\right)\right)\ell\right) \\ &= \left(1+O\left(n^{-\Omega(1)}\right)\right)\Delta k\left(1-\left(1-(1+\varepsilon)^{-\Delta}-O\left(n^{-\Omega(1)}(1+\varepsilon)^{-\Delta}\right)\right)\ell\right),\end{aligned} \tag{3.35}$$

and combining this with the Chernoff bound and Claim 3.18 yields

$$\boldsymbol{B} = \left(1+O\left(n^{-\Omega(1)}\right)\right)\Delta k\cdot(1-(1+\varepsilon)^{-1}n^{-(1-\theta)/\Delta}) \tag{3.36}$$

with probability $1-o(n^{-8})$. Therefore, the probability of an infected individual not belonging to $V_{1--}(\mathscr{G})$ satisfies the following analog of (3.32):

$$\mathbb{P}\left(x \notin V_{1--}(\mathscr{G}_\Delta) \mid x \in V_1(\mathscr{G})\right) = (1+o(k/n))\binom{k\Delta-\boldsymbol{B}}{\Delta}\binom{k\Delta}{\Delta}^{-1} = (1+o(k/n))(1+\varepsilon)^{-\Delta}n^{-(1-\theta)}.$$

Since $2\theta - 1 \le 0$ by assumption, it follows that

$$\mathbb{E}[\boldsymbol{A}] = (1+o(k/n))(1+\varepsilon)^{-\Delta} n^{\theta} n^{-(1-\theta)} \le (1+(k/n))(1+\varepsilon)^{-\Delta}. \tag{3.37}$$

Thus, Lemma 3.17 follows from (3.33), (3.37) and Markov's inequality. □

3.6. **A converse for DD in the sparse regime: Proof of Theorem 3.4.** In accordance with Corollary 2.4, we first provide a lemma bounding the size of $V_{1--}(\mathscr{G}_\Delta)$, the set of infected individuals appearing in at least one test with only easy uninfected individuals.

Lemma 3.19. *For $\theta < 1/2$ and $m = (1-\varepsilon)m_{\mathrm{DD}}(\Delta)$, we have under the random regular design that*

$$\mathbb{E}[|V_{1--}(\mathscr{G}_\Delta)|] = (1+o(1))k\Big(1-\big(1-\exp\big(-(1-\varepsilon)^{-\Delta}(1-1/\Delta)\big)\big)^{\Delta}\Big).$$

Proof. We re-use the notations $\bar{\ell}$ and $\bar{\Gamma}$ in (3.6), but their expressions are modified as follows in accordance with the choice $m = (1-\varepsilon)\Delta k^{1+\frac{(1-\theta)}{\Delta\theta}}$ associated with $\theta < \frac{1}{2}$:

$$\bar{\ell} = (1-\varepsilon)^{-1} n^{-(1-\theta)/\Delta} \qquad \text{and} \qquad \bar{\Gamma} = (1-\varepsilon)^{-1} n^{(1-\theta)(1-1/\Delta)}. \tag{3.38}$$

We additionally recall $\boldsymbol{B}$ from (3.27) as the number of tests featuring exactly one infected individual and no elements of V_{0+}. By the same calculation as in (3.35) and (3.36) with ℓ and $\bar{\Gamma}$ replaced by the values in (3.38), we obtain

$$\boldsymbol{B} = \big(1+O\big(n^{-\Omega(1)}\big)\big)\, k\Delta\big(1-(1-\varepsilon)^{-\Delta} k n^{-1}\big)^{\bar{\Gamma}} = \big(1+O\big(n^{-\Omega(1)}\big)\big)\, k\Delta \exp\big(-(1-\varepsilon)^{-\Delta}(1-1/\Delta)\big) \tag{3.39}$$

with probability at least $1-o(n^{-8})$. Therefore, we can calculate the probability that an infected individual does not belong to $V_{1--}(\mathscr{G}_\Delta)$ via Claim 7.4 as follows:

$$\mathbb{P}(x \notin V_{1--}(\mathscr{G}_\Delta) \mid x \in V_1(\mathscr{G})) = (1+o(1))\frac{\binom{k\Delta-\boldsymbol{B}}{\Delta}}{\binom{k\Delta}{\Delta}} = (1+o(1)\big(1-\exp\big(-(1-\varepsilon)^{-\Delta}(1-1/\Delta)\big)\big)^{\Delta}.$$

Since there are k individuals in $x \in V_1(\mathscr{G})$ by assumption, we obtain

$$\mathbb{E}[|V_1(\mathscr{G}) \setminus V_{1--}(\mathscr{G}_\Delta)|] = (1+o(1))k\big(1-\exp\big(-(1-\varepsilon)^{-\Delta}(1-1/\Delta)\big)\big)^{\Delta} \tag{3.40}$$

and the lemma follows using $|V_{1--}(\mathscr{G}_\Delta)| = k - |V_1(\mathscr{G}) \setminus V_{1--}(\mathscr{G}_\Delta)|$. □

Knowing the expected size of $|V_{1--}(\mathscr{G}_\Delta)|$, Markov's inequality leads to the following.

Corollary 3.20. *Let $\theta < 1/2$ and $m = (1-\varepsilon)m_{\mathrm{DD}}(\Delta)$ and $\Delta = \Theta(1)$. Then, with probability at least*

$$1 - \frac{1-\big(1-\exp\big(-(1-\varepsilon)^{-\Delta}(1-1/\Delta)\big)\big)^{\Delta}}{1-\gamma} \tag{3.41}$$

there are at least γk infected individuals $x \in V_1(\mathscr{G}) \setminus V_{1--}(\mathscr{G}_\Delta)$.

Corollaries 2.4 and 3.20 immediately imply Theorem 3.4, since (3.41) is always positive for sufficiently small γ, and approaches one as $\Delta \to \infty$.

4. Non-Adaptive Group Testing with Γ-Sized Tests

In this section, we formally state and prove our main results concerning non-adaptive group testing Γ-sized tests, namely, a universal lower bound and an algorithmic upper bound that matches the lower bound. Recall that we focus on the regime $\Gamma = \Theta(1)$. Within this section $\mathscr{G}$ denotes an arbitrary non-adaptive pooling scheme with respect to the Γ-sparsity constraint.

4.1. **A universal information-theoretic bound.** The first statement that we prove is an information-theoretic converse that applies to *any* non-adaptive group testing scheme with maximum test size Γ. Denote by

$$m_{\inf,\Gamma} = m_{\inf}(\mathscr{G}) = \max\left\{1 + \left\lfloor \frac{\theta}{1-\theta} \right\rfloor \frac{n}{\Gamma}, 2\frac{n}{\Gamma+1}\right\}, \tag{4.1}$$

which we will show to be the sharp information-theoretic phase transition point. In [20] a lower bound of $(n/\Gamma)(1+o(1))$ was proved, and we see that in the regime $\Gamma = \Theta(1)$, our lower bound improves on this for all $\theta \in (0,1)$.

Theorem 4.1. *Let $\delta > 0$ and $m = (1-\delta)m_{\inf,\Gamma}$. Furthermore, let $\mathscr{G}$ be any non-adaptive pooling scheme (deterministic or randomised) such that each test contains at most $\Gamma = \Theta(1)$ individuals. Then any inference algorithm $\mathscr{A}$ fails in recovering $\boldsymbol{\sigma}$ from $(\hat{\boldsymbol{\sigma}}, \mathscr{G})$*

- *with probability* $1-o(1)$ *if* $\theta/(1-\theta) \notin \mathbb{Z}$,
- *with probability* $\Omega(1)$ *if* $\theta/(1-\theta) \in \mathbb{Z}$.

Thus, even with unlimited computational power, there cannot be any algorithm with a maximum test size of Γ that is able to infer the infected individuals correctly w.h.p. once the number of tests drops below (4.1). The distinction between integer vs. non-integer values of $\theta/(1-\theta)$ arises for technical reasons (e.g., counting the number of nodes with degree at most $\lfloor \theta/(1-\theta) \rfloor$), and we found it difficult to prove a high-probability (rather than constant-probability) failure result in the integer case.

The proof of the universal information-theoretic converse resembles the proof of [12] for the existence of a universal information-theoretic bound for unrestricted non-adaptive group testing, but several modifications are required to handle the test size constraint. Before conducting the proof in detail, we provide a short sketch.

4.2. **Proof of Theorem 4.1.** We start by defining

$$d^+ = 1 + \left\lfloor \frac{\theta}{1-\theta} \right\rfloor \qquad \text{and} \qquad d^- = \left\lfloor \frac{\theta}{1-\theta} \right\rfloor.$$

For the proof, we distinguish two different regimes for θ, as stated in Proposition 4.2 and Proposition 4.6. We start with the following proposition.

Proposition 4.2. *Let* $1/2 \le \theta < 1$*, and let* $\mathscr{G}$ *be an arbitrary pooling scheme with tests of size at most* Γ*. For all* $\delta > 0$ *and* n *sufficiently large, if* $m = (1-\delta)d^+\frac{n}{\Gamma}$*, then*

$$\begin{cases} \mathbb{P}\left(|V_{1+}(\mathscr{G})| > \ln n\right) \ge 1-o(1) & \text{and} \quad \mathbb{P}\left(|V_{0+}(\mathscr{G})| > \ln n\right) \ge 1-o(1), & \text{if } \frac{\theta}{1-\theta} \notin \mathbb{Z} \\ \mathbb{P}\left(|V_{1+}(\mathscr{G})| \ge 1\right) = \Omega(1) & \text{and} \quad \mathbb{P}\left(|V_{0+}(\mathscr{G})| > \ln n\right) \ge 1-o(1), & \text{if } \frac{\theta}{1-\theta} \in \mathbb{Z}. \end{cases}$$

4.2.1. *Proof of Proposition 4.2.* Let $\mathscr{G}$ be an arbitrary pooling scheme such that each test contains at most Γ individuals. We denote by $V(\mathscr{G})$ the set of individuals, and by $F(\mathscr{G})$ the set of tests in $\mathscr{G}$ (by the identification of $\mathscr{G}$ with a bipartite graph). Instead of analysing $(\mathscr{G}, \hat{\boldsymbol{\sigma}})$, similarly to in the Δ-divisible case, we analyse a very similar model that eliminates nuisance dependencies between the infection status of different individuals.

Specifically, let $p = \frac{k-\sqrt{k}\ln n}{n}$, and let $\boldsymbol{\sigma}^*$ be a $\{0,1\}$-valued vector, where every entry is one with probability p. Corollary 3.6 guarantees that if the modified model satisfies

$$\mathbb{P}\left(\left|V_{1+}(\mathscr{G}, \boldsymbol{\sigma}^*)\right| > 2C\right) \ge 1-o(1) \qquad \text{and} \qquad \mathbb{P}\left(\left|V_{0+}(\mathscr{G}, \boldsymbol{\sigma}^*)\right| > 2C\right) \ge 1-o(1),$$

then the original model satisfies

$$\mathbb{P}\left(|V_{1+}(\mathscr{G}, \boldsymbol{\sigma})| > C\right) \ge 1-o(1) \qquad \text{and} \qquad \mathbb{P}\left(|V_{0+}(\mathscr{G}, \boldsymbol{\sigma})| > C\right) \ge 1-o(1).$$

Thus, working with the modified model is sufficient.

For the sake of brevity, we henceforth write $\mathscr{G}_\Gamma$ in place of $(\mathscr{G}, \boldsymbol{\sigma}^*)$, leaving the dependencies on $\boldsymbol{\sigma}^*$ implicit.

We proceed by finding a set of (many) individuals, that have a high probability of being disguised. We will apply the probabilistic method iteratively to create the desired set. Creating this set turns out to be delicate due to the dependencies in an arbitrary pooling scheme. Luckily, it will suffice for our purposes to note that whenever individuals have distance at least 6 in the underlying graph, the events of being disguised are independent [12].

In the following, we denote the set of all disguised individuals by

$$V^+(\mathscr{G}) = V_{0+}(\mathscr{G}) \cup V_{1+}(\mathscr{G}).$$

We first present a claim establishing that we may safely assume that each individual gets tested $\Theta(1)$ times.

Claim 4.3. *Given any pooling scheme* $\mathscr{G}'$ *with* $m = (1-2\varepsilon)d^+\frac{n}{\Gamma}$ *(for some* $\varepsilon > 0$*) such that each test contains at most* $\Gamma = \Theta(1)$ *individuals, there is another pooling scheme* $\mathscr{G}$ *such that each test contains at most* $\Gamma = \Theta(1)$ *individuals with* $m = (1-\varepsilon)d^+\frac{n}{\Gamma}$*, while also satisfying the following:*

- *Each individual is contained in at most* $C = \Theta(1)$ *tests;*
- *Recovery of* $\boldsymbol{\sigma}$ *from* $(\mathscr{G}', \hat{\boldsymbol{\sigma}}')$ *implies recovery from* $(\mathscr{G}, \hat{\boldsymbol{\sigma}})$.

Proof. Given $\mathscr{G}'$ and a constant $C \in \mathbb{N}$, there is $C' \in \mathbb{N}$ such that there are at most n/C individuals of degree at least C' in $\mathscr{G}'$, which is an immediate consequence of m being linear in n (due to $\Gamma = \Theta(1)$). Design $\mathscr{G}$ such that each individual of $\mathscr{G}'$ with degree larger C' gets tested individually (causing n/C additional tests) and all other individuals and tests stay the same as under $\mathscr{G}'$. Clearly, if recovery in $\mathscr{G}'$ was possible, then it is possible in $\mathscr{G}$ as well. Setting $C = \frac{\Gamma}{\varepsilon d^+}$, the claim follows. □

In addition to the fact the we suppose that there are no individuals of (unbounded) large degree, we can prove that there cannot be too many individuals of low degree.

Lemma 4.4. *Let $\mathcal{G}$ be the given pooling scheme and $m \le (1-\delta)d^+\frac{n}{\Gamma}$. If there is a constant $\alpha > 0$ such that the number of individuals of degree at most d^- is αn, then we have the following:*

- $|V_{1+}(\mathcal{G})| > 2\ln n$ *w.h.p. if* $\theta/(1-\theta) \notin \mathbb{Z}$,
- $|V_{1+}(\mathcal{G})| > 0$ *with probability* $\Omega(1)$ *if* $\theta/(1-\theta) \in \mathbb{Z}$.

Proof. Suppose that the number of individuals with degree at most d^- is αn, and recall that $p = \frac{k-\sqrt{k}\ln(n)}{n}$. Without loss of generality, suppose that there are no tests of degree 1 (otherwise, remove them and each connected individual from the testing scheme). Clearly, if the inference of $\boldsymbol{\sigma}$ does not succeed on this manipulated graph, then it cannot succeed in $\mathcal{G}$. Before proceeding, we introduce the following auxiliary result.

Claim 4.5. *If there are αn individuals of degree at most d^-, then there exists $0 < \beta \le \alpha$ such that there are at least βn individuals of degree at most d^- of distance at least 6.*

Proof. This is a consequence of the test degrees and individual degrees both behaving as $\Theta(1)$, due to the assumption $\Gamma = \Theta(1)$ and Claim 4.3. If we select an arbitrary individual with degree at most d^-, remove all individuals within distance 4, and then repeat, then we construct a set of individuals with degree at most d^- and pairwise distances at least 6 in the original graph. Since this procedure starts with αn individuals under consideration and only removes $\Theta(1)$ individuals per step, the set formed has size $\Theta(n)$, as desired. □

Let B be the largest possible subset of individuals satisfying the requirements of Claim 4.5. Thus, B is a set of βn individuals such that for all $x \neq x' \in B$ we have

(B1) $\deg(x) \le d^-$

(B2) $\operatorname{dist}(x, x') \ge 6$.

We analyze a single individual $x \in B$ using the FKG inequality (e.g., see [19, Proposition 1]); as noted in [1, Lemma 4], the events of x being disguised in each of its tests are increasing with respect to $\boldsymbol{\sigma}$, so the FKG inequality yields

$$\mathbb{P}\left(x \in V^+(\mathcal{G})\right) \ge \prod_{a \in \partial x}\left(1-\left(1-p\right)^{\deg(a)-1}\right).$$

Then, by the fact that $\deg(x) \le d^- = O(1)$ within B, Claim 7.5 guarantees that

$$\prod_{a \in \partial x}\left(1-\left(1-p\right)^{\deg(a)-1}\right) \ge Cp^{d^-}$$

for some constant C depending on θ and Γ.

We now turn to the total number of totally disguised individuals in B. As noted above, for two individuals $x, x' \in B$, the events of being totally disguised are independent due to the pairwise distances being at least 6. Furthermore, each such individual is infected independently with probability p under our modified infection model $\boldsymbol{\sigma}^*$. Thus, the number of totally disguised infected individuals $|V_{1+}(\mathcal{G})|$ is dominated by a binomial random variable $\mathrm{Bin}(\beta n, p \cdot Cp^{d^-})$. Since $np \sim k = n^\theta$, the mean of this binomial distribution scales as $\Theta(n^{\theta-(1-\theta)d^-})$. In particular, when $\frac{\theta}{1-\theta}$ is non-integer, the choice $d^- = \left\lfloor \frac{\theta}{1-\theta} \right\rfloor$ ensures that the exponent is positive, and the Chernoff bound gives w.h.p. that

$$|V_{1+}(\mathcal{G})| \ge n^{\Omega(1)}. \tag{4.2}$$

On the other hand, if $\frac{\theta}{1-\theta}$ is integer-valued, then the mean of the binomial is $\Theta(1)$, which is enough to ensure that $|V_{1+}(\mathcal{G})| > 0$ with $\Omega(1)$ probability. Combining these two cases completes the proof of Lemma 4.4. □

As an immediate consequence of Lemma 4.4, in any group testing instance that succeeds w.h.p., there are at most $o(n)$ individuals of degree up to d^-. However, if $m \le (1-\delta)d^+n/\Gamma$ we find at least αn individuals of degree at most d^- (for some α depending on δ) by the handshaking lemma [42, Corollary 1.3], which is a contradiction. Therefore, Proposition 4.2 is a direct consequence of Lemma 4.4, with the claims regarding $|V_{0+}(\mathcal{G})|$ following easily from those regarding $|V_{1+}(\mathcal{G})|$ in the same way as Corollary 3.7. □

We now turn to the sparse regime $\theta < \frac{1}{2}$, establishing the following proposition as a stepping stone to Theorem 4.1.

Proposition 4.6. *Let $0 < \theta < 1/2$, and let $\mathcal{G}$ be an arbitrary pooling scheme with tests of size at most Γ. For all $\delta > 0$ and sufficiently large n, if $m \le (2-\delta)\frac{n}{\Gamma+1}$, then any algorithm (efficient or not) fails at recovering $\boldsymbol{\sigma}$ from $\hat{\boldsymbol{\sigma}}$ and $\mathcal{G}$ w.h.p..*

4.2.2. *Proof of Proposition 4.6.* The proof hinges on a fairly straightforward observation. We can assume without loss of generality that there are no tests containing only one individual (otherwise, we remove them and their corresponding individuals from the testing scheme). By a simple counting argument, there can be only $o(n)$ such tests (since otherwise $m > 2n/\Gamma$, which is a contradiction). Then, another counting argument leads to the fact that the number of individuals of degree 1 is large when $m < 2n/\Gamma$, as stated in the following.

Lemma 4.7. *If $m = (2-\varepsilon)n/\Gamma$, then there are at least εn individuals of degree 1.*

Proof. Denote by αn the number of individuals of degree 1, i.e., $\alpha > 0$ is the proportion of individuals of degree 1. Then the lemma follows by double counting edges (on the individual side and on the test-side):

$$(2-\varepsilon)n = m\Gamma \geq \sum_{a \in F(\mathscr{G})} \deg(a) = \sum_{x \in V(\mathscr{G})} \deg(x) \geq \alpha n + 2(1-\alpha)n.$$

Solving for α yields $\alpha \geq \varepsilon$, and the lemma follows. □

The next lemma shows that there can only be a small number of tests containing more than one individual of degree 1.

Lemma 4.8. *If there is any algorithm recovering $\boldsymbol{\sigma}$ from $\mathscr{G}$ and $\hat{\boldsymbol{\sigma}}$ with $\Omega(1)$ probability, then the number of tests containing more than one individual of degree 1 is below $n/\sqrt{k} = o(n)$.*

Proof. Suppose that at least $n/\sqrt{k}$ tests contain at least 2 individuals of degree 1. By definition, each of these individuals is infected with probability $p \sim k/n$, so by the Chernoff bound, there are at least $\sqrt{k}/\ln n = \omega(1)$ tests containing two individuals of degree one, out of which exactly one is infected. For these tests, the inference algorithm cannot do better than guess which one is the infected one, but then the probability of all guesses being correct is $(1/2)^{\omega(1)} = o(1)$. Therefore, the lemma follows. □

We are now in a position to prove Proposition 4.6. For $m = (2-\varepsilon)n/\Gamma$, we find by Lemma 4.7 that there are at least εn individuals of degree 1. By Lemma 4.8 and the fact that $\Gamma = \Theta(1)$, only $o(n)$ such individuals can be placed together in any tests, and hence, the total number of tests is at least $\varepsilon n - o(n)$. Formally,

$$(2-\varepsilon)n/\Gamma = m \geq \varepsilon n - o(n). \tag{4.3}$$

Solving (4.3) for ε, we find $\varepsilon \leq \frac{2}{\Gamma+1} + o(1)$. Hence,

$$m \geq \left(2 - \frac{2}{\Gamma+1} - o(1)\right)\frac{n}{\Gamma} = 2\frac{n}{\Gamma+1} - o(n),$$

and the proposition follows. □

The universal lower bound in the considered regime is a direct consequence of Proposition 4.2, Proposition 4.6, and Claim 2.3. The proof of Theorem 4.1 is thus complete.

4.3. **Algorithmic bound: Preliminaries and statement of result.** We now turn to the problem of establishing an upper bound, with a suitably-chosen test design and an efficient inference algorithm, that matches the universal lower bound. We start by recalling the definition of $\tilde{\mathscr{G}}_\Gamma$ in Section 2.2.2:

$$\tilde{\mathscr{G}}_\Gamma(\theta) = \begin{cases} \mathscr{G}_\Gamma & \text{if } \theta \geq 1/2 \\ \mathscr{G}_\Gamma^* & \text{otherwise} \end{cases} \tag{4.4}$$

We equip this pooling scheme with the efficient DD algorithm (see Algorithm 1). In the following, we will see that the combination of these tools will lead to information-theoretically optimal performance in the Γ-sparse regime.

Proposition 4.9. *Define*

$$m_{\mathtt{DD}}(\tilde{\mathscr{G}}_\Gamma) = \max\left\{1 + \left\lfloor \frac{\theta}{1-\theta} \right\rfloor \frac{n}{\Gamma}, 2\frac{n}{\Gamma+1}\right\}.$$

For $\Gamma = \Theta(1)$ and $m = (1+\varepsilon)m_{\mathtt{DD}} = (1+\varepsilon)\max\left\{1 + \left\lfloor \frac{\theta}{1-\theta} \right\rfloor \frac{n}{\Gamma}, 2\frac{n}{\Gamma+1}\right\}$, we have

$$\mathbb{P}(\mathscr{A}_{\mathtt{DD}}(\tilde{\mathscr{G}}_\Gamma, \hat{\boldsymbol{\sigma}}, k) = \boldsymbol{\sigma}) = 1 - o(1).$$

The statement for the dense is derived in Theorem 4.10 below, and the sparse case in Theorem 4.18. Combining these two statements with Theorem 4.1 implies that the DD algorithm succeeds with roughly $m_{\mathrm{inf},\Gamma}$ tests, and the achievability and converse results match for all $\theta \in (0,1)$.

4.4. **Algorithmic feasibility I: The configuration model.** We show that the DD algorithm succeeds with roughly $\max\left\{2, 1+\left\lfloor \frac{\theta}{1-\theta} \right\rfloor\right\}$ tests employing the configuration model $\mathscr{G}_\Gamma$.

We define

$$\Delta_{\mathtt{DD}}(\theta) = \max\left\{2, 1+\left\lfloor \frac{\theta}{1-\theta} \right\rfloor\right\} \qquad \text{and} \qquad m_{\mathtt{DD}}(\mathscr{G}_\Gamma) = \Delta_{\mathtt{DD}}(\theta)\frac{n}{\Gamma}, \tag{4.5}$$

representing the achievability bound for DD in $\mathscr{G}_\Gamma$.

Theorem 4.10. *Let $\varepsilon > 0$ and $m \geq m_{\mathtt{DD}}(\mathscr{G}_\Gamma)$. Then* DD *infers $\boldsymbol{\sigma}$ from $(\mathscr{G}_\Gamma, \hat{\boldsymbol{\sigma}})$ correctly w.h.p..*

We stress at this point that Theorem 4.10 gives a performance guarantee for the configuration model with any sparsity level, but it will turn out in due course that for $\theta < \frac{1}{2}$ a different model performs slightly better. Note also that for $\theta \geq \frac{1}{2}$, we can simplify $\max\left\{2, 1+\left\lfloor \frac{\theta}{1-\theta} \right\rfloor\right\} = 1+\left\lceil \frac{\theta}{1-\theta} \right\rceil$.

4.4.1. *Proof of Theorem 4.10.* The proof of Theorem 4.10 hinges on a slightly delicate combinatorial argument. Recall from Figure 1 that V_{1--} consists of those infected individuals that appear in at least one test with only individuals that are removed in the first step of DD (i.e., the easy uninfected individuals V_{0+}). By Corollary 2.4, DD succeeds if and only if $V_1 = V_{1--}$.

Lemma 4.11. *Let $\boldsymbol{A} = |V_1(\mathscr{G}) \setminus V_{1--}(\mathscr{G}_\Gamma)|$ denote the number of infected individuals that are not identified in the second step of* DD*. If $m \geq m_{\mathtt{DD}}$, then $\mathbb{E}[\boldsymbol{A}] = o(1)$.*

The proof of Lemma 4.11, while conceptually not difficult and similar to [11], is technically challenging, as we have to deal with subtle dependencies in the pooling scheme, caused by the mutli-edges given through the configuration model. A heuristic argument with a (false) independence assumption can provide some intuition as follows: In order for an individual x to be part of a test containing no infected individual (besides possibly x itself) is roughly $(1-k/n)^{\Gamma-1}$. For x to be disguised, thus being element of $V_{0+}(\mathscr{G}_\Gamma)$ or $V_{1+}(\mathscr{G}_\Gamma)$, x may not be part of such a test. Hence, the probability of x being disguised would be roughly $\left(1-(1-k/n)^{\Gamma-1}\right)^{\Delta}$ if the associated Δ events were independent (recall that $\Delta = m\Gamma/n$ is the degree of each individual in the random regular design).

To formally deal with the dependencies in the graph, we proceed as follows. Denote by $(\boldsymbol{Y}_1, \dots, \boldsymbol{Y}_m)$ the number of infected individuals in the tests. There are $n\Delta$ edges connected to individuals, out of which exactly $k\Delta$ correspond to infected individuals. Each test chooses exactly Γ individuals without replacement, and hence, the number of individuals in any test follows a hypergeometric distribution. In order to get a handle on this distribution, we introduce a family $(\boldsymbol{X}_1, \dots, \boldsymbol{X}_m)$ of independent binomial variables, such that $\boldsymbol{X}_i \sim \mathrm{Bin}(\Gamma, k/n)$. These variables can accurately describe the local behavior of how many infected individuals belong to test a_i. We define $\mathscr{E}_\Gamma$ to be the event that the overall number of edges containing infected individuals is *correct*, i.e.,

$$\mathscr{E}_\Gamma = \left\{ \sum_{i=1}^{m} \boldsymbol{X}_i = k\Delta \right\}. \tag{4.6}$$

Claim 7.6 implies that $\mathbb{P}(\mathscr{E}_\Gamma) = \Omega((n\Delta)^{-1/2})$. In addition, given $\mathscr{E}_\Gamma$, we find the following.

Lemma 4.12. *Give the event $\mathscr{E}_\Gamma$, the sequences $(\boldsymbol{Y}_1, \dots, \boldsymbol{Y}_n)$ and $(\boldsymbol{X}_1, \dots, \boldsymbol{X}_n)$ are identically distributed with respect to the pooling graph $\mathscr{G}_\Gamma$.*

Proof. By the definition of $\boldsymbol{Y}_i$, we find for any $(y_i)_i$ satisfying $\sum_i y_i = k\Delta$ that

$$\mathbb{P}(\boldsymbol{Y}_i = y_i \,\forall\, i \in [m]) = \binom{k\Delta}{y_1, \dots, y_m}\binom{(n-k)\Delta}{\Gamma - y_1, \dots, \Gamma - y_m}\binom{n\Delta}{\Gamma, \dots, \Gamma}^{-1} = \frac{\prod_{i=1}^{m}\binom{\Gamma}{y_i}}{\binom{n\Delta}{k\Delta}}.$$

where the equality follows by rewriting in terms of factorials and simplifying. Furthermore, given $\sum_i x_i = k\Delta$, we have

$$\mathbb{P}(\boldsymbol{X}_i = x_i \,\forall\, i \in [m] | \mathscr{E}_\Gamma) = \prod_{i=1}^{m}\binom{\Gamma}{x_i}(k/n)^{x_i}(1-k/n)^{\Gamma-x_i}\left(\mathbb{P}(\mathscr{E}_\Gamma)\right)^{-1}.$$

Now, for two sequences $(y_i)_{i\in[m]}$ and $(y_i')_{i\in[m]}$ such that $\sum_{i=1}^m y_i = \sum_{i=1}^m y_i' = k\Delta$, we obtain

$$\frac{\mathbb{P}(\forall\, i \in [m] : Y_i = y_i)}{\mathbb{P}(\forall\, i \in [m] : Y_i = r_i')} = \prod_{i=1}^{m}\frac{\binom{\Gamma}{y_i}}{\binom{\Gamma}{y_i'}} = \frac{\mathbb{P}(\forall\, i \in [m] : Y_i = y_i | \mathscr{E}_\Gamma)}{\mathbb{P}(\forall\, i \in [m] : H_i = y_i' | \mathscr{E}_\Gamma)}.$$

This implies the lemma. □

Thus, we are able to carry out all necessary calculations with respect to $(\boldsymbol{X}_1,...,\boldsymbol{X}_n)$ and transfer the results to the original pooling scheme. For the next step, we need to get a handle on the number of positive and negative tests occurring in this setting. Let $\boldsymbol{m}_0 = \boldsymbol{m}_0(\mathscr{G}_\Gamma,\boldsymbol{\sigma})$ be the number of tests that render a negative result, and let $\boldsymbol{m}_1 = \boldsymbol{m}_1(\mathscr{G}_\Gamma,\boldsymbol{\sigma})$ be the number of tests that render a positive result. Then $\boldsymbol{m}_0$ and $\boldsymbol{m}_1$ are highly concentrated around their means as follows.

Lemma 4.13. *With probability* $1-o(n^{-2})$*, we have*

$$\boldsymbol{m}_0 = m(1-k/n)^\Gamma + O\big(\sqrt{m}\ln(n)\big) \qquad \text{and} \qquad \boldsymbol{m}_1 = m\big(1-(1-k/n)^\Gamma\big) + O\big(\sqrt{m}\ln(n)\big).$$

Proof. Recalling the definitions of $(\boldsymbol{Y}_i)_i$ and $(\boldsymbol{X}_i)_i$ from (4.6), we have

$$\boldsymbol{m}_0 = \sum_{i=1}^m \mathbf{1}\{\boldsymbol{Y}_i = 0\},$$

and we further denote by

$$\boldsymbol{m}_0' = \sum_{i=1}^m \mathbb{1}\{\boldsymbol{X}_i = 0\}$$

the number of negative tests as modelled by the family of independent binomial variables $(\boldsymbol{X}_i)_i$. Clearly, as the $\boldsymbol{X}_i$ are mutually independent,

$$\mathbb{E}\big[\boldsymbol{m}_0'|\mathscr{E}_\Gamma\big] = m\cdot\mathbb{P}(\mathrm{Bin}(\Gamma,k/n)=0) = m\left(1-\frac{k}{n}\right)^\Gamma.$$

Hence, the Chernoff bound (Lemma 7.1) guarantees that

$$\mathbb{P}\big(\big|\boldsymbol{m}_0' - \mathbb{E}(\boldsymbol{m}_0'|\mathscr{E}_\Gamma)\big| > \sqrt{m}\ln(n)\,|\,\Gamma\big) = o(n^{-10})$$

and by combining Lemma 4.12 with Claim 7.6, we conclude that

$$\mathbb{P}\big(|\boldsymbol{m}_0 - \mathbb{E}(\boldsymbol{m}_0)| > \sqrt{m}\ln(n)\,|\,\Gamma\big) = o(n^{-9}).$$

Thus, the first part of the lemma follows. The second part is immediate, as $\boldsymbol{m}_0+\boldsymbol{m}_1 = m$. □

Hence, the above-mentioned naive calculation (assuming independence) can be rigorously justified, and we have established the sizes of the disguised individuals w.h.p..

The following lemma characterizes the high-probability behavior of the number of disguised individuals.

Lemma 4.14. *Given* n *and* $k=n^\theta$ *as well as* $\Gamma=\Theta(1)$*, we have w.h.p. that*

$$|V_{0+}(\mathscr{G}_\Gamma)| = (1+o(1))n\big(1-(1-k/n)^{\Gamma-1}\big)^\Delta \quad \text{and} \quad |V_{1+}(\mathscr{G}_\Gamma)| = (1+o(1))k\big(1-(1-k/n)^{\Gamma-1}\big)^\Delta.$$

Proof. By the definition of $\mathscr{G}_\Gamma$ via the configuration model, Lemma 4.13 guarantees that the total number of edges connected to a positive test is, with probability at least $1-o(n^{-2})$, given by

$$\boldsymbol{m}_1\Gamma = m\Gamma\big(1-(1-k/n)^\Gamma + O\big(n^{-\Omega(1)}\big)\big). \tag{4.7}$$

Let x be an uninfected individual. We can calculate the probability of x belonging to $V_{0+}(\mathscr{G}_\Gamma)$ as follows: Each of the $\Delta=\Theta(1)$ edges that are mapped to x in the configuration model have to be connected to a positive test. Thus, by (4.7), we find

$$\mathbb{P}\left(x\in V_{0+}(\mathscr{G}_\Gamma)\,|\,x\in V_0(\mathscr{G}_\Gamma),\boldsymbol{m}_1\right) = \binom{\boldsymbol{m}_1\Gamma}{\Delta}\binom{m\Gamma}{\Delta}^{-1} = \big(1-(1-k/n)^\Gamma\big)^\Delta + O\left(\frac{\ln(n)}{n^{1/2}}\right),$$

where the equality follows from Claim 7.4. Therefore,

$$\mathbb{E}\left[|V_{0+}(\mathscr{G}_\Gamma)|\right] = \left(1+O\left(\frac{\ln(n)}{n^{3/2}}\right)\right)(n-k)\big(1-(1-k/n)^\Gamma\big)^\Delta = \left(1+O\left(\frac{\ln(n)}{n^{3/2}}\right)\right)n\big(1-(1-k/n)^\Gamma\big)^\Delta. \tag{4.8}$$

By a similar argument, the second moment turns out to be

$$\begin{aligned}\mathbb{E}\left[|V_{0+}(\mathscr{G}_\Gamma)|^2\right] &= \binom{n-k}{2}\binom{m\Gamma(1-(1-k/n)^\Gamma + O\left(n^{1/2}\ln(n)\right)}{2\Delta}\binom{m\Gamma}{2\Delta}^{-1}\\ &= \left(1+O\left(\frac{\ln(n)}{n^{3/2}}\right)\right)n^2\big(1-(1-k/n)^\Gamma\big)^{2\Delta}. \end{aligned} \tag{4.9}$$

Thus, (4.8), (4.9) and Chebychev's inequality lead to the first part of Lemma 4.14.

To prove the second part, namely that $|V_{1+}(\mathscr{G}_\Gamma)| \sim k\big(1-(1-k/n)^\Gamma\big)^\Delta$, we need to be more careful: It can happen that a given individual and a given test are paired more than once in the configuration model. In particular, if an infected individual x is allocated to the same test twice, then an analysis of the type above

would count it as disguised (i.e., $x \in V_{1+}(\mathscr{G}_\Gamma)$) even when it is the only infected individual in the test (i.e., $x \in V_{1-}(\mathscr{G}_\Gamma)$).

To deal with this issue, we again introduce $\boldsymbol{R}_i = (\boldsymbol{R}_i^1, \boldsymbol{R}_i^{0+}, \boldsymbol{R}_i^{0-})$ and $\boldsymbol{H}_i = (\boldsymbol{H}_i^1, \boldsymbol{H}_i^{0+}, \boldsymbol{H}_i^{0-})$, representing partitions of the individual types within a test. Specifically, $\boldsymbol{R}_i = (\boldsymbol{R}_i^1, \boldsymbol{R}_i^{0+}, \boldsymbol{R}_i^{0-})$ denotes the triplet containing the number of infected individuals, disguised uninfected individuals (i.e., in $V_{0+}(\mathscr{G}_\Gamma)$), and non-disguised uninfected individuals (i.e., in $V_{0-}(\mathscr{G}_\Gamma)$) appearing in test i, respectively. By definition, $\boldsymbol{R}_i^{0-} = \Gamma - \boldsymbol{R}_i^{0+} - \boldsymbol{R}_i^1$. Furthermore, given $|V_{0+}(\mathscr{G}_\Gamma)|$, $\boldsymbol{H}_i = (\boldsymbol{H}_i^1, \boldsymbol{H}_i^{0+}, \boldsymbol{H}_i^{0-})$ denotes a sequence of random variables distributed as follows:

$$\boldsymbol{H}_i = (\boldsymbol{H}_i^1, \boldsymbol{H}_i^{0+}, \boldsymbol{H}_i^{0-})_{i\in[\boldsymbol{m}_1]} \sim \mathrm{Mult}_{\geq(1,0,0)}\left(\Gamma, \left(k/n, |V_{0+}(\mathscr{G}_\Gamma)|/n, 1-k/n-|V_{0+}(\mathscr{G}_\Gamma)|/n\right)\right),$$

where $\mathrm{Mult}_{\geq(1,0,0)}$ denotes the multinomial distribution conditioned on the first coordinate being at least one. We additionally introduce

$$\boldsymbol{W} = \sum_{i=1}^{m} \Gamma \mathbf{1}\left\{\boldsymbol{R}_i^1 = 1\right\} \qquad \boldsymbol{W}' = \sum_{i=1}^{m} \Gamma \mathbf{1}\left\{\boldsymbol{H}_i^1 = 1\right\},$$

representing the number of connections to tests that contain exactly one infected individual, first in the original (multi-)graph and then with respect to the independent random variables $(\boldsymbol{H}_i)_i$.

The expectation of $\boldsymbol{W}'$ given $\mathscr{E}_\Gamma$ is straightforward to calculate as

$$\mathbb{E}\left[\boldsymbol{W}' \mid \mathscr{E}_\Gamma\right] = m\Gamma(k/n)(1-k/n)^{\Gamma-1} = k\Delta(1-k/n)^{\Gamma-1}.$$

Therefore, the Chernoff bound of Lemma 7.1 guarantees that the following holds given $\mathscr{E}_\Gamma$ with probability at least $1-o(n^{-10})$:

$$\boldsymbol{W}' = k\Delta(1-k/n)^{\Gamma-1} + O(n^{-\Omega(1)}). \tag{4.10}$$

Combining (4.10) with Claim 7.6 and Lemma 4.12 yields with probability at least $1-o(n^{-9})$ that

$$\boldsymbol{W} = k\Delta(1-k/n)^{\Gamma-1} + O(n^{-\Omega(1)}). \tag{4.11}$$

We are now able to calculate the number of infected individuals that appear only in tests that either contain at least one other infected individual, or in which that individual is included more than once (and hence a multi-edge exists). We denote this number by $\boldsymbol{U}$. An infected individual that contributes to $\boldsymbol{U}$ has to chose its Δ connections from exactly $k\Delta - \boldsymbol{W}$ possibilities, i.e., those connected to an infected individual, but not connected to a test that contains only one infected individual. Applying the result in Claim 7.4 based on Stirling's approximation, we find that

$$\mathbb{E}[\boldsymbol{U}|\boldsymbol{W}] = k\binom{k\Delta-\boldsymbol{W}}{\Delta}\binom{k\Delta}{\Delta}^{-1} = (1+o(1))k\left(1-(1-k/n)^{\Gamma-1}\right)^{\Delta}. \tag{4.12}$$

Furthermore, by an analogous argument, we have

$$\mathbb{E}\left[\boldsymbol{U}^2|\boldsymbol{W}\right] = \binom{k}{2}\binom{k\Delta-\boldsymbol{W}}{2\Delta}\binom{k\Delta}{2\Delta} = (1+o(1))k^2\left(1-(1-k/n)^{\Gamma-1}\right)^{2\Delta}. \tag{4.13}$$

Equations (4.12), (4.13) and Chebyshev's inequality yield w.h.p. that

$$\boldsymbol{U} = (1+o(1))k\left(1-(1-k/n)^{\Gamma-1}\right)^{\Delta}. \tag{4.14}$$

As described above, an infected individual appearing in the same test more than once contributes to $\boldsymbol{U}$, so at this stage we can only state that $\boldsymbol{U} \geq |V_{1+}(\mathscr{G}_\Gamma)|$. To better understand $|V_{1+}(\mathscr{G}_\Gamma)|$, we introduce a new random variable, $\boldsymbol{T}$, that counts the number of individuals occurring in some test at least twice. Clearly,

$$\boldsymbol{U} \geq |V_{1+}(\mathscr{G}_\Gamma)| \geq \boldsymbol{U} - \boldsymbol{T}. \tag{4.15}$$

It is well known [32, Chapter 13.2.1] that the number of multi-edges on a bounded degree multi-graph generated through the configuration model is finite w.h.p., we denote this magnitude with $\boldsymbol{M} = \Theta(1)$. Hence, the probability that a given individual x features a multi-edge is at most $\boldsymbol{M}/n$. Noting that $\mathbb{E}[\boldsymbol{T} \mid \boldsymbol{M}] \leq \boldsymbol{M}$ w.h.p., we find that $\boldsymbol{T}$ is w.h.p. upper bounded by any arbitrarily-slowly growing $\omega(1)$ term. In particular, $\boldsymbol{T} = o(\boldsymbol{U})$ (since (4.14) gives $\boldsymbol{U} = \omega(1)$ w.h.p.), and hence the second part of Lemma 4.14 follows from (4.15). □

Next, we define the event

$$\mathscr{F}_\Gamma = \left\{\boldsymbol{m}_1 = (1+o(1))m\left(1-(1-k/n)^{\Gamma}\right)\right\} \cap \left\{|V_{0+}(\mathscr{G}_\Gamma)| = (1+o(1))n\left(1-(1-k/n)^{\Gamma-1}\right)^{\Delta}\right\}, \tag{4.16}$$

in which the number of positive tests and disguised uninfected individuals behave as expected. By Lemmas 4.13 and 4.14, we have $\mathbb{P}(\mathscr{F}_\Gamma) \geq 1 - o(1)$. We assume without loss of generality that the first $\boldsymbol{m}_1$ tests render a positive result.

Letting

$$\mathscr{D}_\Gamma = \left\{ \sum_{i=1}^{\boldsymbol{m}_1} \boldsymbol{H}_i^1 = k\Delta, \quad \sum_{i=1}^{\boldsymbol{m}_1} \boldsymbol{H}_i^{0+} = |V_{0+}(\mathscr{G}_\Gamma)|\,\Delta \right\}$$

be the event that $\boldsymbol{H} = \sum_{i=1}^{\boldsymbol{m}_1} \boldsymbol{H}_i$ equals its expectation, we have the following analog of Corollary 3.18.

Claim 4.15. *The distribution of $\boldsymbol{R}_i$ equals the distribution of $\boldsymbol{H}_i$ given $\mathscr{D}_\Gamma$ and Γ, and furthermore, $\mathbb{P}(\mathscr{D}_\Gamma) = \Omega(n^{-1})$.*

Proof of Claim 4.15. Let $(r_i)_{i\in[\boldsymbol{m}_1]}$ be a sequence such that $r_i = (r_i^1, r_i^{0+}, r_i^{0-})$ and $\sum_i r_i^1 = k\Delta, \sum_i r_i^{0+} = |V_{0+}(\mathscr{G}_\Gamma)|\,\Delta$, and $r_i^{0-} = \Gamma - r_i^1 - r_i^{0+}$. Let

$$S_1 = k\Delta, \qquad S_{0+} = \Delta\,|V_{0+}(\mathscr{G}_\Gamma)| \qquad \text{and} \qquad S_{0-} = n\Delta - n\Delta(1-(1-k/n)^\Gamma) - k\Delta.$$

By the definition of $\boldsymbol{R}_i$, we have

$$\mathbb{P}(\forall\, i \in [\boldsymbol{m}_1] : \boldsymbol{R}_i = r_i \mid \boldsymbol{m}_1) = \frac{\binom{S_1}{r_1^1 \ldots r_{\boldsymbol{m}_1}^1}\binom{S_{0+}}{r_1^{0+}\ldots r_{\boldsymbol{m}_1}^{0+}}\binom{S_{0-}}{\Gamma - r_1^1 - r_1^{0+} \ldots \Gamma - r_{\boldsymbol{m}_1}^1 - r_{\boldsymbol{m}_1}^{0+}}}{\binom{n\Delta}{\Gamma,\ldots,\Gamma}} = \binom{n\Delta}{S_1, S_{0+}, S_{0-}}^{-1} \prod_{i=1}^{\boldsymbol{m}_1} \binom{\Gamma}{r_i^1, r_i^{0+}, r_i^{0-}}.$$

Letting $(r_i^{'})_{i\in[\boldsymbol{m}_1]}$ be a second sequence as above, it follows that

$$\frac{\mathbb{P}(\forall\, i \in [\boldsymbol{m}_1] : \boldsymbol{R}_i = y_i \mid \boldsymbol{m}_1}{\mathbb{P}(\forall\, i \in [\boldsymbol{m}_1] : \boldsymbol{R}_i = y_i^{'} \mid \boldsymbol{m}_1)} = \prod_{i=1}^{\boldsymbol{m}_1} \frac{\binom{\Gamma}{r_i^1 r_i^{0+} r_i^{0-}}}{\binom{\Gamma}{(r')_i^1 (r')_i^{0+} (r')_i^{0-}}}. \tag{4.17}$$

Furthermore, by the definition of $\boldsymbol{X}$, we have

$$\begin{aligned}
&\frac{\mathbb{P}(\forall\, i \in [\boldsymbol{m}_1] : \boldsymbol{H}_i = r_i \mid \boldsymbol{m}_1, \mathscr{D}_\Gamma)}{\mathbb{P}(\forall\, i \in [\boldsymbol{m}_1] : \boldsymbol{H}_i = r_i^{'} \mid \boldsymbol{m}_1, \mathscr{D}_\Gamma)} \\
&= \frac{(k/n)^{\sum_{i=1}^{\boldsymbol{m}_1} r_i^1} (|V_{0+}(\mathscr{G}_\Gamma)|/n)^{\sum_{i=1}^{\boldsymbol{m}_1} r_i^{0+}} (1 - k/n - |V_{0+}(\mathscr{G}_\Gamma)|/n)^{\sum_{i=1}^{\boldsymbol{m}_1} r_i^{0-}}}{(k/n)^{\sum_{i=1}^{\boldsymbol{m}_1} (r')_i^1} (|V_{0+}(\mathscr{G}_\Gamma)|/n)^{\sum_{i=1}^{\boldsymbol{m}_1} (r')_i^{0+}} (1 - k/n - |V_{0+}(\mathscr{G}_\Gamma)|/n)^{\sum_{i=1}^{\boldsymbol{m}_1} (r')_i^{0-}}} \prod_{i=1}^{\boldsymbol{m}_1} \frac{\binom{\Gamma}{r_i^1, r_i^{0+}, r_i^{0-}}}{\binom{\Gamma}{(r')_i^1, (r')_i^{0+},)(r')_i^{0-}}} \\
&= \prod_{i=1}^{\boldsymbol{m}_1} \frac{\binom{\Gamma}{r_i^1, r_i^{0+}, r_i^{0-}}}{\binom{\Gamma}{(r')_i^1, (r')_i^{0+}, (r')_i^{0-}}}.
\end{aligned} \tag{4.18}$$

The desired claim follows from Equations (4.17) and (4.18). □

We are interested in the number of positive tests that contain exactly one infected individual and no elements of $V_{0+}(\mathscr{G}_\Gamma)$. Therefore, we define

$$\boldsymbol{B} = \sum_{i=1}^{\boldsymbol{m}_1} \mathbf{1}\left\{\boldsymbol{R}_i^1 + \boldsymbol{R}_i^{0+} = 1\right\} \qquad \text{and} \qquad \boldsymbol{B}' = \sum_{i=1}^{\boldsymbol{m}_1} \mathbf{1}\left\{\boldsymbol{H}_i^1 + \boldsymbol{H}_i^{0+} = 1\right\}.$$

Claim 4.16. *We have w.h.p. that*

$$\boldsymbol{B} \leq \Delta k(1 - 2\Gamma n^{-(1-\theta)})$$

Proof of Claim 4.16. We use Claim 4.15 to simulate $\boldsymbol{B}$ through independent random variables as in $\boldsymbol{B}'$. Since $\boldsymbol{B}'$ is a sum of independent multinomial variables, we obtain its expectation by applying (4.16), Lemma 7.3 and Bayes Theorem:

$$\begin{aligned}
\mathbb{E}\left[\boldsymbol{B}' \mid \mathscr{D}_\Gamma, |V_{0+}(\mathscr{G}_\Gamma)|, \boldsymbol{m}_1\right] &= \sum_{i=1}^{\boldsymbol{m}_1} \mathbb{P}\left(\boldsymbol{H}_i = (1,0,\Gamma-1) \mid \mathscr{D}_\Gamma, |V_{0+}(\mathscr{G}_\Gamma)|\right) \\
&= \boldsymbol{m}_1 \Gamma \frac{k/n \cdot (1 - (k + |V_{0+}(\mathscr{G}_\Gamma)|)/n)^{\Gamma-1}}{1 - (1-k/n)^\Gamma} \\
&= \left(1 + O\left(\Gamma \frac{k}{n}\right)\right) \boldsymbol{m}_1 \left(1 - \frac{k + |V_{0+}(\mathscr{G}_\Gamma)|}{n}\right)^{\Gamma-1}
\end{aligned} \tag{4.19}$$

where the last step follows from Lemma 7.5 and $\Gamma = \Theta(1)$. Conditioning on $\mathscr{F}_\Gamma$ defined in (4.16) and using $(1-\varepsilon_n)^{\Gamma-1} = 1-(\Gamma-1)\varepsilon_n(1+o(1)$ for $\varepsilon_n = o(1)$ and $\Gamma - 1 = \Theta(1)$, we obtain

$$\begin{aligned}
\mathbb{E}\left[\boldsymbol{B}' \mid \mathscr{D}_\Gamma, \mathscr{F}_\Gamma\right] &= (1+o(1))\, m\Gamma k/n \cdot \left(1 - \frac{k + n(\Gamma-1)^\Delta (k/n)^\Delta}{n} - o(1)\right)^{\Gamma-1} \\
&= \left(1 + O\left(\Gamma\frac{k}{n}\right)\right) m\Gamma k/n \cdot \left(1 - (\Gamma-1)\frac{k + (\Gamma-1)^\Delta n(k/n)^\Delta}{n}\right) \\
&= \left(1 + O\left(\Gamma\frac{k}{n}\right)\right) m\Gamma k/n \cdot \left(1 - (\Gamma-1)\left(n^{-(1-\theta)} + (\Gamma-1)^\Delta n^{-(1-\theta)\Delta}\right)\right) \\
&= \left(1 + O\left(\Gamma\frac{k}{n}\right)\right) \Delta k \left(1 - (\Gamma-1)n^{-(1-\theta)}\right) \\
&= \Delta k \left(1 + O\left(\Gamma n^{-(1-\theta)}\right)\right), \qquad (4.20)
\end{aligned}$$

where we used $k = n^\theta$ and $\Delta = \frac{m\Gamma}{n}$. Moreover, since $\boldsymbol{B}'$ is a binomial random variable, the Chernoff bound (Lemma 7.1) yield with probability $o(n^{-10})$ that

$$\boldsymbol{B}' \le \Delta k \left(1 + O\left(\Gamma n^{-(1-\theta)}\right)\right).$$

Thus, by Claim 4.15 we have w.h.p. that

$$\boldsymbol{B} \le \Delta k \left(1 + O\left(\Gamma n^{-(1-\theta)}\right)\right). \qquad (4.21)$$

□

We are now in a position to characterize $\boldsymbol{A} = |V_1(\mathscr{G}) \setminus V_{1--}(\mathscr{G}_\Gamma)|$.

Claim 4.17. *Given $\boldsymbol{B} \le \Delta k\left(1 - O\left(\Gamma n^{-(1-\theta)}\right)\right)$, we have for some constant $C > 0$ that*

$$\mathbb{E}\left[\boldsymbol{A} \mid \boldsymbol{B}, \mathscr{F}_\Gamma\right] = k\binom{k\Delta - \boldsymbol{B}}{\Delta}\binom{k\Delta}{\Delta}^{-1} \le k(C\cdot\Gamma)^\Delta n^{-(1-\theta)\Delta} \qquad (4.22)$$

Proof of Claim 4.17. The combinatorial expression follows by adding k probabilities, one per defective item. Each probability is the probability that an infected individual does not belong to V_{1--}, which equals the probability that all of its Δ connections are disjoint from the $k\Delta - \boldsymbol{B}$ connections to tests in which it would have been the only infected individual with no disguised uninfected individuals. The assertion then follows by combining the assumption $\boldsymbol{B} \le \Delta k\left(1 - O\left(\Gamma n^{-(1-\theta)}\right)\right)$ with Claim 7.4. □

Proof of Lemma 4.11. We distinguish between $\theta/(1-\theta) \notin \mathbb{Z}$ and $\theta/(1-\theta) = T \in \mathbb{Z}$, and recall $m_{\mathtt{DD}}$ from (4.5) with $\Delta = \max\left\{2, 1 + \left\lfloor \frac{\theta}{1-\theta} \right\rfloor\right\}$. For simplicity, we assume that the inequality $m \ge m_{\mathtt{DD}}$ holds with equality, but the general case is analogous.

Case A: $\theta/(1-\theta) \notin \mathbb{Z}$**.** In this case, for $m = m_{\mathtt{DD}}$, we have $\Delta = \max\left\{2, 1 + \left\lfloor \frac{\theta}{1-\theta} \right\rfloor\right\} = \max\{2, \lceil \theta/(1-\theta)\rceil\}$. We distinguish the two cases $\theta < 1/2$ and $\theta > 1/2$ as follows:

- **Case A1:** $\theta > 1/2$**.** In this case, we have $\Delta = \lceil \theta/(1-\theta)\rceil$. Defining $\eta = \theta - (1-\theta)\cdot\lceil\theta/(1-\theta)\rceil < 0$, using (4.22) and $\Gamma, \Delta = \Theta(1)$, we find

$$\mathbb{E}\left[\boldsymbol{A} \mid \boldsymbol{B}, \mathscr{F}_\Gamma\right] \le O(1)\, n^{\theta - (1-\theta)\cdot\lceil\theta/(1-\theta)\rceil} = O(n^\eta). \qquad (4.23)$$

- **Case A2:** $\theta < 1/2$**.** In this case, we have $\Delta = 2$, and hence

$$\mathbb{E}\left[\boldsymbol{A} \mid \boldsymbol{B}, \mathscr{F}_\Gamma\right] \le O(1)\Gamma^\Delta n^{3\theta-2} \le o(1). \qquad (4.24)$$

Case B: $\theta/(1-\theta) = T \in \mathbb{Z}$**.** Again, we distinguish the cases $\theta = 1/2$ and $\theta > 1/2$:

- **Case B1:** $\theta > 1/2$**.** We have $\Delta = T + 1$, so by (4.22) and $\Gamma, \Delta = \Theta(1)$, we find

$$\mathbb{E}\left[\boldsymbol{A} \mid \boldsymbol{B}, \mathscr{F}_\Gamma\right] \le O(1)\, n^{\theta - (1-\theta)\cdot(T+1)} = O(n^{-(1-\theta)}), \qquad (4.25)$$

  where the last step uses $1 - \theta \ge \theta$ and $T > 1$.
- **Case B2:** $\theta = 1/2$**.** We have $\Delta = 2$, and hence

$$\mathbb{E}[\boldsymbol{A} \mid \boldsymbol{B}, \mathscr{F}_\Gamma] \le O(n^{-1/2}). \qquad (4.26)$$

Using (4.23)–(4.26) and combining Bayes' rule with the fact that $\mathscr{F}_\Gamma$ occurs w.h.p. completes the proof of Lemma 4.11. □

Thus, we can use the previous findings to prove Theorem 4.10 .

Proof of Theorem 4.10. We can finally estimate the expected size of $\boldsymbol{A} = |V_1(\mathscr{G}) \setminus V_{1--}(\mathscr{G}_\Gamma)|$. By recalling Claim 4.17 and combining it with the fact that $\mathscr{F}_\Gamma$ occurs w.h.p., and using Lemma 4.11, for $\Delta = \Delta_{\mathtt{DD}}(\theta)$ and $m = m_{\mathtt{DD}}$ we get $\mathbb{E}[\boldsymbol{A}] = o(1)$ w.h.p.. The assertion follows by applying Markov's inequality. □

In this subsection, we addressed the case where the test design is formed using the configuration model, and showed that the DD-algorithm is optimal in this regime if applied to the random regular pooling scheme $\mathscr{G}_\Gamma$. However, the analysis of this section is not able to provide a tight bound for the matching design.

4.5. **Algorithmic feasibility II: Matching-based model.** Recall from from Section 2.2.2 that the matching-based model with parameter γ is denoted by $\mathscr{G}^*_\Gamma$. It turns out that the DD algorithm is also optimal for this model.

Theorem 4.18. *If $m \geq 2n/(\Gamma+1)$ and $0 < \theta < 1/2$, DD recovers $\boldsymbol{\sigma}$ from $\mathscr{G}^*_\Gamma$ and $\hat{\boldsymbol{\sigma}}$ w.h.p..*

4.5.1. *Proof of Theorem 4.18.* We prove the theorem for $m = 2n/(\Gamma+1)$ (which implies $\gamma = \frac{2}{\Gamma+1}n$), but the more general case follows analogously (intuitively, a higher number of tests can only help). We analyse the DD algorithm on $\mathscr{G}^*_\Gamma$ in two steps, starting with the regular part of the graph. Denote by $\mathscr{G}^{*,r}_\Gamma$ the $(\Gamma-1,2)$ regular part, in which we select $n-\gamma$ individuals and pool them into two tests each. Denote by $\boldsymbol{\sigma}[\mathscr{G}^{*,r}_\Gamma]$ and $\hat{\boldsymbol{\sigma}}[\mathscr{G}^{*,r}_\Gamma]$ the infection status vector and outcome vector resulting from the regular part alone.

Lemma 4.19. *If $m \geq 2n/(\Gamma+1)$, then DD recovers $\boldsymbol{\sigma}[\mathscr{G}^{*,r}_\Gamma]$ from $(\mathscr{G}^{*,r}_\Gamma, \hat{\boldsymbol{\sigma}}[\mathscr{G}^{*,r}_\Gamma])$ correctly w.h.p..*

This makes direct use of Theorem 4.10, as $\mathscr{G}^{*,r}_\Gamma$ is equally distributed as $\mathscr{G}^r_{\Gamma-1}$. With $\gamma = \frac{2}{\Gamma+1}n$ individuals removed from the population, we find $n' = \frac{\Gamma-1}{\Gamma+1}n$ individuals being tested in $\mathscr{G}_{\Gamma-1}$. Thus, we require at most $m' = 2\frac{n'}{\Gamma-1} = 2\frac{n}{\Gamma+1}$ tests in order for DD to succeed w.h.p.on $\mathscr{G}_{\Gamma-1}$. In the second step below, we will argue that adding the $2\frac{n}{\Gamma+1}$ individuals (one to each test) does not harm the result.

We denote by k' the number of infected individuals under the remaining n' individuals, and let θ' be the value such that $k' = \Theta((n')^{\theta'})$, which is well-defined due to the following.

Claim 4.20. *Under the preceding setup, we have w.h.p.that $\theta' = \theta$.*

Proof. As we remove $\gamma = \frac{2}{\Gamma+1}n$ individuals randomly, the number of infected individuals in the left-over is a hypergeometrically distributed random variable $\boldsymbol{K}' \sim H\left(n,k,n'\right)$. Thus, the Chernoff bound for the hypergeometric distribution guarantees w.h.p. that

$$\boldsymbol{K}' = (1+o(1))kn'/n = (1+o(1))\frac{\Gamma-1}{\Gamma+1}k.$$

Hence, the assertion follows. □

In the second step, we analyse the remaining part of the graph, in which the γ remaining items are placed into one test each.

Lemma 4.21. *If $\theta < 1/2$, then the DD algorithm recovers $\boldsymbol{\sigma}$ w.h.p. from $(\mathscr{G}^*_\Gamma, \hat{\boldsymbol{\sigma}})$ provided that it recovers $\boldsymbol{\sigma}[\mathscr{G}^{*,r}_\Gamma]$ w.h.p. from $(\mathscr{G}^{*,r}_\Gamma, \hat{\boldsymbol{\sigma}}[\mathscr{G}^{*,r}_\Gamma])$.*

Proof. The intuition behind our mildly delicate combinatorial proof is as follows. Essentially, the lemma follows from the observation that in the third neighborhood of each positive test we find exactly one infected individual w.h.p.. Indeed, if $\theta < 1/2$, w.h.p. there are no two infected individuals in a finite neighborhood in $\mathscr{G}^*_\Gamma$, as the expected number of such individuals is bounded by $\sim O\left(n\frac{k^2}{n^2}\right) = o(1)$.

We now proceed more formally. We need to prove that the DD algorithm succeeds w.h.p. on $\mathscr{G}^*_\Gamma$ if it succeeds w.h.p. on the $(\Gamma-1,2)-$regular part. (Note that for $\Gamma = 3$ this regular graph consists of disjoint cycles.) By construction of $\mathscr{G}^*_\Gamma$, there are at most $\gamma = \frac{2}{\Gamma+1}n$ individuals left to be added to $\mathscr{G}_{\Gamma-1}$. Denote the set of these individuals by $X = \{x_1 \dots x_\gamma\}$. As $\gamma \leq m$, there is a matching from X to the the m tests. Therefore, an immediate consequence is that in $\mathscr{G}^*_\Gamma$, every test contains at most Γ individuals and there are individuals of degree 2 and individuals of degree 1.

Suppose the DD algorithm succeeds on the regular part $\mathscr{G}^{*,r}_\Gamma \sim \mathscr{G}^r_{\Gamma-1}$. We distinguish four different cases.

Case A: Connecting to a negative test. Suppose that an individual $x \in X$ connects to a (previously) negative test a. Then, for all $y \in \partial_{\mathscr{G}^{*,r}_\Gamma}(a)$ we have $y \in V_{0-}(\mathscr{G}^{*,r}_\Gamma)$.

- **Case A-1: $\boldsymbol{\sigma}_x = 0$.** If the individual which was added in the second step is uninfected as well, we find $\hat{\boldsymbol{\sigma}}_a(\mathscr{G}^*_\Gamma) = \hat{\boldsymbol{\sigma}}_a(\mathscr{G}^{*,r}_\Gamma) = 0$. Hence, $x \in V_{0-}(\mathscr{G}^*_\Gamma)$, and for all $y \in \partial_{\mathscr{G}^{*,r}_\Gamma}(a)$ we have $y \in V_{0-}(\mathscr{G}^*_\Gamma)$ as well. Therefore, DD succeeds if (and only if) it succeeded on $\mathscr{G}^{*,r}_\Gamma$.

- **Case A-2: $\boldsymbol{\sigma}_x = 1$.** If the individual that was added is infected, things are a bit more complicated. Clearly, $\hat{\boldsymbol{\sigma}}_a(\mathcal{G}_\Gamma^{*,r}) = 0$ but $\hat{\boldsymbol{\sigma}}_a(\mathcal{G}_\Gamma^*) = 1$. Therefore, we need to prove that $x \in V_{1--}(\mathcal{G}_\Gamma^*)$ which holds by definition if and only if for all $y \in \partial_{\mathcal{G}_\Gamma^*}(a)$ we have $y \in V_{0-}(\mathcal{G}_\Gamma^*)$. Given $y \in \partial_{\mathcal{G}_\Gamma^{*,r}}(a)$, the condition $y \in V_{0-}(\mathcal{G}_\Gamma^*)$ holds if the second test that y belongs to is negative as well. As x is connected uniformly at random to a test, this happens as long as there are no two infected individuals of distance at most 4 in $\mathcal{G}_\Gamma^*$. Denote the individuals up to the fourth neighborhood as
$$V_x^4 = \left\{x' \in V(\mathcal{G}_\Gamma^*) : 1 \le \mathrm{dist}_{\mathcal{G}_\Gamma^*}(x, x') \le 4\right\}.$$
Clearly,
$$\left|V_x^4\right| \le 2\Gamma^2 \qquad \text{and, therefore,} \qquad \mathbb{E}\left[\left|V_x^4 \cap V_1(\mathcal{G}_\Gamma^*)\right| \mid \boldsymbol{\sigma}_x = 1\right] \le O\left(\Gamma^2 k/n\right).$$
As x was chosen randomly in the beginning, it is infected with probability $O(k/n)$. Furthermore, because x gets connected uniformly at random to a test, the infection status of x is independent of the infection status of the vertices in its neighborhood given $\mathcal{G}_\Gamma^{*,r}$ (as the graph was constructed independently). Thus, the probability that there is an infected individual $x \in X$ that gets connected to a test such that V_x^4 contains an infected individual is bounded by a union bound by
$$O\left(n \cdot \Gamma^2 (k/n)^2\right) = o(1) \quad (\text{as } \theta < 1/2).$$
Hence, for all $y \in \partial_{\mathcal{G}_\Gamma^*}(a)$ we find $y \in V_{0-}(\mathcal{G}_\Gamma^*)$, and DD succeeds at inferring x as well as all other vertices, given that it succeeded on G.

Case B: Connecting to a positive test. Suppose that an individual $x \in X$ connects to a (previously) positive test a. Therefore, there exists at least one $y \in V_1(\mathcal{G}_\Gamma^{*,r}) \cap \partial_{\mathcal{G}_\Gamma^{*,r}}(a)$. As DD succeeds on G by assumption, this y is element of $V_{1--}(\mathcal{G}_\Gamma^{*,r})$.

- **Case B-1: $\boldsymbol{\sigma}_x = 1$.** Similarly as in Case A-2, the probability that there is $x \in X \cap V_1(\mathcal{G}_\Gamma^*)$ connecting to an already positive test is $o(1)$. Indeed, as adding x to a is independent from the construction of $\mathcal{G}_\Gamma^{*,r}$, the probability that there are two positive individuals in one test is at most $O\left(k^2/n^2\right)$, a union bound shows that the probability that such an x exists is at most $O\left(k^2/n\right) = o(1)$, as $\theta < 1/2$. Hence, w.h.p., Case B-1 does not occur.
- **Case B-2: $\boldsymbol{\sigma}_x = 0$.** We apply a similar argument as the previous case. We first claim that y is not only part of $V_{1--}(\mathcal{G}_\Gamma^{*,r})$ because of a, but that the second test y belongs to (call this test a') consists of y and individuals from $V_{0-}(\mathcal{G}_\Gamma^*)$ (and, thus, of individuals from $V_{0-}(\mathcal{G}_\Gamma^{*,r})$). Suppose that was not the case; then there needs to be an infected individual within the fourth neighborhood of y. As proven in Case A-2, w.h.p., this does not occur. Hence, DD is able to identify y in $\mathcal{G}_\Gamma^*$ given that this was possible in $\mathcal{G}_\Gamma^{*,r}$. It remains to argue that DD classifies x correctly. As $x \in V_0^+(\mathcal{G}_\Gamma^*)$ and DD does not produce false positives by definition (see Corollary 2.4), the declared infection status of x is correct.

Thus, the lemma follows by combining the above cases. □

Now we have all ingredients to proof Theorem 4.18

Proof of Theorem 4.18. By construction, we find that $\mathcal{G}_\Gamma^*$ consists of n individuals and $m = 2n/(\Gamma+1)$ tests. By Lemma 4.19 this m suffices for DD to succeed on the regular part of $\mathcal{G}_\Gamma^*$. Furthermore, Lemma 4.21 guarantees that this is also enough for DD to infer $\boldsymbol{\sigma}$ from $\mathcal{G}_\Gamma^*$ and $\hat{\boldsymbol{\sigma}}$ correctly w.h.p., and the theorem follows. □

4.6. **Putting the pieces together.** Theorem 4.10 proves that DD succeeds on the bi-regular graph $\mathcal{G}_\Gamma$ created by the configuration model using $\max\left\{2, 1 + \left\lfloor \frac{\theta}{1-\theta} \right\rfloor\right\}$ tests. Furthermore, as Theorem 4.18 shows, for $\theta < 1/2$, $\frac{2n}{\Gamma+1}$ tests suffice employing $\mathcal{G}_\Gamma^*$. By the definition of $\tilde{\mathcal{G}}_\Gamma$, we employ $\mathcal{G}_\Gamma^*$ if and only if $\theta < 1/2$, in which case we have $\left\lceil \frac{\theta}{1-\theta} \right\rceil = 1$ and hence $\frac{2}{\Gamma+1} < \left\lceil \frac{\theta}{1-\theta} \right\rceil \frac{1}{\Gamma}$ in the lower bound.

As we established an information-theoretic lower bound on the number of tests $m_{\mathrm{inf},\Gamma}$ for the Γ-sparse group testing such that any attempt with fewer tests will fail (see (4.1)) before, Theorems 4.10 and Theorem 4.18 imply that the DD- Algorithm is information-theoretically optimal when combined with the pooling scheme $\tilde{\mathcal{G}}_\Gamma$.

5. Adaptive Group Testing with Δ-Divisible Individuals

In this section, we turn to adaptive testing strategies in the case of Δ-divisible individuals, and demonstrate that in certain cases the number of tests can be reduced significantly.

Require: Number of individuals n, number of infected individuals k, and divisibility of each individual Δ

Initialize $\widetilde{n} \leftarrow \left(\frac{n}{k}\right)^{\frac{\Delta-1}{\Delta}}$ and the estimate $\widehat{\mathscr{K}} \leftarrow \emptyset$
Arbitrarily group the n individuals into $n/\widetilde{n}$ groups of size $\widetilde{n}$
Test each group and discard any that return negative
Label the remaining groups incrementally as $G_j^{(0)}$, where $j = 1, 2, \ldots$
for $i = 1$ to $\Delta - 1$ **do**
  for each group $G_j^{(i-1)}$ from the previous stage **do**
    Arbitrarily group all individuals in $G_j^{(i-1)}$ into $\widetilde{n}^{1/(\Delta-1)}$ sub-groups of size $\widetilde{n}^{1-i/(\Delta-1)}$
    Test each sub-group and discard any that return a negative outcome
    Label the remaining sub-groups incrementally as $G_j^{(i)}$
Add the individuals from all of the remaining singleton groups $G_j^{(\Delta-1)}$ to $\widehat{\mathscr{K}}$
return $\widehat{\mathscr{K}}$

Algorithm 2: Adaptive algorithm for Δ-divisible individuals

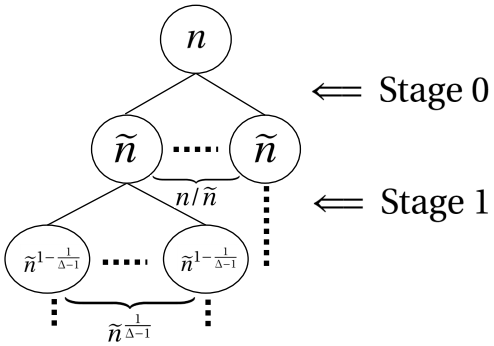

FIGURE 2. Visualization of splitting in the adaptive algorithm.

5.1. **Converse.** Recall that the converse bound proved in Theorem 3.1 already considered adaptive test designs. Thus, any adaptive strategy fails w.h.p.when $m \le (1-\varepsilon)\mathrm{e}^{-1}\Delta k^{1+\frac{(1-\theta)}{\Delta\theta}}$ for fixed $\varepsilon > 0$.

5.2. **Algorithm.** We present an algorithm that can be viewed as an analog of Hwang's binary splitting algorithm [22], instead using *non-binary* splitting in order to ensure that each item is in at most Δ tests. Like with Hwang's algorithm, we assume that the size k of the infected set is known. In the case case that only an upper bound $k_{\max} \ge k$ is known, the same analysis and results apply with $k_{\max}$ in place of k. However, such bounds may somewhat loose, and care should be taken in using initial tests to estimate k as an initial step (e.g., see [13, 17, 8]), as this may use a significant portion of the Δ budget. For clarity, we only consider the case of known k in this section, and leave the case of unknown k to future work (see also [38] for some initial findings).

5.2.1. *Recovering the infected Set.* Our adaptive algorithm is described in Algorithm 2, where we assume for simplicity that $\left(\frac{n}{k}\right)^{1/\Delta}$ is an integer.[6] Using Algorithm 2, we have the following theorem, which is proved throughout the remainder of the subsection. We define

$$m_{\mathrm{ada}}(\Delta) = \Delta k^{1+\frac{1-\theta}{\theta\Delta}}. \tag{5.1}$$

Theorem 5.1. *For $\Delta = o(\ln n)$ and $k = n^\theta$ with $\theta \in (0,1)$, the adaptive algorithm in Algorithm 2 tests each individual at most Δ times and uses at most $m_{\mathrm{ada}}(\Delta)(1+o(1))$ tests to recover the infected set exactly with zero error probability.*

Proof. Similar to Hwang's generalized binary splitting algorithm [22], the idea behind the parameter $\widetilde{n}$ in Algorithm 2 is that when k becomes large, having large groups during the initial splitting stage is wasteful, as it results in each test having a high probability of being positive (not very informative). Hence, we want to find the appropriate group sizes that result in more informative tests to minimize the number of tests. Each

[6]Note that we assume $k = o(n)$ and $\Delta = o\left(\ln\left(\frac{n}{k}\right)\right)$, meaning that $\left(\frac{n}{k}\right)^{1/\Delta} \to \infty$. Hence, the effect of rounding is asymptotically negligible, and is accounted for by the $1 + o(1)$ term in Theorem 5.1.

Require: Number of individuals n, number of infected individuals k, and test size restriction Γ
Initialize infected set $\mathscr{K} \leftarrow \emptyset$
Randomly group n individuals into n/Γ groups of size Γ
for each group G_i where $i \in \mathbb{Z} : i \in [1, n/\Gamma]$ **do**
  while testing G_i returns a positive outcome **do**
    run Algorithm 4 on a copy of G_i, and add its one infected individual output k^* into $\mathscr{K}$
    $G_i \leftarrow G_i \setminus \{k^*\}$
return $\mathscr{K}$

Algorithm 3: Adaptive algorithm for Γ-sparse tests

Require: a group of individuals $\tilde{G}$
while $\tilde{G}_i$ consists of multiple individuals **do**
  Pick half of the individuals in $\tilde{G}$ and call this set $\tilde{G}'$. Perform a single test on $\tilde{G}'$.
  If the test is positive, set $\tilde{G} \leftarrow \tilde{G}'$. Otherwise, set $\tilde{G} \leftarrow \tilde{G} \setminus \tilde{G}'$.
return single individual in $\tilde{G}$

Algorithm 4: Binary splitting

stage (outermost for-loop in Algorithm 2) here refers to the process where all groups of the same sizes are split into smaller groups (e.g., see Figure 2). We let $\tilde{n}$ be the group size at the initial splitting stage of the algorithm. The algorithm first tests $n/\tilde{n}$ groups of size $\tilde{n}$ each,[7] then steadily decrease the sizes of each group down the stages: $\tilde{n} \to \tilde{n}^{1-1/(\Delta-1)} \to \tilde{n}^{1-2/(\Delta-1)} \to \cdots \to 1$ (see Figure 2). Hence, we have $n/\tilde{n}$ groups in the initial splitting and $\tilde{n}^{\frac{1}{\Delta-1}}$ groups in all subsequent splits.

With the above observations, we can derive an upper bound on the total number of tests needed. We have $n/\tilde{n}$ tests in the first stage. Since we have k infected and split into $\tilde{n}^{\frac{1}{\Delta-1}}$ sub-groups in subsequent stages, the number of smaller groups that each stage can produce is at most $k\tilde{n}^{\frac{1}{\Delta-1}}$. This implies that the number of tests conducted at each stage is at most $k\tilde{n}^{\frac{1}{\Delta-1}}$, giving the following bound on m:

$$m \le \frac{n}{\tilde{n}} + (\Delta-1)k\tilde{n}^{\frac{1}{\Delta-1}}. \tag{5.2}$$

We optimize with respect to $\tilde{n}$ by differentiating the upper bound and setting it to zero. This gives $\tilde{n} = \left(\frac{n}{k}\right)^{\frac{\Delta-1}{\Delta}} = n^{\frac{(1-\theta)(\Delta-1)}{\Delta}}$, and substituting $\tilde{n} = \left(\frac{n}{k}\right)^{\frac{\Delta-1}{\Delta}}$ into the general upper bound in (5.2) gives the following upper bound:

$$m \le \frac{n}{(n/k)^{\frac{\Delta-1}{\Delta}}} + (\Delta-1)k\left(\left(\frac{n}{k}\right)^{\frac{\Delta-1}{\Delta}}\right)^{\frac{1}{\Delta-1}} = \Delta k\left(\frac{n}{k}\right)^{\frac{1}{\Delta}} = \Delta k^{1+\frac{1-\theta}{\theta\Delta}}. \tag{5.3}$$

□

Comparisons: We observe that $m_{\text{ada}}(\Delta)$ matches the universal lower bound in Theorem 3.1 to within a factor of e for all $\theta \in (0,1)$. For $\theta < \frac{1}{2}$, we have $m_{\text{ada}}(\Delta) = m_{\text{DD}}(\Delta) = \Delta k^{1+\frac{(1-\theta)}{\Delta\theta}}$, meaning that the best known bounds for the adaptive and non-adaptive settings are identical (though the adaptive algorithm attains *zero* error probability). In contrast, for $\theta > \frac{1}{2}$, we have $m_{\text{DD}}(\Delta) = \Delta k^{1+\frac{1}{\Delta}}$ and $m_{\text{ada}}(\Delta) = \Delta k^{1+\frac{(1-\theta)}{\Delta\theta}}$. The former is significantly higher, and Theorem 3.2 reveals that this limitation is inherent to *any* non-adaptive test design and algorithm. Hence, for $\theta > \frac{1}{2}$, there is a significant gap between the number of tests required by adaptive and non-adaptive algorithms.

6. Adaptive Group Testing with Γ-Sized Tests

Our adaptive algorithm with Γ-sparse tests, shown in Algorithm 3, is again a modification of Hwang's generalized binary splitting algorithm [22], where we initially divide the n individuals into $\frac{n}{\Gamma}$ groups of size Γ, instead of k groups of size $\frac{n}{k}$ as in the original algorithm.

Our main result is stated as follows, in which we define

$$m_{\text{ada}}(\Gamma) = \frac{n}{\Gamma} + k\log_2\Gamma. \tag{6.1}$$

Theorem 6.1. *For any $\Gamma = o\left(\frac{n}{k}\right)$, Algorithm 3 outputs the correct configuration of infection statuses with probability one, while using at most $m_{\text{ada}}(\Gamma)(1+o(1))$ tests, each containing at most Γ items.*

[7] Note that $n/\tilde{n}$ is an integer for our chosen $\tilde{n}$ below, which gives $\frac{n}{\tilde{n}} = k\left(\frac{n}{k}\right)^{1/\Delta}$, and $\left(\frac{n}{k}\right)^{1/\Delta}$ was already assumed to be an integer.

Proof. Let k_i be the number of infected individuals in each of the initial $\frac{n}{\Gamma}$ groups. Note that since $\Gamma = o\left(\frac{n}{k}\right)$ implies $k = o(\frac{n}{\Gamma})$, most groups will not have a infected individual. In the binary splitting stage of the algorithm, we can round the halves in either direction if they are not an integer. Hence, for each of the initial $\frac{n}{\Gamma}$ groups, we take at most $\lceil \log_2 \Gamma \rceil$ adaptive tests to find a infected individual, or one test to confirm that there are no infected individuals. Therefore, for each of the initial $\frac{n}{\Gamma}$ groups, we need $\max\{1, k_i \log_2 \Gamma + O(k_i)\}$ tests to find k_i infected individuals. Summing across all $\frac{n}{\Gamma}$ groups, we need a total of $m = \sum_{i=1}^{n/\Gamma} \max\{1, k_i \log_2 \Gamma + O(k_i)\}$ tests. This has the following upper bound:

$$m \le \frac{n}{\Gamma} + k\log_2\Gamma + O(k) \overset{(a)}{=} \frac{n}{\Gamma}(1+o(1)) + k\log_2\Gamma = m_{\text{ada}}(\Gamma)(1+o(1)), \tag{6.2}$$

where (a) uses $k = o\left(\frac{n}{\Gamma}\right)$. □

If we slightly strengthen the requirement $\Gamma = o\left(\frac{n}{k}\right)$ to $\Gamma = o\left(\frac{n}{k\ln(n/k)}\right)$ (which, in particular, includes the regime $\Gamma = \left(\frac{n}{k}\right)^{1-\Omega(1)}$ studied in [20]), then we have $\frac{n}{\Gamma} = \omega\left(k\ln\left(\frac{n}{k}\right)\right)$ and hence $\frac{n}{\Gamma} = \omega(k\ln\Gamma)$. Thus, we obtain

$$m_{\text{ada}}(\Gamma) = \frac{n}{\Gamma}(1+o(1)). \tag{6.3}$$

This simplified upper bound is tight, due the simple fact that $\frac{n}{\Gamma}(1-o(1))$ tests (of size at most Γ) are needed just to test a fraction $1-o(1)$ of the items at least once each (which is a minimal requirement for recovering $\boldsymbol{\sigma}$ w.h.p.). Formally, this argument reveals the following.

Theorem 6.2. *In the setup of Γ-sparse tests with $k = n^\theta$ for some $\theta \in (0,1)$, any (possibly adaptive) group testing procedure that recovers $\boldsymbol{\sigma}$ w.h.p. must use at least $\frac{n}{\Gamma}(1-o(1))$ tests.*

7. Auxiliary Results

Lemma 7.1 (Corollary 2.3 of [25])**.** *Letting $\phi(t) = (1+t)\ln(1+t) - t$, if $\boldsymbol{X}$ is a binomial random variable, then we have for all $t > 0$ that*

$$\mathbb{P}(|\boldsymbol{X} - \mathbb{E}[\boldsymbol{X}]| \ge t\mathbb{E}[\boldsymbol{X}]) \le \exp(-\phi(t)\mathbb{E}[\boldsymbol{X}]).$$

The following variant of Hoeffding's inequality is very convenient to work with. As the result is not known so broadly, we include a short proof.

Lemma 7.2 (Hoeffding-type Inequality)**.** *Let $\boldsymbol{X}_1, \dots, \boldsymbol{X}_n$ be independent random variables such that $0 \le \boldsymbol{X}_i \le 1$ a.s. and $\mathbb{E}[\boldsymbol{X}_i] = p_i$. Furthermore, denote by $p = n^{-1}\sum_{i=1}^n p_i$ the average expectation, $\boldsymbol{X} = \sum_{i=1}^n \boldsymbol{X}_i$ and let $t \in (0,p)$. Then,*

$$\mathbb{P}\left(\boldsymbol{X} \le (p-t)n\right) \le \exp\left(-nD_{\text{KL}}\left(p-t\|p\right)\right).$$

Proof. Let $\lambda > 0$. Markov's inequality guarantees

$$\mathbb{P}\left(\exp(\lambda\boldsymbol{X}) \le \exp\left(\lambda(p-t)n\right)\right) \le \frac{\mathbb{E}\left[\exp(\lambda\boldsymbol{X})\right]}{\exp\left(\lambda(p-t)n\right)} = \frac{\mathbb{E}\left[\exp\left(\sum_{i=1}^n \lambda\boldsymbol{X}_i\right)\right]}{\exp\left(\lambda(p-t)n\right)}. \tag{7.1}$$

As $x \mapsto \exp(\lambda z)$ is convex, we find

$$\mathbb{E}\left[\exp(\lambda\boldsymbol{X}_i)\right] \le 1 - p_i + p_i\exp(\lambda).$$

Therefore, by the inequality of arithmetic and geometric means,

$$\mathbb{E}\left[\exp(\lambda\boldsymbol{X})\right] \le \left(1-p+p\exp(\lambda)\right)^n. \tag{7.2}$$

By (7.1) and (7.2), we have

$$\mathbb{P}\left(\exp(\lambda\boldsymbol{X}) \le \exp\left(\lambda(p-t)n\right)\right) \le \exp\left(n\left(\ln\left(1-p+p\exp(\lambda)\right) - \lambda(p-t)\right)\right) \tag{7.3}$$

Proceeding as in the standard proof of the Chernoff bound, we find that the r.h.s. of (7.3) is minimised with respect to λ for $\exp(\lambda) = \frac{(1-p)(p-t)}{p(1-(p-t))}$. Therefore, (7.1) gives

$$\mathbb{P}\left(\exp(\lambda\boldsymbol{X}) \le \exp\left(\lambda(p-t)n\right)\right) \le \exp\left(n\ln\left(\left(\frac{p}{p-t}\right)^{p-t}\left(\frac{1-p}{1-(p-t)}\right)^{1-(p-t)}\right)\right) = \exp\left(-nD_{\text{KL}}\left(p-t\|p\right)\right),$$

implying the assertion of the lemma. □

Lemma 7.3 (Stirling Approximation, [30])**.** *We have for $n \to \infty$ that*

$$n! = (1+O(1/n))\sqrt{2\pi n}n^n\exp(-n).$$

Claim 7.4. *Let $n > 0$, $\Delta = \ln^{O(1)} n$ be integers, and let $\alpha \in (0,1)$. Then*

$$\binom{\alpha n}{\Delta}\binom{n}{\Delta}^{-1} = \left(1 + O\left(n^{-\Omega(1)}\right)\right)\alpha^{\Delta}.$$

Proof. By definition, we have

$$\frac{\binom{\alpha n}{\Delta}}{\binom{n}{\Delta}} = \frac{(\alpha n)!(n-\Delta)!}{n!(\alpha n-\Delta)!}.$$

Hence, applying Lemma 7.3 on each factor yields

(7.4)

$$\frac{\binom{\alpha n}{\Delta}}{\binom{n}{\Delta}} = (1+O(n^{-1}))\exp\left(-\alpha n + (n-\Delta) - (\alpha n-\Delta) - n\right)(\alpha n)^{\alpha n}(n-\Delta)^{n-\Delta}(\alpha n-\Delta)^{-(\alpha n-\Delta)}n^{-n}\sqrt{\frac{(\alpha n)(n-\Delta)}{n(\alpha n-\Delta)}}.$$

As $\Delta = \ln^{O(1)} n$, we find that (7.4) equals

$$\frac{\binom{\alpha n}{\Delta}}{\binom{n}{\Delta}} = \left(1 + O\left(n^{-\Omega(1)}\right)\right)(\alpha n)^{\alpha n} n^{n} (\alpha n)^{-(\alpha n-\Delta)} n^{-n} = \left(1 + O\left(n^{-\Omega(1)}\right)\right)\alpha^{\Delta}, \tag{7.5}$$

and the assertion follows. □

We also use the following direct consequence of the binomial expansion.

Claim 7.5. *For any real number $x \geq -1$ and any integer $t \geq 0$ the following holds:*

$$(1+x)^t = 1 + tx + O(t^2x^2).$$

Finally, we state the following useful result relating to Stirling's approximation and the local limit theorem.

Claim 7.6. *[Appendix B1 of [11]] For any $m, \Delta \in \mathbb{N}, \theta \in (0,1)$, $k \sim n^{\theta}$, let $(\boldsymbol{X}_i)_{i\in[m]}$ denote a sequence of independent $\mathrm{Bin}(\Gamma_i, k/n)$ and define*

$$\mathscr{E} = \left\{ \sum_{i\in[m]} \boldsymbol{X}_i = k\Delta \right\}.$$

Then, we have $\mathbb{P}(\mathscr{E}) = \Omega(1/\sqrt{n\Delta})$.

8. Simulations

In Figures 3 and 4, we compare our theoretical findings to empirical results obtained as follows:

- In the non-adaptive case, we fix the number of individuals n, the infection parameter θ, and, depending on the setup considered, the individual degree Δ or test degree Γ. We vary the number m of tests (x-axis), and simulate 10^4 independent trials per parameter set. DD's performance (y-axis) is reported as the fraction of simulations per parameter point that inferred the infected set without errors.
- In the adaptive case, we cannot directly control the number of tests m *a priori*. Instead, we fix the same parameter set as in the non-adaptive case, and carry out 10^6 simulations. We then report the cumulative distribution of tests required, i.e., the y-value corresponding to some m is given as the fraction of runs that required at most m tests.

We observe that the empirical results are consistent with our theoretical thresholds in all cases. The adaptive testing strategies show a particularly rapid transition at $m_{\mathrm{ada}}(\Delta)$ and $m_{\mathrm{ada}}(\Gamma)$ respectively. We find that the non-adaptive DD algorithm requires more tests in comparison to the adaptive schemes, and has a much broader range of transient behaviour. This suggests that *convergence rates* to the first-order asymptotic threshold may reveal an even wider gap between adaptive and non-adaptive designs, in analogy with studies of channel coding [36]. Note that the change of slope in Figure 4 (right) at m=2000 is due to rounding of Δ.

9. Conclusion

We have studied the information-theoretic and algorithmic thresholds of group testing with constraints on the number of items-per-test or test-per-item. For Δ-divisible items, we proved that at least for $\Delta = \omega(1)$, the DD algorithm is asymptotically optimal for $\theta > \frac{1}{2}$, and is optimal to within a factor of e for all $\theta \in (0,1)$, thus significantly improving on existing bounds for the COMP algorithm having suboptimal scaling laws. For Γ-sized tests with $\Gamma = \Theta(1)$, we improved on both the best known upper bounds and lower bounds, established a precise threshold for almost all $\theta \in (0,1)$, and introduced a new randomised test design for $\theta > \frac{1}{2}$. In both

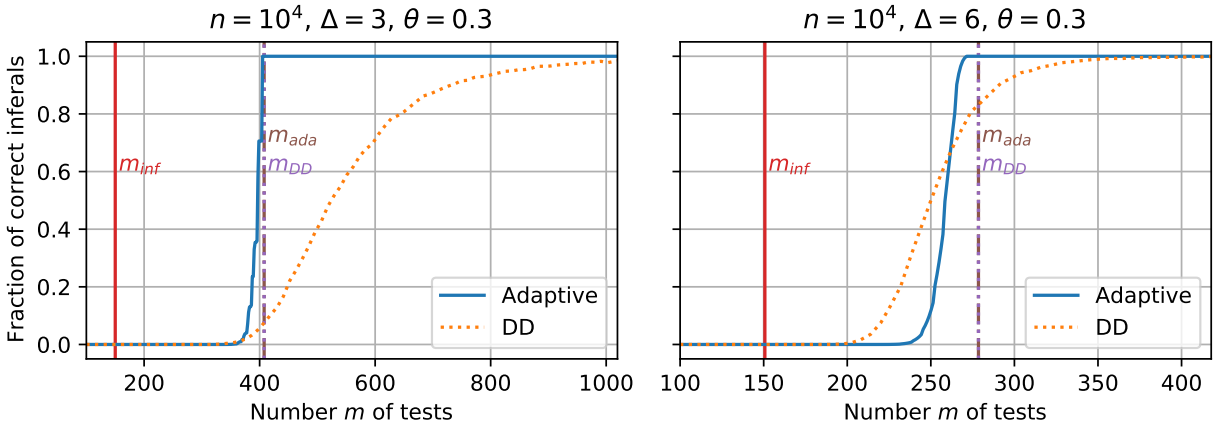

FIGURE 3. Performance of adaptive and non-adaptive Δ-divisible algorithms as function of number of tests.

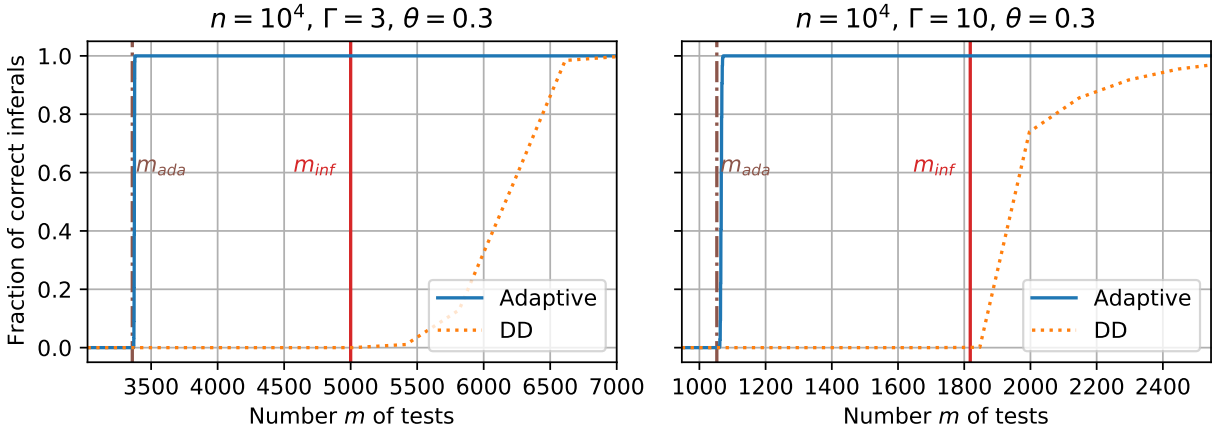

FIGURE 4. Performance of adaptive and non-adaptive Γ-sparse algorithms as function of number of tests.

settings, we additionally provided near-optimal adaptive algorithms, and demonstrated a strict gap between the number of tests for adaptive and non-adaptive designs in broad scaling regimes.

REFERENCES

[1] M. Aldridge. Individual testing is optimal for non-adaptive group testing in the linear regime. *IEEE Transactions on Information Theory*, 65:2058–2061, 2019.

[2] M. Aldridge, L. Baldassini, and O. Johnson. Group testing algorithms: bounds and simulations. *IEEE Transactions on Information Theory*, 60:3671–3687, 2014.

[3] M. Aldridge, O. Johnson, and J. Scarlett. Improved group testing rates with constant column weight designs. *IEEE Transactions on Information Theory*, 65(2):1381–1385, 2016.

[4] M. Aldridge, O. Johnson, and J. Scarlett. *Group testing: an information theory perspective*. Foundations and Trends in Communications and Information Theory, 2019.

[5] R. Ash. *Information Theory*. Dover Publications Inc., New York, 1990.

[6] L. Baldassini, O. Johnson, and M. Aldridge. The capacity of adaptive group testing. *Proc. ISIT*, 1:2676–2680, 2013.

[7] R. Benz, S. Swamidass, and P. Baldi. Discovery of power-laws in chemical space. *Journal of Chemical Information and Modeling*, 48:1138–1151, 2008.

[8] N. Bshouty, V. Bshouty-Hurani, T. Hashem, and O. Sharafy. Adaptive group testing algorithms to estimate the number of defectives. *Algorithmic Learning Theory*, 2018.

[9] C. Chan, P. Che, S. Jaggi, and V. Saligrama. Non-adaptive probabilistic group testing with noisy measurements: near-optimal bounds with efficient algorithms. *49th Annual Allerton Conference on Communication, Control, and Computing*, 1:1832–1839, 2011.

[10] I. Cheong. The experience of south korea with covid-19. *Mitigating the COVID Economic Crisis: Act Fast and Do Whatever It Takes(CEPR Press)*, pages 113–120, 2020.

[11] A. Coja-Oghlan, O. Gebhard, M. Hahn-Klimroth, and P. Loick. Information-theoretic and algorithmic thresholds for group testing. *46th International Colloquium on Automata, Languages, and Programming (ICALP 2019)*, 132(43):1–14, 2019.

[12] A. Coja-Oghlan, O. Gebhard, M. Hahn-Klimroth, and P. Loick. Optimal group testing. *33rd Annual Conference on Learning Theory (COLT 2020)*, 125:1374–1388, 2020.

[13] P. Damaschke and A. Muhammad. Competitive group testing and learning hidden vertex covers with minimum adaptivity. *Disc. Math., Algs. and Apps.*, 2(03):291–311, 2010.
[14] R. Dorfman. The detection of defective members of large populations. *Annals of Mathematical Statistics*, 14:436–440, 1943.
[15] D. Du and F. Hwang. *Combinatorial group testing and its application*. World Scientific, Singapore, 1993.
[16] S. Ciesek E. Seifried. Pool testing of sars-cov-02 samples increases worldwide test capacities many times over, 2020. https://www.bionity.com/en/news/1165636/pool-testing-of-sars-cov-02-samples-increases-worldwide-test-capacities-many-times-over.html, last accessed on 2020-04-08.
[17] M. Falahatgar, A. Jafarpour, A. Orlitsky, V. Pichapati, and A. Suresh. Estimating the number of defectives with group testing. In *IEEE Int. Symp. Inf. Theory*, pages 1376–1380, 2016.
[18] European Centre for Disease Prevention and Control. Surveillance and studies in a pandemic in europe. *ECDC Technical Report*, 2009.
[19] C. Fortuin, P. Kasteleyn, and J. Ginibre. Correlation inequalities on some partially ordered sets. *Communications in Mathematical Physics*, 22:89–103, 1971.
[20] Venkata Gandikota, Elena Grigorescu, Sidharth Jaggi, and Samson Zhou. Nearly optimal sparse group testing. *IEEE Transactions on Information Theory*, 65(5):2760–2773, May 2019.
[21] Y. Gefen, M. Szwarcwort-Cohen, and R. Kishony. Pooling method for accelerated testing of covid-19. *https://www.technion.ac.il/en/2020/03/pooling-method-for-accelerated-testing-of-covid-19/*, 03/26/20.
[22] F. Hwang. A method for detecting all defective members in a population by group testing. *Journal of the American Statistical Association*, 67:605–608, 1972.
[23] H. Inan, K. Kairouz, and A. Ozgur. Sparse group testing codes for low-energy massive random access. *55th Annual Allerton Conference*, 1:658—-665, 2017.
[24] H. A. Inan, P. Kairouz, and A. Ozgur. Sparse combinatorial group testing. *IEEE Transactions on Information Theory*, 66(5):2729–2742, 2020.
[25] S. Janson, T. Luczak, and A. Rucinski. *Random Graphs*. John Wiley and Sons, 2011.
[26] O. Johnson, M. Aldridge, and J. Scarlett. Performance of group testing algorithms with near-constant tests per item. *IEEE Transactions on Information Theory*, 65:707–723, 2018.
[27] H. Kwang-Ming and D. Ding-Zhu. Pooling designs and nonadaptive group testing: important tools for dna sequencing. *World Scientific*, 2006.
[28] A. Macula. A simple construction of d-disjunct matrices with certain constant weights. *Discrete Mathematics*, 162:311–312, 1996.
[29] N. Madhav, B. Oppenheim, M. Gallivan, P. Mulembakani, E. Rubin, and N. Wolfe. Pandemics: Risks, impacts and mitigation. *The World Bank:Disease control priorities*, 9:315–345, 2017.
[30] A. J. Maria. A remark on stirling's formula. *The American Mathematical Monthly*, 72(10):1096, 1965.
[31] R. Mourad, Z. Dawy, and F. Morcos. Designing pooling systems for noisy high-throughput protein-protein interaction experiments using boolean compressed sensing. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 10:1478–1490, 2013.
[32] M. Newman. Networks: An introduction. *Oxford University Press*, 2010.
[33] H. Ngo and D. Du. A survey on combinatorial group testing algorithms with applications to dna library screening. *Discrete Mathematical Problems with Medical Applications*, 7:171–182, 2000.
[34] U.S. Department of Health and Human Services. Pandemic influenza plan. *Planning and Preparedness Ressources*, 2017.
[35] World Health Origanisation. Global surveillance during an influenza pandemic. *Global Influenza Program*, 2009.
[36] Yury Polyanskiy, H Vincent Poor, and Sergio Verdú. Feedback in the non-asymptotic regime. *IEEE Transactions on Information Theory*, 57(8):4903–4925, 2011.
[37] J. Scarlett and V. Cevher. Phase transitions in group testing. *Proceedings of the 27th Annual ACM-SIAM Symposium on Discrete Algorithms(SODA 2016)*, 1:40–53, 2016.
[38] N. Tan and J. Scarlett. Near-optimal sparse adaptive group testing. In *IEEE International Symposium on Information Theory (ISIT)*, 2020.
[39] N. Thierry-Mieg. A new pooling strategy for high-throughput screening: the shifted transversal design. *BMC Bioinformatics*, 7:28, 2006.
[40] L. Wang, X. Li, Y. Zhang, and K. Zhang. Evolution of scaling emergence in large-scale spatial epidemic spreading. *PLoS ONE*, 6, 2011.
[41] L. Wein and S. Zenios. Pooled testing for HIV screening: Capturing the dilution effect. *Operations Research*, 44:543–569, 1996.
[42] B. Wu. The weighted version of the handshaking lemma. *Journal of inequalities and application*, 351, 2014.

OLIVER GEBHARD, gebhard@math.uni-frankfurt.de, GOETHE UNIVERSITY, INSTITUTE OF MATHEMATICS, 10 ROBERT MAYER ST, FRANKFURT 60325, GERMANY.

MAX HAHN-KLIMROTH, hahnklim@math.uni-frankfurt.de, GOETHE UNIVERSITY, INSTITUTE OF MATHEMATICS, 10 ROBERT MAYER ST, FRANKFURT 60325, GERMANY.

OLAF PARCZYK, o.parczyk@lse.ac.uk, LONDON SCHOOL OF ECONOMICS, DEPARTMENT OF MATHEMATICS, HOUGHTON ST, LONDON, WC2A 2AE, UK.

MANUEL PENSCHUCK, mpenschuck@ae.cs.uni-frankfurt.de, GOETHE UNIVERSITY, INSTITUTE OF COMPUTER SCIENCES , 11-15 ROBERT MAYER ST, FRANKFURT 60325, GERMANY.

MAURICE ROLVIEN, rolvien@math.uni-frankfurt.de, GOETHE UNIVERSITY, INSTITUTE OF MATHEMATICS, 10 ROBERT MAYER ST, FRANKFURT 60325, GERMANY.

JONATHAN SCARLETT, scarlett@comp.nus.edu.sg, NATIONAL UNIVERSITY OF SINGAPORE, DEPARTMENT OF COMPUTER SCIENCE & DEPARTMENT OF MATHEMATICS.

NELVIN TAN, nelvintan@u.nus.edu, NATIONAL UNIVERSITY OF SINGAPORE, DEPARTMENT OF COMPUTER SCIENCE.

E. The number of satisfying assignments of random 2-SAT formulas

THE RANDOM 2-SAT PARTITION FUNCTION

DIMITRIS ACHLIOPTAS, AMIN COJA-OGHLAN, MAX HAHN-KLIMROTH, JOON LEE, NOËLA MÜLLER, MANUEL PENSCHUCK, GUANGYAN ZHOU

ABSTRACT. We show that throughout the satisfiable phase the normalised number of satisfying assignments of a random 2-SAT formula converges in probability to an expression predicted by the cavity method from statistical physics. The proof is based on showing that the Belief Propagation algorithm renders the correct marginal probability that a variable is set to 'true' under a uniformly random satisfying assignment. *MSC:* 05C80, 60C05, 68Q87

1. INTRODUCTION

1.1. **Background and motivation.** The random 2-SAT problem was the first random constraint satisfaction problem whose satisfiability threshold could be pinpointed precisely, an accomplishment attained independently by Chvátal and Reed [14] and Goerdt [30] in 1992. The proofs evince the link between the 2-SAT threshold and the percolation phase transition of a random digraph. This connection subsequently enabled Bollobás, Borgs, Chayes, Kim and Wilson [11] to identify the size of the scaling window, which matches that of the giant component phase transition of the Erdős-Rényi random graph [10, 33]. Ramifications and extensions of these results pertain to random 2-SAT formulas with given literal degrees [19], the random MAX 2-SAT problem [20] and the performance of algorithms [45]. But despite the great attention devoted to random 2-SAT over the years, a fundamental question, mentioned prominently in the survey [28], remained conspicuously open: *how many satisfying assignments does a random* 2*-SAT formula typically possess?* While percolation-type arguments have been stretched to derive (rough) bounds [12], the exact answer remained beyond the reach of elementary techniques.

In addition to the mathematical literature, the 2-SAT problem attracted the interest of statistical physicists, who brought to bear a canny but non-rigorous approach called the cavity method [36, 37]. Instead of relying on percolation ideas, the physics *ansatz* seizes upon a heuristic message passing scheme called Belief Propagation. Its purpose is to calculate the marginal probabilities that a random satisfying assignment sets specific variables of the 2-SAT formula to 'true'. According to physics intuition Belief Propagation reveals a far more fine-grained picture than a mere percolation argument possibly could. Indeed, in combination with a functional called the Bethe free entropy, Belief Propagation renders a precise conjecture as to the number of satisfying assignments.

We prove this conjecture. Specifically, we show that for all clause-to-variable densities below the 2-SAT threshold the number of satisfying assignments is determined by the Bethe functional applied to a particular solution of a stochastic fixed point equation that mimics Belief Propagation. The formula that we obtain does not boil down to a simple algebraic expression, which may explain why the problem has confounded classical methods for nearly three decades. Nonetheless, thanks to rapid convergence of the stochastic fixed point iteration, the formula can be evaluated numerically within arbitrary precision. A crucial step towards the main theorem is to verify that Belief Propagation does indeed yield the correct marginals, a fact that may be of independent interest.

By comparison to prior work on Belief Propagation in combinatorics (e.g., [16, 22, 21, 39]), we face the substantial technical challenge of dealing with the 'hard' constraints of the 2-SAT problems, which demands that *all* clauses be satisfied. A second novelty is that in order to prove convergence of Belief Propagation to the correct marginals we need to investigate delicately constructed extremal boundary conditions. Since these depend on the random 2-SAT formula itself, we need to develop means to confront the ensuing stochastic dependencies between the construction of the boundary condition and the subsequent message passing iterations. We proceed to state the main results precisely. An outline of the proofs and a detailed discussion of related work follow in Sections 2 and 3.

Amin Coja-Oghlan's research received support under DFG CO 646/4. Max Hahn-Klimroth has been supported by Stiftung Polytechnische Gesellschaft. Manuel Penschuck's research received support under DFG ME 2088/3-2 and ME 2088/4-2. Guangyan Zhou is supported by National Natural Science Foundation of China, No. 61702019.

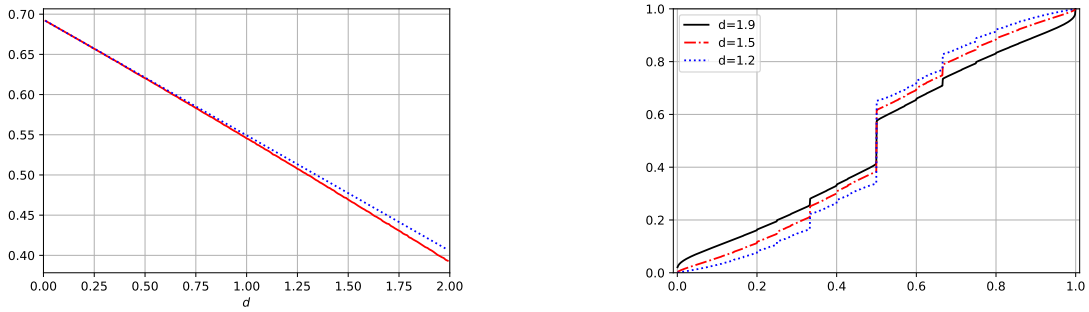

FIGURE 1. *Left:* the red line depicts a numerical approximation to the r.h.s. of (1.2) after 24 iterations of $\mathrm{BP}_d(\cdot)$. The dotted blue line displays the first moment bound. *Right:* the cumulative density functions of numerical approximations to $\mathrm{BP}_d^{24}(\delta_{1/2})$ for various d.

1.2. **The main result.** Let $n > 1$ be an integer, let $d > 0$ be a positive real and let $\boldsymbol{m} \stackrel{\mathrm{d}}{=} \mathrm{Po}(dn/2)$ be a Poisson random variable. Further, let $\boldsymbol{\Phi} = \boldsymbol{\Phi}_n$ be a random 2-SAT formula with Boolean variables $x_1, \ldots, x_n$ and $\boldsymbol{m}$ clauses, drawn uniformly and independently from the set of all $4n(n-1)$ possible clauses with two distinct variables. Thus, each variable appears in d clauses on the average and the satisfiability threshold occurs at $d = 2$. We aim to estimate the number $Z(\boldsymbol{\Phi})$ of satisfying assignments, the *partition function* in physics jargon. More precisely, since $Z(\boldsymbol{\Phi})$ remains exponentially large for all $d < 2$ w.h.p., in order to obtain a well-behaved limit we compute the normalised logarithm $n^{-1}\log Z(\boldsymbol{\Phi})$.

The result comes in terms of the solution to a stochastic fixed point equation on the unit interval. Hence, let $\mathscr{P}(0,1)$ be the set of all Borel probability measures on $(0,1)$, endowed with the weak topology. Further, define an operator $\mathrm{BP}_d : \mathscr{P}(0,1) \to \mathscr{P}(0,1)$, $\pi \mapsto \hat{\pi}$ as follows. With $\boldsymbol{d}^+, \boldsymbol{d}^-$ Poisson variables with mean $d/2$ and $\boldsymbol{\mu}_{\pi,1}, \boldsymbol{\mu}_{\pi,2}, \ldots$ random variables with distribution π, all mutually independent, let $\hat{\pi}$ be the distribution of the random variable

$$\frac{\prod_{i=1}^{\boldsymbol{d}^-}\boldsymbol{\mu}_{\pi,i}}{\prod_{i=1}^{\boldsymbol{d}^-}\boldsymbol{\mu}_{\pi,i} + \prod_{i=1}^{\boldsymbol{d}^+}\boldsymbol{\mu}_{\pi,i+\boldsymbol{d}^-}} \in (0,1). \tag{1.1}$$

Let $\delta_{1/2} \in \mathscr{P}(0,1)$ signify the atom at $1/2$ and write $\mathrm{BP}_d^\ell(\cdot)$ for the ℓ-fold application of the operator BP_d.

Theorem 1.1. *For any $d < 2$ the limit $\pi_d = \lim_{\ell\to\infty}\mathrm{BP}_d^\ell(\delta_{1/2})$ exists and*

$$\lim_{n\to\infty}\frac{1}{n}\log Z(\boldsymbol{\Phi}) = \mathbb{E}\left[\log\left(\prod_{i=1}^{\boldsymbol{d}^-}\boldsymbol{\mu}_{\pi_d,i} + \prod_{i=1}^{\boldsymbol{d}^+}\boldsymbol{\mu}_{\pi_d,i+\boldsymbol{d}^-}\right) - \frac{d}{2}\log\left(1-\boldsymbol{\mu}_{\pi_d,1}\boldsymbol{\mu}_{\pi_d,2}\right)\right] \qquad \textit{in probability.} \tag{1.2}$$

Of course, the fact that the r.h.s. of (1.2) is well-defined is part of the statement of Theorem 1.1.

By construction, the distribution π_d is a solution to the stochastic fixed point equation

$$\pi_d = \mathrm{BP}_d(\pi_d). \tag{1.3}$$

The equation (1.3) is known as the *density evolution* equation in physics lore, while the expression on the r.h.s. of (1.2) is called the Bethe free entropy [34]. Hence, Theorem 1.1 matches the conjecture from [36]. By comparison, Markov's inequality yields the elementary first moment bound

$$\frac{1}{n}\log Z(\boldsymbol{\Phi}) \le \frac{1}{n}\log\mathbb{E}[Z(\boldsymbol{\Phi})] + o(1) = (1-d)\log 2 + \frac{d}{2}\log 3 + o(1) \qquad \text{w.h.p.,} \tag{1.4}$$

which, however, fails to be tight for any $0 < d < 2$ [42]. Furthermore, while (1.2) may appear difficult to evaluate, the proof reveals that the fixed point iteration $\mathrm{BP}_d^\ell(\delta_{1/2})$ converges geometrically (in an appropriate metric). In effect, decent numerical approximations can be obtained; see Figure 1.

For $d < 1$ the random digraph on $\{x_1, \neg x_1, \ldots, x_n, \neg x_n\}$ obtained by inserting for each clause $l_1 \vee l_2$ of $\boldsymbol{\Phi}$ the two directed edges $\neg l_1 \to l_2$, $\neg l_2 \to l_1$ is sub-critical and the distribution π_d is supported on a countable set. In effect, for $d < 1$ the formula (1.2) can be obtained via elementary counting arguments. By contrast, the emergence of a weak giant component for $1 < d < 2$ turns the computation of $Z(\boldsymbol{\Phi})$ into a challenge. Finally, for $d > 2$ the digraph

contains a strongly connected giant component w.h.p. Its long directed cycles likely cause contradictions, which is why satisfying assignments cease to exist.

An asymptotically tight upper bound on $n^{-1}\log Z(\mathbf{\Phi})$ could be obtained via the interpolation method from mathematical physics [29, 42]. We will revisit this point in Section 3. Thus, the principal contribution of Theorem 1.1 is the lower bound on $\log Z(\mathbf{\Phi})$. The best prior lower bound was obtained by Boufkhad and Dubois [12] in 1999 via percolation arguments. However, this bound drastically undershoots the actual value from Theorem 1.1. For instance, for $d = 1.2$, [12] gives $n^{-1}\log Z(\mathbf{\Phi}) \geq 0.072\ldots$, while actually $n^{-1}\log Z(\mathbf{\Phi}) = 0.515\ldots$ w.h.p.

1.3. **Belief Propagation.** To elaborate on the combinatorial meaning of the distribution π_d, we need to look into the Belief Propagation heuristic. Instantiated to 2-SAT, Belief Propagation is a message passing algorithm designed to approximate the marginal probability that a specific variable takes the value 'true' under a random satisfying assignment. While finding satisfying assignments of a given 2-SAT formula is an easy computational task, calculating these marginals is not. In fact, the problem is #P-hard [49]. Nonetheless, we are going to prove that Belief Propagation approximates the marginals well on random formulas w.h.p.

To introduce Belief Propagation, we associate a bipartite graph $G(\mathbf{\Phi})$ with the formula $\mathbf{\Phi}$. One vertex class $V_n = \{x_1, \ldots, x_n\}$ represents the propositional variables, the other class $F_{\boldsymbol{m}} = \{a_1, \ldots, a_{\boldsymbol{m}}\}$ represents the clauses. Each clause a_i is adjacent to the two variables that it contains. We write $\partial v = \partial(\mathbf{\Phi}, v)$ for the set of neighbours of a vertex v of $G(\mathbf{\Phi})$. Moreover, for $\ell \geq 1$ let $\partial^\ell v$ signify the set of all vertices at distance precisely ℓ from v.

Associated with the edges of $G(\mathbf{\Phi})$, the Belief Propagation messages are probability distributions on the Boolean values 'true' and 'false'. To be precise, any adjacent clause/variable pair a, x comes with two messages, one directed from a to x and a reverse one from x to a. Encoding 'true' and 'false' by ± 1, we initialise all messages by

$$\nu^{(0)}_{\mathbf{\Phi},a\to x}(\pm 1) = \nu^{(0)}_{\mathbf{\Phi},x\to a}(\pm 1) = 1/2. \tag{1.5}$$

For $\ell \geq 1$ the messages $\nu^{(\ell)}_{\mathbf{\Phi},a\to x}, \nu^{(\ell)}_{\mathbf{\Phi},x\to a}$ are defined inductively. Specifically, suppose that clause a contains the two variables x, y. Let $r, s \in \{\pm 1\}$ indicate whether x, y appear as positive or negative literals in a. Then for $t = \pm 1$ let

$$\nu^{(\ell)}_{\mathbf{\Phi},a\to x}(t) = \frac{1 - \mathbf{1}\{r \neq t\}\,\nu^{(\ell-1)}_{\mathbf{\Phi},y\to a}(-s)}{1 + \nu^{(\ell-1)}_{\mathbf{\Phi},y\to a}(s)}, \qquad \nu^{(\ell)}_{\mathbf{\Phi},x\to a}(t) = \frac{\prod_{b\in\partial x\setminus\{a\}} \nu^{(\ell)}_{\mathbf{\Phi},b\to x}(t)}{\prod_{b\in\partial x\setminus\{a\}} \nu^{(\ell)}_{\mathbf{\Phi},b\to x}(1) + \prod_{b\in\partial x\setminus\{a\}} \nu^{(\ell)}_{\mathbf{\Phi},b\to x}(-1)}. \tag{1.6}$$

The last expression is deemed to equal $1/2$ if the denominator vanishes (which does not happen if $\mathbf{\Phi}$ is satisfiable). Finally, the Belief Propagation estimate of the marginal of a variable x after ℓ iterations reads

$$\nu^{(\ell)}_{\mathbf{\Phi},x}(t) = \frac{\prod_{a\in\partial x} \nu^{(\ell)}_{\mathbf{\Phi},a\to x}(t)}{\prod_{a\in\partial x} \nu^{(\ell)}_{\mathbf{\Phi},a\to x}(1) + \prod_{a\in\partial x} \nu^{(\ell)}_{\mathbf{\Phi},a\to x}(-1)}, \tag{1.7}$$

again interpreted to yield $1/2$ if the denominator vanishes. For an excellent exposition of Belief Propagation, including the derivation of (1.6)–(1.7), we point to [34, Chapter 14].

The next theorem establishes that (1.7) approximates the true marginals well for large ℓ. In fact, we prove a significantly stronger result. To set the stage, let $S(\mathbf{\Phi})$ be the set of all satisfying assignments of $\mathbf{\Phi}$. Assuming $S(\mathbf{\Phi}) \neq \emptyset$, let

$$\mu_{\mathbf{\Phi}}(\sigma) = \mathbf{1}\{\sigma \in S(\mathbf{\Phi})\}\,/\,Z(\mathbf{\Phi}) \qquad (\sigma \in \{\pm 1\}^{\{x_1,\ldots,x_n\}}) \tag{1.8}$$

be the uniform distribution on $S(\mathbf{\Phi})$. Further, write $\boldsymbol{\sigma}$ for a sample from $\mu_{\mathbf{\Phi}}$. Then for a satisfying assignment $\tau \in S(\mathbf{\Phi})$ and $\ell \geq 1$ the conditional distribution $\mu_{\mathbf{\Phi}}(\cdot \mid \boldsymbol{\sigma}_{\partial^{2\ell} x_1} = \tau_{\partial^{2\ell} x_1}) = \mu_{\mathbf{\Phi}}(\cdot \mid \forall y \in \partial^{2\ell} x_1 : \boldsymbol{\sigma}_y = \tau_y)$ imposes the 'boundary condition' τ on all variables y at distance 2ℓ from x_1. The following theorem shows that Belief Propagation does not just approximate the plain, unconditional marginals well w.h.p., but even the conditional marginals given any conceivable boundary condition. Recall that $\mathbb{P}[Z(\mathbf{\Phi}) > 0] = 1 - o(1)$ for $d < 2$.

Theorem 1.2. *If $d < 2$, then*

$$\lim_{\ell\to\infty} \limsup_{n\to\infty} \mathbb{E}\left[\max_{\tau\in S(\mathbf{\Phi})} \left|\mu_{\mathbf{\Phi}}(\boldsymbol{\sigma}_{x_1} = 1 \mid \boldsymbol{\sigma}_{\partial^{2\ell} x_1} = \tau_{\partial^{2\ell} x_1}) - \nu^{(\ell)}_{\mathbf{\Phi},x_1}(1)\right| \,\middle|\, Z(\mathbf{\Phi}) > 0\right] = 0. \tag{1.9}$$

Since $\nu^{(\ell)}_{\mathbf{\Phi},x_1}$ does not depend on τ, averaging (1.9) on the boundary condition $\tau \in S(\mathbf{\Phi})$ yields

$$\lim_{\ell\to\infty} \limsup_{n\to\infty} \mathbb{E}\left[\left|\mu_{\mathbf{\Phi}}(\boldsymbol{\sigma}_{x_1} = \pm 1) - \nu^{(\ell)}_{\mathbf{\Phi},x_1}(\pm 1)\right| \mid Z(\mathbf{\Phi}) > 0\right] = 0. \tag{1.10}$$

Thus, Belief Propagation approximates the unconditional marginal of x_1 well in the limit of large n and ℓ. Indeed, because the distribution of $\mathbf{\Phi}$ is invariant under permutations of the variables $x_1,\ldots,x_n$, (1.10) implies that the marginals of all but $o(n)$ variables x_i are within $\pm o(1)$ of the Belief Propagation approximation w.h.p.

But thanks to the presence of the boundary condition τ, Theorem 1.2 leads to further discoveries. For a start, applying the triangle inequality to (1.9) and (1.10), we obtain

$$\lim_{\ell\to\infty}\limsup_{n\to\infty}\mathbb{E}\left[\max_{\tau\in S(\mathbf{\Phi})}\left|\mu_{\mathbf{\Phi}}(\boldsymbol{\sigma}_{x_1}=1\mid\boldsymbol{\sigma}_{\partial^{2\ell}x_1}=\tau_{\partial^{2\ell}x_1})-\mu_{\mathbf{\Phi}}(\boldsymbol{\sigma}_{x_1}=1)\right|\,\middle|\,Z(\mathbf{\Phi})>0\right]=0. \tag{1.11}$$

Thus, no discernible shift of the marginal of x_1 is likely to ensue upon imposition of any possible boundary condition τ. The spatial mixing property (1.11) is colloquially known as *Gibbs uniqueness* [32]. Further, (1.11) rules out extensive long-range correlations. Specifically, for any fixed ℓ the first two variables x_1,x_2 likely have distance greater than 4ℓ in $G(\mathbf{\Phi})$. Therefore, (1.11) implies that for all $d<2$,

$$\lim_{n\to\infty}\sum_{s,t\in\{\pm1\}}\mathbb{E}\left[\left|\mu_{\mathbf{\Phi}}(\boldsymbol{\sigma}_{x_1}=s,\boldsymbol{\sigma}_{x_2}=t)-\mu_{\mathbf{\Phi}}(\boldsymbol{\sigma}_{x_1}=s)\cdot\mu_{\mathbf{\Phi}}(\boldsymbol{\sigma}_{x_2}=t)\right|\,\middle|\,Z(\mathbf{\Phi})>0\right]=0. \tag{1.12}$$

Thus, the truth values $\boldsymbol{\sigma}_{x_1},\boldsymbol{\sigma}_{x_2}$ are asymptotically independent. Of course, once again by permutation invariance, (1.12) implies that asymptotic independence extends to all but $o(n^2)$ pairs of variables x_i,x_j w.h.p. The decorrelation property (1.12) is called *replica symmetry* in the physics literature [32].

Finally, we can clarify the combinatorial meaning of the distribution π_d from Theorem 1.1. Namely, π_d is the limit of the empirical distribution of the marginal probabilities $\mu_{\mathbf{\Phi}}(\boldsymbol{\sigma}_{x_i}=1)$.

Corollary 1.3. *For any $0<d<2$ the random probability measure*

$$\pi_{\mathbf{\Phi}}=\frac{1}{n}\sum_{i=1}^{n}\delta_{\mu_{\mathbf{\Phi}}(\boldsymbol{\sigma}_{x_i}=1)} \tag{1.13}$$

converges to π_d weakly in probability.[1]

Thus, the stochastic fixed point equation (1.3) that characterises π_d simply expresses that the marginal probabilities $\mu_{\mathbf{\Phi}}(\boldsymbol{\sigma}_{x_i}=1)$ result from the Belief Propagation recurrence (1.6).

1.4. **Preliminaries and notation.** Throughout we denote by $V_n=\{x_1,\ldots,x_n\}$ the variable set of $\mathbf{\Phi}_n$. Generally, given a 2-SAT formula Φ we write $V(\Phi)$ for the set of variables and $F(\Phi)$ for the set of clauses. The bipartite clause/variable-graph $G(\Phi)$ is defined as in Section 1.3. For a vertex v of $G(\Phi)$ we let $\partial(\Phi,v)$ be the set of neighbours. Where Φ is apparent we just write ∂v. Moreover, $\partial^\ell(\Phi,v)$ or briefly $\partial^\ell v$ stands for the set of vertices at distance exactly ℓ from v. Additionally, $\nabla^\ell(\Phi,v)$ denotes the sub-formula obtained from Φ by deleting all clauses and variables at distance greater than ℓ from v. This sub-formula may contain clauses of length less than two. Further, for a clause a and a variable x of Φ we let $\mathrm{sign}(x,a)=\mathrm{sign}_\Phi(x,a)\in\{\pm1\}$ be the sign with which x appears in a. In addition, we let $S(\Phi)$ be the set of all satisfying assignments of Φ, $Z(\Phi)=|S(\Phi)|$ and, assuming $Z(\Phi)>0$, we let μ_Φ be the probability distribution on $\{\pm1\}^{V(\Phi)}$ that induces the uniform distribution on $S(\Phi)$ as in (1.8). Moreover, $\boldsymbol{\sigma}_\Phi=(\boldsymbol{\sigma}_{\Phi,x})_{x\in V(\Phi)}$ signifies a uniformly random satisfying assignment; we drop Φ where the reference is apparent.

For any Φ we set up Belief Propagation as in (1.5)–(1.7). It is well known that Belief Propagation yields the correct marginals if $G(\Phi)$ is a tree. To be precise, the *depth* of $x\in V(\Phi)$ is the maximum distance between x and a leaf of $G(\Phi)$.

Proposition 1.4 ([34, Theorem 14.1])**.** *If $G(\Phi)$ is a tree and $x\in V(\Phi)$, then for any ℓ greater than or equal to the depth of x we have $\mu_\Phi(\boldsymbol{\sigma}_x=\pm1)=\nu^{(\ell)}_{\Phi,x}(\pm1)$.*

We will encounter the following functions repeatedly. For $\varepsilon>0$ let $\Lambda_\varepsilon(z)=\log(z\vee\varepsilon)$ be the log function truncated at $\log\varepsilon$. Moreover, we need the continuous and mutually inverse functions

$$\psi:\mathbb{R}\to(0,1),\quad z\mapsto(1+\tanh(z/2))/2,\qquad\qquad \varphi:(0,1)\to\mathbb{R},\quad p\mapsto\log(p/(1-p)). \tag{1.14}$$

Let $\mathscr{P}(\mathbb{R})$ be the set of all Borel probability measures on $\mathbb{R}$ with the weak topology. Moreover, for a real $q\geq1$ let $\mathscr{W}_q(\mathbb{R})$ be the set of all $\rho\in\mathscr{P}(\mathbb{R})$ such that $\int_{\mathbb{R}}|x|^q\mathrm{d}\rho(x)<\infty$. We equip this space with the Wasserstein metric

$$W_q(\rho,\rho')=\inf\left\{\left(\int_{\mathbb{R}^2}|x-y|^q\mathrm{d}\gamma(x,y)\right)^{1/q}:\gamma\text{ is a coupling of }\rho,\rho'\right\}, \tag{1.15}$$

[1] That is, for any continuous function $f:[0,1]\to\mathbb{R}$ we have $\lim_{n\to\infty}\mathbb{E}\left|\int_0^1 f(z)\mathrm{d}\pi_d(z)-\int_0^1 f(z)\mathrm{d}\pi_{\mathbf{\Phi}}(z)\right|=0$.

thereby turning $\mathscr{W}_q(\mathbb{R})$ into a complete separable space [9].

For $\rho \in \mathscr{P}(\mathbb{R})$ we denote by $\boldsymbol{\eta}_\rho, \boldsymbol{\eta}_{\rho,1}, \boldsymbol{\eta}_{\rho,2}, \ldots$ random variables with distribution ρ. Similarly, for $\pi \in \mathscr{P}(0,1)$ we let $\boldsymbol{\mu}_\pi, \boldsymbol{\mu}_{\pi,1}, \boldsymbol{\mu}_{\pi,2}, \ldots$ be a sequence of random variables with distribution π. We also continue to let $\boldsymbol{d}$ be a Poisson variable with mean d and $\boldsymbol{d}^+, \boldsymbol{d}^-$ Poisson variables with mean $d/2$. Moreover, $\boldsymbol{s}_1, \boldsymbol{s}_1', \boldsymbol{s}_2, \boldsymbol{s}_2', \ldots \in \{\pm 1\}$ always denote uniformly distributed random variables. All of these random variables are mutually independent as well as independent of any other sources of randomness.

Finally, from here on we tacitly assume that $0 < d < 2$**.**

2. Overview

The proof of Theorem 1.1 proceeds in four steps. First we show that the limit π_d from Theorem 1.1 exists. Subsequently we establish the fact (1.9) that Belief Propagation approximates the conditional marginals well. This will easily imply the convergence of the empirical marginals (1.13) to π_d. Third, building upon these preparations, we will prove that the truncated mean $n^{-1}\mathbb{E}[\log(Z(\boldsymbol{\Phi}) \vee 1)]$ converges to the r.h.s. of (1.2). The truncation is necessary to deal with the (unlikely) event that $Z(\boldsymbol{\Phi}) = 0$. Finally, we will show that $\log(Z(\boldsymbol{\Phi}) \vee 1)$ concentrates about its mean to obtain convergence in probability, thus completing the proof of Theorem 1.1.

2.1. **Step 1: density evolution.** We begin by verifying that the distribution π_d from Theorem 1.1 is well-defined and that π_d satisfies a tail bound.

Proposition 2.1. *The weak limit* $\pi_d = \lim_{\ell\to\infty} \mathrm{BP}_d^\ell(\delta_{1/2})$ *exists and*

$$\mathbb{E}\left[\log^2 \frac{\boldsymbol{\mu}_{\pi_d}}{1-\boldsymbol{\mu}_{\pi_d}}\right] < \infty. \tag{2.1}$$

Moreover, $\boldsymbol{\mu}_{\pi_d}$ *and* $1-\boldsymbol{\mu}_{\pi_d}$ *are identically distributed and*

$$\mathbb{E}\left|\log\left(\prod_{i=1}^{\boldsymbol{d}^-}\boldsymbol{\mu}_{\pi_d,i} + \prod_{i=1}^{\boldsymbol{d}^+}\boldsymbol{\mu}_{\pi_d,i+\boldsymbol{d}^-}\right)\right| < \infty, \qquad \mathbb{E}\left|\log\left(1-\boldsymbol{\mu}_{\pi_d,1}\boldsymbol{\mu}_{\pi_d,2}\right)\right| < \infty. \tag{2.2}$$

The proof of Proposition 2.1, which we carry out in Section 4, is based on a contraction argument. This argument implies that the fixed point iteration converges rapidly to π_d, a fact that can be exploited to obtain numerical estimates. The bounds (2.2) ensure that the expectation on the r.h.s. of (1.2) is well-defined.

2.2. **Step 2: Gibbs uniqueness.** As a next step we verify the Gibbs uniqueness property (1.11). We proceed by way of analysing a multi-type Galton-Watson tree $\boldsymbol{T}$ that mimics the local structure of the graph $G(\boldsymbol{\Phi})$ upon exploration from variable x_1. The Galton-Watson process has five types: variable nodes and four types of clause nodes $(+1,+1), (+1,-1), (-1,+1), (-1,-1)$. The root is a variable node o. Moreover, each variable node spawns independent $\mathrm{Po}(d/4)$ numbers of clauses nodes of each of the four types. Additionally, each clause has a single offspring, which is a variable. The semantics of the clause types is that the first component indicates whether the parent variable appears in the clause positively or negatively. The second component indicates whether the child variable appears as a positive or as a negative literal. Clearly, for $d \le 1$ the tree $\boldsymbol{T}$ is finite with probability one, while infinite trees appear with positive probability for $d > 1$.

Let $\boldsymbol{T}^{(\ell)}$ be the finite tree obtained from $\boldsymbol{T}$ by dropping all nodes at distance greater than ℓ from the root. For even ℓ it will be convenient to view $\boldsymbol{T}^{(\ell)}$ interchangeably as a tree or as a 2-SAT formula. In particular, we write $\partial^{2\ell} o = \partial^{2\ell}(\boldsymbol{T}, o)$ for the set of all variables at distance exactly 2ℓ from o. The following proposition, which is the linchpin of the entire proof strategy, establishes the Gibbs uniqueness property for the tree formula $\boldsymbol{T}^{(2\ell)}$.

Proposition 2.2. *We have*

$$\lim_{\ell\to\infty}\mathbb{E}\left[\max_{\tau\in S(\boldsymbol{T}^{(2\ell)})}\left|\mu_{\boldsymbol{T}^{(2\ell)}}(\boldsymbol{\sigma}_o = 1 \mid \boldsymbol{\sigma}_{\partial^{2\ell}o} = \tau_{\partial^{2\ell}o}) - \mu_{\boldsymbol{T}^{(2\ell)}}(\boldsymbol{\sigma}_o = 1)\right|\right] = 0. \tag{2.3}$$

Thus, w.h.p. no conceivable boundary condition is apt to significantly shift the marginal of the root.

We prove Proposition 2.2 by a subtle contraction argument in combination with a construction of extremal boundary conditions of the tree formula $\boldsymbol{T}^{(2\ell)}$. More specifically, we will construct boundary conditions $\boldsymbol{\sigma}^\pm$ that maximise or minimise the conditional probability

$$\mu_{\boldsymbol{T}^{(2\ell)}}(\boldsymbol{\sigma}_o = 1 \mid \boldsymbol{\sigma}_{\partial^{2\ell}o} = \boldsymbol{\sigma}^\pm_{\partial^{2\ell}o}), \tag{2.4}$$

respectively. Then we will show that the difference of the conditional marginals induced by both these extremal boundary conditions vanishes with probability tending to one as $\ell \to \infty$. The delicate point is that the extremal boundary conditions $\boldsymbol{\sigma}^{\pm}$ depend on the tree $\boldsymbol{T}^{(2\ell)}$. Thus, at first glance it seems that we need to pass the tree twice, once top–down to construct $\boldsymbol{\sigma}^{\pm}$ and then bottom–up to calculate the conditional marginals (2.4). But such an analysis seems untenable because after the top–down pass the tree is exposed and 'no randomness remains' to facilitate the bottom–up phase. Fortunately, we will see that a single stochastic fixed point equation captures both the top–down and the bottom–up phase. This discovery reduces the proof of Proposition 2.2 to showing that the fixed point iteration contracts. The details of this delicate argument can be found in Section 5.

Proposition 2.2 easily implies the Gibbs uniqueness condition (1.11) and thereby Theorem 1.2. A further consequence is the asymptotic independence of the joint truth values of bounded numbers of variables.

Corollary 2.3. *The statement* (1.9) *is true and for any integer $k \geq 2$ we have*

$$\lim_{n\to\infty} \sum_{\sigma\in\{\pm1\}^k} \mathbb{E}\left[\left|\mu_{\boldsymbol{\Phi}}(\boldsymbol{\sigma}_{x_1}=\sigma_1,\ldots,\boldsymbol{\sigma}_{x_k}=\sigma_k) - \prod_{i=1}^{k}\mu_{\boldsymbol{\Phi}}(\boldsymbol{\sigma}_{x_i}=\sigma_i)\right| \mid Z(\boldsymbol{\Phi})>0\right] = 0.$$

2.3. **Step 3: the Aizenman-Sims-Starr scheme.** The aforementioned results pave the way for deriving an expression for the conditional expectation of $\log Z(\boldsymbol{\Phi})$ given that $\boldsymbol{\Phi}$ is satisfiable. Since $\boldsymbol{\Phi}$ is satisfiable w.h.p. for all $d < 2$, an equivalent task is to calculate $\mathbb{E}[\log(Z(\boldsymbol{\Phi})\vee 1)]$. To this end we seize upon a simple but powerful strategy colloquially called the Aizenman-Sims-Starr scheme [5]. Originally proposed in the context of the Sherrington-Kirkpatrick spin glass model, this proof strategy suggests to compute the asymptotic mean of a random variable on a 'system' of size n by carefully estimating the change of that mean upon going to a 'system' of size $n+1$. This difference is calculated by coupling the systems of size n and $n+1$ such that the latter is obtained from the former by a small expected number of local changes.

We apply this idea to the random 2-SAT problem by coupling the random formula $\boldsymbol{\Phi}_n$ with n variables and $\mathrm{Po}(dn/2)$ clauses and the random formula $\boldsymbol{\Phi}_{n+1}$ with $n+1$ variables and $\mathrm{Po}(d(n+1)/2)$ clauses. Roughly speaking, we obtain $\boldsymbol{\Phi}_{n+1}$ from $\boldsymbol{\Phi}_n$ by adding a new variable x_{n+1} along with a few random adjacent clauses that connect x_{n+1} with the variables $x_1,\ldots,x_n$ of $\boldsymbol{\Phi}_n$. Then the information about the joint distribution of the truth values of bounded numbers of variables furnished by Corollaries 1.3 and 2.3 and the tail bound (2.1) will enable us to accurately estimate $\mathbb{E}\big[\log(Z(\boldsymbol{\Phi}_{n+1})\vee 1) - \log(Z(\boldsymbol{\Phi}_n)\vee 1)\big]$.

Needless to say, upon closer inspection matters will emerge to be rather subtle. The main source of complications is that, in contrast to other models in mathematical physics such as the Sherrington-Kirkpatrick model or the Ising model, the 2-SAT problem has hard constraints. Thus, the addition of a single clause could trigger a dramatic drop in the partition function. In fact, in the worst case a single awkward clause could wipe out all satisfying assignments. In Section 6 we will iron out all these difficulties and prove the following.

Proposition 2.4. *We have*

$$\lim_{n\to\infty} \mathbb{E}[\log(Z(\boldsymbol{\Phi}_{n+1})\vee 1)] - \mathbb{E}[\log(Z(\boldsymbol{\Phi}_n)\vee 1)] = \mathbb{E}\left[\log\left(\prod_{i=1}^{\boldsymbol{d}^-}\boldsymbol{\mu}_{\pi_d,i} + \prod_{i=1}^{\boldsymbol{d}^+}\boldsymbol{\mu}_{\pi_d,i+\boldsymbol{d}^-}\right) - \frac{d}{2}\log\left(1-\boldsymbol{\mu}_{\pi_d,1}\boldsymbol{\mu}_{\pi_d,2}\right)\right]. \tag{2.5}$$

We notice that (2.2) guarantees that the r.h.s. of (2.5) is well-defined. As an immediate consequence of Proposition 2.4 we obtain a formula for $\mathbb{E}[\log(Z(\boldsymbol{\Phi})\vee 1)]$.

Corollary 2.5. *For any $d < 2$ we have*

$$\lim_{n\to\infty}\frac{1}{n}\mathbb{E}[\log(Z(\boldsymbol{\Phi})\vee 1)] = \mathbb{E}\left[\log\left(\prod_{i=1}^{\boldsymbol{d}^-}\boldsymbol{\mu}_{\pi_d,i} + \prod_{i=1}^{\boldsymbol{d}^+}\boldsymbol{\mu}_{\pi_d,i+\boldsymbol{d}^-}\right) - \frac{d}{2}\log\left(1-\boldsymbol{\mu}_{\pi_d,1}\boldsymbol{\mu}_{\pi_d,2}\right)\right].$$

Proof. Writing $\mathbb{E}[\log(Z(\boldsymbol{\Phi})\vee 1)]$ as a telescoping sum and applying Proposition 2.4, we obtain

$$\begin{aligned}\lim_{n\to\infty}\frac{1}{n}\mathbb{E}[\log(Z(\boldsymbol{\Phi}_n)\vee 1)] &= \lim_{n\to\infty}\frac{1}{n}\sum_{N=2}^{n-1}\mathbb{E}[\log(Z(\boldsymbol{\Phi}_{N+1})\vee 1)] - \mathbb{E}[\log(Z(\boldsymbol{\Phi}_N)\vee 1)]\\ &= \mathbb{E}\left[\log\left(\prod_{i=1}^{\boldsymbol{d}^-}\boldsymbol{\mu}_{\pi_d,i} + \prod_{i=1}^{\boldsymbol{d}^+}\boldsymbol{\mu}_{\pi_d,i+\boldsymbol{d}^-}\right) - \frac{d}{2}\log\left(1-\boldsymbol{\mu}_{\pi_d,1}\boldsymbol{\mu}_{\pi_d,2}\right)\right],\end{aligned}$$

as desired. □

2.4. **Step 4: concentration.** The final step towards Theorem 1.1 is to show that $\log(Z(\mathbf{\Phi}) \vee 1)$ concentrates about its mean.

Proposition 2.6. *We have* $\lim_{n\to\infty} n^{-1}\mathbb{E}\left|\log(Z(\mathbf{\Phi}) \vee 1) - \mathbb{E}[\log(Z(\mathbf{\Phi}) \vee 1)]\right| = 0$.

Proposition 2.6 does not easily follow from routine arguments such as the Azuma-Hoeffding inequality. Once more the issue is that changing a single clause could alter $\log(Z(\mathbf{\Phi}) \vee 1)$ by as much as $\Theta(n)$. Instead we will resort to another technique from mathematical physics called the interpolation method. The details can be found in Section 7.

Proof of Theorem 1.1. The theorem follows from Proposition 2.1, Corollary 2.5 and Proposition 2.6. □

3. Discussion

The random 2-SAT satisfiability threshold was established mathematically shortly after the experimental work of Cheeseman, Kanefsky and Taylor [13] that triggered the quest for satisfiability thresholds appeared. The second successful example, nearly a decade later, was the random 1-in-k-SAT threshold (to satisfy exactly one literal in each clause), which Achlioptas, Chtcherba, Istrate and Moore pinpointed by analysing the Unit Clause algorithm [2]. In a subsequent landmark contribution Dubois and Mandler determined the 3-XORSAT threshold via the second moment method [27]. Subsequent work extended this result to random k-XORSAT [23, 43]. Finally, the most notable success thus far has been the verification of the '1RSB cavity method' prediction [35] of the random k-SAT threshold for large k due to Ding, Sly and Sun [25], the culmination of a line of work that refined the use of the second moment method [3, 4, 17].

Over the past two decades the general theme of estimating the partition functions of discrete structures has received a great deal of attention; e.g., [8]. With respect to random 2-SAT (and, more generally, k-SAT), Montanari and Shah [39], Panchenko [41] and Talagrand [48] investigated 'soft' versions of the partition function. To be precise, introducing a parameter $\beta > 0$ called the 'inverse temperature', these articles study the random variable

$$Z_\beta(\mathbf{\Phi}) = \sum_{\sigma\in\{\pm1\}^n}\prod_{i=1}^{m}\exp\left(-\beta\mathbf{1}\left\{\sigma \text{ violates clause } a_i\right\}\right). \tag{3.1}$$

Thus, instead of dismissing assignments that fail to satisfy all clauses outright, there is an $\exp(-\beta)$ penalty factor for each violated clause. Talagrand [48] computes $\lim_{n\to\infty} n^{-1}\mathbb{E}[\log Z_\beta(\mathbf{\Phi})]$ for β not exceeding a small but unspecified $\beta_0 > 0$. Panchenko [41] calculates this limit under the assumption $(4\beta \wedge 1)d < 1$. Thus, for $\beta > 1/4$ the result is confined to $d < 1$, in which case the random graph $G(\mathbf{\Phi})$ is sub-critical and both $Z_\beta(\mathbf{\Phi})$ and the actual number $Z(\mathbf{\Phi})$ of satisfying assignments could be calculated via elementary methods. Furthermore, Montanari and Shah [39] obtain $\lim_{n\to\infty} n^{-1}\mathbb{E}[\log Z_\beta(\mathbf{\Phi})]$ for all finite β under the assumption $d < 1.16\ldots$. Although for any fixed formula $\mathbf{\Phi}$ the limit $\lim_{\beta\to\infty} Z_\beta(\mathbf{\Phi})$ is equal to the number of satisfying assignments, it is not possible to interchange the limits $\beta\to\infty$ and $n\to\infty$. Thus, [39, 41] do not yield the the number of actual satisfying assignments even for $d < 1.16\ldots$ or $d < 1$, respectively. Apart from estimating $\mathbb{E}\log Z_\beta(\mathbf{\Phi})$, Montanari and Shah [39] also show that the Belief Propagation message passing scheme approximates the marginals of the Boltzmann distribution that goes with $Z_\beta(\mathbf{\Phi})$ well, i.e., they obtain a 'soft' version of Theorem 1.2 for $d < 1.16\ldots$.

In terms of proof techniques, all three contributions [39, 41, 48] are based on establishing the Gibbs uniqueness property. So is the present paper. But while [39, 41, 48] rely on relatively straightforward contraction arguments, a key distinction is that here we develop a more accurate (and delicate) method for verifying the Gibbs uniqueness property based on the explicit construction of an extremal boundary condition. This is the key to pushing the range of d all the way up to the satisfiability threshold $d = 2$.

Specifically, in order to construct a boundary condition of the random tree $\boldsymbol{T}^{(2\ell)}$ for large ℓ that maximises the conditional probability of observing the truth value +1 at the root we will work our way top–down from the root to level 2ℓ. Exposing the degrees and the signs with which the variables appear, the construction assigns a 'desired' truth value to each variable of the tree so as to nudge the parent variable towards its desired value as much as possible. Subsequently, once this process reaches the bottom level of the tree, we go into reverse gear and study the Belief Propagation messages bottom–up to calculate the conditional marginal of the root. Clearly, analysing this upwards process seems like a tall order because the tree was already exposed during the top-down phase, a challenge that is exacerbated by the presence of hard constraints. Fortunately, in Section 5 we will see how this problem can be transformed into the study of another stochastic fixed point equation that captures the effect of the children's 'nudging' their parents. This fixed point problem is amenable to the contraction method. A spatial

mixing analysis from an extremal boundary condition was previously conducted in by Dembo and Montanari [21] for the Ising model on random graphs. But of course a crucial difference is that in the Ising model the extremal boundary conditions are constant (all-$+1$ and all--1, respectively).

A second novelty of the present work is that we directly deal with the 'hard' 2-SAT problem. Montanari and Shah [39] interpolate on the 'inverse temperature' parameter $\beta > 0$, effectively working their way from smaller to larger β. Because the limits $\beta \to \infty$ and $n \to \infty$ do not commute, this approach does not seem applicable to problems with hard constraints. Furthermore, while Panchenko [40, 41] applies the Aizenman-Sims-Starr scheme to the soft constraint version, the hard problem of counting actual satisfying assignments requires a far more careful analysis. Indeed, adding one clause can shift $\log Z_\beta(\mathbf{\Phi})$ merely by $\pm\beta$. By contrast, a single additional clause could very well reduce the logarithm $\log Z(\mathbf{\Phi})$ of the number of satisfying assignments by as much as $\Omega(n)$, or even render the formula unsatisfiable. A few prior applications of the Aizenman-Sims-Starr scheme to problems with hard constraints exist [7, 15, 16], but these hinge on peculiar symmetry properties that enable an indirect approach via a 'planted' version of the problem in question. The required symmetries for this approach are absent in several important problems, with random satisfiability the most prominent example. Thus, a significant technical contribution of the present work is that we show how to apply the Aizenman-Sims-Starr scheme directly to problems with hard constraints. Among other things, this requires a careful quantification of the probabilities of rare, potentially cataclysmic events in comparison to their impact on $\log Z(\mathbf{\Phi})$. That said, we should point out that [39, 41, 48] actually also deal with the (soft) k-SAT partition function for $k > 2$ for certain regimes of clause/variable densities, while the technique that we develop here does not seem to extend beyond binary problems.

A mathematical physics technique called the interpolation method, first proposed by Guerra for the study of the Sherrington-Kirkpatrick model [31], can be applied to the random k-SAT problem [29, 42] to bound the number of satisfying assignments from above. For $k = 2$ the interpolation method yields the upper bound

$$\frac{1}{n}\log Z(\mathbf{\Phi}) \leq \inf_{\pi\in\mathscr{P}(0,1)} \mathbb{E}\left[\log\left(\prod_{i=1}^{\boldsymbol{d}^-}\boldsymbol{\mu}_{\pi,i} + \prod_{i=1}^{\boldsymbol{d}^+}\boldsymbol{\mu}_{\pi,i+\boldsymbol{d}^-}\right) - \frac{d}{2}\log\left(1-\boldsymbol{\mu}_{\pi,1}\boldsymbol{\mu}_{\pi,2}\right)\right] + o(1) \qquad \text{w.h.p.,} \tag{3.2}$$

for all $0 < d < 2$; we will revisit this bound in Section 7. Since the expression on the r.h.s. coincides with (1.2) for $\pi = \pi_d$, the main contribution of Theorem 1.1 is the matching lower bound on $\log Z(\mathbf{\Phi})$. Furthermore, Abbe and Montanari [1] used the interpolation method to establish the *existence* of a function ϕ such that

$$\lim_{n\to\infty} n^{-1}\log(Z(\mathbf{\Phi})\vee 1) = \phi(d) \qquad \text{in probability} \tag{3.3}$$

for all but a countable number of $d \in (0,2)$. Theorem 1.1 actually determines $\phi(d)$ and shows that convergence holds for *all* $d \in (0,2)$. Clearly, (3.3) implies the concentration bound from Proposition 2.6 for all d outside the countable set. But of course we need concentration for all d, and in Section 7 we will use the upper bound (3.2) to prove this concentration result. As an aside, a conditional concentration inequality for $\log Z(\mathbf{\Phi})$, quoted in [28], was obtained by Sharell [46] (unpublished). But the necessary conditions appear to be difficult to check.

In addition, several prior contributions deal with the combinatorial problem of counting solutions to random CSPs. For problems such as k-NAESAT, k-XORSAT or graph colouring where the first moment provides the correct answer due to inherent symmetry properties, the second moment method and small subgraph conditioning yield very precise information as to the number of solutions [15, 18, 44]. Verifying that the number of solutions is determined by the physicists' 1RSB formula [34], the contribution of Sly, Sun and Zhang [47] on the random regular k-NAESAT problem near its satisfiability threshold [24] deals with an even more intricate scenario.

Finally, returning to random 2-SAT, as an intriguing question for future work determining the precise limiting distribution of $\log Z(\mathbf{\Phi})$ stands out. This random variable has standard deviation $\Omega(\sqrt{n})$ for all $0 < d < 2$ even once we condition on $\boldsymbol{m}$, as is easily seen by re-randomising the signs of the literals in small components. In effect, $\log Z(\mathbf{\Phi})$ is far less concentrated than the partition functions of symmetric random constraint satisfaction problems [15]. May $n^{-1/2}(\log Z(\mathbf{\Phi}) - \mathbb{E}[\log Z(\mathbf{\Phi})])$ be asymptotically normal?

4. Proof of Proposition 2.1

We prove Proposition 2.1 by means of a contraction argument. The starting point is the following observation. For $\ell \geq 0$ let $\pi_d^{(\ell)} = \mathrm{BP}_d^\ell(\delta_{1/2})$ be the probability measure obtained after ℓ iterations of the operator $\mathrm{BP}_d(\cdot)$.

Fact 4.1. *For all $\ell \geq 0$ the random variables $\boldsymbol{\mu}_{\pi_d^{(\ell)}}$ and $1-\boldsymbol{\mu}_{\pi_d^{(\ell)}}$ are identically distributed.*

Proof. This is because $\boldsymbol{d}^-, \boldsymbol{d}^+$ and hence the random variables

$$\left(\prod_{i=1}^{\boldsymbol{d}^-}\boldsymbol{\mu}_{\pi_d^{(\ell-1)},i}, \prod_{i=1}^{\boldsymbol{d}^-}\boldsymbol{\mu}_{\pi_d^{(\ell-1)},i} + \prod_{i=1}^{\boldsymbol{d}^+}\boldsymbol{\mu}_{\pi_d^{(\ell-1)},i+\boldsymbol{d}^-}\right) \quad \text{and} \quad \left(\prod_{i=1}^{\boldsymbol{d}^+}\boldsymbol{\mu}_{\pi_d^{(\ell-1)},i+\boldsymbol{d}^-}, \prod_{i=1}^{\boldsymbol{d}^-}\boldsymbol{\mu}_{\pi_d^{(\ell-1)},i} + \prod_{i=1}^{\boldsymbol{d}^+}\boldsymbol{\mu}_{\pi_d^{(\ell-1)},i+\boldsymbol{d}^-}\right)$$

from (1.1) are identically distributed. □

Due to Fact 4.1 we can rewrite the construction of the sequence $\pi_d^{(\ell)}$ in terms of another operator that is easier to analyse. This operator describes the expression (1.1) in terms of log-likelihood ratios, a simple reformulation that proved useful in the context of Belief Propagation for random satisfiability before [38]. Thus, we define an operator $\mathrm{LL}_d : \mathscr{P}(\mathbb{R}) \to \mathscr{P}(\mathbb{R})$, $\rho \mapsto \hat{\rho}$ by letting $\hat{\rho}$ be the distribution of the random variable

$$\sum_{i=1}^{\boldsymbol{d}} \boldsymbol{s}_i \log\frac{1+\boldsymbol{s}'_i\tanh(\boldsymbol{\eta}_{\rho,i}/2)}{2}. \tag{4.1}$$

Further, let $\rho_d^{(\ell)} = \mathrm{LL}_d^{\ell}(\delta_0) \in \mathscr{P}(\mathbb{R})$ be the result of ℓ iterations of LL_d launched from the atom at zero. We recall the functions ψ, φ from (1.14). For a measure $\rho \in \mathscr{P}(\mathbb{R})$ and a measurable $f : \mathbb{R} \to \mathbb{R}$ let $f(\rho)$ denote the pushforward measure of ρ that assigns mass $\rho(f^{-1}(A))$ to Borel sets $A \subseteq \mathbb{R}$.

Lemma 4.2. *For all $\ell \geq 0$ we have $\pi_d^{(\ell)} = \psi(\rho_d^{(\ell)})$.*

Proof. Since $\psi(\delta_0) = \delta_{1/2}$, the assertion is true for $\ell = 0$. Proceeding by induction, we obtain

$$\begin{aligned}\boldsymbol{\mu}_{\pi_d^{(\ell+1)}} &\stackrel{\mathrm{d}}{=} \frac{\prod_{i=1}^{\boldsymbol{d}^+}\boldsymbol{\mu}_{\pi_d^{(\ell)},i}}{\prod_{i=1}^{\boldsymbol{d}^-}\boldsymbol{\mu}_{\pi_d^{(\ell)},i} + \prod_{i=1}^{\boldsymbol{d}^+}\boldsymbol{\mu}_{\pi_d^{(\ell)},i+\boldsymbol{d}^-}} = \psi\left(\log\frac{\prod_{i=1}^{\boldsymbol{d}^-}\boldsymbol{\mu}_{\pi_d^{(\ell)},i}}{\prod_{i=1}^{\boldsymbol{d}^+}\boldsymbol{\mu}_{\pi_d^{(\ell)},i+\boldsymbol{d}^-}}\right)\\ &= \psi\left(\sum_{i=1}^{\boldsymbol{d}^-}\log\left(\boldsymbol{\mu}_{\pi_d^{(\ell)},i}\right) - \sum_{i=1}^{\boldsymbol{d}^+}\log\left(\boldsymbol{\mu}_{\pi_d^{(\ell)},i+\boldsymbol{d}^-}\right)\right) \stackrel{\mathrm{d}}{=} \psi\left(\sum_{i=1}^{\boldsymbol{d}}\boldsymbol{s}_i\log\boldsymbol{\mu}_{\pi_d^{(\ell)},i}\right) \stackrel{\mathrm{d}}{=} \psi\left(\sum_{i=1}^{\boldsymbol{d}}\boldsymbol{s}_i\log\left(\psi(\boldsymbol{\eta}_{\rho_d^{(\ell)},i})\right)\right).\end{aligned} \tag{4.2}$$

Moreover, since $\boldsymbol{s}_i \in \{\pm 1\}$ is random, it is immediate from (4.1) that $\boldsymbol{\eta}_{\rho_d^{(\ell)},i} \stackrel{\mathrm{d}}{=} -\boldsymbol{\eta}_{\rho_d^{(\ell)},i}$. Consequently, (4.2) yields

$$\boldsymbol{\mu}_{\pi_d^{(\ell+1)}} \stackrel{\mathrm{d}}{=} \psi\left(\sum_{i=1}^{\boldsymbol{d}}\boldsymbol{s}_i\log\left(\psi(\boldsymbol{s}'_i\boldsymbol{\eta}_{\rho_d^{(\ell)},i})\right)\right) \stackrel{\mathrm{d}}{=} \psi(\boldsymbol{\eta}_{\rho_d^{(\ell+1)}}),$$

which completes the induction. □

Due to the continuous mapping theorem, to establish convergence of $(\pi_d^{(\ell)})_{\ell\geq 0}$ it suffices to show that $(\rho_d^{(\ell)})_{\ell\geq 0}$ converges weakly. To this end, we will prove that the operator $\mathrm{LL}_d(\cdot)$ is a contraction.

Lemma 4.3. *If $d < 2$, then LL_d is a contraction on the space $\mathscr{W}_2(\mathbb{R})$.*

Proof. The operator LL_d maps the space $\mathscr{W}_2(\mathbb{R})$ into itself because the derivative of $x \mapsto \log((1+\tanh(x/2))/2)$ is bounded by one in absolute value for all $x \in \mathbb{R}$. To show contraction let $\rho, \rho' \in \mathscr{W}_2(\mathbb{R})$ and consider a sequence of independent random pairs $(\boldsymbol{\eta}_i, \boldsymbol{\eta}'_i)_{i\geq 1}$ such that the $\boldsymbol{\eta}_i$ have distribution ρ and the $\boldsymbol{\eta}'_i$ have distribution ρ'. Because the signs $\boldsymbol{s}_i$ are uniform and independent, we obtain

$$\begin{aligned}W_2(\mathrm{LL}(\rho),\mathrm{LL}(\rho'))^2 &\leq \mathbb{E}\left[\left(\sum_{i=1}^{\boldsymbol{d}}\boldsymbol{s}_i\log\frac{1+\boldsymbol{s}'_i\tanh(\boldsymbol{\eta}_i/2)}{1+\boldsymbol{s}'_i\tanh(\boldsymbol{\eta}'_i/2)}\right)^2\right] = \mathbb{E}\left[\sum_{h,i=1}^{\boldsymbol{d}}\boldsymbol{s}_h\boldsymbol{s}_i\log\frac{1+\boldsymbol{s}'_h\tanh(\boldsymbol{\eta}_h/2)}{1+\boldsymbol{s}'_h\tanh(\boldsymbol{\eta}'_h/2)}\log\frac{1+\boldsymbol{s}'_i\tanh(\boldsymbol{\eta}_i/2)}{1+\boldsymbol{s}'_i\tanh(\boldsymbol{\eta}'_i/2)}\right]\\ &= \mathbb{E}\left[\sum_{i=1}^{\boldsymbol{d}}\log^2\frac{1+\boldsymbol{s}'_i\tanh(\boldsymbol{\eta}_i/2)}{1+\boldsymbol{s}'_i\tanh(\boldsymbol{\eta}'_i/2)}\right] = d\mathbb{E}\left[\log^2\frac{1+\boldsymbol{s}_1\tanh(\boldsymbol{\eta}_1/2)}{1+\boldsymbol{s}_1\tanh(\boldsymbol{\eta}'_1/2)}\right].\end{aligned} \tag{4.3}$$

Further,

$$\log^2\frac{1+\tanh(\boldsymbol{\eta}_1/2)}{1+\tanh(\boldsymbol{\eta}'_1/2)} = \left[\int_{\boldsymbol{\eta}'_1}^{\boldsymbol{\eta}_1}\frac{\partial\log(1+\tanh(z/2))}{\partial z}\mathrm{d}z\right]^2 = \left[\int_{\boldsymbol{\eta}_1\wedge\boldsymbol{\eta}'_1}^{\boldsymbol{\eta}_1\vee\boldsymbol{\eta}'_1}\frac{1-\tanh(z/2)}{2}\mathrm{d}z\right]^2, \tag{4.4}$$

$$\log^2\frac{1-\tanh(\boldsymbol{\eta}_1/2)}{1-\tanh(\boldsymbol{\eta}'_1/2)} = \left[\int_{\boldsymbol{\eta}'_1}^{\boldsymbol{\eta}_1}\frac{\partial\log(1-\tanh(z/2))}{\partial z}\mathrm{d}z\right]^2 = \left[\int_{\boldsymbol{\eta}_1\wedge\boldsymbol{\eta}'_1}^{\boldsymbol{\eta}_1\vee\boldsymbol{\eta}'_1}\frac{1+\tanh(z/2)}{2}\mathrm{d}z\right]^2. \tag{4.5}$$

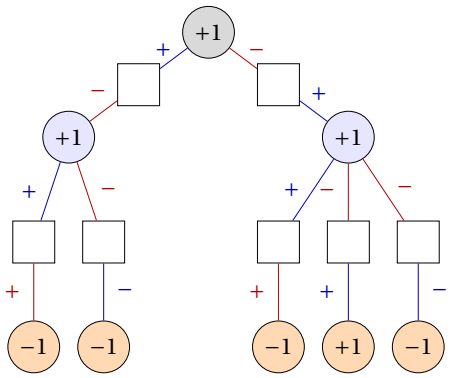

FIGURE 2. The graph $G(\mathbf{\Phi})$ together with extremal boundary condition $\boldsymbol{\sigma}^+$. Variables are indicated by circles and clauses by squares. The labels on the edges illustrate the sign with which variables appears in the clauses. To obtain the extremal boundary condition $\boldsymbol{\sigma}^+$ we proceed top-down. The truth values of the children are chosen so as to nudge the parent variables in the direction provided by $\boldsymbol{\sigma}^+$.

Combining (4.4)–(4.5) and applying the Cauchy-Schwarz inequality, we obtain

$$\mathbb{E}\left[\log^2\frac{1+\boldsymbol{s}_1\tanh(\boldsymbol{\eta}_1/2)}{1+\boldsymbol{s}_1\tanh(\boldsymbol{\eta}_1'/2)}\right] = \frac{1}{2}\mathbb{E}\left[\left[\int_{\boldsymbol{\eta}_1\wedge\boldsymbol{\eta}_1'}^{\boldsymbol{\eta}_1\vee\boldsymbol{\eta}_1'}\frac{1-\tanh(z/2)}{2}\mathrm{d}z\right]^2+\left[\int_{\boldsymbol{\eta}_1\wedge\boldsymbol{\eta}_1'}^{\boldsymbol{\eta}_1\vee\boldsymbol{\eta}_1'}\frac{1+\tanh(z/2)}{2}\mathrm{d}z\right]^2\right]$$
$$\leq \frac{1}{2}\mathbb{E}\left[\left|\boldsymbol{\eta}_1-\boldsymbol{\eta}_1'\right|\int_{\boldsymbol{\eta}_1\wedge\boldsymbol{\eta}_1'}^{\boldsymbol{\eta}_1\vee\boldsymbol{\eta}_1'}\left(\frac{1-\tanh(z/2)}{2}\right)^2+\left(\frac{1+\tanh(z/2)}{2}\right)^2\mathrm{d}z\right] \leq \frac{1}{2}\mathbb{E}\left[\left(\boldsymbol{\eta}_1-\boldsymbol{\eta}_1'\right)^2\right]. \quad (4.6)$$

Finally, (4.3) and (4.6) yield $W_2(\mathrm{LL}(\rho),\mathrm{LL}(\rho'))^2 \leq d\mathbb{E}[(\boldsymbol{\eta}_1-\boldsymbol{\eta}_1')^2]/2$, which implies contraction because $d<2$. □

Proof of Proposition 2.1. Together with the Banach fixed point theorem Lemma 4.3 ensures that the W_2-limit $\rho_d = \lim_{\ell\to\infty}\mathrm{LL}_d^\ell(\delta_0)$ exists. Therefore, Lemma 4.2 implies that the sequence $(\pi_d^{(\ell)})_{\ell\geq 0}$ converges weakly. In addition, since $\rho_d \in \mathscr{W}_2(\mathbb{R})$, Lemma 4.2 also implies the bound (2.1). Finally, to prove (2.2) we apply (2.1) to obtain

$$\mathbb{E}\left|\log\left(\prod_{i=1}^{\boldsymbol{d}^-}\boldsymbol{\mu}_{\pi_d,i}+\prod_{i=1}^{\boldsymbol{d}^+}\boldsymbol{\mu}_{\pi_d,i+\boldsymbol{d}^-}\right)\right| \leq \log(2)-\mathbb{E}\log\prod_{i=1}^{\boldsymbol{d}^-}\boldsymbol{\mu}_{\pi_d,i} \leq \log(2)-\frac{d}{2}\mathbb{E}\log\boldsymbol{\mu}_{\pi_d,1} \leq 2\log(2)+d\mathbb{E}\left|\log\frac{\boldsymbol{\mu}_{\pi_d}}{1-\boldsymbol{\mu}_{\pi_d}}\right|<\infty,$$
$$\mathbb{E}\left|\log(1-\boldsymbol{\mu}_{\pi_d,1}\boldsymbol{\mu}_{\pi_d,2})\right| \leq \mathbb{E}\left|\log(1-\boldsymbol{\mu}_{\pi_d})\right| \leq \mathbb{E}\left|\log\frac{\boldsymbol{\mu}_{\pi_d}}{1-\boldsymbol{\mu}_{\pi_d}}\right|+\log 2<\infty,$$

thereby completing the proof. □

5. PROOF OF PROPOSITION 2.2

5.1. **Outline.** The goal is to prove that the marginal of the root variable o of $\boldsymbol{T}^{(2\ell)}$ remains asymptotically invariant even upon imposition of an arbitrary (feasible) boundary condition on the variables at distance 2ℓ from the root o. A priori, a proof of this statement seems challenging because of the very large number of possible boundary conditions. Indeed, we expect about d^ℓ variables at distance 2ℓ. But a crucial feature of the 2-SAT problem is that we can construct a pair of extremal boundary conditions. One of these maximises the probability that the root is set to one. The other one minimises that probability. As a consequence, instead of inspecting all possible boundary conditions, it suffices to show that the marginals on the root o that these two extremal boundary induce asymptotically coincide with the unconditional marginals. Of course, due to symmetry it actually suffices to consider the 'positive' extremal boundary condition that maximally nudges the root towards +1.

To construct this extremal boundary condition we define a satisfying assignment $\boldsymbol{\sigma}^+$ by working our way down the tree $\boldsymbol{T}^{(2\ell)}$. We begin by defining $\boldsymbol{\sigma}_o^+ = 1$. Further, suppose for $\ell \geq 1$ the values of the variables at distance $2(\ell-1)$ from o have been defined already. Consider a variable $v \in \partial^{2\ell} o$, its parent clause a and the parent variable u of a.

Our aim is to choose $\boldsymbol{\sigma}^+_v$ so as to 'nudge' u towards $\boldsymbol{\sigma}^+_u$ as much as possible. To this end we set $\boldsymbol{\sigma}^+_v$ so as to not satisfy a if setting u to $\boldsymbol{\sigma}^+_u$ satisfies a. Otherwise we pick the value that satisfies a; see Figure 2. In formulas,

$$\boldsymbol{\sigma}^+_v = \mathrm{sign}(a,v)\mathbf{1}\{\mathrm{sign}(a,u) \neq \boldsymbol{\sigma}^+_u\} - \mathrm{sign}(a,v)\mathbf{1}\{\mathrm{sign}(a,u) = \boldsymbol{\sigma}^+_u\}.$$

The following lemma verifies that $\boldsymbol{\sigma}^+$ is extremal, i.e., that imposing the values provided by $\boldsymbol{\sigma}^+$ on the boundary variables $\partial^{2\ell} o$ maximises the probability of the truth value 1 at the root o. The proof can be found in Section 5.2.

Lemma 5.1. *For any integer $\ell \geq 0$ we have* $\max_{\tau \in S(\boldsymbol{T}^{(2\ell)})} \mu_{\boldsymbol{T}^{(2\ell)}}(\boldsymbol{\sigma}_o = 1 \mid \boldsymbol{\sigma}_{\partial^{2\ell} o} = \tau_{\partial^{2\ell} o}) = \mu_{\boldsymbol{T}^{(2\ell)}}(\boldsymbol{\sigma}_o = 1 \mid \boldsymbol{\sigma}_{\partial^{2\ell} o} = \boldsymbol{\sigma}^+_{\partial^{2\ell} o})$.

Lemma 5.1 reduces the task of proving Proposition 2.2 to establishing the following statement.

Proposition 5.2. *We have* $\lim_{\ell\to\infty} \mathbb{E}\left|\mu_{\boldsymbol{T}^{(2\ell)}}(\boldsymbol{\sigma}_o = 1) - \mu_{\boldsymbol{T}^{(2\ell)}}(\boldsymbol{\sigma}_o = 1 \mid \boldsymbol{\sigma}_{\partial^{2\ell} o} = \boldsymbol{\sigma}^+_{\partial^{2\ell} o})\right| = 0$.

In words, the root marginal given the extremal boundary condition $\boldsymbol{\sigma}^+$ asymptotically coincides with the unconditional marginal.

The proof of Proposition 5.2 is delicate because the boundary condition $\boldsymbol{\sigma}^+$ depends on the tree $\boldsymbol{T}^{(2\ell)}$. Indeed, it seems hopeless to confront these dependencies head on by first passing down the tree to construct $\boldsymbol{\sigma}^+$ and to subsequently work up the tree to calculate marginals. To sidestep this problem we devise a quantity that recovers the Markov property of the random tree. Specifically, with each variable node $x \in \partial^{2k} o$, $k > 0$, of $\boldsymbol{T}^{(2\ell)}$ we will associate a carefully defined quantity $\boldsymbol{\eta}^{(\ell)}_x \in \mathbb{R} \cup \{\pm\infty\}$ that gauges how strongly x can nudge its (grand-)parent variable y towards the truth value mandated by $\boldsymbol{\sigma}^+_y$. This random variable $\boldsymbol{\eta}^{(\ell)}_x$ will turn out to be essentially independent of the top $2k$ levels of the tree. In effect, we will discover that the distribution of $\boldsymbol{\eta}^{(\ell)}_o$ can be approximated by the k-fold application of a suitable operator that will turn out to be a W_1-contraction. Taking limits $k, \ell \to \infty$ carefully will then complete the proof.

To facilitate this construction we need to count satisfying assignments of sub-formulas of $\boldsymbol{T}^{(2\ell)}$ subject to certain boundary conditions. Specifically, for a variable x we let $\boldsymbol{T}^{(2\ell)}_x$ be the sub-formula of $\boldsymbol{T}^{(2\ell)}$ comprising x and its progeny. Moreover, for a satisfying assignment $\tau \in S(\boldsymbol{T}^{(2\ell)})$ we let

$$S(\boldsymbol{T}^{(2\ell)}_x, \tau) = \left\{\chi \in S(\boldsymbol{T}^{(2\ell)}_x) : \forall y \in V(\boldsymbol{T}^{(2\ell)}_x) \cap \partial^{2\ell}(\boldsymbol{T}, o) : \chi_y = \tau_y\right\}, \qquad Z(\boldsymbol{T}^{(2\ell)}_x, \tau) = \left|S(\boldsymbol{T}^{(2\ell)}_x, \tau)\right|.$$

In words, $S(\boldsymbol{T}^{(2\ell)}_x, \tau)$ contains all satisfying assignments of $\boldsymbol{T}^{(2\ell)}_x$ that comply with the boundary condition induced by τ. As a final twist, for $t = \pm 1$ we also need the number

$$Z(\boldsymbol{T}^{(2\ell)}_x, \tau, t) = \left|\left\{\chi \in S(\boldsymbol{T}^{(2\ell)}_x, \tau) : \chi_x = t\right\}\right|$$

of satisfying assignments of $\boldsymbol{T}^{(2\ell)}_x$ that agree with τ on the boundary and assign value t to x.

The protagonist of the proof of Proposition 5.2 is the log-likelihood ratio

$$\boldsymbol{\eta}^{(\ell)}_x = \log \frac{Z(\boldsymbol{T}^{(2\ell)}_x, \boldsymbol{\sigma}^+, \boldsymbol{\sigma}^+_x)}{Z(\boldsymbol{T}^{(2\ell)}_x, \boldsymbol{\sigma}^+, -\boldsymbol{\sigma}^+_x)} \in \mathbb{R} \cup \{\pm\infty\} \qquad (x \in V(\boldsymbol{T}^{(2\ell)})), \tag{5.1}$$

with the conventions $\log 0 = -\infty$, $\log\infty = \infty$. Thus, $\boldsymbol{\eta}^{(\ell)}_x$ gauges how likely a random satisfying assignment $\boldsymbol{\sigma}$ of $\boldsymbol{T}^{(2\ell)}_x$ subject to the $\boldsymbol{\sigma}^+$-boundary condition is to set x to its designated value $\boldsymbol{\sigma}^+_x$.

To get a handle on the $\boldsymbol{\eta}^{(\ell)}_x$, we show that these quantities can be calculated by propagating the extremal boundary condition $\boldsymbol{\sigma}^+$ bottom–up toward the root of the tree. Specifically, we consider the operator

$$\mathrm{LL}^+_{\boldsymbol{T}^{(2\ell)}} : (-\infty,\infty]^{V(\boldsymbol{T}^{(2\ell)})} \to (-\infty,\infty]^{V(\boldsymbol{T}^{(2\ell)})}, \qquad \eta \mapsto \hat{\eta}$$

defined as follows. For all $x \in \partial^{2\ell} o$ we set $\hat{\eta}_x = \infty$. Moreover, for a variable $x \in \partial^{2k} o$ with $k < \ell$ with children $a_1, \ldots, a_j$ and grandchildren $y_1 \in \partial a_1 \setminus \{x\}, \ldots, y_j \in \partial a_j \setminus \{x\}$ we define

$$\hat{\eta}_x = -\sum_{i=1}^{j} \boldsymbol{\sigma}^+_x \mathrm{sign}(x, a_i) \log \frac{1 - \boldsymbol{\sigma}^+_x \mathrm{sign}(x, a_i)\tanh(\eta_{y_i}/2)}{2}. \tag{5.2}$$

It may not be apparent that the above sum is well-defined as a $-\infty$ summand might occur. However, the next lemma rules this out and shows that ℓ-fold iteration of $\mathrm{LL}^+_{\boldsymbol{T}^{(2\ell)}}$ from all-$+\infty$ yields $\boldsymbol{\eta}^{(\ell)} = (\boldsymbol{\eta}^{(\ell)}_x)_{x \in V(\boldsymbol{T}^{(2\ell)})}$.

Lemma 5.3. *The operator* $\mathrm{LL}^+_{\boldsymbol{T}^{(2\ell)}}$ *is well-defined and* $\mathrm{LL}^{+(\ell)}_{\boldsymbol{T}^{(2\ell)}}(\infty, \ldots, \infty) = \boldsymbol{\eta}^{(\ell)}$.

We defer the proof of Lemma 5.3 to Section 5.3.

The next aim is to approximate the ℓ-fold iteration of $\mathrm{LL}^+_{\boldsymbol{T}^{(2\ell)}}$, and specifically the distribution of the value $\boldsymbol{\eta}_o^{(\ell)}$ associated with the root, via a non-random operator $\mathscr{P}(\mathbb{R}) \to \mathscr{P}(\mathbb{R})$. To this end we need to cope with the $\pm\infty$-entries of the vector $\boldsymbol{\eta}^{(\ell)}$, a task that we solve by bounding $\boldsymbol{\eta}_x^{(\ell)}$ for variables x near the top of the tree.

Lemma 5.4. *There exist $c = c(d) > 0$ and a sequence $(\varepsilon_k)_{k\geq 1}$ with $\lim_{k\to\infty}\varepsilon_k = 0$ such that for any $k > 0$, $\ell > ck$ we have $\mathbb{P}[\max_{x\in\partial^{2k}o}|\boldsymbol{\eta}_x^{(\ell)}| \leq ck] > 1-\varepsilon_k$.*

The proof of Lemma 5.4, based on a percolation argument, can be found in Section 5.4. We continue to denote by c and $(\varepsilon_k)_k$ the number and the sequence supplied by Lemma 5.4.

Guided by Lemma 5.4 we consider the vector $\bar{\boldsymbol{\eta}}^{(\ell,k)}$ of truncated log-likelihood ratios

$$\bar{\boldsymbol{\eta}}_x^{(\ell,k)} = \begin{cases} -ck & \text{if } x \in \partial^{2k}o \text{ and } \boldsymbol{\eta}_x^{(\ell)} < -ck, \\ ck & \text{if } x \in \partial^{2k}o \text{ and } \boldsymbol{\eta}_x^{(\ell)} > ck, \\ \boldsymbol{\eta}_x^{(\ell)} & \text{otherwise.} \end{cases}$$

Further, let

$$\boldsymbol{\eta}^{(\ell,k)} = \mathrm{LL}^{+(k)}_{\boldsymbol{T}^{(2\ell)}}(\bar{\boldsymbol{\eta}}^{(\ell,k)})$$

be the result of k iterations of $\mathrm{LL}^+_{\boldsymbol{T}^{(2\ell)}}(\cdot)$ starting from $\bar{\boldsymbol{\eta}}^{(\ell,k)}$.

Corollary 5.5. *For any $\ell > ck$ we have $d_{\mathrm{TV}}(\boldsymbol{\eta}_o^{(\ell,k)}, \boldsymbol{\eta}_o^{(\ell)}) < \varepsilon_k$.*

Proof. This follows from Lemma 5.3 and Lemma 5.4, which shows that the truncation is inconsequential with probability at least $1-\varepsilon_k$. □

We are ready to introduce the operator $\mathscr{P}(\mathbb{R}) \to \mathscr{P}(\mathbb{R})$ that mimics $\mathrm{LL}^+_{\boldsymbol{T}^{(2\ell)}}$. Specifically, $\mathrm{LL}^+_d : \mathscr{P}(\mathbb{R}) \to \mathscr{P}(\mathbb{R})$ maps $\rho \in \mathscr{P}(\mathbb{R})$ to the distribution of

$$-\sum_{i=1}^{\boldsymbol{d}} \boldsymbol{s}_i \log \frac{1 - \boldsymbol{s}_i \tanh(\boldsymbol{\eta}_{\rho,i}/2)}{2}. \tag{5.3}$$

We emphasise the subtle difference between (5.3) and (4.1), which involves two independent signs $\boldsymbol{s}_i, \boldsymbol{s}_i'$. The next lemma establishes the connection between the random operator $\mathrm{LL}^+_{\boldsymbol{T}^{(2\ell)}}$ and the operator LL^+_d. Namely, let $\rho^{(\ell,k)}$ be the distribution of $\boldsymbol{\eta}_o^{(\ell,k)}$. Moreover, let $\bar{\rho}^{(\ell-k)}$ be the distribution of

$$\boldsymbol{\eta}_o^{(\ell-k)}\mathbf{1}\{-ck < \boldsymbol{\eta}_o^{(\ell-k)} < ck\} + ck\mathbf{1}\{ck < \boldsymbol{\eta}_o^{(\ell-k)}\} - ck\mathbf{1}\{\boldsymbol{\eta}_o^{(\ell-k)} < -ck\},$$

i.e., the truncation of $\boldsymbol{\eta}_o^{(\ell-k)}$.

Lemma 5.6. *For $\ell > ck$ we have $\rho^{(\ell,k)} = \mathrm{LL}^{+(k)}_d(\bar{\rho}^{(\ell-k)})$.*

We prove Lemma 5.6 in Section 5.5. Recalling φ from (1.14), as in the proof of Proposition 2.1 we let $\rho_d = \varphi(\pi_d)$ be the distribution of the log-likelihood ratio $\log(\boldsymbol{\mu}_{\pi_d}/(1-\boldsymbol{\mu}_{\pi_d}))$.

Lemma 5.7. *The operator LL^+_d is a W_1-contraction with unique fixed point ρ_d.*

The proof of Lemma 5.7 can be found in Section 5.6. Let $(\rho^{(\ell)})_\ell$ be the sequence of distributions of $(\boldsymbol{\eta}_o^{(\ell)})_\ell$. As an immediate consequence we obtain the limit of the sequence $(\rho^{(\ell)})_\ell$. We recall ψ from (1.14).

Corollary 5.8. *The sequence $(\psi(\rho^{(\ell)}))_{\ell\geq 0}$ converges weakly to π_d.*

Proof. This follows from Corollary 5.5, Lemma 5.6, Lemma 5.7 and the continuous mapping theorem. □

Proof of Proposition 5.2. Set $\boldsymbol{\vartheta}_o^{(\ell)} = (\mathrm{LL}^{+(\ell)}_{\boldsymbol{T}^{(2\ell)}}(0,\ldots,0))_o = \log(\mu_{\boldsymbol{T}^{(2\ell)}}(\boldsymbol{\sigma}_o = 1)/\mu_{\boldsymbol{T}^{(2\ell)}}(\boldsymbol{\sigma}_o = -1))$. Then

$$\mu_{\boldsymbol{T}^{(2\ell)}}(\boldsymbol{\sigma}_o = 1) = \psi(\boldsymbol{\vartheta}_o^{(\ell)}) \qquad \text{and} \qquad \mu_{\boldsymbol{T}^{(2\ell)}}(\boldsymbol{\sigma}_o = 1 \mid \boldsymbol{\sigma}_{\partial^{2\ell}o} = \boldsymbol{\sigma}^+_{\partial^{2\ell}o}) = \psi(\boldsymbol{\eta}_o^{(\ell)}).$$

Moreover, Lemma 5.1 shows that $0 \leq \psi(\boldsymbol{\vartheta}_o^{(\ell)}) \leq \psi(\boldsymbol{\eta}_o^{(\ell)}) \leq 1$. Further, Lemma 5.7 implies that $\psi(\boldsymbol{\vartheta}_o^{(\ell)})$ converges weakly to π_d. Finally, Corollary 5.8 implies that $\psi(\boldsymbol{\eta}_o^{(\ell)})$ also converges weakly to π_d, whence

$$\lim_{\ell\to\infty}\mathbb{E}\left|\psi(\boldsymbol{\eta}_o^{(\ell)}) - \psi(\boldsymbol{\vartheta}_o^{(\ell)})\right| = \lim_{\ell\to\infty}\left|\mathbb{E}[\psi(\boldsymbol{\vartheta}_o^{(\ell)})] - \mathbb{E}[\psi(\boldsymbol{\eta}_o^{(\ell)})]\right| = 0,$$

which directly implies the assertion. □

Proof of Proposition 2.2. The proposition follows immediately from Lemma 5.1 and Proposition 5.2. □

5.2. **Proof of Lemma 5.1.** The proof is by induction on the height of the tree. The following claim summarises the main step of the induction.

Claim 5.9. *For all $\ell \geq 0$, all variables x of $\boldsymbol{T}^{(2\ell)}$ and all satisfying assignments $\tau \in S(\boldsymbol{T}^{(2\ell)})$ we have*

$$\frac{Z(\boldsymbol{T}_x^{(2\ell)}, \tau, \boldsymbol{\sigma}_x^+)}{Z(\boldsymbol{T}_x^{(2\ell)}, \tau)} \leq \frac{Z(\boldsymbol{T}_x^{(2\ell)}, \boldsymbol{\sigma}^+, \boldsymbol{\sigma}_x^+)}{Z(\boldsymbol{T}_x^{(2\ell)}, \boldsymbol{\sigma}^+)}. \tag{5.4}$$

Proof. For boundary variables $x \in \partial^{2\ell} o$ there is nothing to show because the r.h.s. of (5.4) equals one. Hence, consider a variable $x \in \partial^{2k} o$ for some $k < \ell$. If $Z(\boldsymbol{T}_x^{(2\ell)}, \tau, \boldsymbol{\sigma}_x^+) = 0$, then (5.4) is trivially satisfied. Hence, assume that $Z(\boldsymbol{T}_x^{(2\ell)}, \tau, \boldsymbol{\sigma}_x^+) > 0$. Let $a_1^+, \ldots, a_g^+$ be the children (clauses) of x with $\mathrm{sign}(x, a_i^+) = \boldsymbol{\sigma}_x^+$. Also let $y_1, \ldots, y_g$ be the children (variables) of $a_1^+, \ldots, a_g^+$. Similarly, let $a_1^-, \ldots, a_h^-$ be the children of x with $\mathrm{sign}(x, a_i^-) = -\boldsymbol{\sigma}_x^+$ and let $z_1, \ldots, z_h$ be their children. We claim that for all $\tau \in S(\boldsymbol{T}^{(2\ell)})$,

$$Z(\boldsymbol{T}_x^{(2\ell)}, \tau, \boldsymbol{\sigma}_x^+) = \prod_{i=1}^{g} Z(\boldsymbol{T}_{y_i}^{(2\ell)}, \tau) \prod_{i=1}^{h} Z(\boldsymbol{T}_{z_i}^{(2\ell)}, \tau, \boldsymbol{\sigma}_{z_i}^+), \quad Z(\boldsymbol{T}_x^{(2\ell)}, \tau, -\boldsymbol{\sigma}_x^+) = \prod_{i=1}^{g} Z(\boldsymbol{T}_{y_i}^{(2\ell)}, \tau, -\boldsymbol{\sigma}_{y_i}^+) \prod_{i=1}^{h} Z(\boldsymbol{T}_{z_i}^{(2\ell)}, \tau). \tag{5.5}$$

For setting x to $\boldsymbol{\sigma}_x^+$ satisfies $a_1^+, \ldots, a_g^+$; hence, arbitrary satisfying assignments of the sub-trees $\boldsymbol{T}_{y_i}^{(2\ell)}$ can be combined, which explains the first product. By contrast, upon assigning x the value $\boldsymbol{\sigma}_x^+$ we need to assign the variables z_i the values $\boldsymbol{\sigma}_{z_i}^+$ so that they satisfy the clauses a_i^-. This leaves us with $Z(\boldsymbol{T}_{z_i}^{(2\ell)}, \tau, \boldsymbol{\sigma}_{z_i}^+)$ possible satisfying assignments of the sub-trees $\boldsymbol{T}_{z_i}^{(2\ell)}$; hence the second product, and we obtain the left equation. A similar argument yields the right one. Dividing the two expressions from (5.5) and invoking the induction hypothesis (for $k+1$), we obtain

$$\begin{aligned}\frac{Z(\boldsymbol{T}_x^{(2\ell)}, \tau, -\boldsymbol{\sigma}_x^+)}{Z(\boldsymbol{T}_x^{(2\ell)}, \tau, \boldsymbol{\sigma}_x^+)} &= \prod_{i=1}^{g} \frac{Z(\boldsymbol{T}_{y_i}^{(2\ell)}, \tau, -\boldsymbol{\sigma}_{y_i}^+)}{Z(\boldsymbol{T}_{y_i}^{(2\ell)}, \tau)} \cdot \prod_{i=1}^{h} \frac{Z(\boldsymbol{T}_{z_i}^{(2\ell)}, \tau)}{Z(\boldsymbol{T}_{z_i}^{(2\ell)}, \tau, \boldsymbol{\sigma}_{z_i}^+)} \\ &\geq \prod_{i=1}^{g} \frac{Z(\boldsymbol{T}_{y_i}^{(2\ell)}, \boldsymbol{\sigma}^+, -\boldsymbol{\sigma}_{y_i}^+)}{Z(\boldsymbol{T}_{y_i}^{(2\ell)}, \boldsymbol{\sigma}^+)} \cdot \prod_{i=1}^{h} \frac{Z(\boldsymbol{T}_{z_i}^{(2\ell)}, \boldsymbol{\sigma}^+)}{Z(\boldsymbol{T}_{z_i}^{(2\ell)}, \boldsymbol{\sigma}^+, \boldsymbol{\sigma}_{z_i}^+)} = \frac{Z(\boldsymbol{T}_x^{(2\ell)}, \boldsymbol{\sigma}^+, -\boldsymbol{\sigma}_x^+)}{Z(\boldsymbol{T}_x^{(2\ell)}, \boldsymbol{\sigma}^+, \boldsymbol{\sigma}_x^+)},\end{aligned}$$

completing the induction. □

Proof of Lemma 5.1. The assertion follows by applying Claim 5.9 to $x = o$. □

5.3. **Proof of Lemma 5.3.** To show that $\mathrm{LL}^+_{\boldsymbol{T}^{(2\ell)}}$ is well defined we verify that, in the notation of (5.2), $\hat{\eta}_x \in (-\infty, \infty]$ for all x. Indeed, in the expression on the r.h.s. of (5.2) a $\pm\infty$ summand can arise only from variables y_i with $\eta_{y_i} = \infty$. But the definition of $\boldsymbol{\sigma}^+$ ensures that such y_i either render a zero summand if $\boldsymbol{\sigma}_x^+ \mathrm{sign}(x, a_i) = -1$, or a $+\infty$ summand if $\boldsymbol{\sigma}_x^+ \mathrm{sign}(x, a_i) = 1$. Thus, the sum is well-defined and $\hat{\eta}_x \in (-\infty, \infty]$.

Further, to verify the identity $\boldsymbol{\eta}^{(\ell)} = \mathrm{LL}^{+(\ell)}_{\boldsymbol{T}^{(2\ell)}}(\infty, \ldots, \infty)$, consider a variable x of $\boldsymbol{T}^{(2\ell)}$. Let $a_1^+, \ldots, a_g^+$ be its children with $\mathrm{sign}(a_i^+, x) = \boldsymbol{\sigma}_x^+$, let $y_1, \ldots, y_g$ be their children, let $a_1^-, \ldots, a_h^-$ be the children of x with $\mathrm{sign}(a_i^-, x) = -\boldsymbol{\sigma}_x^+$ and let $z_1, \ldots, z_h$ be their children. Then (1.14) and (5.5) yield

$$\boldsymbol{\eta}_x^{(\ell)} = -\sum_{i=1}^{g} \log \frac{Z(\boldsymbol{T}_{y_i}^{(2\ell)}, \boldsymbol{\sigma}^+, -\boldsymbol{\sigma}_{y_i}^+)}{Z(\boldsymbol{T}_{y_i}^{(2\ell)}, \boldsymbol{\sigma}^+)} + \sum_{i=1}^{h} \log \frac{Z(\boldsymbol{T}_{z_i}^{(2\ell)}, \boldsymbol{\sigma}^+, \boldsymbol{\sigma}_{z_i}^+)}{Z(\boldsymbol{T}_{z_i}^{(2\ell)}, \boldsymbol{\sigma}^+)} = -\sum_{i=1}^{g} \log \frac{1 - \tanh(\boldsymbol{\eta}_{y_i}^{(\ell)}/2)}{2} + \sum_{i=1}^{h} \log \frac{1 + \tanh(\boldsymbol{\eta}_{z_i}^{(\ell)}/2)}{2}.$$

The assertion follows because $\mathrm{sign}(x, a_i^+)\boldsymbol{\sigma}_x^+ = 1$ and $\mathrm{sign}(x, a_i^-)\boldsymbol{\sigma}_x^+ = -1$.

5.4. **Proof of Lemma 5.4.** The goal is to prove that for variables some distance away from level 2ℓ of $\boldsymbol{T}^{(2\ell)}$ the counts $Z(\boldsymbol{T}_x^{(2\ell)}, \boldsymbol{\sigma}^+, \pm 1)$ are roughly of the same order of magnitude. Approaching this task somewhat indirectly, we begin by tracing the logical implications of imposing a specific value $s = \pm 1$ on a variable x of the (possibly infinite) tree $\boldsymbol{T}$. Clearly, upon setting x to the value s a child (clause) a of x will be satisfied iff x appears in a with sign s. In effect, all clauses a with $\mathrm{sign}(a, x) \neq s$ need to be satisfied by their second variable y, a grandchild of x. Thus, we impose the value $\mathrm{sign}(a, y)$ on y and recurse down the tree. Let $\boldsymbol{T}_{x,s}$ denote the sub-tree of $\boldsymbol{T}$ comprising x and all other variables on which this process imposes specific values as well as all clauses that contain two such variables. Clearly, for every leaf y of $\boldsymbol{T}_{x,s}$ the values imposed on y happens to satisfy all child clauses of y in $\boldsymbol{T}$. Let $\boldsymbol{N}_{x,s} \in [1, \infty]$ be the number of variables in $\boldsymbol{T}_{x,s}$. The next lemma shows that the impact of a boundary condition on the marginal of x can be bounded in terms of $\boldsymbol{N}_{x,s}$.

Claim 5.10. *Let $s \in \{\pm 1\}$. If $x \in \partial^{2k} o$ satisfies $\boldsymbol{N}_{x,s} < \ell - k$ then $Z(\boldsymbol{T}_x^{(2\ell)}, \tau) \le 2^{\boldsymbol{N}_{x,s}} Z(\boldsymbol{T}_x^{(2\ell)}, \tau, s)$.*

Proof. The construction of the implication tree $\boldsymbol{T}_{x,s}$ imposes a truth value σ_y on each variable y of the tree that y must inevitably take if x gets assigned s. Thus, $\boldsymbol{T}_{x,s}$ comes with a satisfying assignment $\sigma \in S(\boldsymbol{T}_{x,s})$ with $\sigma_x = s$. For any leaf y of $\boldsymbol{T}_{x,s}$ every child clause a of y in the super-tree $\boldsymbol{T}$ will be automatically satisfied by setting y to σ_y (because otherwise a would have been included in $\boldsymbol{T}_{x,s}$). Hence, all the clauses of $\boldsymbol{T}$ that are children of the leaves of $\boldsymbol{T}_{x,s}$ are satisfied by σ. Moreover, because $\boldsymbol{N}_{x,s} < \ell - k$, any leaf y of $\boldsymbol{T}_{x,s}$ has distance less than 2ℓ from o. Thus, the assignment σ does not clash with the boundary condition τ. As a consequence, for any $\chi \in S(\boldsymbol{T}_x^{(2\ell)}, \tau)$ we obtain another satisfying assignment $\chi' \in S(\boldsymbol{T}_x^{(2\ell)}, \tau)$ by letting

$$\chi'_z = \begin{cases} \sigma_z & \text{if } z \in V(\boldsymbol{T}_{x,s}), \\ \chi_z & \text{otherwise.} \end{cases}$$

Moreover, under the map $\chi \mapsto \chi'$ the number of inverse images of any assignment χ' is bounded by the total number $2^{\boldsymbol{N}_{x,s}}$ of different truth assignments of the variables $V(\boldsymbol{T}_{x,s})$. Therefore, $Z(\boldsymbol{T}_x^{(2\ell)}, \tau) \le 2^{\boldsymbol{N}_{x,s}} Z(\boldsymbol{T}_x^{(2\ell)}, \tau, s)$. □

As a next step we bound the random variable $\boldsymbol{N}_{x,s}$.

Claim 5.11. *There exists a number $\alpha = \alpha(d) > 0$ such that $\mathbb{P}\left[\boldsymbol{N}_{o,s} \ge u\right] \le \exp(-u\alpha)/\alpha$ for all $u \ge 0$, $s \in \{\pm 1\}$.*

Proof. In the construction of $\boldsymbol{T}_{o,s}$ we only propagate along clauses in which the parent variable is forced to take a value that fails to satisfy the clause. Since the signs are uniformly random, the number of such child clauses has distribution $\mathrm{Po}(d/2)$. Therefore, $\boldsymbol{N}_{o,s}$ is bounded by the total progeny of a Galton-Watson process with $\mathrm{Po}(d/2)$ offspring. The assertion therefore follows from the tail bound for such processes (e.g., [6, eq. (11.7)]). □

As a final preparation toward the proof of Lemma 5.4 we need a bound on the size of the $2k$-th level of $\boldsymbol{T}$.

Claim 5.12. *We have $\lim_{k\to\infty} \mathbb{P}\left[|\partial^{2k} o| > 2d^k + k\right] = 0$.*

Proof. Since every clause of $\boldsymbol{T}$ has precisely one child, the size of level $2k$ of $\boldsymbol{T}$ coincides with the size of the k-th level of a $\mathrm{Po}(d)$ Galton-Watson tree. Therefore, the assertion follows from standard tail bounds for Galton-Watson processes (e.g., [6, eq. (11.7)]). □

Proof of Lemma 5.4. Claim 5.11 ensures that for a large enough constant $c = c(d) > 0$ and all large enough k,

$$\mathbb{P}\left(\boldsymbol{N}_{o,\pm 1} \ge ck\right) \le (2d)^{-k}. \tag{5.6}$$

Combining (5.6) with Claim 5.12 and using the union bound, we obtain a sequence $\varepsilon_k \to 0$ such that

$$\mathbb{P}\left(\forall x \in \partial^{2k} o : \boldsymbol{N}_{x,\pm 1} < ck\right) \ge 1 - \varepsilon_k. \tag{5.7}$$

Further, if $x \in \partial^{2k} o$ satisfies $\boldsymbol{N}_{x,\pm 1} < ck$ and $\ell > (1+c)k$, Claim 5.10 ensures that for all $x \in \partial^{2k} o$,

$$\left|\boldsymbol{\eta}_x^{(\ell)}\right| \le \log \frac{Z(\boldsymbol{T}_x^{(2\ell)}, \boldsymbol{\sigma}^+)}{Z(\boldsymbol{T}_x^{(2\ell)}, \boldsymbol{\sigma}^+, 1)} + \log \frac{Z(\boldsymbol{T}_x^{(2\ell)}, \boldsymbol{\sigma}^+)}{Z(\boldsymbol{T}_x^{(2\ell)}, \boldsymbol{\sigma}^+, -1)} \le \boldsymbol{N}_{x,1} + \boldsymbol{N}_{x,-1} < 2ck. \tag{5.8}$$

Combining (5.7) and (5.8) completes the proof. □

5.5. **Proof of Lemma 5.6.** A straightforward induction shows that for any $p \in \mathscr{P}(\mathbb{R})$ the result $p^{(k)} = \mathrm{LL}_d^{+(k)}(p)$ of the k-fold application of LL_d^+ coincides with the distribution of the root value of the random operator $\mathrm{LL}^{+(k)}_{\boldsymbol{T}^{(2k)}}$ applied to a vector $(\boldsymbol{\eta}_x)_{x \in V(\boldsymbol{T}^{(2k)})}$ of independent samples from p. Indeed, for $k = 1$ the claim is immediate from the definitions. Moreover, for the inductive step we notice that the k-fold application of LL_d^+ comes down to applying LL_d^+ once to the outcome of the $(k-1)$-fold application. By the induction hypothesis,

$$p^{(k-1)} = \left(\mathrm{LL}^{+(k-1)}_{\boldsymbol{T}^{(2(k-1))}}(\boldsymbol{\eta}_x)_x\right)_o.$$

Finally, applying LL_d^+ to $p^{(k-1)}$ implies the assertion because the first layer of $\boldsymbol{T}^{(2k)}$ is independent of the subtrees rooted at the grandchildren $\partial^2 o$ of the root, which are distributed as independent random trees $\boldsymbol{T}^{(2(k-1))}$. The lemma follows from applying this identity to $p = \bar{\rho}^{(\ell-k)}$.

5.6. **Proof of Lemma 5.7.** The operator LL_d^+ maps the space $\mathscr{W}_1(\mathbb{R})$ into itself because the derivative of $x \mapsto \log((1-\tanh(x/2))/2)$ is bounded by one in absolute value for all $x \in \mathbb{R}$. We proceed to show that $\mathrm{LL}_d^+ : \mathscr{W}_1(\mathbb{R}) \to \mathscr{W}_1(\mathbb{R})$ is a contraction. Thus, consider a sequence of independent random pairs $(\boldsymbol{\eta}_i, \boldsymbol{\eta}_i')_{i\geq 1}$ with $\boldsymbol{\eta}_i \stackrel{\mathrm{d}}{=} \rho$, $\boldsymbol{\eta}_i' \stackrel{\mathrm{d}}{=} \rho'$. Then

$$W_1(\mathrm{LL}_d^+(\rho), \mathrm{LL}_d^+(\rho')) \leq \mathbb{E}\left|\sum_{i=1}^{\boldsymbol{d}} \boldsymbol{s}_i \log\frac{1-\boldsymbol{s}_i\tanh(\boldsymbol{\eta}_i/2)}{1-\boldsymbol{s}_i\tanh(\boldsymbol{\eta}_i'/2)}\right| \leq d\mathbb{E}\left|\log\frac{1-\boldsymbol{s}_1\tanh(\boldsymbol{\eta}_1/2)}{1-\boldsymbol{s}_1\tanh(\boldsymbol{\eta}_1'/2)}\right|.$$

Since the function $z \mapsto \log(1+\tanh(z/2))$ is monotonically increasing, we obtain

$$\left|\log\frac{1+\tanh(\boldsymbol{\eta}_1/2)}{1+\tanh(\boldsymbol{\eta}_1'/2)}\right| = \left|\int_{\boldsymbol{\eta}_1'}^{\boldsymbol{\eta}_1} \frac{\partial\log(1+\tanh(z/2))}{\partial z}\mathrm{d}z\right| = \int_{\boldsymbol{\eta}_1\wedge\boldsymbol{\eta}_1'}^{\boldsymbol{\eta}_1\vee\boldsymbol{\eta}_1'} \frac{1-\tanh(z/2)}{2}\mathrm{d}z,$$
$$\left|\log\frac{1-\tanh(\boldsymbol{\eta}_1/2)}{1-\tanh(\boldsymbol{\eta}_1'/2)}\right| = \left|\int_{\boldsymbol{\eta}_1'}^{\boldsymbol{\eta}_1} \frac{\partial\log(1-\tanh(z/2))}{\partial z}\mathrm{d}z\right| = \int_{\boldsymbol{\eta}_1\wedge\boldsymbol{\eta}_1'}^{\boldsymbol{\eta}_1\vee\boldsymbol{\eta}_1'} \frac{1+\tanh(z/2)}{2}\mathrm{d}z.$$

Hence, $W_1(\mathrm{LL}_d^+(\rho), \mathrm{LL}_d^+(\rho')) \leq d\mathbb{E}\left|\boldsymbol{\eta}_1 - \boldsymbol{\eta}_1'\right|/2$ and therefore $W_1(\mathrm{LL}_d^+(\rho), \mathrm{LL}_d^+(\rho')) \leq dW_1(\rho,\rho')/2$.

Finally, we observe that ρ_d is a fixed point of LL_d^+. Indeed, Proposition 2.1 implies that $\boldsymbol{\eta}^{\rho_d}$ and $-\boldsymbol{\eta}^{\rho_d}$ are identically distributed. Therefore, if $\boldsymbol{s}_i, \boldsymbol{s}_i' \in \{\pm 1\}$ are uniform and independent, we obtain

$$\boldsymbol{s}_i \log\Big(\Big(1-\boldsymbol{s}_i\tanh(\boldsymbol{\eta}_{\rho_d,i}/2)\Big)/2\Big) \stackrel{\mathrm{d}}{=} \boldsymbol{s}_i \log\Big(\Big(1+\boldsymbol{s}_i'\tanh(\boldsymbol{\eta}_{\rho_d,i}/2)\Big)/2\Big).$$

Hence, recalling the definitions (4.1) and (5.3) of the operators, we see that $\mathrm{LL}_d^+(\rho_d) = \mathrm{LL}_d(\rho_d) = \rho_d$.

5.7. **Proof of Theorem 1.2.** Consider the sub-formula $\nabla^{2\ell}(\boldsymbol{\Phi}, x_1)$ of $\boldsymbol{\Phi}$ obtained by deleting all clauses and variables at distance greater than 2ℓ from x_1. By design, we can couple $\nabla^{2\ell}(\boldsymbol{\Phi}, x_1)$ and $\boldsymbol{T}^{(2\ell)}$ such that both coincide w.h.p. Therefore, since any satisfying assignment of $\boldsymbol{\Phi}$ induces a satisfying assignment of $\boldsymbol{T}^{(2\ell)}$, Proposition 2.2 implies the Gibbs uniqueness property (1.11). Furthermore, because Proposition 1.4 shows that Belief Propagation correctly computes the root marginal $\mu_{\boldsymbol{T}^{(2\ell)}}(\boldsymbol{\sigma}_o = 1)$, (1.9) follows from (1.11).

5.8. **Proof of Corollary 1.3.** Let $\pi_d^{(\ell)} = \mathrm{BP}^{(\ell)}(\delta_{1/2})$. Thanks to Proposition 2.1 it suffices to prove that

$$\lim_{\ell\to\infty}\limsup_{n\to\infty}\mathbb{E}[W_1(\pi_{\boldsymbol{\Phi}}, \pi_d^{(\ell)})] = 0. \tag{5.9}$$

Hence, fix $\varepsilon > 0$, pick a large $\ell = \ell(\varepsilon) > 0$ and a larger $L = L(\ell) > 0$. A routine second moment calculation shows that for any possible outcome T of $\boldsymbol{T}^{(2\ell)}$ the number X_T of variables x_i of $\boldsymbol{\Phi}$ such that $\nabla^{2\ell}(\boldsymbol{\Phi}, x_i) = T$ satisfies $X_T = n\mathbb{P}\left[\boldsymbol{T}^{(2\ell)} = T\right] + o(n)$ w.h.p. Hence, w.h.p. $\boldsymbol{\Phi}$ admits a coupling $\gamma_{\boldsymbol{\Phi}}$ of $\boldsymbol{T}^{(2\ell)}$ and a uniform variable $\boldsymbol{i}$ on $[n]$ such that $\gamma(\{\nabla^{2\ell}(\boldsymbol{\Phi}, x_{\boldsymbol{i}}) = \boldsymbol{T}^{(2\ell)}\}) = 1 - o(1)$. Further, Theorem 1.2 implies that given $\nabla^{2\ell}(\boldsymbol{\Phi}, x_{\boldsymbol{i}}) = \boldsymbol{T}^{(2\ell)}$ we have

$$\mathbb{P}\left[\left|\mu_{\boldsymbol{\Phi}}(\boldsymbol{\sigma}_{x_i} = 1) - \mu_{\boldsymbol{T}^{(2\ell)}}(\boldsymbol{\tau}_o = 1)\right| > \varepsilon\right] < \varepsilon, \tag{5.10}$$

provided ℓ is large enough. Finally, Lemma 1.4 implies together with a straightforward induction on ℓ that $\pi_d^{(\ell)}$ is the distribution of $\mu_{\boldsymbol{T}^{(2\ell)}}(\boldsymbol{\tau}_o = 1)$. Therefore, (5.9) follows from (5.10).

5.9. **Proof of Corollary 2.3.** Fix $\varepsilon > 0$ and pick a small $\xi = \xi(\varepsilon) > 0$ and large $\ell = \ell(\xi) > 0$. Since k is fixed independently of n, Theorem 1.2 shows that w.h.p.

$$\sum_{i=1}^{k}\max_{\tau\in S(\boldsymbol{\Phi})}\left|\mu_{\boldsymbol{\Phi}}(\boldsymbol{\sigma}_{x_i} = 1 \mid \boldsymbol{\sigma}_{\partial^{2\ell}x_i} = \tau_{\partial^{2\ell}x_i}) - \mu_{\boldsymbol{\Phi},x_i}^{(\ell)}(1)\right| < \xi. \tag{5.11}$$

Further, the smallest pairwise distance between $x_1,\ldots,x_n$ exceeds 4ℓ w.h.p. Therefore, we can draw a sample $\boldsymbol{\sigma}$ from $\mu_{\boldsymbol{\Phi}}$ in two steps. First, draw $\boldsymbol{\sigma}'$ from $\mu_{\boldsymbol{\Phi}}$. Then, independently re-sample assignments of all the variables in $\nabla^{2\ell-2}(\boldsymbol{\Phi}, x_i)$ from $\mu_{\boldsymbol{\Phi}}(\cdot|\boldsymbol{\sigma}'_{\partial^{2\ell}x_i})$ for $i = 1,\ldots,k$. The resulting assignment $\boldsymbol{\sigma}''$ has distribution $\mu_{\boldsymbol{\Phi}}$ and the values $\boldsymbol{\sigma}''_{x_i}$, $i \in [k]$, are mutually independent given $\boldsymbol{\sigma}'$. Finally, since (5.11) shows that conditioning on the boundary conditions $\boldsymbol{\sigma}'_{\partial^{2\ell}x_i}$ is inconsequential w.h.p., we obtain the assertion by taking $\varepsilon \to 0$ sufficiently slowly.

6. Proof of Proposition 2.4

6.1. **Outline.** The proof is based on a natural coupling of the random formulas $\mathbf{\Phi}_n$ and $\mathbf{\Phi}_{n+1}$ with n and $n+1$ variables, respectively. Specifically, let

$$\boldsymbol{m}' \stackrel{\mathrm{d}}{=} \mathrm{Po}(dn/2 - d/2), \qquad \mathbf{\Delta}'' \stackrel{\mathrm{d}}{=} \mathrm{Po}(d/2), \qquad \mathbf{\Delta}''' \stackrel{\mathrm{d}}{=} \mathrm{Po}(d) \tag{6.1}$$

be independent random variables. Moreover, let $\mathbf{\Phi}'$ be a random formula with n variables and $\boldsymbol{m}'$ clauses, chosen independently and uniformly from the set of all $4n(n-1)$ possible clauses. Then obtain $\mathbf{\Phi}''$ from $\mathbf{\Phi}'$ by adding another $\mathbf{\Delta}''$ uniformly random and independent clauses. Moreover, obtain $\mathbf{\Phi}'''$ from $\mathbf{\Phi}'$ by adding one variable x_{n+1} along with $\mathbf{\Delta}'''$ clauses, chosen uniformly and independently from the set of all $8n$ possible clauses that contain x_{n+1} and another variable from the set $\{x_1, \ldots, x_n\}$.

Fact 6.1. *We have $\mathbf{\Phi}'' \stackrel{d}{=} \mathbf{\Phi}_n$ and $\mathbf{\Phi}''' \stackrel{d}{=} \mathbf{\Phi}_{n+1}$; therefore,*

$$\mathbb{E}[\log(Z(\mathbf{\Phi}_{n+1}) \vee 1)] - \mathbb{E}[\log(Z(\mathbf{\Phi}_n) \vee 1)] = \mathbb{E}\left[\log \frac{Z(\mathbf{\Phi}''') \vee 1}{Z(\mathbf{\Phi}') \vee 1}\right] - \mathbb{E}\left[\log \frac{Z(\mathbf{\Phi}'') \vee 1}{Z(\mathbf{\Phi}') \vee 1}\right]. \tag{6.2}$$

Hence, the proof of Proposition 2.4 boils down to establishing the following two statements.

Proposition 6.2. *We have* $\lim_{n\to\infty} \mathbb{E}\log \frac{Z(\mathbf{\Phi}'') \vee 1}{Z(\mathbf{\Phi}') \vee 1} = \frac{d}{2}\mathbb{E}\left[\log\left(1 - \boldsymbol{\mu}_{\pi_d,1}\boldsymbol{\mu}_{\pi_d,2}\right)\right]$.

Proposition 6.3. *We have* $\lim_{n\to\infty} \mathbb{E}\log \frac{Z(\mathbf{\Phi}''') \vee 1}{Z(\mathbf{\Phi}') \vee 1} = \mathbb{E}\left[\log\left(\sum_{\sigma\in\{\pm1\}} \prod_{i=1}^{\boldsymbol{d}} \left(1 - \mathbf{1}\{\sigma \neq \boldsymbol{s}_i\}\boldsymbol{\mu}_{\pi_d,i}\right)\right)\right]$.

Further, to prove Propositions 6.2 and 6.3 we 'just' need to understand the impact of a bounded expected number of 'local' changes (such as adding a random clause) on the partition function.

The proof strategy sketched in the previous paragraph is known as the Aizenman-Sims-Starr scheme. The technique was originally deployed to study the Sherrington-Kirkpatrick spin glass model [5], but has since found various applications to models on sparse random graphs (e.g., [16, 40]). By comparison to prior applications, the difficulty here is that we apply this technique to a model with hard constraints. In effect, while typically the addition of a single clause will only reduce the number of satisfying assignments by a bounded factor, occasionally a much larger change might ensue. For instance, for any $0 < d < 2$ there is a small but non-zero probability that a single additional clause might close a 'bicycle', i.e., a sequence of clauses that induce an implication chain $x_i \to \cdots \to \neg x_i \to \cdots \to x_i$. Thus, a single unlucky clause might wipe out all satisfying assignments.

Suppose we wish to roughly estimate the change in the number of satisfying assignments upon going from $\mathbf{\Phi}'$ to $\mathbf{\Phi}'''$. Clearly $Z(\mathbf{\Phi}''') \leq 2Z(\mathbf{\Phi}')$ because we only add one new variable. But of course $Z(\mathbf{\Phi}''')$ might be much smaller than $Z(\mathbf{\Phi}')$. To obtain a bound, consider the new clauses $b_1, \ldots, b_{\mathbf{\Delta}'''}$ that were added along with x_{n+1} and let $y_1, \ldots, y_{\mathbf{\Delta}'''}$ be the variables of $\mathbf{\Phi}'$ where the new clauses attach. Define an assignment $\chi : Y = \{y_1, \ldots, y_{\mathbf{\Delta}'''}\} \to \{\pm1\}$ by letting $\chi_{y_i} = \mathrm{sign}(y_i, b_i)$; thus, χ satisfies the b_i. Further, let

$$S(\mathbf{\Phi}', \chi) = \left\{\sigma \in S(\mathbf{\Phi}') : \forall y \in Y : \sigma_y = \chi_y\right\}, \qquad Z(\mathbf{\Phi}', \chi) = |S(\mathbf{\Phi}', \chi)|$$

be the set and the number of satisfying assignments of $\mathbf{\Phi}'$ that coincide with χ on Y. Because each $\sigma \in S(\mathbf{\Phi}', \chi)$ already satisfies all the new clauses regardless of the value assigned to x_{n+1}, we obtain $Z(\mathbf{\Phi}''') \geq 2Z(\mathbf{\Phi}', \chi)$. Hence, it seems that we just need to lower bound $Z(\mathbf{\Phi}', \chi)$.

To this end we could employ a process similar to the one that we applied in Section 5.4 to the tree $\boldsymbol{T}$. Generally, let $Y \subseteq \{x_1, \ldots, x_n\}$ be a set of variables and let $\chi \in \{\pm1\}^Y$ be an assignment. The following process, known as the Unit Clause Propagation algorithm [26], chases the implications of imposing the assignment χ on Y:

> while $\mathbf{\Phi}'$ possesses a clause a that has exactly one neighbouring variable $z \in \partial a$ on which the value $-\mathrm{sign}(z, a)$ has been imposed, impose the value $\mathrm{sign}(a, z')$ on the second variable $z' \in \partial a \setminus \{z\}$ of a.

Let $\mathscr{I}_\chi$ be the set of variables on which the process has imposed a value upon termination (including the initial set Y). Unfortunately, it is possible that $\mathbf{\Phi}'$ contains a clause a on whose both variables z, z' the 'wrong' values $-\mathrm{sign}(a, z), -\mathrm{sign}(a, z')$ got imposed. In other words, Unit Clause might be left with contradictions. If such a clause exists we let $\boldsymbol{I}_\chi = n$. Otherwise we set $\boldsymbol{I}_\chi = |\mathscr{I}_\chi|$. We obtain the following lower bound on $Z(\mathbf{\Phi}', \chi)$.

Fact 6.4. *We have $Z(\mathbf{\Phi}') \leq 2^{\boldsymbol{I}_\chi}(Z(\mathbf{\Phi}', \chi) \vee 1)$.*

Proof. The inequality is trivially satisfied if $Z(\mathbf{\Phi}')=0$ or $\boldsymbol{I}_\chi=n$. Hence, we may assume that $Z(\mathbf{\Phi}')>0$ and that Unit Clause did not run into a contradiction. Consequently, Unit Clause produced an assignment χ^* of the variables $\mathscr{I}_\chi$ that satisfies all clauses a of $\mathbf{\Phi}'$ with $\partial a\cap\mathscr{I}_\chi\neq\emptyset$. Hence, for any satisfying assignment $\sigma\in S(\mathbf{\Phi}')$ we obtain another satisfying assignment $\hat\sigma\in S(\mathbf{\Phi}',\chi)$ by letting $\hat\sigma=\chi^*_x\mathbf{1}\{x\in\mathscr{I}_\chi\}+\sigma_x\mathbf{1}\{x\notin\mathscr{I}_\chi\}$, i.e., we overwrite the variables in $\mathscr{I}_\chi$ according to χ^*. Clearly, under the map $\sigma\mapsto\hat\sigma$ an assignment $\hat\sigma\in S(\mathbf{\Phi}',\chi)$ has at most $2^{\boldsymbol{I}_\chi}$ inverse images. □

Hence, we need an upper bound on $\boldsymbol{I}_\chi$, which will be proven at the end of Section 6.2.

Lemma 6.5. *There exists $C=C(d)>0$ such that for every set $Y\subseteq\{x_1,\ldots,x_n\}$ of size $|Y|\leq\log^2 n$ and any $\chi\in\{\pm1\}^Y$ we have $\mathbb{E}[\boldsymbol{I}_\chi]\leq C|Y|^2$.*

Unfortunately, this first moment bound does not quite suffice for our purposes. Indeed, Lemma 6.5 allows for the possibility that $\boldsymbol{I}_\chi=n$ with probability $\Omega(1/n)$. In combination with Fact 6.4 this rough bound would lead to error terms that eclipse the 'main' terms displayed in Propositions 6.2 and 6.3. But we cannot hope for a much better bound on $\boldsymbol{I}_\chi$. Indeed, $\mathbb{P}\left[\boldsymbol{I}_\chi=n\right]=\Omega(1/n)$ because the graph $G(\mathbf{\Phi}')$ likely contains a few short cycles and if Y contains a variable on a short cycle, then there is a $\Omega(1)$ probability that Unit Clause will cause a contradiction.

Hence, we need to be more circumspect. Previously we aimed for an assignment χ that satisfied *all* the new clauses $b_1,\ldots,b_{\mathbf{\Delta}'''}$ added upon going to $\mathbf{\Phi}'''$. But we still have the new variable x_{n+1} at our disposal to at least satisfy a single clause b_i. Hence, we can afford to start Unit Clause from an assignment χ' that differs from χ on a single variable. Thus, for a set Y of variables and $\chi\in\{\pm1\}^Y$ we define

$$\boldsymbol{A}_\chi=\min\left\{\boldsymbol{I}_{\chi'}:\chi'\in\{\pm1\}^Y,\ \sum_{y\in Y}\mathbf{1}\{\chi_y\neq\chi'_y\}\leq1\right\}. \tag{6.3}$$

Lemma 6.6. *There exists $C'=C'(d)>0$ such that for every set $Y\subseteq\{x_1,\ldots,x_n\}$ of size $|Y|\leq\log^2 n$ and any $\chi\in\{\pm1\}^Y$ we have $\mathbb{E}[\boldsymbol{A}_\chi^2]\leq C'|Y|^4$.*

This second moment bound significantly improves over Lemma 6.5. For instance, Lemma 6.6 implies that the probability of an enormous drop $Z(\mathbf{\Phi}''')\leq\exp(-\Omega(n))Z(\mathbf{\Phi}')$ is bounded by $O(n^{-2})$. Once more this estimate is about tight because there is an $\Omega(n^{-2})$ probability that a single new clause closes a bicycle. As we shall see, with a bit of care the bound from Lemma 6.6 suffices to prove Propositions 6.2 and 6.3. Yet Lemma 6.5 has its uses, too, as it implies the following vital tail bound.

Corollary 6.7. *We have* $\limsup_{n\to\infty}\mathbb{E}\left[n\wedge\left|\log\frac{\mu_{\mathbf{\Phi}'}(\boldsymbol{\sigma}_{x_1}=1)}{\mu_{\mathbf{\Phi}'}(\boldsymbol{\sigma}_{x_1}=-1)}\right|\mid Z(\mathbf{\Phi}')>0\right]<\infty$.

We proceed to study Unit Clause Propagation in order to prove Lemmas 6.5, 6.6 and Corollary 6.7. Then we will prove Propositions 6.2 and 6.3, which imply Proposition 2.4.

6.2. **Unit Clause Propagation.** To avoid dependencies we consider a binomial model $\mathbf{\Phi}^\dagger$ of a random 2-SAT formula with variables $x_1,\ldots,x_n$, where each of the $4\binom{n}{2}$ possible (unordered) 2-clauses is present with probability

$$p=d/(4n)+n^{-4/3} \tag{6.4}$$

independently. We define a random variable $\boldsymbol{A}^\dagger_\chi$ on $\mathbf{\Phi}^\dagger$ in perfect analogy to $\boldsymbol{A}_\chi$. Since the choice (6.4) of p ensures that $\mathbf{\Phi}^\dagger$ and $\mathbf{\Phi}'$ can be coupled so that the former has more clauses than the latter with probability $1-o(n^{-2})$, it suffices to analyse $\boldsymbol{A}^\dagger_\chi$. Moreover, thanks to symmetry it suffices to prove Lemmas 6.5 and 6.6 under the assumption that the initial set of variables is $Y=\{x_1,\ldots,x_\ell\}$, $\ell\leq\log^2 n$.

At first glance investigating $\boldsymbol{A}^\dagger_\chi$ appears to be complicated by the fact that (6.3) takes the minimum over all possible χ'. To sidestep this issue we will investigate a 'comprehensive' propagation process whose progeny encompasses all the unit clauses that may result from any χ'. In its first round this process pursues for each variable x_i, $i\leq\ell$, the Unit Clauses created by imposing either of the two possible truth values on x_i. The effect will be the imposition of truth values on all variables at distance two from Y. Subsequently we trace Unit Clause Propagation from the values imposed on the variables in $\partial^2 Y$. Hence, the difficulty of considering all χ' as in (6.3) disappears because the first step disregards χ.

To deal with possible contradictions the process will actually operate on literals rather than variables. Throughout each literal will belong to one of three possible categories: unexplored, explored, or finished. Initially the 2ℓ literals $x_1,\neg x_1,\ldots,x_\ell,\neg x_\ell$ qualify as explored and all others as unexplored. Formally, we let $\mathscr{E}_0=\{x_1,\neg x_1,\ldots,x_\ell,\neg x_\ell\}$,

$\mathscr{U}_0 = \{x_{\ell+1}, \neg x_{\ell+1}, \ldots, x_n, \neg x_n\}$ and $\mathscr{F}_0 = \emptyset$. Further, for $t \geq 0$ we construct $\mathscr{E}_{t+1}, \mathscr{U}_{t+1}, \mathscr{F}_{t+1}$ as follows. If $\mathscr{E}_t = \emptyset$, the process has terminated and we set $\mathscr{E}_{t+1} = \mathscr{E}_t, \mathscr{U}_{t+1} = \mathscr{U}_t, \mathscr{F}_{t+1} = \mathscr{F}_t$. Otherwise, pick a literal $l_{t+1} \in \mathscr{E}_t$ and let $\mathscr{E}'_{t+1}$ be the set of all literals $l' \in \mathscr{U}_t$ such that $\mathbf{\Phi}^\dagger$ features the clause $\neg l_{t+1} \vee l'$. Further, let

$$\mathscr{U}_{t+1} = \mathscr{U}_t \setminus \mathscr{E}'_{t+1}, \qquad \mathscr{E}_{t+1} = (\mathscr{E}_t \cup \mathscr{E}'_{t+1}) \setminus \{l_{t+1}\}, \qquad \mathscr{F}_{t+1} = \mathscr{F}_t \cup \{l_{t+1}\}.$$

Finally, the set $\mathscr{F}_\infty = \bigcup_{t \geq 1} \mathscr{F}_t$ contains all literals upon which Unit Clause could impose the value 'true' from any initial assignment χ. A contradiction might result only if $x_i, \neg x_i \in \mathscr{F}_\infty$ for some $i > \ell$.

Claim 6.8. *For all $T > 8\ell/(2-d)$ we have $\mathbb{P}[|\mathscr{F}_\infty| > T] \leq \exp(-dT/36)$.*

Proof. Let $t \geq 0$. Given $|\mathscr{U}_t|$ and $|\mathscr{E}_t|$ we have

$$\boldsymbol{X}_{t+1} = |\mathscr{E}_{t+1}| - |\mathscr{E}_t| + \mathbf{1}\{\mathscr{E}_t \neq \emptyset\} \stackrel{d}{=} \mathrm{Bin}\left(|\mathscr{U}_t| \mathbf{1}\{|\mathscr{E}_t| \geq 0\}, p\right).$$

Moreover, given $|\mathscr{U}_t|$ and $|\mathscr{E}_t|$ let $\boldsymbol{Y}_{t+1} \stackrel{d}{=} \mathrm{Bin}\left(2n - |\mathscr{U}_t| \mathbf{1}\{|\mathscr{E}_t| \geq 0\}, p\right)$ be independent of $\boldsymbol{X}_{t+1}$ and everything else, and set $\boldsymbol{X}^{\geq}_{t+1} = \boldsymbol{X}_{t+1} + \boldsymbol{Y}_{t+1}$. Then $(\boldsymbol{X}^{\geq}_t)_{t \geq 1}$ is an i.i.d. sequence of $\mathrm{Bin}(2n, p)$ random variables such that $\boldsymbol{X}^{\geq}_t \geq \boldsymbol{X}_t$ for all t. Hence, for any $T \geq 1$,

$$\mathbb{P}[|\mathscr{F}_\infty| > T] = \mathbb{P}[|\mathscr{E}_T| > 0] \leq \mathbb{P}\left[\sum_{t=1}^{T} \boldsymbol{X}_t > T - 2\ell\right] \leq \mathbb{P}\left[\sum_{t=1}^{T} \boldsymbol{X}^{\geq}_t > T - 2\ell\right] = \mathbb{P}\left[\mathrm{Bin}(2nT, p) > T - 2\ell\right]. \tag{6.5}$$

Further, the Chernoff bound shows that for $T > 8\ell/(2-d)$ (and n large enough),

$$\mathbb{P}\left[\mathrm{Bin}(2Tdn, p) > T - 2\ell\right] \leq \exp\left(-\min\left\{(d - n^{-4/3}), (d - n^{-4/3})^2\right\} \frac{2nTp}{3}\right) \leq \exp\left(-\frac{dT}{36}\right), \tag{6.6}$$

Combining (6.5) and (6.6) completes the proof. □

Let $\mathbf{\Phi}^*$ be the sub-formula of $\mathbf{\Phi}^\dagger$ comprising all variables x such that $x \in \mathscr{F}_\infty$ or $\neg x \in \mathscr{F}_\infty$ along with all clauses a that contain two such variables. Let $\boldsymbol{n}^*$ be the number of variables of $\mathbf{\Phi}^*$ and let $\boldsymbol{m}^*$ be the number of clauses.

Claim 6.9. *We have $\mathbb{P}[\boldsymbol{m}^* \geq \boldsymbol{n}^* - \ell + 1] \leq O(\ell^2/n)$ and $\mathbb{P}[\boldsymbol{m}^* > \boldsymbol{n}^* - \ell + 1] \leq O(\ell^4/n^2)$.*

Proof. We set up a graph representing the literals involved in the exploration process and the clauses that contain such literals. Specifically, let $\neg\mathscr{F}_\infty = \{\neg l : l \in \mathscr{F}_\infty\}$ contain all negations of literals in $\mathscr{F}_\infty$. Moreover, let $\mathscr{G}$ be the graph whose vertices are the literals $\mathscr{F}_\infty \cup \neg\mathscr{F}_\infty$ as well as all clauses a of $\mathbf{\Phi}^\dagger$ that consist of two literals from $\mathscr{F}_\infty \cup \neg\mathscr{F}_\infty$. Let $\mathscr{C}_\infty$ be the set of such clauses a. For each clause $a \in \mathscr{C}_\infty$ the graph $\mathscr{G}$ contains two edges joining a and its two constituent literals. The graph $G(\mathbf{\Phi}^*)$ that we are ultimately interested in results from $\mathscr{G}$ by contracting pairs of inverse literals $l, \neg l \in \mathscr{F}_\infty \cup \neg\mathscr{F}_\infty$.

A large excess $\boldsymbol{m}^* - \boldsymbol{n}^*$ can either caused by the presence of atypically many clauses in $\mathscr{G}$ or by excess pairs of inverse litetals that get contracted. We first address the gain in clauses due to inclusion of $\neg\mathscr{F}_\infty$ and all induced clauses. The exploration process discovers each literal $\lambda \in \mathscr{F}_\infty \setminus \{x_1, \neg x_1, \ldots, x_\ell, \neg x_\ell\}$ via a clause $\neg l_t \vee \lambda$, where $\neg l_t \in \mathscr{E}_{t-1}$. Thus, $|\mathscr{C}_\infty| \geq |\mathscr{F}_\infty| - 2\ell$. Hence, the random variable $\boldsymbol{X} = |\mathscr{C}_\infty| - |\mathscr{F}_\infty| + 2\ell$ accounts for the number of excess clauses that are present among the literals $\mathscr{F}_\infty \cup \neg\mathscr{F}_\infty$ but that were not probed by the process. We highlight that $\boldsymbol{X}$ also counts clauses that contain two literals from the seed set $\{x_1, \neg x_1, \ldots, x_\ell, \neg x_\ell\}$. Because clauses appear in $\mathbf{\Phi}^\dagger$ independently with probability $p = O(d/n)$, we obtain the bounds

$$\mathbb{P}[\boldsymbol{X} \geq 1 \,|\, |\mathscr{F}_\infty|] \leq O(|\mathscr{F}_\infty|^2/n), \qquad \mathbb{P}[\boldsymbol{X} \geq 2 \,|\, |\mathscr{F}_\infty|] \leq O(|\mathscr{F}_\infty|^4/n^2). \tag{6.7}$$

Secondly, we investigate the loss in nodes due to contraction. Hence, $\boldsymbol{n}^* = |\mathscr{F}_\infty \cup \neg\mathscr{F}_\infty|/2$. By construction, the seeds $x_1, \neg x_1, \ldots, x_\ell, \neg x_\ell$ come in pairs. Let $\boldsymbol{X}' = \frac{1}{2}|\mathscr{F}_\infty \cap \neg\mathscr{F}_\infty| - \ell$ count the number of excess inverse literal pairs that we need to contract. Since the process is oblivious to the identities of the variables underlying the literals, given its size the set $\mathscr{F}_\infty \setminus \{x_1, \neg x_1, \ldots, x_\ell, \neg x_\ell\}$ is a uniformly random subset of the set $\{x_i, \neg x_i : \ell < i \leq n\}$ of non-seed literals. Therefore, a routine balls-into-bins argument shows that

$$\mathbb{P}[\boldsymbol{X}' \geq 1 \,|\, |\mathscr{F}_\infty|] \leq O(|\mathscr{F}_\infty|^2/n), \qquad \mathbb{P}[\boldsymbol{X}' \geq 2 \,|\, |\mathscr{F}_\infty|] \leq O(|\mathscr{F}_\infty|^4/n^2). \tag{6.8}$$

Finally, in order to estimate $\boldsymbol{m}^* - \boldsymbol{n}^*$ we consider four separate cases.

Case 1: $\boldsymbol{X} = \boldsymbol{X}' = 0$: Since $\boldsymbol{X} = 0$ the graph $\mathscr{G}$ is a forest with 2ℓ components rooted at $x_1, \neg x_1, \ldots, x_\ell, \neg x_\ell$. Moreover, since $\boldsymbol{X}' = 0$ we have $\mathscr{F}_\infty \cap \neg\mathscr{F}_\infty = \{x_1, \neg x_1, \ldots, x_\ell, \neg x_\ell\}$. Therefore, $G(\boldsymbol{\Phi}^*)$ is obtained from $\mathscr{G}$ by identifying the pairs $x_i, \neg x_i$ for $i = 1, \ldots, \ell$. Hence, $G(\boldsymbol{\Phi}^*)$ is a forest with ℓ components, and thus

$$\boldsymbol{m}^* = \boldsymbol{n}^* - \ell. \tag{6.9}$$

Case 2: $\boldsymbol{X} = 1$, $\boldsymbol{X}' = 0$: Obtain $\hat{\mathscr{G}}$ from $\mathscr{G}$ by adding one new vertex r whose neighbours are $x_1, \neg x_1, \ldots, x_\ell, \neg x_\ell$. Then $\hat{\mathscr{G}}$ is unicyclic because $\boldsymbol{X} = 1$. Let $\tilde{\mathscr{G}}$ be the graph obtained from $\hat{\mathscr{G}}$ by deleting the vertex r along with one (arbitrary) clause a from the cycle of $\hat{\mathscr{G}}$. Then $\tilde{\mathscr{G}}$ is a forest with 2ℓ components. Therefore, by the same token as in Case 1, $G(\boldsymbol{\Phi}^* - a)$ is a forest with ℓ components. Hence, $G(\boldsymbol{\Phi}^*)$, obtained by inserting clause a into $G(\boldsymbol{\Phi}^* - a)$, either contains a single cycle or consists of exactly $\ell - 1$ components. Thus, by (6.7)

$$\boldsymbol{m}^* \le \boldsymbol{n}^* - \ell + 1, \qquad \mathbb{P}\left[\boldsymbol{X} = 1, \boldsymbol{X}' = 0 \,|\, |\mathscr{F}_\infty|\right] = O(|\mathscr{F}_\infty|^2/n). \tag{6.10}$$

Case 3: $\boldsymbol{X} = 0$, $\boldsymbol{X}' = 1$: The graph $\hat{\mathscr{G}}$, defined as in Case 2, is a tree because $\boldsymbol{X} = 0$. Suppose $(\mathscr{F}_\infty \cap \neg\mathscr{F}_\infty) \setminus \{x_1, \neg x_1, \ldots, x_\ell, \neg x_\ell\} = \{y, \neg y\}$. Let a be a clause on the unique path from y to $\neg y$ in $\hat{\mathscr{G}}$. Then the same argument as in Case 1 shows that $G(\boldsymbol{\Phi}^* - a)$ is a forest with ℓ components. Therefore, $G(\boldsymbol{\Phi}^*)$ either contains a unique cycle or has precisely $\ell - 1$ components. Consequently, (6.8) yields

$$\boldsymbol{m}^* \le \boldsymbol{n}^* - \ell + 1, \qquad \mathbb{P}\left[\boldsymbol{X} = 0, \boldsymbol{X}' = 1 \,|\, |\mathscr{F}_\infty|\right] = O(|\mathscr{F}_\infty|^2/n). \tag{6.11}$$

Case 4: $\boldsymbol{X} + \boldsymbol{X}' \ge 2$: In this case we do not have a bound on $\boldsymbol{m}^* - \boldsymbol{n}^*$, but we claim that

$$\mathbb{P}\left[\boldsymbol{X} + \boldsymbol{X}' \ge 2 \,|\, |\mathscr{F}_\infty|\right] = O(|\mathscr{F}_\infty|^4/n^2). \tag{6.12}$$

Indeed, (6.7) and (6.8) readily imply that $\mathbb{P}\left[\boldsymbol{X} \vee \boldsymbol{X}' \ge 2 \,|\, |\mathscr{F}_\infty|\right] = O(|\mathscr{F}_\infty|^4/n^2)$. Further, since $\boldsymbol{X}$ is independent of $\boldsymbol{X}'$ given $\mathscr{F}_\infty$, (6.7) and (6.8) also yield the bound $\mathbb{P}\left[\boldsymbol{X} = \boldsymbol{X}' = 1 \,|\, |\mathscr{F}_\infty|\right] = O(|\mathscr{F}_\infty|^4/n^2)$.

The assertion follows by combining (6.9)–(6.12) with Claim 6.8. □

Claim 6.10. *For all $\chi \in \{\pm 1\}^{\{x_1, \ldots, x_\ell\}}$ we have $\boldsymbol{A}_\chi^\dagger \le |\mathscr{F}_\infty| \mathbf{1}\{\boldsymbol{m}^* \le \boldsymbol{n}^* - \ell + 1\} + n\mathbf{1}\{\boldsymbol{m}^* > \boldsymbol{n}^* - \ell + 1\}$.*

Proof. The graph $G(\boldsymbol{\Phi}^*)$ consists of at most ℓ components (one for each of the initial variables $x_1, \ldots, x_\ell$). Hence, $\boldsymbol{m}^* \ge \boldsymbol{n}^* - \ell$ and $G(\boldsymbol{\Phi}^*)$ is acyclic if $\boldsymbol{m}^* = \boldsymbol{n}^* - \ell$. Moreover, if $G(\boldsymbol{\Phi}^*)$ is acyclic then $\boldsymbol{A}_\chi^\dagger \le |\mathscr{F}_\infty|$ by construction.

Thus, we are left to consider the case $\boldsymbol{m}^* = \boldsymbol{n}^* - \ell + 1$. Then $\boldsymbol{\Phi}^*$ contains a clause a such that $G(\boldsymbol{\Phi}^* - a)$ is a forest with ℓ components rooted at $x_1, \ldots, x_\ell$. Assume without loss that $a = x_{n-1} \vee x_n$. Then by construction we have $\{x_{n-1}, \neg x_{n-1}\} \cap \mathscr{F}_\infty \neq \emptyset$ and $\{x_n, \neg x_n\} \cap \mathscr{F}_\infty \neq \emptyset$. Further, unless $\neg x_{n-1}, \neg x_n \in \mathscr{F}_\infty$ we have $\boldsymbol{A}_\chi \le \boldsymbol{I}_\chi \le |\mathscr{F}_\infty|$ as in the first case. Hence, assume that $\neg x_{n-1}, \neg x_n \in \mathscr{F}_\infty$. Let $i \in [\ell]$ be such that x_n belongs to the connected component of x_i in $G(\boldsymbol{\Phi}^* - a)$ and obtain χ' from χ by flipping the value assigned to x_i. Because $G(\boldsymbol{\Phi}^* - a)$ is a forest, we conclude that $\boldsymbol{A}_\chi^\dagger \le \boldsymbol{I}_\chi \wedge \boldsymbol{I}_{\chi'} \le |\mathscr{F}_\infty|$. □

Proof of Lemma 6.6. The choice of the clause probability p ensures that $\boldsymbol{A}_\chi^\dagger$ stochastically dominates $\boldsymbol{A}_\chi$. Therefore, the assertion follows from Claims 6.8–6.10. □

Proof of Lemma 6.5. The choice of the clause probability p and the construction of the set $\mathscr{F}_\infty$ guarantee that $\boldsymbol{I}_\chi$ is stochastically dominated by the random variable $|\mathscr{F}_\infty| \mathbf{1}\{\boldsymbol{m}^* \le \boldsymbol{n}^* - \ell\} + n\mathbf{1}\{\boldsymbol{m}^* > \boldsymbol{n}^* - \ell\}$. Hence, Claims 6.8–6.10 imply the desired bound. □

Proof of Corollary 6.7. Let $Y = \{x_1\}$ and $\chi^+_{x_1} = 1$, $\chi^-_{x_1} = -1$. Assume that $\boldsymbol{\Phi}'$ is satisfiable. Then Fact 6.4 implies that

$$n \wedge \left|\log \frac{\mu_{\boldsymbol{\Phi}'}(\boldsymbol{\sigma}_{x_1} = 1)}{\mu_{\boldsymbol{\Phi}'}(\boldsymbol{\sigma}_{x_1} = -1)}\right| \le \boldsymbol{I}_{\chi^-} + \boldsymbol{I}_{\chi^+}.$$

Therefore, the assertion follows from Lemma 6.5. □

6.3. **Proof of Proposition 6.2.** Let $c_1, \ldots, c_{\boldsymbol{\Delta}''}$ be the new clauses added to $\boldsymbol{\Phi}''$ and let $\boldsymbol{Y} = \{\boldsymbol{y}_1, \boldsymbol{z}_1, \ldots, \boldsymbol{y}_{\boldsymbol{\Delta}''}, \boldsymbol{z}_{\boldsymbol{\Delta}''}\}$ be the set of variables that occur in these clauses. We begin by deriving the following rough bound.

Lemma 6.11. *We have* $\mathbb{E}\left[\log^2 \frac{Z(\boldsymbol{\Phi}'') \vee 1}{Z(\boldsymbol{\Phi}') \vee 1}\right] = O(1)$.

Proof. If $\mathbf{\Phi}'$ is unsatisfiable then so is $\mathbf{\Phi}''$ and thus $(Z(\mathbf{\Phi}'')\vee 1)/(Z(\mathbf{\Phi}')\vee 1)=1$. Hence, we may assume that $Z(\mathbf{\Phi}')\geq 1$. If $|\boldsymbol{Y}|=2\boldsymbol{\Delta}''$, the new clauses attach to disjoint sets of variables. Consider the truth value assignment $\boldsymbol{\chi}\in\{\pm 1\}^{\boldsymbol{Y}}$ that satisfies both literals in each of the clauses $c_1,\ldots,c_{\boldsymbol{\Delta}''}$. Fact 6.4 shows that

$$Z(\mathbf{\Phi}'')\vee 1\geq Z(\mathbf{\Phi}',\boldsymbol{\chi})\vee 1\geq 2^{-\boldsymbol{A}_{\boldsymbol{\chi}}}Z(\mathbf{\Phi}'). \tag{6.13}$$

Combining (6.13) with Lemma 6.6 and recalling that $\boldsymbol{\Delta}''\stackrel{\mathrm{d}}{=}\mathrm{Po}(d/2)$, we obtain

$$\mathbb{E}\left[\mathbf{1}\left\{Z(\mathbf{\Phi}')\geq 1,\ |\boldsymbol{Y}|=2\boldsymbol{\Delta}''\right\}\log^2\frac{Z(\mathbf{\Phi}'')\vee 1}{Z(\mathbf{\Phi}')\vee 1}\right]\leq\mathbb{E}\left[\mathbf{1}\left\{Z(\mathbf{\Phi}')\geq 1,\ |\boldsymbol{Y}|=2\boldsymbol{\Delta}''\right\}\boldsymbol{A}_{\boldsymbol{\chi}}^2\right]=O(1). \tag{6.14}$$

Next, consider the event $|\boldsymbol{Y}|=2\boldsymbol{\Delta}''-1$. Because $c_1,\ldots,c_{\boldsymbol{\Delta}''}$ are drawn independently, we have

$$\mathbb{P}\left[|\boldsymbol{Y}|=2\boldsymbol{\Delta}''-1\mid\boldsymbol{\Delta}''\right]\leq O((\boldsymbol{\Delta}'')^2/n). \tag{6.15}$$

Moreover, because the signs of the clauses $c_1,\ldots,c_{\boldsymbol{\Delta}''}$ are independent of $\mathbf{\Phi}'$, given $|\boldsymbol{Y}|=2\boldsymbol{\Delta}-1$ there exists an assignment $\boldsymbol{\chi}\in\{\pm 1\}^{\boldsymbol{Y}}$, stochastically independent of $\mathbf{\Phi}'$, that satisfies $c_1,\ldots,c_{\boldsymbol{\Delta}''}$. Fact 6.4 yields $Z(\mathbf{\Phi}'')\vee 1\geq Z(\mathbf{\Phi}',\boldsymbol{\chi})\geq 2^{-\boldsymbol{I}_{\boldsymbol{\chi}}}Z(\mathbf{\Phi}')$. Therefore, since $\log((Z(\mathbf{\Phi}'')\vee 1)/(Z(\mathbf{\Phi}')\vee 1))\leq n$, Lemma 6.5 and (6.15) imply

$$\mathbb{E}\left[\mathbf{1}\left\{Z(\mathbf{\Phi}')\geq 1,\ |\boldsymbol{Y}|=2\boldsymbol{\Delta}''-1\right\}\log^2\frac{Z(\mathbf{\Phi}'')\vee 1}{Z(\mathbf{\Phi}')\vee 1}\right]\leq n\mathbb{E}\left[\mathbf{1}\left\{|\boldsymbol{Y}|=2\boldsymbol{\Delta}''-1\right\}\boldsymbol{I}_{\boldsymbol{\chi}}\right]=O(1). \tag{6.16}$$

Finally, consider the event $|\boldsymbol{Y}|<2\boldsymbol{\Delta}''-1$. Due to the independence of $c_1,\ldots,c_{\boldsymbol{\Delta}''}$, this event occurs with probability $O(n^{-2})$. Hence, the deterministic bound $(Z(\mathbf{\Phi}'')\vee 1)/(Z(\mathbf{\Phi}')\vee 1)\geq 2^{-n}$ implies

$$\mathbb{E}\left[\mathbf{1}\left\{Z(\mathbf{\Phi}')\geq 1,\ |\boldsymbol{Y}|<2\boldsymbol{\Delta}''-1\right\}\log^2\frac{Z(\mathbf{\Phi}'')\vee 1}{Z(\mathbf{\Phi}')\vee 1}\right]=O(1). \tag{6.17}$$

The assertion follows from (6.14), (6.16) and (6.17). □

Lemma 6.12. *There exists a number $K>0$ such that for every $\varepsilon>0$ we have*

$$\limsup_{n\to\infty}\mathbb{E}\left[\left(\sum_{i=1}^{\boldsymbol{\Delta}''}\Lambda_\varepsilon(1-\mu_{\mathbf{\Phi}'}(\boldsymbol{\sigma}_{\boldsymbol{y}_i}=-\operatorname{sign}(c_i,\boldsymbol{y}_i))\mu_{\mathbf{\Phi}'}(\boldsymbol{\sigma}_{\boldsymbol{z}_i}=-\operatorname{sign}(c_i,\boldsymbol{z}_i))\right)^2\mid Z(\mathbf{\Phi}')>0\right]\leq K.$$

Proof. Since $\boldsymbol{\Delta}"\stackrel{\mathrm{d}}{=}\mathrm{Po}(d/2)$ and the pair $(\boldsymbol{y}_1,\boldsymbol{z}_1)$ is uniformly random, due to Cauchy-Schwarz it suffices to prove $\limsup_{n\to\infty}\mathbb{E}\left[\Lambda_\varepsilon(1-\mu_{\mathbf{\Phi}'}(\boldsymbol{\sigma}_{x_1}=1)\mu_{\mathbf{\Phi}'}(\boldsymbol{\sigma}_{x_2}=1))^2\mid Z(\mathbf{\Phi}')>0\right]\leq K$ for every $\varepsilon>0$. We observe that

$$\begin{aligned}\limsup_{n\to\infty}\mathbb{E}\left[\Lambda_\varepsilon(1-\mu_{\mathbf{\Phi}'}(\boldsymbol{\sigma}_{x_1}=1)\mu_{\mathbf{\Phi}'}(\boldsymbol{\sigma}_{x_2}=1))^2\mid Z(\mathbf{\Phi}')>0\right]&\leq\limsup_{n\to\infty}\mathbb{E}\left[\Lambda_\varepsilon(1-\mu_{\mathbf{\Phi}'}(\boldsymbol{\sigma}_{x_1}=1))^2\mid Z(\mathbf{\Phi}')>0\right]\\&=\limsup_{n\to\infty}\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^n\Lambda_\varepsilon(1-\mu_{\mathbf{\Phi}'}(\boldsymbol{\sigma}_{x_i}=1))^2\mid Z(\mathbf{\Phi}')>0\right].\end{aligned} \tag{6.18}$$

Moreover, $\mathbf{\Phi}'$ has $\boldsymbol{m}'\stackrel{\mathrm{d}}{=}\mathrm{Po}(dn/2-d/2)$ clauses, while $\mathbf{\Phi}=\mathbf{\Phi}_n$ has $\boldsymbol{m}\stackrel{\mathrm{d}}{=}\mathrm{Po}(dn/2)$ clauses. Since $d_{\mathrm{TV}}(\boldsymbol{m}',\boldsymbol{m})=o(1)$, the formulas $\mathbf{\Phi}'$, $\mathbf{\Phi}$ can be coupled such that both coincide w.h.p. Hence, for any fixed $\varepsilon>0$ we have

$$\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^n\Lambda_\varepsilon(1-\mu_{\mathbf{\Phi}'}(\boldsymbol{\sigma}_{x_i}=1))^2\mid Z(\mathbf{\Phi}')\right]=\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^n\Lambda_\varepsilon(1-\mu_{\mathbf{\Phi}}(\boldsymbol{\sigma}_{x_i}=1))^2\mid Z(\mathbf{\Phi}')\right]+o(1). \tag{6.19}$$

Further, since for every $\varepsilon>0$ the function $u\in[0,1]\mapsto\Lambda_\varepsilon(1-u)^2$ is continuous, Corollary 1.3 implies that

$$\frac{1}{n}\sum_{i=1}^n\Lambda_\varepsilon(1-\mu_{\mathbf{\Phi}}(\boldsymbol{\sigma}_{x_i}=1))^2\quad\stackrel{n\to\infty}{\longrightarrow}\quad\mathbb{E}\left[\Lambda_\varepsilon(1-\boldsymbol{\mu}_{\pi_d})^2\right]\qquad\text{in probability.} \tag{6.20}$$

Since (2.1) shows that $\mathbb{E}\left[\Lambda_\varepsilon(1-\boldsymbol{\mu}_{\pi_d})^2\right]\leq\mathbb{E}\left[\log^2(1-\boldsymbol{\mu}_{\pi_d})\right]<\infty$, the assertion follows from (6.18)–(6.20). □

Lemma 6.13. *For any $\delta>0$ there exists $\varepsilon>0$ such that*

$$\limsup_{n\to\infty}\left|\mathbb{E}\left[\log\frac{Z(\mathbf{\Phi}'')\vee 1}{Z(\mathbf{\Phi}')\vee 1}\right]-\frac{d}{2}\mathbb{E}\left[\Lambda_\varepsilon\left(1-\mu_{\mathbf{\Phi}',x_1}(\boldsymbol{s}_1)\mu_{\mathbf{\Phi}',x_2}(\boldsymbol{s}_2)\right)\mid Z(\mathbf{\Phi}')\right]\right|<\delta.$$

Proof. Choose small enough $\xi=\xi(\delta)>\eta=\eta(\xi)>\varepsilon=\varepsilon(\eta)>0$, assume that $n>n_0(\varepsilon)$ is sufficiently large and let $(\gamma_n)_n$ be a sequence of positive reals, depending on ξ and η, that tends to zero sufficiently slowly. Let $\mathscr{E}=\mathscr{E}_n$ be the event that the following five statements hold.

E1: $Z(\mathbf{\Phi}')>0$.

E2: $|\boldsymbol{Y}| = 2\boldsymbol{\Delta}''$.
E3: $\boldsymbol{\Delta}'' < \xi^{-1/4}$.
E4: for all $y \in \boldsymbol{Y}$ and all $s \in \{\pm 1\}$ we have $\mu_{\boldsymbol{\Phi}'}(\boldsymbol{\sigma}_y = s) < 1 - 2\eta$.
E5: $\sum_{\sigma \in \{\pm 1\}^{\boldsymbol{Y}}} \left|\mu_{\boldsymbol{\Phi}}(\forall y \in \boldsymbol{Y} : \boldsymbol{\sigma}_y = \sigma_y) - \prod_{y \in \boldsymbol{Y}} \mu_{\boldsymbol{\Phi}}(\boldsymbol{\sigma}_y = \sigma_y)\right| < \gamma_n$.

The first two events **E1**, **E2** occur with probability $1 - o(1)$ as $n \to \infty$. Moreover, $\mathbb{P}[\mathbf{E3}] > 1 - \xi$ if ξ is small enough. Further, since Corollary 1.3 shows that $\pi_{\boldsymbol{\Phi}}$ converges to π_d weakly in probability, the tail bound (2.1) implies that $\mathbb{P}\left[\mathbf{E4} \mid \boldsymbol{\Delta}'' < \xi^{-1/4}\right] > 1 - \xi$, provided that η is small enough. Additionally, Corollary 2.3 implies $\mathbb{P}[\mathbf{E5} \mid \mathbf{E1}\text{–}\mathbf{E4}] = 1 - o(1)$ if $\gamma_n \to 0$ slowly enough. Consequently,

$$\mathbb{P}[\mathscr{E}] > 1 - 4\xi. \tag{6.21}$$

Combining Lemma 6.11, (6.21) and the Cauchy-Schwarz inequality, we obtain

$$\left|\mathbb{E}\left[(1 - \mathbf{1}\mathscr{E})\log\frac{Z(\boldsymbol{\Phi}'')}{Z(\boldsymbol{\Phi}')}\right]\right| \le \delta/3 + o(1). \tag{6.22}$$

Similarly, by Lemma 6.12, (6.21) and Cauchy-Schwarz,

$$\left|\mathbb{E}\left[(1 - \mathbf{1}\mathscr{E})\sum_{i=1}^{\boldsymbol{\Delta}''} \Lambda_\varepsilon\left(1 - \mu_{\boldsymbol{\Phi}'}(\boldsymbol{\sigma}_{\boldsymbol{y}_i} = -\operatorname{sign}(\boldsymbol{y}_i, c_i))\mu_{\boldsymbol{\Phi}'}(\boldsymbol{\sigma}_{\boldsymbol{z}_i} = -\operatorname{sign}(\boldsymbol{z}_i, c_i))\right)\right]\right| \le \delta/3 + o(1). \tag{6.23}$$

Further, because the distribution of $\boldsymbol{\Phi}'$ is invariant under permutations of the variables $x_1, \ldots, x_n$ and $\mathbb{E}[\boldsymbol{\Delta}''] = d/2$,

$$\begin{aligned}&\mathbb{E}\left[\sum_{i=1}^{\boldsymbol{\Delta}''} \Lambda_\varepsilon\left(1 - \mu_{\boldsymbol{\Phi}'}(\boldsymbol{\sigma}_{\boldsymbol{y}_i} = -\operatorname{sign}(\boldsymbol{y}_i, c_i))\mu_{\boldsymbol{\Phi}'}(\boldsymbol{\sigma}_{\boldsymbol{z}_i} = -\operatorname{sign}(\boldsymbol{z}_i, c_i))\right) \mid Z(\boldsymbol{\Phi}') > 0\right] \\ &\quad = \frac{d}{2}\mathbb{E}\left[\Lambda_\varepsilon\left(1 - \mu_{\boldsymbol{\Phi}'}(\boldsymbol{\sigma}_{x_1} = \boldsymbol{s}_1)\mu_{\boldsymbol{\Phi}'}(\boldsymbol{\sigma}_{x_2} = \boldsymbol{s}_2)\right) \mid Z(\boldsymbol{\Phi}') > 0\right].\end{aligned} \tag{6.24}$$

Moreover, on the event $\mathscr{E}$ we have

$$\begin{aligned}\frac{Z(\boldsymbol{\Phi}'')}{Z(\boldsymbol{\Phi}')} &= \sum_{\sigma \in \{\pm 1\}^{\boldsymbol{Y}}} \mathbf{1}\left\{\sigma \text{ satisfies } c_1, \ldots, c_{\boldsymbol{\Delta}''}\right\} \mu_{\boldsymbol{\Phi}'}\left(\forall y \in \boldsymbol{Y} : \boldsymbol{\sigma}_y = \sigma_y\right) \\ &= \sum_{\sigma \in \{\pm 1\}^{\boldsymbol{Y}}} \mathbf{1}\left\{\sigma \text{ satisfies } c_1, \ldots, c_{\boldsymbol{\Delta}''}\right\} \prod_{y \in \boldsymbol{Y}} \mu_{\boldsymbol{\Phi}'}\left(\boldsymbol{\sigma}_y = \sigma_y\right) + o(1) \qquad \text{[due to } \mathbf{E3},\mathbf{E5}\text{]} \\ &= \prod_{i=1}^{\boldsymbol{\Delta}''}\left(1 - \mu_{\boldsymbol{\Phi}'}\left(\boldsymbol{\sigma}_{\boldsymbol{y}_i} = -\operatorname{sign}(\boldsymbol{y}_i, c_i)\right)\mu_{\boldsymbol{\Phi}'}\left(\boldsymbol{\sigma}_{\boldsymbol{z}_i} = -\operatorname{sign}(\boldsymbol{z}_i, c_i)\right)\right) + o(1).\end{aligned}$$

Therefore, by **E4**

$$\begin{aligned}\mathbb{E}\left[\mathbf{1}\mathscr{E}\log\frac{Z(\boldsymbol{\Phi}'')}{Z(\boldsymbol{\Phi}')}\right] &= \mathbb{E}\left[\mathbf{1}\mathscr{E}\sum_{i=1}^{\boldsymbol{\Delta}''}\log\left(1 - \mu_{\boldsymbol{\Phi}'}\left(\boldsymbol{\sigma}_{\boldsymbol{y}_i} = -\operatorname{sign}(\boldsymbol{y}_i, c_i)\right)\mu_{\boldsymbol{\Phi}'}\left(\boldsymbol{\sigma}_{\boldsymbol{z}_i} = -\operatorname{sign}(\boldsymbol{z}_i, c_i)\right)\right)\right] + o(1) \\ &= \mathbb{E}\left[\mathbf{1}\mathscr{E}\sum_{i=1}^{\boldsymbol{\Delta}''}\Lambda_\varepsilon\left(1 - \mu_{\boldsymbol{\Phi}'}\left(\boldsymbol{\sigma}_{\boldsymbol{y}_i} = -\operatorname{sign}(\boldsymbol{y}_i, c_i)\right)\mu_{\boldsymbol{\Phi}'}\left(\boldsymbol{\sigma}_{\boldsymbol{z}_i} = -\operatorname{sign}(\boldsymbol{z}_i, c_i)\right)\right)\right] + o(1).\end{aligned} \tag{6.25}$$

Finally, the assertion follows from (6.22)–(6.25). □

Proof of Proposition 6.2. Proposition 2.1 shows that $\boldsymbol{\mu}_{\pi_d,1}$ and $1 - \boldsymbol{\mu}_{\pi_d,1}$ are identically distributed. Since Λ_ε is continuous and bounded, Corollary 1.3 therefore implies that

$$\begin{aligned}\lim_{n\to\infty}\mathbb{E}\left[\Lambda_\varepsilon\left(1 - \mu_{\boldsymbol{\Phi}',x_1}(\boldsymbol{s}_1)\mu_{\boldsymbol{\Phi}',x_2}(\boldsymbol{s}_2)\right)\right] &= \mathbb{E}\left[\Lambda_\varepsilon\left(1 - \left(\frac{1-\boldsymbol{s}_1}{2} + \boldsymbol{s}_1\boldsymbol{\mu}_{\pi_d,1}\right)\left(\frac{1-\boldsymbol{s}_2}{2} + \boldsymbol{s}_2\boldsymbol{\mu}_{\pi_d,2}\right)\right)\right] \\ &= \mathbb{E}\left[\Lambda_\varepsilon\left(1 - \boldsymbol{\mu}_{\pi_d,1}\boldsymbol{\mu}_{\pi_d,2}\right)\right].\end{aligned} \tag{6.26}$$

for every $\varepsilon > 0$. Further, since $\Lambda_\varepsilon(1 - \boldsymbol{\mu}_{\pi_d,1}\boldsymbol{\mu}_{\pi_d,2})$ decreases monotonically to $\log(1 - \boldsymbol{\mu}_{\pi_d,1}\boldsymbol{\mu}_{\pi_d,2})$ as $\varepsilon \to 0$, the monotone convergence theorem and (2.2) yield

$$\lim_{\varepsilon\to 0}\mathbb{E}\left[\Lambda_\varepsilon\left(1 - \boldsymbol{\mu}_{\pi_d,1}\boldsymbol{\mu}_{\pi_d,2}\right)\right] = \mathbb{E}\log\left(1 - \boldsymbol{\mu}_{\pi_d,1}\boldsymbol{\mu}_{\pi_d,2}\right). \tag{6.27}$$

Combining (6.26) and (6.27) and Lemma 6.13 completes the proof. □

6.4. **Proof of Proposition 6.3.** The steps that we follow are analogous to the ones from the proof of Proposition 6.2. Recall that $\mathbf{\Phi}'''$ is obtained from $\mathbf{\Phi}'$ by adding one variable x_{n+1} along with random adjacent clauses $b_1,\ldots,b_{\mathbf{\Delta}'''}$, where $\mathbf{\Delta}'''$ is a Poisson variable with mean d. Let $\boldsymbol{y}_1,\ldots,\boldsymbol{y}_{\mathbf{\Delta}'''} \in \{x_1,\ldots,x_n\}$ be the variables of $\mathbf{\Phi}'$ where the new clauses attach and let $\boldsymbol{Y} = \{\boldsymbol{y}_1,\ldots,\boldsymbol{y}_{\mathbf{\Delta}'''}\}$. We begin with the following L_2-bound.

Lemma 6.14. *We have* $\limsup_{n\to\infty} \mathbb{E}\left[\log^2 \frac{Z(\mathbf{\Phi}''')\vee 1}{Z(\mathbf{\Phi}')\vee 1}\right] < \infty$.

Proof. If $\mathbf{\Phi}'$ is unsatisfiable, then so is $\mathbf{\Phi}'''$ and thus $(Z(\mathbf{\Phi}''')\vee 1)/(Z(\mathbf{\Phi}')\vee 1) = 1$. Hence, we may assume that $Z(\mathbf{\Phi}') \geq 1$. We now consider three scenarios. First, suppose that $|\boldsymbol{Y}| = \mathbf{\Delta}'''$, i.e., the new clauses attach to distinct variables of $\mathbf{\Phi}'$. Then define an assignment $\boldsymbol{\chi} \in \{\pm 1\}^{\boldsymbol{Y}}$ by setting each $y \in \boldsymbol{Y}$ to the value that satisfies the unique clause among $b_1,\ldots,b_{\mathbf{\Delta}'''}$ in which y occurs. We claim that

$$Z(\mathbf{\Phi}''')\vee 1 \geq 2^{-\boldsymbol{A}_{\boldsymbol{\chi}}} Z(\mathbf{\Phi}'). \tag{6.28}$$

Indeed, if $\boldsymbol{\chi}' \in \{\pm 1\}^{\boldsymbol{Y}}$ differs from $\boldsymbol{\chi}$ on only one variable, then we can always satisfy all clauses $b_1,\ldots,b_{\mathbf{\Delta}'''}$ by setting x_{n+1} appropriately. Therefore, (6.28) follows from Fact 6.4 and the definition (6.3) of $\boldsymbol{A}_{\boldsymbol{\chi}}$. Combining (6.28) with Lemma 6.6, we obtain

$$\mathbb{E}\left[\mathbf{1}\left\{Z(\mathbf{\Phi}')\geq 1,\ |\boldsymbol{Y}| = \mathbf{\Delta}'''\right\}\log^2\frac{Z(\mathbf{\Phi}''')\vee 1}{Z(\mathbf{\Phi}')\vee 1}\right] \leq \mathbb{E}\left[\mathbf{1}\left\{|\boldsymbol{Y}| = \mathbf{\Delta}'''\right\}\boldsymbol{A}_{\boldsymbol{\chi}}^2\right] = O(1). \tag{6.29}$$

Second, consider the case $|\boldsymbol{Y}| = \mathbf{\Delta}''' - 1$. Because $b_1,\ldots,b_{\mathbf{\Delta}'''}$ are drawn independently, we have

$$\mathbb{P}\left[|\boldsymbol{Y}| = \mathbf{\Delta}''' - 1 \mid \mathbf{\Delta}'''\right] = O((\mathbf{\Delta}''')^2/n). \tag{6.30}$$

Further, there exists an assignment $\boldsymbol{\chi} \in \{\pm 1\}^{\boldsymbol{Y}}$ under which all but one of the clauses $b_1,\ldots,b_{\mathbf{\Delta}'''}$ are satisfied. This assignment is independent of $\mathbf{\Phi}'$ because the signs of $b_1,\ldots,b_{\mathbf{\Delta}'''}$ are. Since we can use the new variable x_{n+1} to satisfy the last clause as well, Fact 6.4 implies the bound $(Z(\mathbf{\Phi}''')\vee 1)/Z(\mathbf{\Phi}') \geq 2^{-\boldsymbol{I}_{\boldsymbol{\chi}}}$. Therefore, Lemma 6.5 and (6.30) yield

$$\begin{aligned}\mathbb{E}\left[\mathbf{1}\left\{Z(\mathbf{\Phi}')\geq 1,\ |\boldsymbol{Y}| = \mathbf{\Delta}''' - 1\right\}\log^2\frac{Z(\mathbf{\Phi}'')\vee 1}{Z(\mathbf{\Phi}')\vee 1}\right] &\leq \mathbb{E}\left[\mathbf{1}\left\{Z(\mathbf{\Phi}')\geq 1,\ |\boldsymbol{Y}| = \mathbf{\Delta}''' - 1\right\}\boldsymbol{I}_{\boldsymbol{\chi}}^2\right]\\ &\leq n\mathbb{E}\left[\mathbf{1}\left\{|\boldsymbol{Y}| = \mathbf{\Delta}''' - 1\right\}\boldsymbol{I}_{\boldsymbol{\chi}}\right] = O(1).\end{aligned} \tag{6.31}$$

Finally, because $b_1,\ldots,b_{\mathbf{\Delta}'''}$ are drawn independently, the event $\{|\boldsymbol{Y}| < \mathbf{\Delta}''' - 1\}$ has probability $O(n^{-2})$. Therefore, the deterministic bound $(Z(\mathbf{\Phi}''')\vee 1)/(Z(\mathbf{\Phi}')\vee 1) \geq 2^{-n}$ ensures that

$$\mathbb{E}\left[\mathbf{1}\left\{Z(\mathbf{\Phi}')\geq 1,\ |\boldsymbol{Y}| < \mathbf{\Delta}''' - 1\right\}\log^2\frac{Z(\mathbf{\Phi}''')\vee 1}{Z(\mathbf{\Phi}')\vee 1}\right] = O(1). \tag{6.32}$$

The assertion follows from (6.29), (6.31) and (6.32). □

Lemma 6.15. *There exists $K > 0$ such that for every $\varepsilon > 0$ we have*

$$\limsup_{n\to\infty}\mathbb{E}\left[\Lambda_\varepsilon\left(\sum_{s\in\{\pm 1\}}\prod_{i=1}^{\mathbf{\Delta}'''}\left(1-\mathbf{1}\left\{s\neq\operatorname{sign}(x_{n+1},b_i)\right\}\mu_{\mathbf{\Phi}'}(\boldsymbol{\sigma}_{\boldsymbol{y}_i} = -\operatorname{sign}(\boldsymbol{y}_i,b_i))\right)\right)^2 \mid Z(\mathbf{\Phi}')>0\right] \leq K.$$

Proof. Since $\mathbf{\Delta}''' \stackrel{\mathrm{d}}{=} \mathrm{Po}(d/2)$, $\boldsymbol{y}_1,\ldots,\boldsymbol{y}_{\mathbf{\Delta}'''}$ and the signs $\operatorname{sign}(b_i,\boldsymbol{y}_i)$ are uniformly random, we obtain

$$\begin{aligned}&\mathbb{E}\left[\Lambda_\varepsilon\left(\sum_{s\in\{\pm 1\}}\prod_{i=1}^{\mathbf{\Delta}'''}\left(1-\mathbf{1}\left\{s\neq\operatorname{sign}(x_{n+1},b_i)\right\}\mu_{\mathbf{\Phi}'}(\boldsymbol{\sigma}_{\boldsymbol{y}_i} = -\operatorname{sign}(\boldsymbol{y}_i,b_i))\right)\right)^2 \mid Z(\mathbf{\Phi}')>0\right]\\ &\leq 1+\mathbb{E}\left[\Lambda_\varepsilon\left(\prod_{i=1}^{\mathbf{\Delta}'''}\mu_{\mathbf{\Phi}'}(\boldsymbol{\sigma}_{\boldsymbol{y}_i}=1)\right)^2 \mid Z(\mathbf{\Phi}')>0\right] \leq 1+d\mathbb{E}\left[\Lambda_\varepsilon\left(\mu_{\mathbf{\Phi}'}(\boldsymbol{\sigma}_{\boldsymbol{y}_1}=1)\right)^2 \mid Z(\mathbf{\Phi}')>0\right].\end{aligned} \tag{6.33}$$

Further, the formulas $\mathbf{\Phi}'$, $\mathbf{\Phi}$ can be coupled such that both coincide w.h.p. (cf. the proof of Lemma 6.12). Therefore, Corollary 1.3 implies that for every $\varepsilon > 0$,

$$\mathbb{E}\left[\Lambda_\varepsilon\left(\mu_{\mathbf{\Phi}'}(\boldsymbol{\sigma}_{\boldsymbol{y}_1}=1)\right)^2 \mid Z(\mathbf{\Phi}')>0\right] = \mathbb{E}\left[\Lambda_\varepsilon\left(\boldsymbol{\mu}_{\pi_{\mathbf{\Phi}}}\right)^2 \mid Z(\mathbf{\Phi})>0\right] + o(1) = \mathbb{E}\left[\Lambda_\varepsilon\left(\boldsymbol{\mu}_{\pi_d}^2\right)\right] + o(1) \leq \mathbb{E}\left[\log^2\boldsymbol{\mu}_{\pi_d}\right] + o(1). \tag{6.34}$$

Since (2.1) implies that $\mathbb{E}\left[\log^2\boldsymbol{\mu}_{\pi_d}\right] < \infty$, the assertion follows from (6.33)–(6.34). □

Lemma 6.16. *For any $\delta > 0$ there exists $\varepsilon_0 > 0$ such that for every $0 < \varepsilon < \varepsilon_0$,*

$$\left|\mathbb{E}\left[\log\frac{Z(\mathbf{\Phi}''')\vee 1}{Z(\mathbf{\Phi}')\vee 1}\right] - \mathbb{E}\left[\Lambda_\varepsilon\left(\sum_{s\in\{\pm1\}}\prod_{i=1}^{\boldsymbol{d}}\left(1-\mathbf{1}\{s\neq \boldsymbol{s}_i\}\mu_{\mathbf{\Phi}'}(\boldsymbol{\sigma}_{x_i}=\boldsymbol{s}_i')\right)\right)\mid Z(\mathbf{\Phi}')>0\right]\right| < \delta + o(1).$$

Proof. Choose small enough $\xi = \xi(\delta) > \eta = \eta(\xi) > \varepsilon = \varepsilon(\eta) > 0$, assume that $n > n_0(\varepsilon)$ is sufficiently large and let $(\gamma_n)_n$ be a sequence of numbers $\gamma_n > 0$ that tends to zero slowly. Let $\mathscr{E} = \mathscr{E}_n$ be the event that the following five statements are satisfied.

E1: $Z(\mathbf{\Phi}') > 0$.
E2: $|\boldsymbol{Y}| = \boldsymbol{\Delta}'''$.
E3: $\boldsymbol{\Delta}''' < \xi^{-1/4}$.
E4: for all $y \in \boldsymbol{Y}$ we have $\mu_{\mathbf{\Phi}'}(\boldsymbol{\sigma}_y = 1) \vee \mu_{\mathbf{\Phi}'}(\boldsymbol{\sigma}_y = -1) < 1 - 2\eta$.
E5: $\sum_{\sigma\in\{\pm1\}^Y}\left|\mu_{\mathbf{\Phi}}(\forall y\in\boldsymbol{Y}:\boldsymbol{\sigma}_y=\sigma_y) - \prod_{y\in\boldsymbol{Y}}\mu_{\mathbf{\Phi}}(\boldsymbol{\sigma}_y=\sigma_y)\right| < \gamma_n$.

As in the proof of Lemma 6.13 we obtain $\mathbb{P}[\mathscr{E}] > 1 - 4\xi$. Hence, Lemmas 6.14 and 6.15 and the Cauchy-Schwarz inequality yield

$$\left|\mathbb{E}\left[(1-\mathbf{1}_{\mathscr{E}})\log\frac{Z(\mathbf{\Phi}''')}{Z(\mathbf{\Phi}')}\right]\right| \le \delta/3 + o(1), \tag{6.35}$$

$$\left|\mathbb{E}\left[(1-\mathbf{1}_{\mathscr{E}})\Lambda_\varepsilon\left(\sum_{s\in\{\pm1\}}\prod_{i=1}^{\boldsymbol{\Delta}'''}\left(1-\mathbf{1}\left\{s\neq\operatorname{sign}(x_{n+1},b_i)\right\}\mu_{\mathbf{\Phi}'}(\boldsymbol{\sigma}_{\boldsymbol{y}_i}=-\operatorname{sign}(\boldsymbol{y}_i,b_i))\right)\right)\mid Z(\mathbf{\Phi}')>0\right]\right| \le \delta/3 + o(1). \tag{6.36}$$

Moreover, because the distribution of $\mathbf{\Phi}'$ is invariant under variable permutations,

$$\begin{aligned}&\mathbb{E}\left[\Lambda_\varepsilon\left(\sum_{s\in\{\pm1\}}\prod_{i=1}^{\boldsymbol{\Delta}'''}\left(1-\mathbf{1}\left\{s\neq\operatorname{sign}(x_{n+1},b_i)\right\}\mu_{\mathbf{\Phi}'}(\boldsymbol{\sigma}_{\boldsymbol{y}_i}=-\operatorname{sign}(\boldsymbol{y}_i,b_i))\right)\right)\mid Z(\mathbf{\Phi}')>0\right]\\ &\quad=\mathbb{E}\left[\Lambda_\varepsilon\left(\sum_{s\in\{\pm1\}}\prod_{i=1}^{\boldsymbol{d}}\left(1-\mathbf{1}\{s\neq\boldsymbol{s}_i\}\mu_{\mathbf{\Phi}'}(\boldsymbol{\sigma}_{x_i}=\boldsymbol{s}_i')\right)\right)\mid Z(\mathbf{\Phi}')>0\right]+o(1).\end{aligned} \tag{6.37}$$

Further, on $\mathscr{E}$ we obtain

$$\begin{aligned}\frac{Z(\mathbf{\Phi}''')}{Z(\mathbf{\Phi}')} &= \sum_{\sigma\in\{\pm1\}^{\boldsymbol{Y}\cup\{x_{n+1}\}}}\mathbf{1}\left\{\sigma\text{ satisfies }b_1,\ldots,b_{\boldsymbol{\Delta}'''}\right\}\mu_{\mathbf{\Phi}'}\left(\forall y\in\boldsymbol{Y}:\boldsymbol{\sigma}_y=\sigma_y\right)\\ &= \sum_{\sigma\in\{\pm1\}^{\boldsymbol{Y}\cup\{x_{n+1}\}}}\mathbf{1}\left\{\sigma\text{ satisfies }b_1,\ldots,b_{\boldsymbol{\Delta}'''}\right\}\prod_{y\in\boldsymbol{Y}}\mu_{\mathbf{\Phi}'}\left(\boldsymbol{\sigma}_y=\sigma_y\right)+o(1) \qquad \text{[due to \textbf{E3},\textbf{E5}]}\\ &= \sum_{s\in\{\pm1\}}\prod_{\substack{i\in[\boldsymbol{\Delta}'''],\\ \operatorname{sign}(x_{n+1},b_i)=-s}}\mu_{\mathbf{\Phi}'}(\boldsymbol{\sigma}_{\boldsymbol{y}_i}=\operatorname{sign}(\boldsymbol{y}_i,b_i))\ ;\end{aligned} \tag{6.38}$$

to elaborate, in the last step s represents the value assigned to x_{n+1} and the product ensures that the clauses b_i in which x_{n+1} occurs with sign $-s$ are satisfied by assigning their second variable $\boldsymbol{y}_i$ the value $\operatorname{sign}(\boldsymbol{y}_i, b_i)$. Further, (6.38), **E3** and **E4** yield

$$\begin{aligned}\mathbb{E}\left[\mathbf{1}_{\mathscr{E}}\log\frac{Z(\mathbf{\Phi}''')}{Z(\mathbf{\Phi}')}\right] &= \mathbb{E}\left[\mathbf{1}_{\mathscr{E}}\log\left(\sum_{s\in\{\pm1\}}\prod_{i=1}^{\boldsymbol{\Delta}'''}\left(1-\mathbf{1}\left\{\operatorname{sign}(x_{n+1},b_i)=-s\right\}\mu_{\mathbf{\Phi}'}(\boldsymbol{\sigma}_{\boldsymbol{y}_i}=-\operatorname{sign}(\boldsymbol{y}_i,b_i))\right)\right)\right]+o(1)\\ &= \mathbb{E}\left[\mathbf{1}_{\mathscr{E}}\Lambda_\varepsilon\left(\sum_{s\in\{\pm1\}}\prod_{i=1}^{\boldsymbol{\Delta}'''}\left(1-\mathbf{1}\left\{\operatorname{sign}(x_{n+1},b_i)=-s\right\}\mu_{\mathbf{\Phi}'}(\boldsymbol{\sigma}_{\boldsymbol{y}_i}=-\operatorname{sign}(\boldsymbol{y}_i,b_i))\right)\right)\right]+o(1)\end{aligned} \tag{6.39}$$

Finally, the assertion follows from (6.35), (6.36), (6.37) and (6.39). □

Proof of Proposition 6.2. Because $\boldsymbol{\mu}_{\pi_d,1} \stackrel{\mathrm{d}}{=} 1 - \boldsymbol{\mu}_{\pi_d,1}$ by Proposition 2.1, Corollary 1.3 shows that for every $\varepsilon > 0$,

$$\lim_{n\to\infty}\mathbb{E}\left[\Lambda_\varepsilon\left(\sum_{s\in\{\pm1\}}\prod_{i=1}^{\boldsymbol{d}}\left(1-\mathbf{1}\{s\neq\boldsymbol{s}_i\}\mu_{\mathbf{\Phi}'}(\boldsymbol{\sigma}_{x_i}=\boldsymbol{s}_i')\right)\mid Z(\mathbf{\Phi}')>0\right)\right] = \mathbb{E}\left[\Lambda_\varepsilon\left(\sum_{s\in\{\pm1\}}\prod_{i=1}^{\boldsymbol{d}}\left(1-\mathbf{1}\{s\neq\boldsymbol{s}_i\}\boldsymbol{\mu}_{\pi_d,i}\right)\right)\right]. \tag{6.40}$$

Further, the dominated convergence theorem and (2.2) yield

$$\lim_{\varepsilon\to0}\mathbb{E}\left[\Lambda_\varepsilon\left(\sum_{s\in\{\pm1\}}\prod_{i=1}^{\boldsymbol{d}}\left(1-\mathbf{1}\{s\neq\boldsymbol{s}_i\}\boldsymbol{\mu}_{\pi_d,i}\right)\right)\right] = \mathbb{E}\log\left(\sum_{s\in\{\pm1\}}\prod_{i=1}^{\boldsymbol{d}}\left(1-\mathbf{1}\{s\neq\boldsymbol{s}_i\}\boldsymbol{\mu}_{\pi_d,i}\right)\right). \tag{6.41}$$

To complete the proof we combine (6.40), (6.41) and Lemma 6.13. □

7. Proof of Proposition 2.6

Tools such as Azuma's inequality do not apply to the number $Z(\mathbf{\Phi})$ of satisfying assignments because adding or removing even a single clause could change $Z(\mathbf{\Phi})$ by an exponential factor. Therefore, we prove Proposition 2.6 by way of a 'soft' version of the random 2-SAT problem. Specifically, for a real $\beta > 0$ we define $Z_\beta(\mathbf{\Phi})$ via (3.1). Thus, instead of dismissing assignments $\sigma \notin S(\mathbf{\Phi})$ outright, we charge an $\exp(-\beta)$ penalty factor for each violated clause. Because the constraints are soft, showing that $\log Z_\beta(\mathbf{\Phi})$ concentrates is a cinch.

Lemma 7.1. *For all $t, \beta > 0$ we have* $\mathbb{P}\left[\left|\log Z_\beta(\mathbf{\Phi}) - \mathbb{E}[\log Z_\beta(\mathbf{\Phi})]\right| > t \mid \boldsymbol{m}\right] \le 2\exp\left(-\frac{t^2}{2\boldsymbol{m}\beta^2}\right)$.

Proof. Since adding or removing a single clause can alter $Z_\beta(\mathbf{\Phi})$ by at most a factor $\exp(\pm\beta)$, the assertion follows from Azuma's inequality. □

The following statement, whose proof relies on the interpolation method from mathematical physics, will enable us to link the random variables $\log Z_\beta(\mathbf{\Phi})$ and $\log Z(\mathbf{\Phi})$. For a probability measure $p \in \mathscr{P}(0,1)$ and $\beta > 0$ let

$$\begin{aligned}\mathfrak{B}_\beta(p) = \mathbb{E}\left[\log \sum_{s=\pm1}\prod_{i=1}^{\boldsymbol{d}}\left(1 - \mathbf{1}\{\boldsymbol{s}_i \neq s\}\frac{1-\exp(-\beta)}{2}\left(1 - \boldsymbol{s}_i' + 2\boldsymbol{s}_i'\boldsymbol{\mu}_{p,i}\right)\right)\right] \\ - \frac{d}{2}\mathbb{E}\left[\log\left(1 - \frac{1-\exp(-\beta)}{4}\left(1-\boldsymbol{s}_1+2\boldsymbol{s}_1\boldsymbol{\mu}_{p,1}\right)\left(1-\boldsymbol{s}_2+2\boldsymbol{s}_2\boldsymbol{\mu}_{p,2}\right)\right)\right].\end{aligned} \tag{7.1}$$

These two expectations exist and are finite because $0 \le \beta < \infty$. (More precisely, their absolute values are bounded by $\log 2 + \beta d$ and β, respectively.)

Lemma 7.2 ([42, Theorem 1])**.** *For any $p \in \mathscr{P}(0,1)$ and any $0 \le \beta < \infty$ we have* $\lim_{n\to\infty}\frac{1}{n}\mathbb{E}\log Z_\beta(\mathbf{\Phi}) \le \mathfrak{B}_\beta(p)$.

Combining Lemmas 7.1 and 7.2, we obtain the following bound for 'hard' 2-SAT.

Corollary 7.3. *For any $\beta > 0$ we have* $\lim_{n\to\infty}\mathbb{P}\left[\log Z(\mathbf{\Phi}) > n\mathfrak{B}_\beta(\pi_d) + n^{2/3}\right] = 0$.

Proof. We have $Z_\beta(\mathbf{\Phi}) \ge Z(\mathbf{\Phi})$ and Lemmas 7.1 and 7.2 imply $\lim_{n\to\infty}\mathbb{P}\left[\log Z_\beta(\mathbf{\Phi}) > n\mathfrak{B}_\beta(\pi_d) + n^{2/3}\right] = 0$. □

Proof of Proposition 2.6. We begin by observing that the limit $\lim_{\beta\to\infty}\mathfrak{B}_\beta(\pi_d)$ exists and is finite. First, there is the pointwise and monotone convergence of the integrands:

$$\log \sum_{s=\pm1}\prod_{i=1}^{\boldsymbol{d}}\left(1 - \mathbf{1}\{\boldsymbol{s}_i \neq s\}\frac{1-\exp(-\beta)}{2}\left(1 - \boldsymbol{s}_i' + 2\boldsymbol{s}_i'\boldsymbol{\mu}_{\pi_d,i}\right)\right) \overset{\beta\to\infty}{\longrightarrow} \log \sum_{s=\pm1}\prod_{i=1}^{\boldsymbol{d}}\left(1 - \frac{\mathbf{1}\{\boldsymbol{s}_i \neq s\}}{2}\left(1 - \boldsymbol{s}_i' + 2\boldsymbol{s}_i'\boldsymbol{\mu}_{\pi_d,i}\right)\right), \tag{7.2}$$

$$\log\left(1 - \frac{1-\exp(-\beta)}{4}\left(1-\boldsymbol{s}_1+2\boldsymbol{s}_1\boldsymbol{\mu}_{\pi_d,1}\right)\left(1-\boldsymbol{s}_2+2\boldsymbol{s}_2\boldsymbol{\mu}_{\pi_d,2}\right)\right) \overset{\beta\to\infty}{\longrightarrow} \log\left(1 - \frac{1}{4}\left(1-\boldsymbol{s}_1+2\boldsymbol{s}_1\boldsymbol{\mu}_{\pi_d,1}\right)\left(1-\boldsymbol{s}_2+2\boldsymbol{s}_2\boldsymbol{\mu}_{\pi_d,2}\right)\right). \tag{7.3}$$

Further, since $\boldsymbol{\mu}_{\pi_d} \overset{\mathrm{d}}{=} 1 - \boldsymbol{\mu}_{\pi_d}$ by Proposition 2.1 and because $1 - \boldsymbol{s} + 2\boldsymbol{s}\boldsymbol{\mu}_{\pi_d}$ equals either $2\boldsymbol{\mu}_{\pi_d}$ or $2(1-\boldsymbol{\mu}_{\pi_d})$, we obtain

$$\log \sum_{s=\pm1}\prod_{i=1}^{\boldsymbol{d}}\left(1 - \frac{\mathbf{1}\{\boldsymbol{s}_i \neq s\}}{2}\left(1 - \boldsymbol{s}_i' + 2\boldsymbol{s}_i'\boldsymbol{\mu}_{\pi_d,i}\right)\right) \overset{d}{=} \log\left(\prod_{i=1}^{\boldsymbol{d}^-}\boldsymbol{\mu}_{\pi_d,i} + \prod_{i=1}^{\boldsymbol{d}^+}\boldsymbol{\mu}_{\pi_d,i+\boldsymbol{d}^-}\right), \tag{7.4}$$

$$\frac{d}{2}\log\left(1 - \frac{1}{4}\left(1-\boldsymbol{s}_1+2\boldsymbol{s}_1\boldsymbol{\mu}_{\pi_d,1}\right)\left(1-\boldsymbol{s}_2+2\boldsymbol{s}_2\boldsymbol{\mu}_{\pi_d,2}\right)\right) \overset{d}{=} \frac{d}{2}\log\left(1 - \boldsymbol{\mu}_{\pi_d,1}\boldsymbol{\mu}_{\pi_d,2}\right). \tag{7.5}$$

Moreover, Proposition 2.1 shows that the monotone limits are integrable and therefore an application of the monotone convergence theorem to (7.2) and (7.3), followed by the simplifications (7.4), (7.4), yields the identity

$$\lim_{\beta\to\infty}\mathfrak{B}_\beta(\pi_d) = \mathbb{E}\left[\log\left(\prod_{i=1}^{\boldsymbol{d}^-}\boldsymbol{\mu}_{\pi_d,i} + \prod_{i=1}^{\boldsymbol{d}^+}\boldsymbol{\mu}_{\pi_d,i+\boldsymbol{d}^-}\right) - \frac{d}{2}\log\left(1 - \boldsymbol{\mu}_{\pi_d,1}\boldsymbol{\mu}_{\pi_d,2}\right)\right] = \mathfrak{B}_\infty(\pi_d) < \infty.$$

Further, Corollary 2.5 shows that $\mathfrak{B}_\infty(\pi_d) = \lim_{n\to\infty} n^{-1}\mathbb{E}[\log(Z(\mathbf{\Phi}) \vee 1)]$. Therefore, Corollary 7.3 implies that

$$\mathbb{P}\left[n^{-1}\log(Z(\mathbf{\Phi}) \vee 1) > \mathfrak{B}_\infty(\pi_d) + \varepsilon\right] = o(1) \qquad \text{for any } \varepsilon > 0. \tag{7.6}$$

To complete the proof, we upper bound

$$n^{-1}\mathbb{E}\left|\log(Z(\mathbf{\Phi})\vee 1)-\mathbb{E}[\log(Z(\mathbf{\Phi})\vee 1)]\right|\leq \mathbb{E}\left|n^{-1}\log(Z(\mathbf{\Phi})\vee 1)-\mathfrak{B}_{\infty}(\pi_d)\right|+\left|\mathfrak{B}_{\infty}(\pi_d)-\mathbb{E}[\log(Z(\mathbf{\Phi})\vee 1)]\right|. \quad (7.7)$$

Due to Corollary 2.5, the second term on the r.h.s. of (7.7) tends to zero. On the other hand, (7.6) and Corollary 2.5 yield that for any $\varepsilon>0$,

$$\mathbb{E}\left|n^{-1}\log(Z(\mathbf{\Phi})\vee 1)-\mathfrak{B}_{\infty}(\pi_d)\right|\leq \mathbb{E}\left[\mathfrak{B}_{\infty}(\pi_d)-n^{-1}\log(Z(\mathbf{\Phi})\vee 1)\right]+2\varepsilon+o(1)=2\varepsilon+o(1),$$

as desired. □

Acknowledgment. We thank Andreas Galanis and Leslie Goldberg for helpful discussions.

REFERENCES

[1] E. Abbe, A. Montanari: On the concentration of the number of solutions of random satisfiability formulas. Random Structures and Algorithms **45** (2014) 362–382.
[2] D. Achlioptas, A. Chtcherba, G. Istrate, C. Moore: The phase transition in 1-in-k SAT and NAE 3-SAT. Proc. 12th SODA (2001) 721–722.
[3] D. Achlioptas, C. Moore: Random k-SAT: two moments suffice to cross a sharp threshold. SIAM Journal on Computing **36** (2006) 740–762.
[4] D. Achlioptas, Y. Peres: The threshold for random k-SAT is $2^k\ln 2-O(k)$. Journal of the AMS **17** (2004) 947–973.
[5] M. Aizenman, R. Sims, S. Starr: An extended variational principle for the SK spin-glass model. Phys. Rev. B **68** (2003) 214403.
[6] N. Alon, J. Spencer: The probabilistic method. Wiley (2016).
[7] P. Ayre, A. Coja-Oghlan, P. Gao, N. Müller: The satisfiability threshold for random linear equations. Combinatorica, in press.
[8] A. Barvinok: Combinatorics and complexity of partition functions. Springer (2016).
[9] V. Bogachev, A. Kolesnikov: The Monge-Kantorovich problem: achievements, connections, and perspectives. Russian Mathematical Surveys **67** (2012) 785–890.
[10] B. Bollobás: The evolution of random graphs. Transactions of the AMS **286** (1984) 257–274.
[11] B. Bollobás, C. Borgs J. Chayes, J. Kim, D. Wilson: The scaling window of the 2-SAT transition. Random Structures and Algorithms **18** (2001) 201–256.
[12] Y. Boufkhad, O. Dubois: Length of prime implicants and number of solutions of random CNF formulae. Theoretical Computer Science **215** (1999) 1–30.
[13] P. Cheeseman, B. Kanefsky, W. Taylor: Where the *really* hard problems are. Proc. IJCAI (1991) 331–337.
[14] V. Chvátal, B. Reed: Mick gets some (the odds are on his side). Proc. 33th FOCS (1992) 620–627.
[15] A. Coja-Oghlan, T. Kapetanopoulos, N. Müller: The replica symmetric phase of random constraint satisfaction problems. Combinatorics, Probability and Computing, in press.
[16] A. Coja-Oghlan, F. Krzakala, W. Perkins, L. Zdeborová: Information-theoretic thresholds from the cavity method. Advances in Mathematics **333** (2018) 694–795.
[17] A. Coja-Oghlan, K. Panagiotou: The asymptotic k-SAT threshold. Advances in Mathematics **288** (2016) 985–1068.
[18] A. Coja-Oghlan, N. Wormald: The number of satisfying assignments of random regular k-SAT formulas. Combinatorics, Probability and Computing **27** (2018) 496–530.
[19] C. Cooper, A. Frieze, G. Sorkin: Random 2-SAT with prescribed literal degrees. Algorithmica **48** (2007) 249–265.
[20] D. Coppersmith, D. Gamarnik, M. Hajiaghayi, G. Sorkin: Random MAX SAT, random MAX CUT, and their phase transitions. Random Structures and Algorithms **24** (2004) 502–545.
[21] A. Dembo, A. Montanari: Ising models on locally tree-like graphs. Annals of Applied Probability **20** (2010) 565–592.
[22] A. Dembo, A. Montanari, N. Sun: Factor models on locally tree-like graphs. Annals of Probability **41** (2013) 4162–4213.
[23] M. Dietzfelbinger, A. Goerdt, M. Mitzenmacher, A. Montanari, R. Pagh, M. Rink: Tight thresholds for cuckoo hashing via XORSAT. arXiv:0912.0287 (2009).
[24] J. Ding, A. Sly, N. Sun: Satisfiability threshold for random regular NAE-SAT. Communications in Mathematical Physics **341** (2016) 435–489.
[25] J. Ding, A. Sly, N. Sun: Proof of the satisfiability conjecture for large k. Proc. 47th STOC (2015) 59–68.
[26] W. Dowling, J. Gallier: Linear-time algorithms for testing the satisfiability of propositional Horn formulae. Journal of Logic Programming **1** (1984) 267–284.
[27] O. Dubois, J. Mandler: The 3-XORSAT threshold. Proc. 43rd FOCS (2002) 769–778.
[28] W. Fernandez de la Vega: Random 2-SAT: results and problems. Theoretical Computer Science **265** (2001) 131–146.
[29] S. Franz, M. Leone: Replica bounds for optimization problems and diluted spin systems. J. Stat. Phys. **111** (2003) 535–564.
[30] A. Goerdt: A threshold for unsatisfiability. J. Comput. Syst. Sci. **53** (1996) 469–486
[31] F. Guerra: Broken replica symmetry bounds in the mean field spin glass model. Comm. Math. Phys. **233** (2003) 1–12.
[32] F. Krzakala, A. Montanari, F. Ricci-Tersenghi, G. Semerjian, L. Zdeborová: Gibbs states and the set of solutions of random constraint satisfaction problems. Proc. National Academy of Sciences **104** (2007) 10318–10323.
[33] T. Łuczak: Component behavior near the critical point of the random graph process. Random Structures and Algorithms **1** (1990) 287–310.
[34] M. Mézard, A. Montanari: Information, physics and computation. Oxford University Press (2009).
[35] M. Mézard, G. Parisi, R. Zecchina: Analytic and algorithmic solution of random satisfiability problems. Science **297** (2002) 812–815.
[36] R. Monasson, R. Zecchina: The entropy of the k-satisfiability problem. Phys. Rev. Lett. **76** (1996) 3881.
[37] R. Monasson, R. Zecchina: Statistical mechanics of the random K-SAT model. Phys. Rev. E **56** (1997) 1357–1370.
[38] A. Montanari, F. Ricci-Tersenghi, G. Semerjian: Solving constraint satisfaction problems through Belief Propagation-guided decimation. Proc. 45th Allerton (2007).

[39] A. Montanari, D. Shah: Counting good truth assignments of random k-SAT formulae. Proc. 18th SODA (2007) 1255–1264.
[40] D. Panchenko: Spin glass models from the point of view of spin distributions. Annals of Probability **41** (2013) 1315–1361.
[41] D. Panchenko: On the replica symmetric solution of the K-sat model. Electron. J. Probab. **19** (2014) #67.
[42] D. Panchenko, M. Talagrand: Bounds for diluted mean-fields spin glass models. Probab. Theory Relat. Fields **130** (2004) 319–336.
[43] B. Pittel, G. Sorkin: The satisfiability threshold for k-XORSAT. Combinatorics, Probability and Computing **25** (2016) 236–268.
[44] F. Rassmann: On the number of solutions in random graph k-colouring. Combinatorics, Probability and Computing **28** (2019) 130–158.
[45] A. Scott, G. Sorkin: Solving sparse random instances of Max Cut and Max 2-CSP in linear expected time. Combinatorics, Probability and Computing **15** (2006) 281–315.
[46] A. Sharell: Concentration of the number of solutions to a random 2-CNF formula. Manuscript (2000).
[47] A. Sly, N. Sun, Y. Zhang: The number of solutions for random regular NAE-SAT. Proc. 57th FOCS (2016) 724–731.
[48] M. Talagrand: The high temperature case for the random K-sat problem. Probab. Theory Related Fields **119** (2001) 187–212.
[49] L. Valiant: The complexity of enumeration and reliability problems. SIAM Journal on Computing **8** (1979) 410–421.

DIMITRIS ACHLIOPTAS, optas@di.uoa.gr, UNIVERSITY OF ATHENS, DEPARTMENT OF COMPUTER SCIENCE & TELECOMMUNICATIONS, PANEPISTIMIOPOLIS, ILISSIA, ATHENS 15784, GREECE.

AMIN COJA-OGHLAN, acoghlan@math.uni-frankfurt.de, GOETHE UNIVERSITY, MATHEMATICS INSTITUTE, 10 ROBERT MAYER ST, FRANKFURT 60325, GERMANY.

MAX HAHN-KLIMROTH, hahnklim@math.uni-frankfurt.de, GOETHE UNIVERSITY, MATHEMATICS INSTITUTE, 10 ROBERT MAYER ST, FRANKFURT 60325, GERMANY.

JOON LEE, lee@math.uni-frankfurt.de, GOETHE UNIVERSITY, MATHEMATICS INSTITUTE, 10 ROBERT MAYER ST, FRANKFURT 60325, GERMANY.

NOËLA MÜLLER, nmueller@math.uni-frankfurt.de, GOETHE UNIVERSITY, MATHEMATICS INSTITUTE, 10 ROBERT MAYER ST, FRANKFURT 60325, GERMANY.

MANUEL PENSCHUCK, manuel@ae.cs.uni-frankfurt.de, GOETHE UNIVERSITY, COMPUTER SCIENCE INSTITUTE, 11–15 ROBERT MAYER ST, FRANKFURT 60325, GERMANY.

GUANGYAN ZHOU, gyzhou76@gmail.com, SCHOOL OF MATHEMATICS AND STATISTICS, BEIJING TECHNOLOGY AND BUSINESS UNIVERSITY, BEIJING 100048, CHINA.

F. The cut metric for probability distributions

THE CUT METRIC FOR PROBABILITY DISTRIBUTIONS

AMIN COJA-OGHLAN, MAX HAHN-KLIMROTH*

ABSTRACT. Guided by the theory of graph limits, we investigate a variant of the cut metric for limit objects of sequences of discrete probability distributions. Apart from establishing basic results, we introduce a natural operation called *pinning* on the space of limit objects and show how this operation yields a canonical cut metric approximation to a given probability distribution akin to the weak regularity lemma for graphons. We also establish the cut metric continuity of basic operations such as taking product measures. *MSc: 60C05, 60B10*

1. INTRODUCTION AND RESULTS

1.1. **Background and motivation.** The theory of graph limits clearly qualifies as one of the great recent success of modern combinatorics [6, 7, 32, 34]. Exhibiting a complete metric space of limit objects of sequences of finite graphs, the theory strikes a link between combinatorics and analysis. In fact, the notion of graphon convergence unifies several combinatorially meaningful concepts, such as convergence of subgraph counts or with respect to the cut metric. In effect, combinatorial ideas admit neat analytic interpretations. For instance, the Szemerédi regularity lemma yields the compactness of the graphon space [35].

While sequences of graphs occur frequently in combinatorics (e.g., in the theory of random graphs), sequences of probability distributions on increasingly large discrete domains play no less prominent a role in the mathematical sciences. For instance, such sequences are the bread and butter of mathematical physics. A classical example is the Ising model on a d-dimensional integer lattice of side length n, a model of ferromagnetism. The Ising model renders a probability measure, the so-called Boltzmann distribution, on the space $\{-1,+1\}^{[n]^d}$ that captures the distribution of the magnetic spins of the n^d vertices. The objective is to extract properties of this probability distribution in the limit of large n such as the nature of correlations. While mathematical physics has a purpose-built theory of limits of probability measures on lattices [23], this theory fails to cover other classes of important statistical mechanics models, such as mean-field models that 'live' on random graphs [38]. Additionally, in statistics and data science sequences of discrete probability distributions arise naturally, e.g., as the empirical distributions of samples as more data are acquired.

The purpose of this paper is to show how the theory of graph limits can be adapted and extended to obtain a coherent theory of limits of probability distributions on discrete cubes. First cursory steps were already taken in an earlier contribution [14]. For instance, a probabilistic version of the cut metric was defined in that paper. Moreover, Austin [4], Diaconis and Janson [19] and (later) Panchenko [43] pointed out the connection between the theory of graph limits and the Aldous-Hoover representation [1, 25, 29]. But thus far a complete and concise disquisition has been lacking. We therefore develop the basics of a cut-norm based limiting theory for probability measures, including the completeness and compactness of the space of limiting objects, a kernel representation, a sampling theorem and a discussion of the connection with exchangeable arrays. Some of the proofs rely on arguments similar to the ones used in the theory of graph limits, and none of them will come as a gross surprise to experts. In fact, a few statements (such as the compactness of the space of limiting objects) already appeared in [14], albeit without detailed proofs, and a few others (such as the characterisation of exchangeable arrays) are generalisations of results from [19]. But here we present unified proofs of these basic results in full generality to provide a coherent and mostly self-contained treatment that, we hope, will facilitate applications.

Additionally, and this constitutes the main technical novelty of the paper, we present a new construction of regular partitions for limit objects of discrete probability distributions that constitutes a continuous generalisation of the *pinning operation* for discrete probability distributions from [11, 40, 45]. The result provides an approximation akin to the graphon version of the Frieze-Kannan regularity lemma [21]. The pinning operation merely involves a purely mechanical reweighting of the probability distribution. The 'obliviousness' of the operation was critical

Supported by Stiftung Polytechnische Gesellschaft Frankfurt am Main.

to work on spin glass models on random graphs and on inference problems [11, 12, 13, 14]. We show that a similarly oblivious procedure carries over naturally to the space of limit objects. The proof, which hinges on a delicate analysis of cut norm approximations, constitutes the main technical achievement of the paper.

1.2. **Results.** We proceed to set out the main concepts and to state the main results of the paper. A detailed account of related work follows in Section 1.3. The cut metric is a mainstay of the theory of graph limits. An adaptation for probability measures was suggested in [13, 14]. Let us thus begin by recalling this construction.

1.2.1. *The cut metric.* Let $\Omega \neq \emptyset$ be a finite set and let $n \geq 1$ be an integer. Further, for probability distributions μ, ν on the discrete cube Ω^n let $\Gamma(\mu,\nu)$ be the set of all couplings of μ,ν, i.e., all probability distributions γ on the product space $\Omega^n \times \Omega^n$ with marginal distributions μ,ν. Additionally, let $\mathbb{S}_n$ be the set of all permutations $[n] \to [n]$. Following [13], we define the *(weak) cut distance* of μ,ν as

$$\Delta_{\boxtimes}(\mu,\nu) = \inf_{\substack{\gamma\in\Gamma(\mu,\nu)\\ \varphi\in\mathbb{S}_n}} \sup_{\substack{S\subset\Omega^n\times\Omega^n\\ X\subset[n]\\ \omega\in\Omega}} \frac{1}{n}\left|\sum_{\substack{(\sigma,\tau)\in S\\ x\in X}} \gamma(\sigma,\tau)\left(\mathbf{1}\{\sigma_x=\omega\}-\mathbf{1}\{\tau_{\varphi(x)}=\omega\}\right)\right|. \tag{1.1}$$

The idea is that we first get to align μ,ν as best as possible by choosing a suitable coupling γ along with a permutation φ of the n coordinates. Then an adversary comes along and points out the largest remaining discrepancy. Specifically, the adversary picks an event $S \subset \Omega^n \times \Omega^n$ under the coupling, a set $X \subset [n]$ of coordinates and an element $\omega \in \Omega$ and reads off the discrepancy of the frequency of ω on S, X. It is easily verified that (1.1) defines a pre-metric on the space $\mathscr{L}_n = \mathscr{L}_n(\Omega)$ of probability distribution on Ω^n. Thus, $\Delta_{\boxtimes}(\cdot,\cdot)$ is symmetric and satisfies the triangle inequality. But distinct μ,ν need not satisfy $\Delta_{\boxtimes}(\mu,\nu) > 0$. Hence, to obtain a metric space $\mathfrak{L}_n = \mathfrak{L}_n(\Omega)$ we identify any $\mu,\nu \in \mathscr{L}_n$ with $\Delta_{\boxtimes}(\mu,\nu) = 0$.

Following [14], we embed the spaces $\mathscr{L}_n$ into a joint space $\mathscr{L}$. Specifically, let $\mathscr{P}(\Omega)$ be the space of all probability distributions on Ω. We identify $\mathscr{P}(\Omega)$ with the standard simplex in $\mathbb{R}^n$ and thus endow $\mathscr{P}(\Omega)$ with the Euclidean topology and the corresponding Borel algebra. Further, let $\mathscr{S}$ be the space of all measurable maps $\sigma : [0,1] \to \mathscr{P}(\Omega)$, $\sigma \mapsto \sigma_x$, up to equality (Lebesgue-)almost everywhere. We equip $\mathscr{S}$ with the L_1-metric

$$D_1(\sigma,\tau) = \sum_{\omega\in\Omega}\int_0^1 |\sigma_x(\omega)-\tau_x(\omega)|\,\mathrm{d}x \qquad (\sigma,\tau\in\mathscr{S})$$

and the corresponding Borel algebra, thus obtaining a complete, separable metric space. The space $\mathscr{L}$ is defined as the space of all probability measures on $\mathscr{S}$.

Much as in the discrete case, for probability distributions μ,ν on $\mathscr{S}$ we let $\Gamma(\mu,\nu)$ be the space of all couplings of μ,ν, i.e., probability distributions γ on $\mathscr{S}\times\mathscr{S}$ with marginals μ,ν. Moreover, let $\mathbb{S}$ be the space of all measurable bijections $\varphi : [0,1] \to [0,1]$ such that both φ and its inverse φ^{-1} map the Lebesgue measure to itself.[1] Then the *cut distance* of μ,ν is defined by the expression

$$D_{\boxtimes}(\mu,\nu) = \inf_{\substack{\gamma\in\Gamma(\mu,\nu)\\ \varphi\in\mathbb{S}}} \sup_{\substack{S\subset\mathscr{S}\times\mathscr{S}\\ X\subset[0,1]\\ \omega\in\Omega}} \left|\int_S\int_X \left(\sigma_x(\omega)-\tau_{\varphi(x)}(\omega)\right)\mathrm{d}x\mathrm{d}\gamma(\sigma,\tau)\right|, \tag{1.2}$$

where, of course, S, X range over measurable sets. Thus, as in the discrete case we first align μ,ν as best as possible by choosing a coupling and a suitable 'permutation' φ. Then the adversary puts their finger on the largest remaining discrepancy. One easily verifies that (1.2) defines a pre-metric on $\mathscr{L}$. Thus, identifying any μ,ν with $D_{\boxtimes}(\mu,\nu) = 0$, we obtain a metric space $\mathfrak{L}$. The points of this space we call Ω-*laws*.

Theorem 1.1. *The metric space $\mathfrak{L}$ is compact.*

Theorem 1.1 was already stated in [14], but no detailed proof was included. We will give a full proof based on a novel analytic argument in Section 3.

What is the connection between the spaces $\mathfrak{L}_n$ and the 'limiting space' $\mathfrak{L}$? As pointed out in [14], a probability distribution μ on Ω^n naturally induces an Ω-law. Indeed, we represent each $\sigma \in \Omega^n$ by a step function $\dot\sigma : [0,1] \to \mathscr{P}(\Omega)$ whose value on the interval $[(i-1)/n, i/n)$ is just the atom $\delta_{\sigma_i} \in \mathscr{P}(\Omega)$ for each $i \in [n]$. (This construction

[1] We recall that on a standard Borel space the inverse map φ^{-1} is measurable as well, see Lemma 2.2.

is somewhat similar to the one proposed for 'decorated graphs' in [33].) Then we let $\dot\mu \in \mathscr{L}$ be the distribution of $\dot\sigma \in \mathscr{S}$ for σ chosen from μ; in symbols,

$$\dot\mu = \sum_{\sigma\in\Omega^n} \mu(\sigma)\delta_{\dot\sigma} \in \mathscr{L}.$$

Thus, we obtain a map $\mathscr{L}_n \to \mathscr{L}$, $\mu \mapsto \dot\mu$. The definition of the cut metric guarantees that $D_{\boxtimes}(\dot\mu,\dot\nu) = 0$ if $\Delta_{\boxtimes}(\mu,\nu) = 0$. Consequently, the map $\mu \mapsto \dot\mu$ induces a map $\mathfrak{L}_n \to \mathfrak{L}$. The following statement shows that this map is in fact an embedding, and that therefore the space $\mathfrak{L}$ unifies all the spaces $\mathfrak{L}_n$, $n \ge 1$.

Theorem 1.2. *There exists a function $\mathfrak{d} : [0,1] \to [0,1]$ with $\mathfrak{d}^{-1}(0) = \{0\}$ such that for all $n \ge 1$ and all $\mu,\nu \in \mathfrak{L}_n$ we have $\mathfrak{d}(\Delta_{\boxtimes}(\mu,\nu)) \le D_{\boxtimes}(\dot\mu,\dot\nu) \le \Delta_{\boxtimes}(\mu,\nu)$.*

We will see a few examples of convergence in the cut metric momentarily. But let us first explore a convenient representation of the space $\mathfrak{L}$.

Remark 1.3. *The definition of the space $\mathfrak{L}$ is based on measurable functions $\sigma : [0,1] \to \mathscr{P}(\Omega)$. Of course, one could replace the unit interval by another atomless probability space, and this may be natural/convenient in some situations. (The definition of the product and direct sum in Section 1.2.6 below could be quoted as a case a point.) But the use of the unit interval is without loss of generality (see Lemma 2.3 below).*

1.2.2. *The kernel representation.* As in the case of graph limits, Ω-laws can naturally be represented by functions on the unit square that we call kernels. To be precise, let $\mathscr{K}$ be the set of all measurable maps $\kappa : [0,1]^2 \to \mathscr{P}(\Omega)$, $(s,x) \mapsto \kappa_{s,x}$, up to equality almost everywhere. For $\kappa,\kappa' \in \mathscr{K}$ we define, with S, X ranging over measurable sets,

$$D_{\boxtimes}(\kappa,\kappa') = \inf_{\varphi,\varphi'\in\mathbb{S}} \sup_{S,X\subset[0,1]} \left| \int_S \int_X \left(\kappa_{s,x}(\omega) - \kappa'_{\varphi(s),\varphi'(x)}(\omega)\right) \mathrm{d}x\mathrm{d}s \right|. \tag{1.3}$$
$$\omega\in\Omega$$

As before (1.3) defines a pre-metric on $\mathscr{K}$. We obtain a metric space $\mathfrak{K}$ by identifying $\kappa,\kappa' \in \mathscr{K}$ with $D_{\boxtimes}(\kappa,\kappa') = 0$.

There is a natural map $\mathscr{K} \to \mathscr{L}$. Namely, for a kernel κ and $s \in [0,1]$ let $\kappa_s : [0,1] \to \mathscr{P}(\Omega)$ be the measurable map $x \mapsto \kappa_{s,x}$. This map belongs to the space $\mathscr{S}$. Thus, κ induces a probability distribution μ^κ on $\mathscr{S}$, namely the distribution of $\kappa_{\boldsymbol{s}}$ for a uniformly random $\boldsymbol{s} \in [0,1]$. The definition of the cut distance guarantees that $D_{\boxtimes}(\mu^\kappa,\mu^{\kappa'}) = 0$ if $D_{\boxtimes}(\kappa,\kappa') = 0$. Therefore, as pointed out in [14], the map $\kappa \mapsto \mu^\kappa$ induces a map $\mathfrak{K} \to \mathfrak{L}$.

Theorem 1.4. *The map $\mathfrak{K} \to \mathfrak{L}$ induced by $\kappa \mapsto \mu^\kappa$ is an isometric bijection.*

Thus, any Ω-law μ can be represented by an Ω-kernel, which we denote by κ^μ.

Example 1.5. *With $\Omega = \{0,1\}$ let $\mu^{(n)} \in \mathscr{L}_n$ be uniformly distributed over all $\sigma \in \{0,1\}^n$ with even parity. In symbols,*

$$\mu^{(n)}(\sigma) = 2^{1-n}\mathbf{1}\left\{\sum_{i=1}^n \sigma_i \equiv 0 \mod 2\right\}.$$

Similarly, let $\nu^{(n)}$ be uniformly distributed on the set of $\sigma \in \{0,1\}^n$ with odd parity. Then $\mu^{(n)},\nu^{(n)}$ have total variation distance one for all n because they are supported on disjoint subsets of $\{0,1\}^n$. Nevertheless, in the cut distance both sequences $(\mu^{(n)})_n, (\nu^{(n)})_n$ converge to the common limit $\mu = \delta_u \in \mathscr{L}$ supported on $u : [0,1] \to \mathscr{P}(\{0,1\})$, $x \mapsto (1/2,1/2)$. Specifically, we claim that

$$\Delta_{\boxtimes}(\mu^{(n)},\nu^{(n)}) = O(n^{-1}), \qquad D_{\boxtimes}(\dot\mu^{(n)},\mu) = O(n^{-1/2}). \tag{1.4}$$

To verify the first bound, consider the following coupling $\gamma^{(n)}$: choose the first $n-1$ bits $\boldsymbol{\sigma}_1,\dots,\boldsymbol{\sigma}_{n-1} \in \{0,1\}$ uniformly and independently and choose $\boldsymbol{\sigma}_n \in \{0,1\}$ so that $\sum_{i=1}^n \boldsymbol{\sigma}_i \equiv 0 \mod 2$. Then $\gamma^{(n)} \in \mathscr{P}(\Omega^n \times \Omega^n)$ is the distribution of $((\boldsymbol{\sigma}_1,\dots,\boldsymbol{\sigma}_n),(\boldsymbol{\sigma}_1,\dots,1-\boldsymbol{\sigma}_n))$. In effect, under $\gamma^{(n)}$ the two n-bit vectors differ in exactly one position, whence the first part of (1.4) *follows from* (1.1)*. The second bound in* (1.4) *follows from the central limit theorem.*

Example 1.6. *Let $\mu^{(n)}$ be the probability distribution on $\{0,1\}^n$ induced by the following experiment. First, pick $\boldsymbol{s} \in [0,1]$ uniformly at random. Then, given $\boldsymbol{s}$, obtain $\boldsymbol{\sigma} \in \{0,1\}^n$ by letting $\boldsymbol{\sigma}_i = 1$ with probability $i\boldsymbol{s}/n$ independently for each $i \in [n]$. In formulas,*

$$\mu^{(n)}(\sigma) = \int_0^1 \prod_{i=1}^n \left(\frac{is}{n}\right)^{\sigma_i} \left(1 - \frac{is}{n}\right)^{1-\sigma_i} \mathrm{d}s.$$

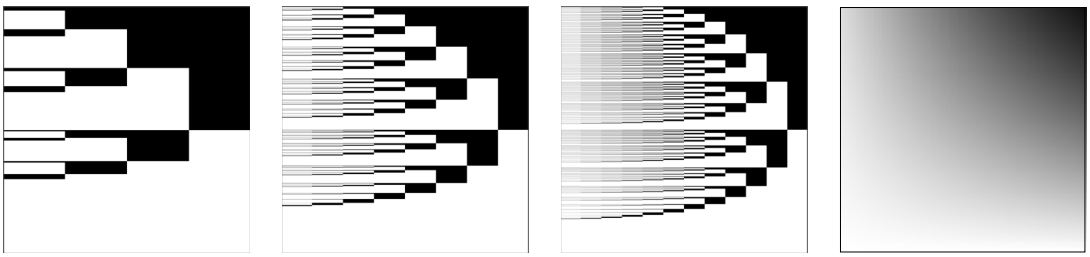

FIGURE 1. The maps $(s,x) \in [0,1]^2 \mapsto \kappa^{(n)}_{s,x}(1)$ for $n = 4, 8, 12$ and the limiting kernel $(s,x) \mapsto \kappa_{s,x}(1)$ from Example 1.6.

Kernel representations $\kappa^{(n)}$ of $\mu^{(n)}$ are displayed in Figure 1 for some values of n. The sequence $\kappa^{(n)}$ converges to the kernel $\kappa : [0,1]^2 \to \mathscr{P}(\{0,1\})$ defined by $\kappa_{s,x}(1) = sx$, $\kappa_{s,x}(0) = 1 - sx$.

Example 1.7. *The Curie-Weiss model is an (extremely) simple model of ferromagnetism. The vertices of a complete graph of order n correspond to iron atoms that can take one of two possible magnetic spins ± 1. Energetically it is beneficial for atoms to be aligned and the impact of the energetic term is governed by a temperature parameter $T > 0$. To be precise, the* Boltzmann distribution *$\mu^{(n)}$ on $\{\pm 1\}^n$ defined by*

$$\mu^{(n)}_T(\sigma) \propto \exp\left(\frac{T}{n} \sum_{1 \le i < j \le n} \sigma_i \sigma_j\right)$$

captures the distribution of spin configurations at a given temperature. The Curie-Weiss model is completely understood mathematically and it is well known that a phase transition occurs at $T = 1$. In the framework of the cut distance, this phase transition manifests itself in the different limits that the sequence $(\mu^{(n)}_T)_n$ converges to. Specifically, the kernel κ_T representing the limit reads

$$\kappa_T : (s,x) \in [0,1]^2 \mapsto (1/2, 1/2) \qquad \text{for } T \le 1,$$

$$\kappa_T : (s,x) \in [0,1]^2 \mapsto \begin{cases} ((1+m_T)/2, (1-m_T)/2) & \text{if } s \le 1/2, \\ ((1-m_T)/2, (1+m_T)/2) & \text{if } s > 1/2 \end{cases} \qquad \text{for } T > 1,$$

where $0 < m_T < 1$ is the unique zero of $m_T/T - \ln(1+m_T)/2 + \ln(1-m_T)/2$ for $T > 1$.

1.2.3. *Counting and sampling.* In the theory of graph limits convergence with respect to the cut metric is equivalent to convergence of subgraph counts. We are going to derive a similar equivalence for Ω-laws. In fact, we are going to derive an extension of this result that links the cut metric to the theory of exchangeable arrays. We recall that a probability distribution Ξ on the space $\Omega^{\mathbb{N}\times\mathbb{N}}$ of infinite Ω-valued arrays is *exchangeable* if the following is true. If $\boldsymbol{X}^\Xi = (\boldsymbol{X}^\Xi(i,j))_{i,j\ge 1} \in \Omega^{\mathbb{N}\times\mathbb{N}}$ is drawn randomly from Ξ, then for any integer n and for any permutations $\varphi, \psi : [n] \to [n]$ the random $n \times n$-arrays

$$(\boldsymbol{X}^\Xi(i,j))_{i,j\in[n]} \qquad \text{and} \qquad (\boldsymbol{X}^\Xi(\varphi(i),\psi(j)))_{i,j\in[n]}$$

are identically distributed. Let $\mathfrak{X} = \mathfrak{X}(\Omega)$ denote the set of all exchangeable distributions. Since the product space $\Omega^{\mathbb{N}\times\mathbb{N}}$ is compact by Tychonoff's theorem, endowed with the weak topology $\mathfrak{X}$ is a compact, separable space.

A kernel $\kappa \in \mathfrak{K}$ naturally induces an exchangeable distribution. Specifically, let $\boldsymbol{s}_1, \boldsymbol{x}_1, \boldsymbol{s}_2, \boldsymbol{x}_2, \ldots \in [0,1]$ be mutually independent uniformly distributed random variables. We obtain a random array $\boldsymbol{X}^\kappa \in \Omega^{\mathbb{N}\times\mathbb{N}}$ by drawing independently for any $i, j \in \mathbb{N}$ an element $\boldsymbol{X}^\kappa(i,j) \in \Omega$ from the distribution $\kappa_{\boldsymbol{s}_i,\boldsymbol{x}_j} \in \mathscr{P}(\Omega)$. Clearly, the distribution Ξ^κ of $\boldsymbol{X}^\kappa$ is exchangeable. By extension, a probability distribution π on $\mathfrak{K}$ induces an exchangeable distribution as well. Indeed, with $\kappa^\pi \in \mathfrak{K}$ drawn from π, we let $\Xi^\pi \in \mathfrak{X}$ be the distribution of the random array $\boldsymbol{X}^\pi \in \Omega^{\mathbb{N}\times\mathbb{N}}$ obtained by first drawing κ^π independently of the $(\boldsymbol{s}_k, \boldsymbol{x}_\ell)_{k,l\ge 1}$ and then drawing each entry $\boldsymbol{X}^\pi(i,j)$ from $\kappa^\pi_{\boldsymbol{s}_i,\boldsymbol{x}_j}$. We equip the space $\mathscr{P}(\mathfrak{K})$ of probability measures on $\mathfrak{K}$ with the weak topology.

Theorem 1.8. *The map $\mathscr{P}(\mathfrak{K}) \to \mathfrak{X}, \pi \mapsto \Xi^\pi$ is a homeomorphism.*

For the special case $\Omega = \{0,1\}$ Theorem 1.8 is the directed graph version of [19, Theorem 5.3].

For $\mu \in \mathfrak{L}$ let us write $\boldsymbol{X}^\mu$ for the exchangeable array $\boldsymbol{X}^{\kappa^\mu}$ induced by a kernel representation of μ. Suppose that $(\mu_N)_{N\geq 1}$ is a sequence of Ω-laws that converges to $\mu \in \mathfrak{L}$. Then Theorem 1.8 shows that for any $n \geq 1$ and for any $\tau = (\tau_{i,j})_{i,j\in[n]} \in \Omega^{n\times n}$,

$$\lim_{N\to\infty} \mathbb{P}\left[\forall i,j \in [n] : \boldsymbol{X}^{\mu_N}(i,j) = \tau_{i,j}\right] = \mathbb{P}\left[\forall i,j \in [n] : \boldsymbol{X}^{\mu}(i,j) = \tau_{i,j}\right]. \tag{1.5}$$

Conversely, if $\mu_N, \mu \in \mathfrak{L}$ are such that (1.5) holds for all n, τ, then Theorem 1.8 implies that $\lim_{N\to\infty} D_\boxtimes(\mu_N, \mu) = 0$. Thus, with $\Omega^{n\times n}$-matrices replacing subgraphs, Theorem 1.8, provides the probabilistic counterpart of the equivalence of subgraph counting and graphon convergence [32, Theorem 11.5].

Additionally, the theory of graph limits shows that a large enough random graph obtained from a graphon by sampling is close to the original graphon in the cut metric. There is a corresponding statement in the realm of probability distributions as well. Specifically, for an integer $n \geq 1$ let $\mu_n \in \mathscr{P}(\Omega^n)$ be the discrete probability distribution defined by

$$\mu_n(\sigma) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}\left\{\forall j \in [n] : \boldsymbol{X}^\mu(i,j) = \sigma_j\right\} \qquad (\sigma \in \Omega^n).$$

In words, μ_n is the empirical distribution of the rows of $(\boldsymbol{X}^\mu(i,j))_{i,j\in[n]}$. Strictly speaking, being dependent on the random coordinates $(\boldsymbol{s}_i, \boldsymbol{x}_j)_{i,j\geq 1}$, μ_n is a *random* probability distribution on Ω^n. The following theorem supplies a probabilistic version of the sampling theorem for graphons [6, Lemma 4.4].

Theorem 1.9. *There exists $c = c(\Omega) > 0$ such that for all $n > 1$ and all $\mu \in \mathfrak{L}$ we have $\mathbb{E}\left[D_\boxtimes(\mu, \mu_n)\right] \leq c/\sqrt{\log n}$.*

The following theorem implies that the dependence on n in Theorem 1.9 is best possible, apart from the value of the constant c.

Theorem 1.10. *There is a constant $c > 0$ such that for any $\varepsilon > 0$ there exists $\mu \in \mathfrak{L}$ such that $D_\boxtimes(\mu, \nu) \geq \varepsilon$ for all $\nu \in \mathscr{L}$ whose support contains at most $\exp(c/\varepsilon^2)$ configurations.*

1.2.4. *Extremality.* Among all the probability measures on the discrete domain Ω^n, the product measures are clearly the simplest. We will therefore be particularly interested in distributions that are close to product measures in the cut metric. To this end, for a probability measure μ on Ω^n we let

$$\bar{\mu}_i(\sigma) = \sum_{\tau\in\Omega^n} \mathbf{1}\{\tau_i = \sigma\}\,\mu(\tau) \qquad \text{for } \sigma \in \Omega, \text{ and} \qquad \bar{\mu} = \bigotimes_{i=1}^{n} \bar{\mu}_i.$$

Thus, $\bar{\mu}_i \in \mathscr{P}(\Omega)$ is the marginal distribution of the ith coordinate under the measure μ, and $\bar{\mu}$ is the product measure with the same marginals as μ. Then $\Delta_\boxtimes(\mu, \bar{\mu})$ gauges how 'similar' μ is to a product measure. To be precise, since the cut metric is quite weak, a 'small' value of $\Delta_\boxtimes(\mu, \bar{\mu})$ need not imply that μ behaves like a product measure in every respect. For instance, the entropy of μ might be much smaller than that of $\bar{\mu}$. But if $\Delta_\boxtimes(\mu, \bar{\mu})$ is small, then (1.5) implies that the joint distribution of a bounded number of randomly chosen coordinates of μ is typically close to a product measure in total variation distance.

A similar measure of proximity to a product distribution is meaningful on the space of Ω-laws as well. Formally, for $\mu \in \mathscr{L}$ define $\bar{\mu} \in \mathscr{L}$ as the atom concentrated on the single function

$$[0,1] \to \mathscr{P}(\Omega), \qquad x \mapsto \int_{\mathscr{S}} \sigma_x \mathrm{d}\mu(\sigma). \tag{1.6}$$

Since $D_\boxtimes(\bar{\mu}, \bar{\nu}) = 0$ whenever $D_\boxtimes(\mu, \nu) = 0$, (1.6) induces a map $\mu \in \mathfrak{L} \mapsto \bar{\mu} \in \mathfrak{L}$. The laws $\bar{\mu}$ with $\mu \in \mathfrak{L}$ represent the generalisation of discrete product measures. Since each $\mu \in \mathfrak{L}$ is represented by a distribution on $\mathscr{S}$ that places all the probability mass on a single point, we call the laws $\bar{\mu}$ *extremal*. Moreover, $\mu \in \mathfrak{L}$ is called *ε-extremal* if $D_\boxtimes(\mu, \bar{\mu}) < \varepsilon$. The following result summarises basic properties of extremal laws and of the map $\mu \mapsto \bar{\mu}$.

Theorem 1.11. *For all $\mu, \nu \in \mathscr{L}$ we have*

$$D_\boxtimes(\bar{\mu}, \bar{\nu}) \leq D_\boxtimes(\mu, \nu) \qquad \textit{and} \tag{1.7}$$

$$D_\boxtimes(\bar{\mu}, \bar{\nu}) \leq \max_{\omega\in\Omega} \int_0^1 \left| \int_{\mathscr{S}} \sigma_x \mathrm{d}\mu(\sigma) - \int_{\mathscr{S}} \sigma_x \mathrm{d}\nu(\sigma) \right| \mathrm{d}x \leq 2 D_\boxtimes(\bar{\mu}, \bar{\nu}). \tag{1.8}$$

Furthermore, the set of extremal laws is a closed subset of $\mathfrak{L}$.

1.2.5. *Pinning.* The regularity lemma constitutes one of the most powerful tools of modern combinatorics. In a nutshell, the lemma shows that any graph can be approximated by a mixture of a bounded number of 'simple' graphs, namely quasi-random bipartite graphs. We will present a corresponding result for probability measures, respectively laws. Specifically, we will show that any law can be approximated by a mixture of a small number of extremal laws. Indeed, we will show that actually this approximation can be obtained by a simple, mechanical procedure called 'pinning'. This is in contrast to the proof of the graphon regularity lemma, where the regular partition results from a delicate construction that involves tracking a potential function.

To describe the pinning procedure, consider $\mu \in \mathscr{L}$, $\theta \geq 1$, $x_1, \ldots, x_\theta \in [0,1]$ and $\tau \in \Omega^\theta$. Then we define

$$z_\mu(\tau, x_1, \ldots, x_\theta) = \int_{\mathscr{S}} \prod_{i=1}^{\theta} \sigma_{x_i}(\tau_i) \mathrm{d}\mu(\sigma).$$

Further, assuming that $z_\mu(\tau, x_1, \ldots, x_\theta) > 0$, we define a reweighted probability distribution $\mu_{\tau\downarrow x_1,\ldots,x_\theta}$ by

$$\mathrm{d}\mu_{\tau\downarrow x_1,\ldots,x_\theta}(\sigma) = \frac{1}{z_\mu(\tau, x_1, \ldots, x_\theta)} \prod_{i=1}^{\theta} \sigma_{x_i}(\tau_i) \mathrm{d}\mu(\sigma); \tag{1.9}$$

Thus, $\mu_{\tau\downarrow x_1,\ldots,x_\theta}$ is obtained by reweighting μ according to the 'reference configuration' τ, evaluated at the coordinates $x_1, \ldots, x_\theta$. For completeness we also let $\mu_{\tau\downarrow x_1,\ldots,x_\theta} = \mu$ if $z_\mu(\tau, x_1, \ldots, x_\theta) = 0$.

The effect of this reweighting procedure becomes particularly interesting if the reference configuration and the coordinates are chosen randomly. Specifically, let $\hat{\boldsymbol{x}}_1, \hat{\boldsymbol{x}}_2, \ldots \in [0,1]$ be uniform and mutually independent. Further, for an integer $\theta \geq 1$ draw $\hat{\boldsymbol{\tau}} = \hat{\boldsymbol{\tau}}^\mu \in \Omega^\theta$ from the distribution

$$\mathbb{P}\left[\hat{\boldsymbol{\tau}}^\mu = \tau \mid \hat{\boldsymbol{x}}_1, \ldots, \hat{\boldsymbol{x}}_\theta\right] = \frac{\boldsymbol{z}_\mu(\tau)}{\boldsymbol{z}_\mu}, \qquad \text{where} \quad \boldsymbol{z}_\mu(\tau) = \int_0^1 \prod_{i=1}^{\theta} \sigma_{\hat{\boldsymbol{x}}_i}(\tau_i) \mathrm{d}\mu(\sigma), \qquad \boldsymbol{z}_\mu = \sum_{\tau \in \Omega^\theta} \boldsymbol{z}_\mu(\tau). \tag{1.10}$$

Equivalently, and perhaps more intuitively, we can describe the choice of $\hat{\boldsymbol{\tau}}$ as follows. First, draw $\boldsymbol{\tau} \in \mathscr{S}$ from the distribution μ; then pick $\hat{\boldsymbol{\tau}}$ from the product measure $\boldsymbol{\tau}_{\boldsymbol{x}_1} \otimes \cdots \otimes \boldsymbol{\tau}_{\boldsymbol{x}_\theta} \in \mathscr{P}(\Omega^\theta)$. Now, having drawn the 'reference vector' $\hat{\boldsymbol{\tau}}$, we obtain the reweighted distribution $\mu_{\hat{\boldsymbol{\tau}}\downarrow\theta} = \mu_{\hat{\boldsymbol{\tau}}\downarrow\hat{\boldsymbol{x}}_1,\ldots,\hat{\boldsymbol{x}}_\theta}$ as defined in (1.9). Clearly, (1.10) guarantees that $\boldsymbol{z}_\mu(\hat{\boldsymbol{\tau}}) > 0$ almost surely. Finally, we define

$$\mu_{\downarrow\theta} = \mathbb{E}\left[\overline{\mu_{\hat{\boldsymbol{\tau}}\downarrow\theta}} \mid \boldsymbol{x}_1, \ldots, \boldsymbol{x}_\theta\right] \in \mathscr{L}.$$

Hence, $\mu_{\downarrow\theta}$ weights each possible outcome according to the probability of its reference configuration $\hat{\boldsymbol{\tau}}$. The discrete version of the operation $\mu \mapsto \mu_{\downarrow\theta}$ for $\mu \in \mathscr{L}_n$ was introduced in [13]. Following the terminology from that paper, we refer to the map $\mu \mapsto \mu_{\hat{\boldsymbol{\tau}}\downarrow\theta}$ as the *pinning operation*. The term is explained by the fact that in the discrete case, each of the products on the r.h.s. of (1.9) is either one or zero.

The next theorem shows that pinning furnishes a probabilistic equivalent of weak regular graphon partitions. To state this result, we observe that the pinning construction is well-defined on the space $\mathfrak{L}$ as well. To be precise, if $\mu, \nu \in \mathfrak{L}$ have cut distance zero, then $\mu_{\hat{\boldsymbol{\sigma}}\downarrow\theta}, \nu_{\hat{\boldsymbol{\sigma}}\downarrow\theta}$ are identically distributed, and so are $\mu_{\downarrow\theta}$ and $\nu_{\downarrow\theta}$. Consequently, we can apply the pinning operation directly to elements of the space $\mathfrak{L}$.

Theorem 1.12. *Let $0 < \varepsilon < 1$, let $\mu \in \mathfrak{L}$ and draw $0 \leq \boldsymbol{\theta} = \boldsymbol{\theta}(\varepsilon) \leq 64\varepsilon^{-8}\log|\Omega|$ uniformly and independently of everything else. Then $\mathbb{P}\left[\mu_{\hat{\boldsymbol{\tau}}\downarrow\boldsymbol{\theta}} \text{ is } \varepsilon\text{-extremal}\right] \geq 1-\varepsilon$ and $\mathbb{E}[D_{\boxtimes}(\mu, \mu_{\downarrow\boldsymbol{\theta}})] < \varepsilon$.*

Hence, the law $\mu_{\downarrow\boldsymbol{\theta}}$, a mixture of no more than $|\Omega|^{\boldsymbol{\theta}}$ extremal laws, likely provides an ε-approximation to μ.

1.2.6. *Continuity and overlaps.* There are certain natural operations on probability measures and, by extension, laws that turn out to be continuous with respect to the cut metric. First, we consider the construction of the product measure. For discrete measures $\mu, \nu \in \mathscr{L}_n(\Omega)$ we can view their product $\mu \otimes \nu$ as a probability distribution on $(\Omega \times \Omega)^n$ such that for any $\sigma_1, \tau_1, \ldots, \sigma_n, \tau_n \in \Omega$,

$$\mu \otimes \nu\left(\begin{pmatrix}\sigma_1\\ \tau_1\end{pmatrix}, \ldots, \begin{pmatrix}\sigma_n\\ \tau_n\end{pmatrix}\right) = \mu(\sigma_1, \ldots, \sigma_n)\nu(\tau_1, \ldots, \tau_n).$$

We extend this construction to laws by way of the kernel representation. To this end, let $\Lambda : [0,1] \to [0,1] \times [0,1]$, $x \mapsto (\Lambda_1(x), \Lambda_2(x))$ be a measurable bijection that maps the Lebesgue measure on $[0,1]$ to the Lebesgue measure

on $[0,1]^2$ such that, conversely, Λ^{-1} maps the Lebesgue measure on $[0,1]^2$ to the Lebesgue measure on $[0,1]$.[2] Following [14], for measurable maps $\kappa,\kappa' : [0,1]^2 \to \mathscr{P}(\Omega)$ we introduce

$$\kappa \otimes \kappa' : [0,1]^2 \to \mathscr{P}(\Omega^2), \qquad (s,x) \in [0,1]\times[0,1] \mapsto \kappa_{\Lambda_1(s),x} \otimes \kappa'_{\Lambda_2(s),x} \in \mathscr{P}(\Omega^2).$$

For any kernels $\kappa,\kappa',\kappa'',\kappa'''$ such that $D_{\boxtimes}(\kappa,\kappa'') = D_{\boxtimes}(\kappa',\kappa''') = 0$ we clearly have $D_{\boxtimes}(\kappa\otimes\kappa',\kappa''\otimes\kappa''') = 0$. Thus, the $\otimes$-operation is well defined on the kernel space $\mathfrak{K}$. Hence, due to Theorem 1.4 the construction extends to laws, i.e., given Ω-laws μ,ν we obtain an Ω^2-law $\mu\otimes\nu$. Furthermore, it is easy to see that for any $\mu,\nu\in\mathscr{L}_n(\Omega)$ the Ω^2-law representing the product measure $\mu\otimes\nu$ is precisely the $\otimes$-product of the laws $\dot\mu,\dot\nu$ representing μ,ν.

Theorem 1.13. *The map $(\mu,\nu)\in\mathfrak{L}(\Omega)\mapsto\mu\otimes\nu\in\mathfrak{L}(\Omega^2)$ is continuous.*

There is a second fundamental operation on distributions/laws that resembles the operation of obtaining a $n\times n$-rank one matrix from two vectors of length n. Specifically, for vectors $\sigma,\tau\in\Omega^{[n]}$ let $\sigma\oplus\tau\in(\Omega^2)^{[n]\times[n]}$ be the vector with entries $(\sigma\oplus\tau)_{ij} = (\sigma_i,\tau_j)$ for all $i,j\in[n]$. Additionally, for distributions $\mu,\nu\in\mathscr{L}_n(\Omega)$ let $\mu\oplus\nu$ be the distribution of the pair $\boldsymbol{\sigma}^\mu\oplus\boldsymbol{\tau}^\nu$ with $\boldsymbol{\sigma}^\mu,\boldsymbol{\tau}^\nu\in\Omega^n$ chosen from μ,ν, respectively.

We extend the $\oplus$-operation to kernels as follows. For $\kappa,\kappa' : [0,1]^2\to\mathscr{P}(\Omega)$ let

$$\kappa\oplus\kappa' : [0,1]^2\to\mathscr{P}(\Omega^2), \qquad (s,x)\mapsto\kappa_{s,\Lambda_1(x)}\otimes\kappa_{s,\Lambda_2(x)}.$$

It is easy to see that for $\kappa,\kappa',\kappa'',\kappa'''$ with $D_{\boxtimes}(\kappa,\kappa'') = D_{\boxtimes}(\kappa',\kappa''') = 0$ we have $D_{\boxtimes}(\kappa\oplus\kappa',\kappa''\oplus\kappa''') = 0$. Hence, the $\oplus$-operation is well-defined on the space $\mathfrak{K}$ and thus, due to Theorem 1.4, on the space $\mathfrak{L}$ as well. Moreover, for discrete measures $\mu,\nu\in\mathscr{P}(\Omega^n)$ one verifies immediately that the law representing $\mu\oplus\nu$ coincides with $\dot\mu\oplus\dot\nu$.

Theorem 1.14. *The map $\mathfrak{L}(\Omega)\to\mathfrak{L}(\Omega^2)$, $(\mu,\nu)\mapsto\mu\oplus\nu$ is continuous.*

Theorems 1.13 and 1.14 immediately imply the continuity of further functionals that play a fundamental role in mathematical physics. Specifically, let $\sigma_1,\dots,\sigma_n\in\mathscr{S}$. For $\sigma_1,\dots,\sigma_n\in\mathscr{S}$ and $\omega_1,\dots,\omega_n\in\Omega$ we define

$$R_{\omega_1,\dots,\omega_n}(\sigma_1,\dots,\sigma_n) = \int_0^1\prod_{i=1}^n\sigma_{i,x}(\omega_i)\mathrm{d}x.$$

Furthermore, for $\mu\in\mathscr{L}$ and $\ell\geq 1$ we define

$$R_{\ell,\omega_1,\dots,\omega_n}(\mu) = \int_{\mathscr{S}}\cdots\int_{\mathscr{S}} R_{\omega_1,\dots,\omega_n}(\sigma_1,\dots,\sigma_n)^\ell\,\mathrm{d}\mu(\sigma_1)\cdots\mathrm{d}\mu(\sigma_n).$$

Additionally, let $R_{\ell,n}(\mu) = (R_{\ell,\omega_1,\dots,\omega_n}(\mu))_{\omega_1,\dots,\omega_n\in\Omega}$. In physics jargon, the arrays $R_{\ell,n}(\mu)$ are known as *multi-overlaps* of μ. Since $R_{\ell,n}(\mu) = R_{\ell,n}(\nu)$ if $D_{\boxtimes}(\mu,\nu) = 0$, the multi-overlaps are well-defined on the space $\mathfrak{L}$ of laws.

Corollary 1.15. *The functions $\mu\in\mathfrak{L}\mapsto R_{\ell,n}(\mu)$ with $\ell,n\geq 1$ are continuous.*

1.3. **Discussion and related work.** Borgs, Chayes, Lovász, Sós, Szegedy and Vesztergombi launched the theory of (dense) graph limits in a series of important and influential articles [6, 7, 33, 34, 35]. Lovász [32] provides a unified account of the state of the art up to about 2012. Moreover, Janson [28] gives an excellent account of the measure-theoretic foundations of the theory of graph limits and some of its generalisation.

Given the many areas of application where sequences of probability measures on increasingly large discrete cubes appear, the most prominent example being perhaps the study of Boltzmann distributions in mathematical physics, it is unsurprising that attempts have been made to construct limiting objects for such sequences. The theory of Gibbs measures embodies the classical, physics-inspired approach to this task [23]. Here the aim is to construct and classify all possible 'infinite-volume' limits of Boltzmann distributions defined on spatial structures such as trees or lattices. The limiting objects are called *Gibbs measures*. A fundamental question, whose ramifications extend from the study of phase transitions in physics to the computational complexity of counting and sampling, is whether there is a unique Gibbs measure that satisfies all the finite-volume conditional equations (e.g, [22, 46, 47]). However, since the theory of Gibbs measures is confined to systems with an underlying lattice-like geometry, numerous applications are beyond its reach. For instance, Marinari et al. [38] argued that the classical theory of Gibbs measures does not provide an appropriate framework for the study of (diluted) mean-field models such as the Sherrington-Kirkpatrick model, the Viana-Bray model or the hardcore model on a sparse random graph. Further examples of 'non-spatial' sequences of distributions abound in computer science, statistics and data science.

[2] The existence of such a Λ follows from Lemma 2.3 below.

Panchenko [43, 44] employed the more abstract Aldous-Hoover representation of exhangeable arrays in his work on mean-field models [1, 25]. Kallenberg's monograph [29] provides the definite treatment of this abstract theory. Furthermore, Austin [4] extends and generalises the concept of exchangeable arrays and discusses applications to the Viana-Bray spin glass model. The close relationship between the theory of graph limits and exchangeable arrays was first noticed by Diaconis and Janson [19]. Their [19, Theorem 9.1] is essentially a directed graph version of Theorem 1.8 in the special case $\Omega = \{0,1\}$. Moreover, the appendix of Panchenko's monograph [43] also contains a proof of the Aldous-Hoover representation theorem via graph limits.

Although the connection between genuinely probabilistic constructions such as the Aldous-Hoover representation and graph limits was noticed in prior work [4, 19, 43], those contributions stopped short of working out a fully-fledged adaptation of the theory of graph limits to a limit theory for probability measures on discrete cubes. A prior article by Coja-Oghlan, Perkins and Skubch [14] made a first cursory attempt at filling this gap and already contained the definition (1.2) of the cut metric and of the space $\mathfrak{L}$ of laws. Additionally, the compactness of the space $\mathfrak{L}$ (Theorem 1.1) and a weaker version of the kernel representation (Theorem 1.4) were stated in [14], although no detailed proofs were given. Furthermore, a definition similar to the discrete cut metric (1.1) was devised in [13] and a statement similar to Theorem 1.14 was previously proved by Coja-Oghlan and Perkins [12, Proposition A.2]. Finally, versions of the pinning operation for discrete probability measures appeared in [11, 40, 45] and recently Eldad [20] devised an extension to subspaces of $\mathbb{R}^n$, i.e., to the case of spins that need not take discrete values.

The contribution of the present paper is that we expressly and explicitly adapt and extend the concepts of the theory of graph limits to the context of probability distributions on increasing sequences of discrete cubes. We present in a unified way the proofs of the most important basic facts such as the relationship between the discrete and the continuous cut metric (Theorem 1.2), the kernel representation (Theorem 1.4), the sampling theorem (Theorem 1.9) and the continuity of product measures (Theorems 1.13 and 1.14). The proofs of these results are based on extensions and adaptations of techniques from the theory of graph limits. Moreover, we present a self-contained derivation of the representation theorem for exchangeable arrays (Theorem 1.8). The added value by comparison to prior work [14, 19] is that here we present detailed, unified proofs that operate directly in the probabilistic setting, rather than by extensive allusion to the graphon space. Additionally, we present a self-contained proof of the compactness result (Theorem 1.1). While the argument set out in, e.g., [32, Chapter 9] could be adapted to the probabilistic setting, we present a different argument based on analytic techniques that might be of independent interest. But the main technical novelty is certainly the pinning theorem (Theorem 1.12) that generalises the discrete version from [11]. The proof is delicate and uses many of the other, more basic results.

The pinning operation from Theorem 1.12 is somewhat reminiscent of Tao's construction of regular partitions [49] and of the construction of Lovász and Szegedy [35]. For example, Tao's construciton of a regular partition is based on sampling a number θ of vertices of a graph G and then partitioning the remaining vertices into 2^θ classes according to their adjacencies with the reference vertices. The discrete version pinning operation from [11, 45] proceeds similarly; see Theorem 4.1 below, except that the number of pinned coordinates $\boldsymbol{\theta}$ is chosen randomly, rather than deligently given G. The same is true of the number of pinned coordinates in Theorem 1.12, which additionally yields a continuous version applicable to general Ω-laws.

Finally, there have been several further related contributions that extend the classical (dense) theory of graph limits as set out in [32]. Just as the classical theory, these extensions partly have a probabilistic component as they incorporate random graphs. For example, the important L^p-theory of sparse graph convergence covers limit objects of exchangeable sparse graphs [9]. A further contribution pertinent to sparse random graphs is the work of Crane and Dempsey [17] and Cai, Campbell and Broderick [15] on edge-exchangeability. Moreover, the articles [8, 50] deal with graphexes, which are limit objects of random geometric graphs. Further important extensions of the theory of graph limits include the work of Nešetřil and Ossona de Mendez [41] on convergence of sparse graphs that satisfy first order formulas, the paper of Hoppen, Kohayakawa, Moreira, Rath and Sampaio [26] on sequences of permutations (permutons), the article by Coregliano and Razborov [16] on limits on dense combinatorial objects, and Janson's work [27] on limits of posets. Some of these contributions, as well as Austin's work [3, 4] involve stronger versions of exchangeability than the classical de Finetti or Aldous-Hoover notions of exchangeability.

1.4. **Outline.** After presenting the necessary background and notation in Section 2, in Section 3 we will prove the basic facts about laws and the cut metric stated above. Specifically, Section 3 contains the proofs of Theorems 1.1,

1.4, 1.8, 1.9, 1.11, 1.13 and 1.14. Subsequently, in Section 4 we prove Theorem 1.12, which constitutes the main technical contribution of the paper. Finally, in Section 4.4 we establish Theorem 1.2.

2. Preliminaries

2.1. **Measure theory.** Throughout the paper we continue to denote by λ the Lebesgue measure on the Euclidean space $\mathbb{R}^k$; the reference to k will always be clear from the context. For the convenience of the reader we collect a few basic facts from measure theory that we will need. The first lemma follows from the Isomorphism Theorem, see e.g. [30, Sec. 15.B].

Lemma 2.1. *Suppose that $\mathscr{E} = (X, \mathscr{A}, \mu)$ is a standard Borel space equipped with a probability measure μ. Then there exists a measurable map $f : [0,1] \to X$ that maps the Lebesgue measure to μ.*

Lemma 2.2 (Theorem 3.2 of [37])**.** *Suppose that $\mathscr{E} = (X, \mathscr{A})$, $\mathscr{E}' = (X', \mathscr{A}')$ are standard Borel spaces and that $f : X \to X'$ is a measurable bijection. Then its inverse f^{-1} is measurable.*

Lemma 2.3 (Theorem A.7 of [28])**.** *If $(\mathscr{X}, \mu)$ is an atomless Borel probability space and λ is the Lebesgue measure, then there is a measure preserving bijection of $(\mathscr{X}, \mu)$ to $([0,1], \lambda)$.*

The following is the Riesz-Markov-Kakutani representation theorem [24].

Lemma 2.4. *Suppose that $\mathscr{E}_0$ is a compact metric space and that $\varphi : C(\mathscr{E}_0) \to \mathbb{R}$ is a positive linear functional on the space of continuous functions $C(\mathscr{E}_0)$ on $\mathscr{E}_0$. Moreover, assume that $\varphi(\mathbf{1}) = 1$. Then there exists a unique probability measure μ on $\mathscr{E}_0$ such that $\varphi(f) = \int_{\mathscr{E}_0} f \, \mathrm{d}\mu$ for all $f \in C(\mathscr{E}_0)$.*

We will need Lemma 2.4 in Section 3.4 to prove the completeness of the space of laws with respect to the cut metric.

Additionally, in several places throughout the paper we will need the following metric on probability measures. Suppose that $(\mathscr{E}, D)$ is a complete separable metric space and that D is bounded. Then the space $\mathscr{P}(\mathscr{E})$ of probability measures on $\mathscr{E}$ equipped with the *Wasserstein metric*

$$\mathscr{D}(\mu, \nu) = \inf\left\{ \int_{\mathscr{E} \times \mathscr{E}} D(x,y) \mathrm{d}\gamma(x,y) : \gamma \in \Gamma(\mu, \nu) \right\}, \tag{2.1}$$

where we recall that $\Gamma(\mu, \nu)$ is the set of all couplings of μ, ν, also is complete and separable. The Wasserstein metric induces the weak topology on $\mathscr{P}(\mathscr{E})$ [51, Theorem 6.9]. The definition (2.1) extends to $\mathscr{E}$-valued random variables $\boldsymbol{X}, \boldsymbol{Y}$, for which we define

$$\mathscr{D}(\boldsymbol{X}, \boldsymbol{Y}) = \inf\left\{ \int_{\mathscr{E} \times \mathscr{E}} D(x,y) \mathrm{d}\gamma(x,y) : \gamma \in \Gamma(\boldsymbol{X}, \boldsymbol{Y}) \right\},$$

with $\Gamma(\boldsymbol{X}, \boldsymbol{Y})$ denoting the set of all couplings of $\boldsymbol{X}, \boldsymbol{Y}$. We will frequently be working with the Wasserstein metric $\mathscr{D}_{\boxtimes}(\cdot, \cdot)$ induced by the cut metric on $\mathfrak{L}$ or $\mathfrak{K}$.

2.2. **Variations on the cut metric.** When we defined the cut metric $D_{\boxtimes}(\mu, \nu)$ in (1.2) we allowed for a coupling of μ, ν as well as a 'coordinate permutation' $\varphi \in \mathbb{S}$. Sometimes the latter is not desirable. Therefore, for $\mu, \nu \in \mathscr{L}$ we define the *strong cut distance* as

$$D_{\square}(\mu, \nu) = \inf_{\gamma \in \Gamma(\mu,\nu)} \sup_{\substack{S \subset \mathscr{S} \times \mathscr{S} \\ X \subset [0,1] \\ \omega \in \Omega}} \left| \int_S \int_X (\sigma_x(\omega) - \tau_x(\omega)) \, \mathrm{d}x \mathrm{d}\gamma(\sigma, \tau) \right| \tag{2.2}$$

with S, X ranging over measurable sets. It is easily verified that $D_{\square}(\cdot, \cdot)$ is a pre-metric on $\mathscr{L}$. Analogously, for $\mu, \nu \in \mathscr{P}(\Omega^n)$ let

$$\Delta_{\square}(\mu, \nu) = \inf_{\gamma \in \Gamma(\mu,\nu)} \sup_{\substack{S \subset \Omega^n \times \Omega^n \\ X \subset [n] \\ \omega \in \Omega}} \frac{1}{n} \left| \sum_{\substack{(\sigma,\tau) \in S \\ x \in X}} \gamma(\sigma, \tau) \left(\mathbf{1}\{\sigma_x = \omega\} - \mathbf{1}\{\tau_x = \omega\} \right) \right|. \tag{2.3}$$

Similarly, we will be led to consider several variants of the kernel cut metric from (1.3). Specifically, let $\mathscr{K}_{\mathbb{R}} = \mathscr{K}_{\mathbb{R}}(\Omega)$ be the set of all maps $\kappa, \kappa' : [0,1]^2 \to \mathbb{R}^\Omega$ such that the functions $(s,x) \in [0,1]^2 \mapsto \kappa_{s,x}(\omega)$ belong to $L^1([0,1]^2, \mathbb{R})$ for all $\omega \in \Omega$, up to equality almost everywhere. Then for $\kappa, \kappa' \in \mathscr{K}_{\mathbb{R}}$ we define

$$D_{\boxtimes}(\kappa,\kappa') = \inf_{\varphi,\psi\in\mathbb{S}} \sup_{\substack{S,X\subset[0,1]\\ \omega\in\Omega}} \left|\int_S\int_X \left(\kappa_{s,x}(\omega) - \kappa'_{\varphi(s),\psi(x)}(\omega)\right) \mathrm{d}x\mathrm{d}s\right|,$$

$$D_{\boxslash}(\kappa,\kappa') = \inf_{\varphi\in\mathbb{S}} \sup_{\substack{S,X\subset[0,1]\\ \omega\in\Omega}} \left|\int_S\int_X \left(\kappa_{s,x}(\omega) - \kappa'_{\varphi(s),x}(\omega)\right) \mathrm{d}x\mathrm{d}s\right|,$$

$$D_{\square}(\kappa,\kappa') = \sup_{\substack{S,X\subset[0,1]\\ \omega\in\Omega}} \left|\int_S\int_X \left(\kappa_{s,x}(\omega) - \kappa'_{s,x}(\omega)\right) \mathrm{d}x\mathrm{d}s\right|,$$

$$D_{\blacksquare}(\kappa,\kappa') = \inf_{\varphi\in\mathbb{S}} \sup_{\substack{S,X\subset[0,1]\\ \omega\in\Omega}} \left|\int_S\int_X \left(\kappa_{s,x}(\omega) - \kappa'_{\varphi(s),\varphi(x)}(\omega)\right) \mathrm{d}x\mathrm{d}s\right|.$$

Thus, $D_{\boxtimes}(\cdot,\cdot)$ is the natural extension of (1.3) to $\mathscr{K}_{\mathbb{R}}$, $D_{\boxslash}(\cdot,\cdot)$ is the kernel version of (2.2), $D_{\square}(\cdot,\cdot)$ represents the strongest variant of the cut metric that does not allow for any measure-preserving transformations, and $D_{\blacksquare}(\cdot,\cdot)$ is the graphon cut metric as studied in [33]. We also recall the graphon cut (pre-)metric for L^1-functions $\kappa, \kappa' : [0,1]^2 \to \mathbb{R}$ from [32], which is defined as

$$D_{\blacksquare}(\kappa,\kappa') = \inf_{\varphi\in\mathbb{S}} \sup_{S,X\subset[0,1]} \left|\int_S\int_X \left(\kappa_{s,x} - \kappa'_{\varphi(s),\varphi(x)}\right) \mathrm{d}x\mathrm{d}s\right|.$$

The different variants of the cut metric are related as follows. For a measurable map $\varphi : [0,1] \to [0,1]$ and $\kappa \in \mathscr{K}_{\mathbb{R}}$ define $\kappa_\varphi, \kappa^\varphi \in \mathscr{K}_{\mathbb{R}}$ by letting $\kappa_{\varphi\, s,x} = \kappa_{s,\varphi(x)}$ and $\kappa^\varphi_{s,x} = \kappa_{\varphi(s),x}$, respectively. Then

$$D_{\boxtimes}(\kappa,\kappa') = \inf_{\psi\in\mathbb{S}} D_{\boxslash}(\kappa,\kappa'_\psi), \qquad\qquad D_{\boxslash}(\kappa,\kappa') = \inf_{\varphi\in\mathbb{S}} D_{\square}(\kappa,\kappa'^\varphi). \tag{2.4}$$

As a consequence, for all $\kappa, \kappa' \in \mathscr{K}_{\mathbb{R}}$ we have

$$D_{\boxtimes}(\kappa,\kappa') \le D_{\boxslash}(\kappa,\kappa') \le D_{\square}(\kappa,\kappa') \qquad \text{and} \qquad D_{\blacksquare}(\kappa,\kappa') \ge D_{\boxtimes}(\kappa,\kappa'). \tag{2.5}$$

For a function $W : (s,x) \mapsto W_{s,x}$ defined on $[0,1]^2$ we define the *transpose* $W^\dagger : (s,x) \mapsto W_{x,s}$. We call W *symmetric* if $W = W^\dagger$. For $\kappa \in \mathscr{K}_{\mathbb{R}}$ we define a family $(\kappa^{(\omega)})_{\omega\in\Omega}$ of symmetric functions defined by

$$\kappa^{(\omega)}_{s/2,(1+x)/2} = \kappa_{s,x}(\omega), \qquad \kappa^{(\omega)}_{(1+s)/2,x/2} = \kappa_{x,s}(\omega), \qquad \kappa^{(\omega)}_{s/2,x/2} = \kappa^{(\omega)}_{(1+s)/2,(1+x)/2} = 0. \tag{2.6}$$

We can interpret $\kappa^{(\omega)}_{s,x}$ as the edge weight in a bipartite graph with vertex set $[0,1]$. We stress, that in (2.6) the ordering of s and x is quite delicate.

Lemma 2.5. *For all $\kappa \in \mathscr{K}_{\mathbb{R}}$ we have $D_{\square}(\kappa,\kappa') = 2\max_{\omega\in\Omega} D_{\square}(\kappa^{(\omega)}, \kappa'^{(\omega)})$.*

Proof. Given $\omega \in \Omega$ and $S, X \subset [0,1]$ let $T = \{(1+s)/2 : s \in S\} \cup \{x/2 : x \in X\}$, $Y = \{(1+x)/2 : x \in X\} \cup \{s/2 : s \in S\}$. Then by construction

$$2\left|\int_T\int_Y \left(\kappa^{(\omega)}_{s,x} - \kappa'^{(\omega)}_{s,x}\right) \mathrm{d}s\mathrm{d}x\right| = \left|\int_S\int_X \left(\kappa_{s,x}(\omega) - \kappa'_{s,x}(\omega)\right) \mathrm{d}s\mathrm{d}x\right|. \tag{2.7}$$

Hence, $D_{\square}(\kappa,\kappa') \le 2\max_{\omega\in\Omega} D_{\square}(\kappa^{(\omega)}, \kappa'^{(\omega)})$. Regarding the converse bound, we may assume by symmetry that $T, Y \subset [0,1]$ satisfy $T = 1 - Y$. Indeed, the choice $T = 1 - Y$ incorporates that the upper right part of the kernel is the transposed lower left part. Therefore, letting $S = \{2t-1 : t \in T \cap [1/2,1]\}$, $X = \{2t : t \in T \cap [0,1/2]\}$ we again obtain (2.7), and thus $D_{\square}(\kappa,\kappa') \ge 2\max_{\omega\in\Omega} D_{\square}(\kappa^{(\omega)}, \kappa'^{(\omega)})$. □

Remark 2.6. *Clearly, the cut metric from the theory of graph limits $D_{\blacksquare}(\cdot,\cdot)$ can be bounded from below by the present definition $D_{\boxtimes}(\cdot,\cdot)$, as one must apply the same measure-preserving transformation on both axes. On the other hand, for $\Omega = \{0,1\}$, once we turn kernels $\kappa, \kappa' \in \mathfrak{K}$ into 'bipartite graphons' via* (2.6)*, we find directly*

$$D_{\blacksquare}(\kappa^{(1)}, \kappa'^{(1)}) \le \frac{1}{2} D_{\boxtimes}(\kappa,\kappa').$$

The converse bound does not hold for any constant as can be seen as follows. Let $\kappa_{s,x} = \mathbf{1}\{s < 1/2\}$ *and* $\kappa'_{s,x} = \kappa_{x,s}$*. By choosing the measure preserving map* $\varphi(x) = 1 - x$*, we get*

$$D_{\blacksquare}(\kappa, \kappa') \le \sup_{S,X \subset [0,1]} \left| \int_S \int_X \left(\kappa_{s,x} - \kappa'_{\varphi(s),\varphi(x)} \right) \mathrm{d}x \mathrm{d}s \right| = 0.$$

But as κ' *represents the law* ν *supported only on* δ_σ *with* $\sigma_x = \mathbf{1}\{x \le 1/2\}$*, whilst* κ' *is the uniform distribution over the two configurations* $\sigma_1 = 1$ *and* $\sigma_0 = 0$*, we can bound* $D_{\boxtimes}(\kappa, \kappa') \ge 1/4$*.*

For $\kappa, \kappa' \in L^1([0,1]^2, \mathbb{R})$ we define

$$\|\kappa\|_\square = \sup_{S,X \subset [0,1]} \left| \int_S \int_X \kappa_{s,x} \mathrm{d}s \mathrm{d}x \right|, \qquad D_\square(\kappa, \kappa') = \left\| \kappa - \kappa' \right\|_\square = \sup_{S,X \subset [0,1]} \left| \int_S \int_X \left(\kappa_{s,x} - \kappa'_{\varphi(s),x} \right) \mathrm{d}x \mathrm{d}s \right|. \tag{2.8}$$

Then $\|\cdot\|_\square$ is a norm on $L^1([0,1]^2, \mathbb{R})$. Analogously, for a matrix $A \in \mathbb{R}^{n \times n}$ we define

$$\|A\|_\square = \frac{1}{n^2} \max_{S,X \subset [n]} \left| \sum_{s \in S} \sum_{x \in X} A_{sx} \right|. \tag{2.9}$$

We need the following 'sampling lemma' for the cut norm.

Lemma 2.7 ([32, Lemma 10.6])**.** *Suppose that* $\kappa : [0,1]^2 \to [-1,1]$ *is symmetric. Let* $\boldsymbol{x}_1, \dots, \boldsymbol{x}_k$ *be independently and uniformly chosen from* $[0,1]$*. Denote by* $\kappa[k] \in [-1,1]^{k \times k}$ *the matrix with entries* $\kappa_{i,j}[k] = \kappa_{\boldsymbol{x}_i, \boldsymbol{x}_j}$*. Then*

$$\mathbb{P}\left[\|\kappa[k]\|_\square \le \|\kappa\|_\square + 8/k^{1/4} \right] \ge 1 - 4\exp(-\sqrt{k}/10).$$

2.3. **The L_1-metric.** We define a subspance of $\mathscr{K}_\mathbb{R}$ by letting

$$\mathscr{K}_1 = \left\{ \kappa \in \mathscr{K}_\mathbb{R} : 0 \le \kappa_{s,x}(\omega) \le 1 \right\}.$$

Similarly, we let $\mathscr{S}_1$ be the space of all measurable functions $\sigma : [0,1] \to [0,1]^\Omega$. Further, we denote the L_1-metric on $\mathscr{K}_1$ and $\mathscr{S}_1$ by $D_1(\cdot,\cdot)$. Thus,

$$D_1(\kappa, \kappa') = \sum_{\omega \in \Omega} \int_0^1 \int_0^1 \left| \kappa_{s,x}(\omega) - \kappa'_{s,x}(\omega) \right| \mathrm{d}x \mathrm{d}s \qquad (\kappa, \kappa' \in \mathscr{K}_1),$$

and similarly for $\mathscr{S}_1$.

2.4. **Regularity.** For a kernel $\kappa \in \mathscr{K}_\mathbb{R}$ and partitions $S = (S_1, \dots, S_k)$, $X = (X_1, \dots, X_\ell)$ of the unit interval into pairwise disjoint measurable subsets define $\kappa^{S,X} \in \mathscr{K}$ by

$$\kappa^{S,X}_{s,x}(\omega) = \sum_{i \in [k] : \lambda(S_i) > 0} \sum_{j \in [\ell] : \lambda(X_j) > 0} \frac{\mathbf{1}\{(s,x) \in S_i \times X_j\}}{\lambda(S_i)\lambda(X_j)} \int_{S_i} \int_{X_i} \kappa_{t,y}(\omega) \mathrm{d}y \mathrm{d}t.$$

In words, $\kappa^{S,X}_{s,x}$ is the conditional expectation of $\kappa_{\boldsymbol{s},\boldsymbol{x}}$ given the σ-algebra generated by the rectangles $S_i \times X_j$. If the two partitions S, X are identical, we write κ^S instead of $\kappa^{S,X}$. We use similar notation for maps $\kappa : [0,1]^2 \to \mathbb{R}$. The following fact is a kernel variant of the well-known Frieze-Kannan regularity lemma.

Lemma 2.8 ([32, Corollary 9.13])**.** *For every symmetric* $\kappa : [0,1]^2 \to [0,1]$ *and every* $k \ge 1$ *there exists a partition* $S = (S_1, \dots, S_k)$ *of* $[0,1]$ *into pairwise disjoint measurable sets such that* $D_\square(\kappa, \kappa^S) \le 2/\sqrt{\log k}$*.*

This notion of regularity is robust with respect to refining the partition.

Lemma 2.9 ([32, Lemma 9.12])**.** *Let* $\kappa : [0,1]^2 \to [0,1]$ *be symmetric and* $\kappa' : [0,1]^2 \to [0,1]$ *be a symmetric step function and denote by* S *a partition of* $[0,1]$ *into a finite number of meausrable sets on which* κ' *is constant. Then* $D_\square(\kappa, \kappa^S) \le 2D_\square(\kappa, \kappa')$*.*

Applying Lemma 2.9 to the step function κ^R for a partition R that refines a partition S of $[0,1]$, we obtain the following corollary.

Corollary 2.10. *Let* R, S *be partitions of* $[0,1]$ *such that* R *refines* S*. Then* $D_\square(\kappa, \kappa^R) \le 2D_\square(\kappa, \kappa^S)$*.*

Proof. This follows from Lemma 2.9 because κ^S is constant on the partition classes of R. □

3. Fundamentals

This section contains the proofs of the basic facts, namely the compactness of the space of Ω-laws (Theorem 1.1), the isometric property of the kernel representation (Theorem 1.4), the sampling theorem (Theorem 1.9), the comparison of the discrete and the continuous cut metric (Theorem 1.2), the continuity statements from Theorems 1.13 and 1.14 and the connection to exchangeable arrays (Theorem 1.8). We begin with the proof of Theorem 1.4.

3.1. **Proof of Theorem 1.4.** Any measurable map $f:[0,1]\to\mathscr{S}$, $s\mapsto f_s$ induces a kernel $\kappa^f:[0,1]^2\to\mathscr{P}(\Omega)$, $(s,x)\mapsto f_{s,x}\in\mathscr{P}(\Omega)$. Moreover, f maps the Lebesgue measure on $[0,1]$ to a probability distribution $\mu^f\in\mathscr{L}$.

Lemma 3.1. *Suppose that $f,g:[0,1]\to\mathscr{S}$ are measurable. Then $D_{\square}(\mu^f,\mu^g)\le D_{\square}(\kappa^f,\kappa^g)$.*

Proof. Fix $\omega\in\Omega$ and $\varphi\in\mathbb{S}$. The construction of κ^f,κ^g guarantees that with $\boldsymbol{s}\in[0,1]$ chosen uniformly at random, the distribution γ of the pair $(\kappa^f_{\boldsymbol{s}},\kappa^g_{\varphi(\boldsymbol{s})})\in\mathscr{S}\times\mathscr{S}$ is a coupling of μ^f,μ^g. We now claim that

$$\sup_{T\subset\mathscr{S}^2,X\subset[0,1]}\left|\int_T\int_X(\sigma_x(\omega)-\tau_x(\omega))\,\mathrm{d}x\mathrm{d}\gamma(\sigma,\tau)\right|\le\sup_{S,X\subset[0,1]}\left|\int_S\int_X\left(\kappa^f_{s,x}(\omega)-\kappa^g_{s,\varphi(x)}(\omega)\right)\mathrm{d}x\mathrm{d}s\right|. \tag{3.1}$$

Indeed, fix measurable $T\subset\mathscr{S}^2$ and $X\subset[0,1]$ and let $S=\left\{s\in[0,1]:(\kappa^f_s,\kappa^g_{\varphi(s)})\in T\right\}$. Then by the construction of γ,

$$\int_T\int_X(\sigma_x(\omega)-\tau_x(\omega))\,\mathrm{d}x\mathrm{d}\gamma(\sigma,\tau)=\int_S\int_X\left(\kappa^f_{s,x}(\omega)-\kappa^g_{s,\varphi(x)}(\omega)\right)\mathrm{d}x\mathrm{d}s,$$

whence (3.1) follows. Finally, since (3.1) holds for all φ,ω, we conclude that $D_{\square}(\mu^f,\mu^g)\le D_{\square}(\kappa^f,\kappa^g)$. □

The following lemma establishes the converse of Lemma 3.1 for functions that take only finitely many values.

Lemma 3.2. *Suppose that $f,g:[0,1]\to\mathscr{S}$ are measurable maps whose images $f([0,1]),g([0,1])\subset\mathscr{S}$ are finite sets. Then $D_{\square}(\kappa^f,\kappa^g)\le D_{\square}(\mu^f,\mu^g)$.*

Proof. Suppose that $f([0,1])=\{\sigma_1,\ldots,\sigma_k\}$ and $g([0,1])=\{\tau_1,\ldots,\tau_\ell\}$. Moreover, let V_i be the set of all $s\in[0,1]$ such that $f(s)=\sigma_i$ and let W_j be the set of all $s\in[0,1]$ such that $g(s)=\tau_j$. In addition, let $v_i=\lambda(V_i)$, $w_j=\lambda(W_j)$. Then

$$\mu^f=\sum_{i=1}^k v_i\delta_{\sigma_i},\qquad\qquad\mu^g=\sum_{j=1}^\ell w_j\delta_{\tau_j}.$$

Consequently, any coupling γ of μ^f,μ^g induces a coupling $\Gamma\in\mathscr{P}([k]\times[\ell])$ of the probability distributions $(v_1,\ldots,v_k)$ and $(w_1,\ldots,w_\ell)$. To turn Γ into a measure-preserving map $[0,1]\to[0,1]$ we partition any sets V_i,W_j into pairwise disjoint measurable subsets $(V_{i,h})_{h\in[\ell]}$ and $(W_{h,j})_{h\in[k]}$, respectively, such that for all i,j,h,

$$\lambda(V_{i,h})=g(i,h),\qquad\qquad\lambda(W_{h,j})=g(h,j).$$

Then by Lemma 2.3 for any i,j there exists a bijection $\varphi_{i,j}:V_{i,j}\to W_{i,j}$ such that both $\varphi_{i,j}$ and $\varphi_{i,j}^{-1}$ are measurable and preserve the Lebesgue measure. Piecing these maps together, we obtain the bijection

$$\varphi:[0,1]\to[0,1],\qquad\qquad s\mapsto\sum_{(i,j)\in[k]\times[\ell]}\mathbf{1}\left\{s\in V_{i,j}\right\}\varphi_{i,j}(s).$$

Both φ and φ^{-1} are measurable and preserve the Lebesgue measure, i.e., $\varphi\in\mathbb{S}$. Moreover, for any sets $S,X\subset[0,1]$ and any $\omega\in\Omega$ we have

$$\int_S\int_X\left(\kappa^f_{s,x}(\omega)-\kappa^g_{\varphi(s),x}(\omega)\right)\mathrm{d}x\mathrm{d}s=\sum_{i=1}^k\sum_{j=1}^\ell\lambda(S\cap V_{ij})\int_X\left(\sigma_{i,x}(\omega)-\tau_{i,x}(\omega)\right)\mathrm{d}x. \tag{3.2}$$

Hence, (3.2) is extremised by sets S such that for all i,j either $V_{ij}\subset S$ or $S\cap V_{ij}=\emptyset$. For such a set S let $T=T(S)$ contain all pairs (i,j) such that $V_{ij}\subset S$. Then (3.2) yields

$$\left|\int_S\int_X\left(\kappa^f_{s,x}(\omega)-\kappa^g_{\varphi(s),x}(\omega)\right)\mathrm{d}x\mathrm{d}s\right|=\left|\sum_{(i,j)\in T}\Gamma(i,j)\int_X\left(\sigma_{i,x}(\omega)-\tau_{i,x}(\omega)\right)\mathrm{d}x\right|\le\sup_{U\subset\mathscr{S}^2}\left|\int_U\int_X(\sigma_x(\omega)-\tau_x(\omega))\,\mathrm{d}x\mathrm{d}\gamma(\sigma,\tau)\right|. \tag{3.3}$$

Since (3.3) holds for all S,X,ω,γ, the assertion follows. □

Corollary 3.3. *Let $f, g : [0,1] \to \mathscr{S}$ be measurable. Then $D_{\square}(\kappa^f, \kappa^g) \le D_{\square}(\mu^f, \mu^g)$.*

Proof. Because $\mathscr{S}$ is a convex subset of the separable Banach space $L^1([0,1], \mathbb{R}^\Omega)$, the measurable maps f, g are pointwise limits of sequences $(f_n)_{n\ge1}$, $(g_n)_{n\ge1}$ of measurable functions $f_n, g_n : [0,1] \to \mathscr{S}$ whose images are finite sets. Moreover, Lemma 3.2 implies that

$$D_{\square}(\mu^{f_n}, \nu^{f_n}) \ge D_{\square}(\kappa^{f_n}, \kappa^{g_n}) \qquad \text{for all } n \ge 1. \tag{3.4}$$

Further, for all $\omega \in \Omega$ and $S, X \subset [0,1]$ we have

$$\left| \int_S \int_X \left(\kappa^{f_n}_{s,x}(\omega) - \kappa^{f}_{s,x}(\omega) \right) \mathrm{d}s\mathrm{d}x \right| \le \int_0^1 \int_0^1 \left| \kappa^{f_n}_{s,x}(\omega) - \kappa^{f}_{s,x}(\omega) \right| \mathrm{d}s\mathrm{d}x. \tag{3.5}$$

Because $f_n \to f$ pointwise, the r.h.s. of (3.5) vanishes as $n \to \infty$. Consequently,

$$\lim_{n\to\infty} D_{\square}(\kappa^{f_n}, \kappa^f) = 0, \qquad \text{and similarly} \qquad \lim_{n\to\infty} D_{\square}(\kappa^{g_n}, \kappa^g) = 0. \tag{3.6}$$

Combining (3.6) with Lemma 3.1, we conclude that

$$\lim_{n\to\infty} D_{\square}(\mu^{f_n}, \mu^f) = 0, \qquad \lim_{n\to\infty} D_{\square}(\nu^{f_n}, \nu^f) = 0. \tag{3.7}$$

Finally, the assertion follows from (3.4), (3.6), (3.7) and the triangle inequality. □

Corollary 3.4. *For all $\kappa, \kappa' \in \mathscr{K}$ we have $D_{\square}(\mu^\kappa, \mu^{\kappa'}) = D_{\square}(\kappa, \kappa')$.*

Proof. This is an immediate consequence of Lemma 3.1 and Corollary 3.3. □

Proof of Theorem 1.4. Corollary 3.4 and (2.4) show that the map $\mathfrak{K} \to \mathfrak{L}, \kappa \mapsto \mu^\kappa$ is an isometry. Moreover, Lemma 2.1 implies that this map is surjective. Thus, because $\mathfrak{K}, \mathfrak{L}$ are metric spaces, $\kappa \mapsto \mu^\kappa$ is an isometric bijection. □

3.2. **Proof of Theorem 1.9.** We begin by extending Lemma 2.8 to (not necessarily symmetric) kernels $\kappa \in \mathscr{K}$.

Lemma 3.5. *There is $c = c(\Omega) > 0$ such that for any $\varepsilon \in (0,1)$, $\kappa \in \mathscr{K}$ there exist partitions $S = (S_1, \ldots, S_k)$, $X = (X_1, \ldots, X_\ell)$ of the unit interval into measurable subsets such that $k + \ell \le \exp(c/\varepsilon^2)$ and $D_{\square}(\kappa, \kappa^{S,X}) < \varepsilon$.*

Proof. Let $\ell = \lceil \exp(c'/\varepsilon^2) \rceil$ for a large enough $c' = c'(\Omega)$. Applying Lemma 2.8 to the kernels $\kappa^{(\omega)}$ from (2.6), we obtain partitions $T^{(\omega)} = (T_1^{(\omega)}, \ldots, T_\ell^{(\omega)})$ of $[0,1]$ such that

$$D_{\square}\left(\kappa^{(\omega)}, \kappa^{(\omega)\,T^{(\omega)}}\right) < \varepsilon/4. \tag{3.8}$$

Let $T = (T_1, \ldots, T_k)$ be the coarsest common refinement of all the partitions $T^{(\omega)}$ and of the partition $\{[0,1/2), [1/2,1]\}$. Then

$$|T| \le 2\ell^{|\Omega|}. \tag{3.9}$$

Moreover, (3.8) and Corollary 2.10 imply that

$$D_{\square}(\kappa^{(\omega)}, \kappa^{(\omega)\,T}) < \varepsilon/2 \qquad \text{for all } \omega \in \Omega. \tag{3.10}$$

Further, let $S' = (S'_1, \ldots, S'_K)$ comprise all partition classes $T_i \subset [0,1/2]$ and let $X' = (X'_1, \ldots, X'_L)$ be the partition of $[1/2,1]$ consisting of all the classes $T_i \subset [1/2,1]$. Finally, let $S_i = \{2s : s \in S'_i\}$ and $X_i = \{2x - 1 : x \in X'_i\}$. Then the partitions $S = (S_1, \ldots, S_K)$ and $X = (X_1, \ldots, X_L)$ satisfy $D_{\square}(\kappa, \kappa^{S,X}) < \varepsilon$ by Lemma 2.5. The desired bound on the total number $K + L$ of classes of S, X follows from (3.9). □

For a kernel κ and an integer n obtain κ_n as follows. Draw $\boldsymbol{x}_1, \boldsymbol{s}_1, \ldots, \boldsymbol{x}_n, \boldsymbol{s}_n \in [0,1]$ uniformly and independently and let κ_n be the kernel representing the matrix $(\kappa_{\boldsymbol{s}_i, \boldsymbol{x}_j})_{i,j}$. Additionally, obtain $\hat{\kappa}_n \in \Omega^{n\times n}$ by letting $\hat{\kappa}_{n,i,j} = \omega$ with probability $\kappa_{\boldsymbol{s}_i, \boldsymbol{x}_j}(\omega)$ independently for all i, j. We identify $\hat{\kappa}_n$ with its kernel representation. Moreover, we notice that $\hat{\kappa}_n$ coincides with the $n \times n$ upper left sub-matrix of $\boldsymbol{X}^\kappa$ from Section 1.2.3.

Lemma 3.6. *Let $\kappa, \kappa' \in \mathscr{K}$. With probability $1 - \exp(-\Omega(\sqrt{n}))$ we have $D_{\square}\left(\kappa_n, \kappa'_n\right) = O(D_{\square}(\kappa, \kappa') + n^{-1/4})$.*

Proof. Let $\bar\kappa, \bar\kappa'$ be the symmetric kernel representations of κ and κ' respectively given via (2.6). Sample $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_{2n}$ points in $[0,1]$ uniformly and independently at random. Denote by $\mathscr{B}$ the event that $\left|i : \boldsymbol{y}_i \le \frac{1}{2}\right| = n$ and assume, given $\mathscr{B}$, that without loss $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n \le 1/2$ and $\boldsymbol{y}_{n+1}, \ldots, \boldsymbol{y}_{2n} \ge 1/2$. Denote by $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n = 2\boldsymbol{y}_1, \ldots, 2\boldsymbol{y}_n$ and by $\boldsymbol{s}_1, \ldots, \boldsymbol{s}_n = 2\boldsymbol{y}_{n+1} - 1, \ldots, 2\boldsymbol{y}_{2n} - 1$. Clearly, $(\boldsymbol{x}_1, \boldsymbol{s}_1), \ldots, (\boldsymbol{x}_n, \boldsymbol{s}_n)$ are independent uniform samples from $[0,1]^2$.

Now, let $\tilde\kappa : [0,1]^2 \to [-1,1]$ be the kernel representing the matrix $(\bar\kappa_{\boldsymbol{y}_i, \boldsymbol{y}_j} - \bar\kappa'_{\boldsymbol{y}_i, \boldsymbol{y}_j})_{i,j \in [2n]}$. Applying Lemma 2.7 to $\tilde\kappa^{(\omega)}$, we obtain

$$\mathbb{P}\left[\left\|\tilde\kappa^{(\omega)}\right\|_\square \le \left\|\bar\kappa - \bar\kappa'\right\|_\square + 8n^{-1/4}\right] \ge 1 - 4\exp(-\sqrt{n}/10) \qquad (\omega \in \Omega). \tag{3.11}$$

Given $\mathscr{B}$, we translate $\tilde\kappa, \bar\kappa, \bar\kappa'$ back into kernels via (2.6) and apply Lemma 2.5, thus

$$D_\square\left(\kappa_n, \kappa'_n\right) \le 4 \max_{\omega \in \Omega} \left\|\tilde\kappa^{(\omega)}\right\|_\square.$$

Hence, the assertion follows from (3.11) and the fact that $\mathbb{P}(\mathscr{B}) = \Omega\left(n^{-1/2}\right)$. □

Lemma 3.7. *We have* $\mathbb{E}[D_\square(\kappa_n, \hat\kappa_n)] = O(n^{-1/2})$.

Proof. We adapt the simple argument from the proof of [32, Lemma 10.11] for our purposes. Letting $X_{i,j,\omega} = \mathbf{1}\left\{\hat\kappa_{n,i,j} = \omega\right\}$, we have $\mathbb{E}[X_{i,j,\omega}] = \kappa_{n,i,j}(\omega)$. Furthermore, because both $\kappa_n, \hat\kappa_n$ are kernel representations of $n \times n$ matrices, the supremum

$$\sup_{\omega \in \Omega, S, X \subset [0,1]} \left|\int_S \int_X \left(\kappa_{n,s,x}(\omega) - \hat\kappa_{n,s,x}(\omega)\right) \mathrm{d}x \mathrm{d}s\right|$$

is attained at sets S, X that are unions of intervals $[(i-1)/n, i/n)$ with $i \in [n]$. Hence,

$$D_\square(\kappa_n, \hat\kappa_n) = \sup_{\omega \in \Omega, S, X \subset [0,1]} \left|\int_S \int_X \left(\kappa_{n,s,x}(\omega) - \hat\kappa_{n,s,x}(\omega)\right) \mathrm{d}x \mathrm{d}s\right| = n^{-2} \max_{\omega \in \Omega, I, J \subset [n]} \left|\sum_{i \in I} \sum_{j \in J} X_{i,j,\omega} - \mathbb{E}[X_{i,j,\omega}]\right|. \tag{3.12}$$

Now, for any ω, I, J the random variable $\sum_{i \in I} \sum_{j \in J} X_{i,j,\omega}$ is a sum of $|I \times J|$ independent Bernoulli variables. Therefore, Azuma's inequality yields

$$\mathbb{P}\left[\left|\sum_{i \in I} \sum_{j \in J} X_{i,j,\omega} - \mathbb{E}[X_{i,j,\omega}]\right| > 10n^{3/2}\right] \le \exp(-10n). \tag{3.13}$$

Since (3.13) holds for any specific I, J, ω, the assertion follows from the union bound and (3.12). □

Proof of Theorem 1.9. Lemma 3.5 yields partitions $X = (X_1, \ldots, X_\ell)$, $S = (S_1, \ldots, S_\ell)$ of $[0,1]$ with $\ell \le n^{1/4}$ such that

$$D_\square(\kappa, \kappa^{S,X}) = O(\log^{-1/2} n). \tag{3.14}$$

Applying Lemma 3.6 to κ and $\kappa^{S,X}$, we obtain

$$\mathbb{E}\left[D_\square(\kappa_n, \kappa_n^{S,X})\right] = O\left(D_\square(\kappa, \kappa^{S,X}) + n^{-1/4}\right). \tag{3.15}$$

In addition, we claim that

$$\mathbb{E}\left[D_\boxtimes(\kappa^{S,X}, \kappa_n^{S,X})\right] = O(n^{-1/4} \log n). \tag{3.16}$$

To see this, let

$$N_h = \{i \in [n] : \boldsymbol{x}_i \in X_h\}, \qquad M_h = \left\{j \in [n] : \boldsymbol{s}_j \in S_h\right\} \qquad (h \in [\ell]).$$

Since N_h, M_h are binomial variables, the Chernoff bound shows that with probability $1 - o(1/n)$,

$$\max_{h \in \ell} ||N_h| - n\lambda(X_h)| \le \sqrt{n} \log n, \qquad \max_{h \in \ell} ||M_h| - n\lambda(S_h)| \le \sqrt{n} \log n. \tag{3.17}$$

Let

$$\mathscr{N}_h = \bigcup_{i \in N_h} [(i-1)/n, i/n), \qquad \mathscr{M}_h = \bigcup_{i \in M_h} [(i-1)/n, i/n).$$

Providing that the bounds (3.17) hold, we can construct $\varphi, \psi \in \mathbb{S}$ such that for all $h \in [n]$,

$$\lambda\left(\varphi(\mathscr{N}_h) \triangle X_h\right) \le n^{-1/2} \log n, \qquad \lambda\left(\psi(\mathscr{M}_h) \triangle S_h\right) \le n^{-1/2} \log n. \tag{3.18}$$

Furthermore, by construction we have $\kappa^{S,X}_{\varphi(s),\psi(x)} = \kappa^{S,X}_{s,x,n}$ if there exist $h, h' \in [\ell]$ such that $x \in \mathscr{N}_h$, $\varphi(x) \in X_h$ and $s \in \mathscr{M}_h$, $\psi(x) \in S_h$. Therefore, (3.18) implies that for all $T, Y \subset [0,1]$, $\omega \in \Omega$,

$$\begin{aligned}\left|\int_T\int_Y \left(\kappa^{S,X}_{n,y,t}(\omega) - \kappa^{S,X}_{\psi(y),\varphi(t)}(\omega)\right)\mathrm{d}y\mathrm{d}t\right| &\le \sum_{h,h'=1}^{\ell}\left|\int_{T\cap\mathscr{M}_h}\int_{Y\cap\mathscr{N}_{h'}}\left(\kappa^{S,X}_{n,y,t}(\omega) - \kappa^{S,X}_{\psi(y),\varphi(t)}(\omega)\right)\mathrm{d}y\mathrm{d}t\right| \\ &\le \sum_{h,h'=1}^{\ell}\lambda(T\cap\mathscr{M}_h\triangle\psi^{-1}(S_h))\lambda(Y\cap\mathscr{N}_{h'}) + \lambda(T\cap\mathscr{M}_h)\lambda(Y\cap\mathscr{N}_{h'}\triangle\psi^{-1}(X_{h'})) \\ &= O(\ell n^{-1/2}\log n) = O(n^{-1/4}\log n),\end{aligned}$$

whence (3.16) follows. Combining (3.14), (3.15) and (3.16), we see that

$$\mathbb{E}\left[D_{\boxtimes}(\kappa,\kappa_n)\right] = O(\log^{-1/2} n). \tag{3.19}$$

Finally, (3.19), Lemma 3.7 and Theorem 1.4 imply the assertion. □

3.3. **Proof of Theorems 1.13 and 1.14.** For measurable $k, k' : [0,1]^3 \to [0,1]^\Omega$ we let

$$D_\square(k,k') = \sup_{S\subset[0,1],X\subset[0,1]^2,\omega\in\Omega}\left|\int_S\int_X\left(k_{s,x,y}(\omega) - k'_{s,x,y}(\omega)\right)\mathrm{d}x\mathrm{d}y\mathrm{d}s\right|.$$

Then $D_\square(\cdot,\cdot)$ defines a pre-metric. Further, for measurable $\kappa,\kappa' : [0,1]^2 \to [0,1]^\Omega$ we define

$$\kappa \oplus' \kappa' : [0,1]^3 \to [0,1]^\Omega, \qquad (s,x,y) \mapsto \kappa_{s,x} \otimes \kappa'_{s,y}.$$

We will derive Theorem 1.14 from the following statement.

Proposition 3.8. *The map $(\kappa,\kappa') \mapsto \kappa \oplus' \kappa'$ is $D_\square$-continuous.*

Proof. Given $\varepsilon > 0$ choose a small $\delta = \delta(\varepsilon) > 0$. Suppose that $D_\square(\kappa,\kappa') < \delta$. Due to the triangle inequality, to establish continuity it suffices to show that for every $\kappa'' : [0,1]^2 \to [0,1]^\Omega$,

$$D_\square(\kappa\oplus'\kappa'',\kappa'\oplus'\kappa'') = \sup_{\substack{S\subset[0,1]\\X\subset[0,1]^2\\\omega,\omega'\in\Omega}}\left|\int_S\int_X\left(\kappa_{s,x}(\omega)\kappa''_{s,y}(\omega') - \kappa'_{s,x}(\omega)\kappa''_{s,y}(\omega')\right)\mathrm{d}x\mathrm{d}y\mathrm{d}s\right| < \varepsilon. \tag{3.20}$$

Thus, consider measurable X, S and fix $\omega,\omega' \in \Omega$. To estimate the last integral consider $y \in [0,1]$ and let $X_y = \{x \in [0,1] : (x,y) \in X\} \subset [0,1]$. Moreover, let $T_1,\ldots,T_\ell$ be a decomposition of S into pairwise disjoint measurable sets such that for all $j \in [\ell]$ we have

$$t_{j,*} \le t^*_j + \varepsilon/4, \qquad \text{where} \qquad t_{j,*} = \inf_{s\in T_j}\kappa''_{s,y}(\omega'), \qquad t^*_j = \sup_{s\in T_j}\kappa''_{s,y}(\omega').$$

Since $\kappa''_{s,y}(\omega') \in [0,1]$, we may assume that $\ell \le 4/\varepsilon$. Furthermore,

$$\begin{aligned}\left|\int_{X_y}\int_S\left(\kappa_{s,x}(\omega) - \kappa'_{s,x}(\omega)\right)\kappa''_{s,y}(\omega')\mathrm{d}x\mathrm{d}s\right| &\le \sum_{j=1}^{\ell}\left|\int_{X_y}\int_{S\cap T_j}\left(\kappa_{s,x}(\omega) - \kappa'_{s,x}(\omega)\right)\kappa''_{s,y}(\omega')\mathrm{d}x\mathrm{d}s\right| \\ &\le \frac{\varepsilon}{4} + \sum_{j=1}^{\ell}t^*_j\left|\int_{X_y}\int_{S\cap T_j}\left(\kappa_{s,x}(\omega) - \kappa'_{s,x}(\omega)\right)\mathrm{d}x\mathrm{d}s\right| \\ &\le \frac{\varepsilon}{4} + 2\ell D_\square(\kappa,\kappa') \le \frac{\varepsilon}{4} + 2\ell\delta < \varepsilon/2.\end{aligned}$$

Since this estimate holds for all $y \in [0,1]$, we obtain

$$\left|\int_S\int_X\left(\kappa_{s,x}(\omega)\kappa''_{s,y}(\omega') - \kappa'_{s,x}(\omega)\kappa''_{s,y}(\omega')\right)\mathrm{d}x\mathrm{d}y\mathrm{d}s\right| \le \int_0^1\left|\int_{X_y}\int_S\left(\kappa_{s,x}(\omega)\kappa''_{s,y}(\omega') - \kappa'_{s,x}(\omega)\kappa''_{s,y}(\omega')\right)\mathrm{d}s\mathrm{d}x\right|\mathrm{d}y < \frac{\varepsilon}{2}$$

for all S, X, ω, ω'. Thus, we obtain (3.20). □

Proof of Theorem 1.14. Theorem 1.14 follows from Proposition 3.8 and (2.4). □

We use a similar argument to prove Theorem 1.13. Specifically, for $\kappa,\kappa' : [0,1]^2 \to [0,1]^\Omega$ define

$$\kappa \otimes' \kappa' : [0,1]^3 \to [0,1]^\Omega, \qquad (s,t,x) \mapsto \kappa_{s,x} \otimes \kappa'_{t,x}.$$

Proposition 3.9. *The map $(\kappa,\kappa')\mapsto \kappa\otimes'\kappa'$ is $D_\square$-continuous.*

Proof. The definition of $D_\square(\cdot,\cdot)$ ensures that the map $\kappa\mapsto\kappa^\dagger$, where $\kappa^\dagger_{s,x}=\kappa_{x,s}$, is continuous. Therefore, the assertion follows from Proposition 3.8. □

Proof of Theorem 1.13. Theorem 1.13 follows immediately from Proposition 3.9 and (2.4). □

3.4. **Proof of Theorem1.1.** We begin by proving that the space $\mathscr{K}$ is complete with respect to $D_\square(\cdot,\cdot)$, the strongest version of the cut metric.

Lemma 3.10. *The space $\mathscr{K}$ equipped with the $D_\square(\cdot,\cdot)$ metric is complete.*

Proof. Suppose that $(\kappa_n)_{n\geq1}$ is a Cauchy sequence. Then for any measurable $S,X\subset[0,1]$ and any $\omega\in\Omega$ the sequence $\int_S\int_X\kappa_{n,s,x}(\omega)\mathrm{d}x\mathrm{d}s$ is Cauchy as well. Therefore, because any continuous function $f:[0,1]^2\to\mathbb{R}$, $(s,x)\mapsto f_{s,x}$ is uniformly continous, the limit

$$\lim_{n\to\infty}\int_0^1\int_0^1 f_{s,x}\kappa_{n,s,x}(\omega)\mathrm{d}x\mathrm{d}s$$

exists for every $\omega\in\Omega$. Indeed, the map

$$f\mapsto\lim_{n\to\infty}\int_0^1\int_0^1 f_{s,x}\kappa_{n,s,x}(\omega)\mathrm{d}x\mathrm{d}s$$

defines a positive linear functional on the space of all continuous functions $[0,1]^2\to\mathbb{R}$. Hence, by the Riesz representation theorem (Lemma 2.4) there exists a unique measure μ_ω on $[0,1]^2$ such that

$$\mu_\omega(S\times X)=\lim_{n\to\infty}\int_S\int_X\kappa_{n,s,x}(\omega)\mathrm{d}x\mathrm{d}s. \tag{3.21}$$

Indeed, the condition (3.21) ensures that μ_ω is absolutely continuous with respect to the Lebesgue measure. Therefore, the Radon-Nikodym theorem yields an L^1-function $(s,x)\in[0,1]^2\mapsto\kappa_{s,x}(\omega)\in\mathbb{R}_{\geq0}$ such that

$$\mu_\omega(Y)=\int_Y\kappa_{s,x}(\omega)\mathrm{d}s\mathrm{d}x \qquad\qquad \text{for all measurable } Y\subset[0,1]^2. \tag{3.22}$$

We claim that κ is a kernel, i.e., that $\sum_{\omega\in\Omega}\kappa_{s,x}(\omega)=1$ for almost all s,x. Indeed, combining (3.21) and (3.22) yields

$$\int_S\int_X 1\mathrm{d}x\mathrm{d}s=\sum_{\omega\in\Omega}\mu_\omega(S\times X)=\sum_{\omega\in\Omega}\int_S\int_X\kappa_{s,x}(\omega)\mathrm{d}x\mathrm{d}s=\int_S\int_X\sum_{\omega\in\Omega}\kappa_{s,x}(\omega)\mathrm{d}x\mathrm{d}s. \tag{3.23}$$

Since the rectangles $S\times X$ generate the Borel algebra on $[0,1]^2$, (3.23) implies that $\sum_{\omega\in\Omega}\kappa_{s,x}(\omega)=1$ almost everywhere.

Finally, (3.21) and (3.22) show that $\lim_{n\to\infty}D_\square(\kappa_n,\kappa)=0$. Indeed, given $\varepsilon>0$ consider a large enough n and let S,X,ω be such that

$$D_\square(\kappa_n,\kappa)<\varepsilon+\left|\int_S\int_X\big(\kappa_{s,x}(\omega)-\kappa_{n,s,x}(\omega)\big)\mathrm{d}x\mathrm{d}s\right|. \tag{3.24}$$

Equations (3.21) and (3.22) show that for large enough $N>n$,

$$\left|\int_S\int_X\big(\kappa_{s,x}(\omega)-\kappa_{N,s,x}(\omega)\big)\mathrm{d}x\mathrm{d}s\right|<\varepsilon \tag{3.25}$$

Combining (3.24) and (3.25) and recalling that $(\kappa_n)_n$ is $D_\square$-Cauchy, for large enough n we obtain

$$D_\square(\kappa_n,\kappa)<\varepsilon+\left|\int_S\int_X\big(\kappa_{s,x}(\omega)-\kappa_{N,s,x}(\omega)\big)\mathrm{d}x\mathrm{d}s\right|+D_\square(\kappa_n,\kappa_N)<3\varepsilon.$$

Hence, κ_n converges to κ. □

Corollary 3.11. *The space $\mathscr{K}$ equipped with the $D_\boxslash(\cdot,\cdot)$ metric is complete.*

Proof. We adapt a well know proof that a quotient of a Banach space with respect to a linear subspace is complete [10, Theorem 1.12.14]. Thus, suppose that $(\kappa_n)_n$ is a $D_\boxslash(\cdot,\cdot)$-Cauchy sequence. There exists a subsequence $(\kappa_{n_\ell})_\ell$ such that $D_\boxslash(\kappa_{n_\ell},\kappa_{n_{\ell+1}})<2^{-\ell}$ for all ℓ. Hence, passing to this subsequence, we may assume that $(\kappa_n)_n$ satisfies

$$D_\boxslash(\kappa_n,\kappa_{n+1})<2^{-n} \qquad\qquad \text{for all } n. \tag{3.26}$$

We are now going to construct a sequence $(k_n)_n$ of maps $[0,1]^2 \to \mathscr{P}(\Omega)$ such that $D_{\boxslash}(\kappa_n, k_n) = 0$ for all n and

$$D_{\square}(k_n, k_{n+1}) < 2^{-n} \qquad \text{for all } n. \tag{3.27}$$

We let k_1 be any kernel such that $D_{\boxslash}(k_1, \kappa_1) = 0$ and proceed by induction. Having constructed $k_1, \dots, k_n$ already, we observe that the definition of $D_{\boxslash}(\cdot,\cdot)$ ensures that

$$D_{\boxslash}(\kappa_n, \kappa_{n+1}) = D_{\boxslash}(k_n, \kappa_{n+1}) = \inf\left\{D_{\square}(k_n, k) : k : [0,1]^2 \to \mathscr{P}(\Omega), D_{\boxslash}(\kappa_n, k) = 0\right\}.$$

Therefore, (3.26) implies that there is $k_{n+1} : [0,1]^2 \to \mathscr{P}(\Omega)$ with $D_{\boxslash}(\kappa_{n+1}, k) = 0$ such that $D_{\square}(k_n, k_{n+1}) < 2^{-n}$. Thus, we obtain a sequence $(k_n)_n$ satisfying (3.27). Finally, any sequence $(k_n)_n$ that satisfies (3.27) is $D_{\square}(\cdot,\cdot)$-Cauchy. Therefore, Lemma 3.10 shows that $(k_n)_n$ has a limit k. Since $D_{\boxslash}(k_n, \kappa_n) = 0$, we conclude that

$$\lim_{n\to\infty} D_{\boxslash}(\kappa_n, k) = 0,$$

i.e., $(\kappa_n)_n$ converges to k. □

Corollary 3.12. *The spaces $\mathfrak{K}$ and $\mathfrak{L}$ equipped with the $D_{\boxtimes}(\cdot,\cdot)$ metric are complete and separable.*

Proof. To establish the completeness of $\mathfrak{K}$ we repeat the same argument as in the proof of Corollary 3.11. The completeness of $\mathfrak{L}$ then follows from Theorem 1.4. Moreover, Theorem 3.5 shows that the set of laws with finite support is dense in $\mathfrak{L}$. Hence, to prove the separability of $\mathfrak{L}$ it suffices to observe that the space $\mathscr{S}$ is separable, which it is because the set of all finite linear combinations of indicator functions $x \mapsto \mathbf{1}\{a < x < b\}$ with $a, b \in \mathbb{Q}$ is dense in $L^1([0,1], \mathbb{R})$. Finally, Theorem 1.4 implies that $\mathfrak{K}$ is separable as well. □

We denote by $\mathscr{P}(\mathfrak{L})$ the space of probability distributions on the Polish space $\mathfrak{L}$, endowed with the topology of weak convergence. As we saw in Section 2.1, this topology is metrised by the Wasserstein metric

$$\mathscr{D}_{\boxtimes}(\rho, \rho') = \inf\left\{\int_{\mathfrak{L}\times\mathfrak{L}} D_{\boxtimes}(\mu,\nu)\mathrm{d}g(\mu,\nu) : g \in \Gamma(\rho,\rho')\right\} \qquad (\rho, \rho' \in \mathscr{P}(\mathfrak{L})).$$

We begin by proving that $\mathscr{P}(\mathfrak{L})$ is compact. To this end we will construct a continuous map from another compact space onto $\mathscr{P}(\mathfrak{L})$. Specifically, recall that $\Omega^{\mathbb{N}\times\mathbb{N}}$ is a compact Polish space with respect to the product topology. The space $\mathscr{P}(\Omega^{\mathbb{N}\times\mathbb{N}})$ equipped with the weak topology is therefore compact as well. Further, the space $\mathfrak{X} \subset \mathscr{P}(\Omega^{\mathbb{N}\times\mathbb{N}})$ of exchangeable distributions is closed with respect to the weak topology, and therefore compact.

To construct a continuous map $\mathfrak{X} \to \mathscr{P}(\mathfrak{L})$, $\xi \mapsto \rho^\xi$ we are going to take a pointwise limit of maps $\mathfrak{X} \to \mathscr{P}(\mathfrak{L})$, $\xi \mapsto \rho^{\xi,n}$. Given $\xi \in \mathfrak{X}$ and $n \geq 1$ we define $\rho^{\xi,n}$ as follows. Draw $\boldsymbol{X}^\xi = (\boldsymbol{X}^\xi_{i,j})_{i,j\geq 1} \in \Omega^{\mathbb{N}\times\mathbb{N}}$ from ξ. Then define a probability distribution on Ω^n by letting

$$\boldsymbol{\mu}^{\xi,n}_*(\sigma) = \frac{1}{n}\sum_{i=1}^n \prod_{j=1}^n \mathbf{1}\left\{\sigma_j = \boldsymbol{X}^\xi_{i,j}\right\} \qquad (\sigma \in \Omega^n). \tag{3.28}$$

Thus, $\boldsymbol{\mu}^{\xi,n}_*$ is the empirical distribution of the rows of the top-left $n \times n$ minor of $\boldsymbol{X}^\xi$. Finally, let $\boldsymbol{\mu}^{\xi,n} = \dot{\boldsymbol{\mu}}^{\xi,n}_* \in \mathfrak{L}$ be the law induced by this discrete distribution and let $\rho^{\xi,n} \in \mathscr{P}(\mathfrak{L})$ be the distribution of $\boldsymbol{\mu}^{\xi,n}$ (with respect to the choice of $\boldsymbol{X}^\xi$).

Lemma 3.13. *For every $\xi \in \mathfrak{X}$ the limit $\rho^\xi = \lim_{n\to\infty} \rho^{\xi,n}$ exists and the map $\xi \mapsto \rho^\xi$ is continuous.*

Proof. Let $\xi \in \mathfrak{X}$. Since $\mathscr{P}(\mathfrak{L})$ is complete, to establish the existence of the limit we just need to prove that the sequence $(\rho^{\xi,n})_n$ is Cauchy. To this end it suffices to verify the following condition:

$$\exists f : \mathbb{N} \to \mathbb{N}\ \forall \varepsilon > 0\ \exists n_0 = n_0(\varepsilon) > 0\ \forall n > n_0, N > f(n) : \mathscr{D}_{\boxtimes}(\rho^{\xi,n}, \rho^{\xi,N}) < \varepsilon. \tag{3.29}$$

Indeed, if (3.29) is satisfied, then there exists a subsequence $(\rho^{\xi,n_M})_M$ such that $\mathscr{D}_{\boxtimes}(\rho^{\xi,n_M}, \rho^{\xi,n_{M+1}}) < 2^{-M}$ for all M. In particular, the subsequence is Cauchy. Because $\mathscr{P}(\mathfrak{L})$ is complete, this subsequence thus has a limit $\rho^\star$, and (3.29) ensures that the entire sequence $(\rho^{\xi,n})_n$ converges to $\rho^\star$ as well.

To verify (3.29) let $\varepsilon > 0$, pick a large $n = n(\varepsilon) > 0$ and choose $N = N(n)$ large enough. We aim to prove that

$$\mathscr{D}_{\boxtimes}(\rho^{\xi,n}, \rho^{\xi,N}) < \varepsilon. \tag{3.30}$$

To this end, we couple $\rho^{\xi,n}, \rho^{\xi,N}$ by drawing $\boldsymbol{X}^\xi \in \Omega^{\mathbb{N}\times\mathbb{N}}$ from ξ and letting g be the distribution of the pair $(\boldsymbol{\mu}^{\xi,n}, \boldsymbol{\mu}^{\xi,N})$. By definition of the Wasserstein metric, to establish (3.30) it suffices to show that

$$\mathbb{E}\left[D_{\boxtimes}(\boldsymbol{\mu}^{\xi,n}, \boldsymbol{\mu}^{\xi,N})\right] < \varepsilon \tag{3.31}$$

But (3.31) follows from Theorem 1.9. Indeed, the construction (3.28) ensures that $\boldsymbol{\mu}^{\xi,n}$ is the empirical distribution of rows of the upper left $n\times n$-block of $\boldsymbol{X}^\xi$, while $\boldsymbol{\mu}^{\xi,N}$ is the empirical distribution of the rows of the $N\times N$-upper left block. Due to the exchangeability of ξ, the distribution of the upper left $n\times n$-block is identical to the distribution of a random $n\times n$-minor of the matrix $\boldsymbol{X}^\xi$. Therefore, assuming that $N\gg n^2$ so that upon sub-sampling n out of N indices no index is chosen twice, in the notation of Theorem 1.9 we have $d_{\mathrm{TV}}(\boldsymbol{\mu}^{\xi,n},\boldsymbol{\mu}^{\xi,N}_{n,n})<\varepsilon/2$, whence we obtain (3.31) and thus (3.30). Hence, the limit $\rho^\xi=\lim_{n\to\infty}\rho^{\xi,n}$ exists for all ξ.

To show continuity fix $\varepsilon>0$ and let $\xi,\eta\in\mathfrak{X}$. Due to (3.31) there exists $n=n(\varepsilon)>0$ independent of ξ,η such that

$$\mathscr{D}_{\boxtimes}(\rho^\xi,\rho^{\xi,n})<\varepsilon/4,\qquad\qquad \mathscr{D}_{\boxtimes}(\rho^\eta,\rho^{\eta,n})<\varepsilon/4. \tag{3.32}$$

Since $\mathfrak{X}$ is equipped with the weak topology, any $\xi>0$ admits a neighbourhood U such that for all $\eta\in U$,

$$\sum_{X\in\Omega^{n\times n}}\left|\mathbb{P}\left[\forall i,j\in[n]:\boldsymbol{X}^\xi_{i,j}=X_{i,j}\right]-\mathbb{P}\left[\forall i,j\in[n]:\boldsymbol{X}^\eta_{i,j}=X_{i,j}\right]\right|<\varepsilon/8.$$

Hence, the upper left $n\times n$-corners of $\boldsymbol{X}^\xi,\boldsymbol{X}^\eta$ have total variation distance at most $\varepsilon/4$. In effect, there is a coupling of $(\boldsymbol{X}^\xi_{i,j})_{i,j\in[n]},(\boldsymbol{X}^\eta_{i,j})_{i,j\in[n]}$ under which both these random $n\times n$-matrices coincide with probability at least $1-\varepsilon/4$. Clearly, this coupling extends to a coupling of the measures $\boldsymbol{\mu}^{\xi,n}$, $\boldsymbol{\mu}^{\eta,n}$ such that $\mathbb{E}[D_{\boxtimes}(\boldsymbol{\mu}^{\xi,n},\boldsymbol{\mu}^{\eta,n})]\le\varepsilon/4$. Consequently, $\mathscr{D}_{\boxtimes}(\rho^{\xi,n},\rho^{\eta,n})\le\varepsilon/4$. Combining this bound with (3.32), we conclude that $\mathscr{D}_{\boxtimes}(\rho^\xi,\rho^\eta)<\varepsilon$ for all $\eta\in U$, whence $\xi\mapsto\rho^\xi$ is continuous. □

As a next step we are going to embed the space $\mathfrak{L}$ into $\mathfrak{X}$. For a given law $\mu\in\mathfrak{L}$ let ξ^μ be the distribution of $\boldsymbol{X}^\mu$.

Lemma 3.14. *The map $\mu\mapsto\xi^\mu$ is continuous and $\rho^{\xi^\mu}=\delta_\mu$.*

Proof. Due to Theorem 1.4 it suffices to show that the map $\kappa\in\mathfrak{K}\mapsto\xi^\kappa$, where ξ^κ is the distribution of $\boldsymbol{X}^\kappa$, is continuous. Combining Theorems 1.14 and 1.13, we conclude that the map $\mathfrak{K}_\Omega\to\mathfrak{K}_{\Omega^{[n]\times[n]}}$, $\kappa\mapsto(\kappa^{\oplus n})^{\otimes n}$ is continuous, where we iterate the $\oplus$ and the $\otimes$ operations n times. For the sake of clarity, let us spell out the precise meaning of iterating these operations. The definition of the $\oplus$-operation extends to kernels κ,κ' that take values in $\mathscr{P}(\Omega)$, $\mathscr{P}(\Omega')$ for different sets Ω,Ω' by simply viewing κ,κ' as $\Omega\cup\Omega'$-kernels. With this extension it makes sense to iterate the $\oplus$-operation; notice that $\kappa^{\oplus h}$ takes values in Ω^h. We define $\kappa^{\otimes h}$ analogously. Finally, combining these two operations we obtain $\kappa^{\oplus n\otimes n}=(\kappa^{\oplus n})^{\otimes n}$, which is an $\Omega^{n\times n}$-kernel. Furthermore, for any $\sigma\in\Omega^{[n]\times[n]}$ the map

$$\mathfrak{K}_{\Omega^{[n]\times[n]}}\to[0,1],\qquad\qquad k\mapsto\int_0^1\int_0^1 k(\sigma)\mathrm{d}s\mathrm{d}x$$

is continuous by the definition of the cut metric. Therefore, being a concatenations of continuous maps, the functions

$$\mathscr{T}_\sigma:\mathfrak{K}\to[0,1],\qquad\qquad \kappa\mapsto\int_0^1\int_0^1\kappa^{\oplus n\otimes n}_{s,x}(\sigma)\mathrm{d}s\mathrm{d}x$$

are continuous as well. Since $\mathfrak{X}$ carries the weak toplogy, the continuity of the maps $\mathscr{T}_\sigma$ implies the continuity of the map $\kappa\mapsto\xi^\kappa$. □

Corollary 3.15. *The map $\mathfrak{X}\to\mathscr{P}(\mathfrak{L})$, $\xi\mapsto\rho^\xi$ is surjective.*

Proof. Suppose that $\mathfrak{p}\in\mathscr{P}(\mathfrak{L})$. With $\nu\mapsto\xi^\nu$ the measurable map from Lemma 3.14, we define $\xi^{\mathfrak{p}}=\int_{\mathfrak{L}}\delta_{\xi^\mu}\mathrm{d}\mathfrak{p}(\mu)$. Then $\rho^{\xi^{\mathfrak{p}}}=\mathfrak{p}$. □

Corollary 3.16. *The space $\mathfrak{L}$ is compact.*

Proof. The space $\mathfrak{X}$ is compact as it is the space of probability measures on the compact Polish space $\Omega^{\mathbb{N}\times\mathbb{N}}$. Since Lemma 3.13 and Corollary 3.15 render a continuous surjective map $\mathfrak{X}\to\mathscr{P}(\mathfrak{L})$ and a continuous image of a compact space is compact, the space $\mathscr{P}(\mathfrak{L})$ is compact. To finally conclude that $\mathfrak{L}$ is compact as well, consider a sequence $(\mu_n)_{n\ge1}$ in $\mathfrak{L}$. Because $\mathscr{P}(\mathfrak{L})$ is compact, the sequence $(\delta_{\mu_n})_{n\ge1}$ possesses a convergent subsequence $(n_\ell)_{\ell\ge1}$. Let π be the limit of that subsequence. Consider a point ν in the support of π and let $(U_k)_{k\ge0}$ be a sequence of open neighbourhoods of ν such that $U_{k+1}\subset U_k$ for all k and $\bigcap_{k\ge1}U_k=\{\nu\}$. By Urysohn's lemma there

are continuous functions $f_k : \mathfrak{L} \to [0,1]$ such that f_k takes the value one on U_k and the value 0 outside U_{k-1}. Now, for all $k \geq 1$ we have

$$0 < \int f_k \mathrm{d}\pi = \lim_{\ell\to\infty} \int f_k \mathrm{d}\mu_{n_\ell} \leq \lim_{\ell\to\infty} \mathbf{1}\{\mu_{n_\ell} \in U_{k-1}\}.$$

Hence, $\mu_{n_\ell} \in U_{k-1}$ for almost all ℓ. Consequently, $\nu = \lim_{\ell\to\infty} \mu_{n_\ell}$. Thus, the metric space $\mathfrak{L}$ is sequentially compact and therefore compact. □

Of course the second part of the proof above merely establishes the well known fact that the mapping $\mathfrak{L} \to \mathscr{P}(\mathfrak{L})$, $\mu \mapsto \delta_\mu$ is a homeomorphic embedding onto a closed subspace. We included the brief argument for the sake of completeness.

Proof of Theorem1.1. The theorem follows from Corollaries 3.12 and 3.16. □

3.5. **Proof of Theorem 1.11.** Let $\mu, \nu \in \mathscr{L}$. Toward the proof of (1.7) let

$$X^+(\omega) = \left\{ x \in [0,1] : \int_{\mathscr{S}} \sigma_x(\omega) \mathrm{d}\mu(\sigma) - \int_{\mathscr{S}} \sigma_x(\omega) \mathrm{d}\nu(\sigma) \geq 0 \right\}, \qquad X^-(\omega) = [0,1] \setminus X^+.$$

Since μ, ν are atoms concentrated on the pure state (1.6), respectively, we obtain

$$\begin{aligned} D_{\boxslash}(\bar{\mu}, \bar{\nu}) &= \max_{\omega\in\Omega} \left| \int_{X^+(\omega)} \int_{\mathscr{S}} \sigma_x(\omega) \mathrm{d}\mu(\sigma) - \int_{\mathscr{S}} \sigma_x(\omega) \mathrm{d}\nu(\sigma) \right| \vee \left| \int_{X^-(\omega)} \int_{\mathscr{S}} \sigma_x(\omega) \mathrm{d}\mu(\sigma) - \int_{\mathscr{S}} \sigma_x(\omega) \mathrm{d}\nu(\sigma) \right| \\ &= \max_{\omega\in\Omega} \left| \int_{X^+(\omega)} \int_{\mathscr{S}\times\mathscr{S}} (\sigma_x(\omega) - \tau_x(\omega)) \, \mathrm{d}(\mu\otimes\nu)(\sigma,\tau) \right| \vee \left| \int_{X^-(\omega)} \int_{\mathscr{S}\times\mathscr{S}} (\sigma_x(\omega) - \tau_x(\omega)) \, \mathrm{d}(\mu\otimes\nu)(\sigma,\tau) \right| \leq D_{\boxslash}(\mu,\nu), \end{aligned} \tag{3.33}$$

whence (1.7) is immediate. Moreover, the first part of (1.8) follows from (3.33), while the second part is immediate from the triangle inequality.

3.6. **Proof of Theorem 1.8.** The product topology on $\Omega^{\mathbb{N}\times\mathbb{N}}$ is the weakest topology under which all the functions

$$T_\sigma : \Omega^{\mathbb{N}\times\mathbb{N}} \to \{0,1\}, \qquad (X_{i,j})_{i,j\geq 1} \mapsto \prod_{i,j=1}^{n} \mathbf{1}\{X_{i,j} = \sigma_{i,j}\} \qquad \big(n \geq 1, \sigma \in \Omega^{[n]\times[n]}\big).$$

are continuous. Equivalently, the product topology is induced by the metric

$$D_{\max} : \Omega^{\mathbb{N}\times\mathbb{N}} \times \Omega^{\mathbb{N}\times\mathbb{N}} \to [0,1], \qquad (X,Y) \mapsto 2^{-\max\{n\geq 0 : \forall i,j \leq n : X_{i,j} = Y_{i,j}\}}. \tag{3.34}$$

Hence, the weak topology on $\mathfrak{X} \subset \mathscr{P}(\Omega^{\mathbb{N}\times\mathbb{N}})$ is induced by the corresponding Wasserstein metric $\mathscr{D}_{\max}(\cdot,\cdot)$.

As a first step we are going to show that the map $\pi \mapsto \Xi^\pi$ is $(\mathscr{D}_{\boxtimes}, \mathscr{D}_{\max})$-continuous. Indeed, assume that $\pi, \pi' \in \mathscr{P}(\mathfrak{K})$ satisfy $\mathscr{D}_{\boxtimes}(\pi,\pi') < \delta$ for a small enough $\delta = \delta(\varepsilon) > 0$. Then the corresponding coupling shows together with Lemma 3.14 and Theorem 1.4 that $\mathscr{D}_{\max}(\cdot,\cdot) < \varepsilon$. Furthermore, $\pi \mapsto \Xi^\pi$ is one-to-one because it can be inverted via Corollary 3.15. Moreover, Lemma 3.13 implies that the map $\pi \mapsto \Xi^\pi$ is surjective, as the inverse image of $\xi \in \mathfrak{X}$ is just $\rho^\xi \in \mathscr{P}(\mathfrak{L}) \cong \mathscr{P}(\mathfrak{K})$. Thus, we know that $\mathscr{P}(\mathfrak{K}) \to \mathfrak{X}$, $\pi \mapsto \Xi^\pi$ is a continuous bijection. Finally, since $\mathscr{P}(\mathfrak{K})$ is compact and the continuous image of a compact set is compact, the map $\pi \mapsto \Xi^\pi$ is open and thus a homeomorphism. □

3.7. **Proof of Theorem 1.10.** For a bipartite graph $G = (U,V,E)$ with $|U| = |V| = n$, and a partition $P = (S_1 \ldots S_l, V_1 \ldots V_k)$, denote by G^P the weighted bipartite graph on vertex set $([l],[k])$ s.t. the weight of edge ij is given by $\mathrm{d}(S_i, V_j)$.

Theorem 3.17 ([18], Theorem 7.1)**.** *There exists $\varepsilon > 0, n \in \mathbb{N}$ and a bipartite graph $G = (U,V,E)$ with $|U| = |V| = n$ s.t. every partition $P = (S_1 \ldots S_l, V_1 \ldots V_k)$ of (U,V) satisfying $D_{\blacksquare}(G, G^P) \leq \varepsilon$ requires at least $l = \exp\big(\Theta(\varepsilon^{-2})\big)$ parts, independently of k.*

Proof of Theorem 1.10. Let G be a graph given by the previous theorem and let κ_G be the corresponding graphon. Denote by κ a kernel consisting of κ_G and its transposed graphon given by (2.6) in the special case $\Omega = \{0,1\}$. Denote by $\mu = \mu(G) \in \mathscr{L}$ the corresponding law given by Theorem 1.4. Assume there is $\nu \in \mathscr{L}$ with support of size less then $l = \exp\big(\Theta(\varepsilon^{-2})\big)$ and $D_{\boxtimes}(\mu,\nu) < \frac{\varepsilon}{2}$. Then ν induces a partition K of $[0,1]$ into at most l parts s.t. $D_{\boxtimes}(\kappa,\kappa^\nu) = D_{\boxtimes}(\kappa,\kappa^K) \leq \frac{\varepsilon}{2}$ which implies that there is a partition S and a graphon κ^S s.t. $D_{\blacksquare}(\kappa_G,\kappa^S) \leq \varepsilon$. As κ_G and κ_S are by definition embeddings of (finite) graphs into the space of graphons, this is a contradiction to Theorem 3.17. □

4. The pinning operation

In this section we prove Theorem 1.12. We begin by investigating a discrete version of the pinning operation, which played a key role in recent work on random factor graphs [11]. The discrete version of the pinning theorem, Theorem 4.1 below, was already established as [11, Lemma 3.5]. In Section 4.1 we give a shorter proof, based on an argument form [45]. Moreover, in Section 4.2 we show by a somewhat delicate argument that the pinning operation is continuous with respect to the cut metric. Finally, in Section 4.3 we complete the proof of Theorem 1.12.

4.1. Discrete pinning.

For a probability measure $\mu \in \mathscr{L}_n$ and a set $I \subset [n]$ we denote by μ_I the joint distribution of the coordinates $i \in I$. Thus, μ_I is the probability distribution on Ω^I defined by

$$\mu_I(\sigma) = \sum_{\tau \in \Omega^n} \mathbf{1}\{\forall i \in I : \tau_i = \sigma_i\}\,\mu(\tau).$$

Where $I = \{i_1, \ldots, i_t\}$ is given explicitly, we use the shorthand $\mu_I = \mu_{i_1,\ldots,i_t}$.

Theorem 4.1. *For every $\varepsilon > 0$ for all large enough n and all $\mu \in \mathscr{L}_n$ the following is true. Draw and integer $0 \le \boldsymbol{\theta} \le \lceil \log|\Omega|/\varepsilon^2 \rceil$ uniformly random and let $\boldsymbol{I} \subset [n]$ be a random set of size $\boldsymbol{\theta}$. Additionally, draw $\hat{\boldsymbol{\sigma}}$ from μ independently of $\boldsymbol{\theta}, \boldsymbol{I}$. Let*

$$\hat{\mu} = \mu[\cdot \mid \{\sigma \in \Omega^n : \forall i \in \boldsymbol{I} : \sigma_i = \hat{\boldsymbol{\sigma}}_i\}]. \tag{4.1}$$

Then

$$\sum_{1 \le i < j \le n} \mathbb{E}\left\|\hat{\mu}_{i,j} - \hat{\mu}_i \otimes \hat{\mu}_j\right\|_{\mathrm{TV}} \le \varepsilon n^2. \tag{4.2}$$

Apart from [11, Lemma 3.5], statements related to Theorem 4.1 were previously obtained by Montanari [40] and Raghavendra and Tan [45]. To be precise, [40, Theorem 2.2] deals with the special case of the discrete pinning operation for graphical channels and the number $\boldsymbol{\theta}$ of pinned coordinates scales linearly with the dimension n. The original proof of Theorem 4.1 in [11] was based on a generalisation of Montanari's argument. Moreover, [45, Lemma 4.5] asserted the existence of $T = T(\mu, \varepsilon) > 0$ such that

$$\sum_{1 \le i < j \le n} \mathbb{E}\left[\left\|\hat{\mu}_{i,j} - \hat{\mu}_i \otimes \hat{\mu}_j\right\|_{\mathrm{TV}} \mid \boldsymbol{\theta} = T\right] \le \varepsilon n^2,$$

rather than showing that a random $\boldsymbol{\theta}$ does the trick. But at second glance the proof given in [45], which is significantly simpler than the one from [11], actually implies Theorem 4.1.

For completeness we include the short proof of Theorem 4.1 via the argument from [45]. We need a few concepts from information theory. Let X, Y, Z be random variables that take values in finite domains. We recall that the *conditional mutual information* of X, Y given Z is defined as

$$\mathscr{I}(X, Y \mid Z) = \sum_{x,y,z} \mathbb{P}\left[X = x, Y = y, Z = z\right] \log \frac{\mathbb{P}\left[X = x, Y = y \mid Z = z\right]}{\mathbb{P}\left[X = x \mid Z = z\right]\mathbb{P}\left[Y = y \mid Z = z\right]},$$

with the conventions $0 \log 0 = 0$, $0 \log \frac{0}{0} = 0$ and with the sum ranging over all possible values x, y, z of X, Y, Z, respectively. Moreover, the *conditional entropy* of X given Y reads

$$\mathscr{H}(X \mid Y) = \sum_{x,y} \mathbb{P}\left[X = x, Y = y\right] \log \mathbb{P}\left[X = x \mid Y = y\right].$$

We also recall the basic identity

$$\mathscr{I}(X, Y \mid Z) = \mathscr{H}(X \mid Z) - \mathscr{H}(X \mid Y, Z). \tag{4.3}$$

Finally, *Pinsker's inequality* provides that for any two probability distribution μ, ν on a finite set $\mathscr{X}$,

$$d_{\mathrm{TV}}(\mu, \nu) \le \sqrt{D_{\mathrm{KL}}\left(\mu \| \nu\right)/2}, \qquad \text{where} \qquad D_{\mathrm{KL}}\left(\mu \| \nu\right) = \sum_{x \in \mathscr{X}} \mu(x) \log \frac{\mu(x)}{\nu(x)} \tag{4.4}$$

signifies the Kullback-Leibler divergence. The proof of the following lemma is essentially identical to the proof of [45, Lemma 4.5].

Lemma 4.2. *Let $\mu \in \mathscr{P}(\Omega^n)$ and let $\boldsymbol{\sigma} \in \Omega^n$ be a sample drawn from μ. Let $\boldsymbol{i}, \boldsymbol{i}', \boldsymbol{i}_1, \ldots \in [n]$ be uniformly distributed and mutually independent as well as independent of $\boldsymbol{\sigma}$. Then for any integer T we have*

$$\sum_{\theta=0}^{T} \mathscr{I}(\boldsymbol{\sigma}_{\boldsymbol{i}}, \boldsymbol{\sigma}_{\boldsymbol{i}'} \mid \boldsymbol{i}, \boldsymbol{i}', \boldsymbol{i}_1, \ldots, \boldsymbol{i}_\theta, \boldsymbol{\sigma}_{\boldsymbol{i}_1}, \ldots, \boldsymbol{\sigma}_{\boldsymbol{i}_\theta}) \leq \log|\Omega|.$$

Proof. Due to (4.3), for every $\theta \geq 0$,

$$\begin{aligned}\mathscr{I}(\boldsymbol{\sigma}_{\boldsymbol{i}}, \boldsymbol{\sigma}_{\boldsymbol{i}'} \mid \boldsymbol{i}, \boldsymbol{i}', \boldsymbol{i}_1, \ldots, \boldsymbol{i}_\theta, \boldsymbol{\sigma}_{\boldsymbol{i}_1}, \ldots, \boldsymbol{\sigma}_{\boldsymbol{i}_\theta}) &= \mathscr{H}(\boldsymbol{\sigma}_{\boldsymbol{i}} \mid \boldsymbol{i}, \boldsymbol{i}', \boldsymbol{i}_1, \ldots, \boldsymbol{i}_\theta, \boldsymbol{\sigma}_{\boldsymbol{i}_1}, \ldots, \boldsymbol{\sigma}_{\boldsymbol{i}_\theta}) - \mathscr{H}(\boldsymbol{\sigma}_{\boldsymbol{i}} \mid \boldsymbol{i}, \boldsymbol{i}', \boldsymbol{i}_1, \ldots, \boldsymbol{i}_\theta, \boldsymbol{\sigma}_{\boldsymbol{i}_1}, \ldots, \boldsymbol{\sigma}_{\boldsymbol{i}_\theta}, \boldsymbol{\sigma}_{\boldsymbol{i}'}) \\ &= \mathscr{H}(\boldsymbol{\sigma}_{\boldsymbol{i}} \mid \boldsymbol{i}, \boldsymbol{i}_1, \ldots, \boldsymbol{i}_\theta, \boldsymbol{\sigma}_{\boldsymbol{i}_1}, \ldots, \boldsymbol{\sigma}_{\boldsymbol{i}_\theta}) - \mathscr{H}(\boldsymbol{\sigma}_{\boldsymbol{i}} \mid \boldsymbol{i}, \boldsymbol{i}_1, \ldots, \boldsymbol{i}_\theta, \boldsymbol{i}_{\theta+1}, \boldsymbol{\sigma}_{\boldsymbol{i}_1}, \ldots, \boldsymbol{\sigma}_{\boldsymbol{i}_{\theta+1}}).\end{aligned}$$

Summing on $\theta = 1, \ldots, T$, we obtain

$$\sum_{\theta=0}^{T} \mathscr{I}(\boldsymbol{\sigma}_{\boldsymbol{i}}, \boldsymbol{\sigma}_{\boldsymbol{i}'} \mid \boldsymbol{i}, \boldsymbol{i}', \boldsymbol{i}_1, \ldots, \boldsymbol{i}_\theta, \boldsymbol{\sigma}_{\boldsymbol{i}_1}, \ldots, \boldsymbol{\sigma}_{\boldsymbol{i}_\theta}) = \mathscr{H}(\boldsymbol{\sigma}_{\boldsymbol{i}} \mid \boldsymbol{i}) - \mathscr{H}(\boldsymbol{\sigma}_{\boldsymbol{i}} \mid \boldsymbol{i}, \boldsymbol{i}_1, \ldots, \boldsymbol{i}_{T+1}, \boldsymbol{\sigma}_{\boldsymbol{i}_1}, \ldots, \boldsymbol{\sigma}_{\boldsymbol{i}_{T+1}}).$$

The desired bound follows because $\mathscr{H}(\boldsymbol{\sigma}_i) \leq \log|\Omega|$ and $\mathscr{H}(\boldsymbol{\sigma}_i \mid \boldsymbol{\sigma}_{\boldsymbol{i}_1}, \ldots, \boldsymbol{\sigma}_{\boldsymbol{i}_{T+1}}) \geq 0$. □

Now let $T > 0$ be an integer and draw $0 \leq \boldsymbol{\theta} \leq T$ uniformly at random and construct $\hat{\mu}$ as in Theorem 4.1. Then as an immediate consequence of Lemma 4.2 we obtain the following bound, where, of course, the expectation refers to the choice of $\hat{\mu}$ and the independently chosen and uniform $\boldsymbol{i}, \boldsymbol{i}'$.

Corollary 4.3. *We have* $\mathbb{E}\left[D_{\mathrm{KL}}\left(\hat{\mu}_{\boldsymbol{i},\boldsymbol{i}'} \| \hat{\mu}_{\boldsymbol{i}} \otimes \hat{\mu}_{\boldsymbol{i}'}\right)\right] \leq (\log|\Omega|)/T$.

Proof. Keeping the notation from Lemma 4.2, we let $\boldsymbol{I} = (\boldsymbol{i}, \boldsymbol{i}', \boldsymbol{i}_1, \ldots, \boldsymbol{i}_\theta)$ and $\boldsymbol{\Sigma} = (\boldsymbol{\sigma}_{\boldsymbol{i}_1}, \ldots, \boldsymbol{\sigma}_{\boldsymbol{i}_\theta})$. Recalling the definition (4.1) of $\hat{\mu}$, we find

$$\begin{aligned}\mathscr{I}(\boldsymbol{\sigma}_{\boldsymbol{i}}, \boldsymbol{\sigma}_{\boldsymbol{i}'} \mid \boldsymbol{I}, \boldsymbol{S}) &= \mathbb{E}\left[\sum_{\omega,\omega' \in \Omega} \mathbb{P}\left[\boldsymbol{\sigma}_{\boldsymbol{i}} = \omega, \boldsymbol{\sigma}_{\boldsymbol{i}'} = \omega' \mid \boldsymbol{I}, \boldsymbol{\Sigma}\right] \log \frac{\mathbb{P}\left[\boldsymbol{\sigma}_{\boldsymbol{i}} = \omega, \boldsymbol{\sigma}_{\boldsymbol{i}'} = \omega' \mid \boldsymbol{I}, \boldsymbol{\Sigma}\right]}{\mathbb{P}\left[\boldsymbol{\sigma}_{\boldsymbol{i}} = \omega \mid \boldsymbol{I}, \boldsymbol{S}\right] \mathbb{P}\left[\boldsymbol{\sigma}_{\boldsymbol{i}'} = \omega' \mid \boldsymbol{I}, \boldsymbol{\Sigma}\right]}\right] \\ &= \mathbb{E}\Bigg[\sum_{\sigma \in \Omega^n} \mu(\sigma) \sum_{\omega,\omega' \in \Omega} \mu\left(\boldsymbol{\sigma}_{\boldsymbol{i}} = \omega, \boldsymbol{\sigma}_{\boldsymbol{i}'} = \omega' \mid \boldsymbol{\Sigma} = (\sigma_{\boldsymbol{i}_1}, \ldots, \sigma_{\boldsymbol{i}_\theta})\right) \\ &\qquad \log \frac{\mu\left(\boldsymbol{\sigma}_{\boldsymbol{i}} = \omega, \boldsymbol{\sigma}_{\boldsymbol{i}'} = \omega' \mid \boldsymbol{\Sigma} = (\sigma_{\boldsymbol{i}_1}, \ldots, \sigma_{\boldsymbol{i}_\theta})\right)}{\mu\left(\boldsymbol{\sigma}_{\boldsymbol{i}} = \omega \mid \boldsymbol{\Sigma} = (\sigma_{\boldsymbol{i}_1}, \ldots, \sigma_{\boldsymbol{i}_\theta})\right) \mu\left(\boldsymbol{\sigma}_{\boldsymbol{i}'} = \omega' \mid \boldsymbol{\Sigma} = (\sigma_{\boldsymbol{i}_1}, \ldots, \sigma_{\boldsymbol{i}_\theta})\right)}\Bigg] \\ &= \mathbb{E}\left[D_{\mathrm{KL}}\left(\hat{\mu}_{\boldsymbol{i},\boldsymbol{i}'} \| \hat{\mu}_{\boldsymbol{i}} \otimes \hat{\mu}_{\boldsymbol{i}'}\right)\right]\end{aligned}$$

Hence, the assertion follows from Lemma 4.2. □

Proof of Theorem 4.1. Applying Pinsker's inequality (4.4), Jensen's inequality and Corollary 4.3, we find

$$\mathbb{E}\left\|\hat{\mu}_{\boldsymbol{i},\boldsymbol{i}'} - \hat{\mu}_{\boldsymbol{i}} \otimes \hat{\mu}_{\boldsymbol{i}'}\right\|_{\mathrm{TV}} \leq \mathbb{E}\sqrt{D_{\mathrm{KL}}\left(\hat{\mu}_{\boldsymbol{i},\boldsymbol{i}'} \| \hat{\mu}_{\boldsymbol{i}} \otimes \hat{\mu}_{\boldsymbol{i}'}\right)/2} \leq \sqrt{\mathbb{E}\left[D_{\mathrm{KL}}\left(\hat{\mu}_{\boldsymbol{i},\boldsymbol{i}'} \| \hat{\mu}_{\boldsymbol{i}} \otimes \hat{\mu}_{\boldsymbol{i}'}\right)\right]/2} \leq \sqrt{\frac{\log|\Omega|}{2T}},$$

whence the desired bound follows if $T \geq (\log|\Omega|)/(2\varepsilon^2)$. □

Finally, the following lemma clarifies the bearing that the bound (4.2) has on the cut metric. The lemma is an improved version of [13, Lemma 2.9]. Following [5] we say that $\mu \in \mathscr{L}_n$ is *ε-symmetric* if

$$\sum_{1 \leq i < i' \leq n} \left\|\mu_{i,i'} - \mu_i \otimes \mu_{i'}\right\|_{\mathrm{TV}} < \varepsilon n^2.$$

Lemma 4.4. *For any $\varepsilon > 0$ and every finite set Ω there exists $n_0 > 0$ s.t. for every $n \geq n_0$ every $\varepsilon^2/4$-symmetric $\mu \in \mathscr{L}_n$ satisfies $\Delta_{\square}(\mu, \bigotimes_{i=1}^n \mu_i) < \varepsilon$.*

Proof. Let $\delta = \varepsilon^2/4$. Since $\mu \otimes \bar{\mu}$ is a coupling of μ and $\bar{\mu}$ it suffices to show that for any set $I \subset [n]$ and every $\omega \in \Omega$,

$$\sup_{S \subset \Omega^{2n}} \left|\sum_{(\sigma,\tau) \in S} \sum_{i \in I} \mu(\sigma)\bar{\mu}(\tau)\left(\mathbf{1}\{\sigma_i = \omega\} - \mathbf{1}\{\tau_i = \omega\}\right)\right| \leq \varepsilon n. \tag{4.5}$$

Let $X(\sigma) = X(\sigma, I, \omega) = \sum_{i \in I} \mathbf{1}\{\sigma_i = \omega\}$ and denote by $\bar{X}$ its expectation with respect to μ, that is $\bar{X} = \langle X(\sigma), \mu \rangle$. Because μ is δ-symmetric we can bound the second moment of X as follows:

$$\left\langle X(\sigma)^2, \mu \right\rangle = \left\langle \sum_{i,j \in I} \mathbf{1}\{\sigma_i = \sigma_j = \omega\}, \mu \right\rangle = \sum_{i,j \in I} \mu_{ij}(\omega,\omega) \leq \left(\sum_{i,j \in I : i \neq j} \mu_i(\omega)\mu_j(\omega) + \sum_{i \in I} \mu_i(\omega)\right) + \delta n^2 \leq \bar{X}(1 + \bar{X}) + \delta n^2,$$

Hence,

$$\left\langle X(\sigma)^2,\mu\right\rangle - \bar{X}^2 \le \bar{X} + \delta n^2 \le |I| + \delta n^2. \tag{4.6}$$

Let $\alpha \in (0,1)$ and $P(\alpha) = \mu\left\{\left|X(\sigma) - \bar{X}\right| \ge \alpha n\right\}$. Then Chebyshev's inequality and (4.6) yield $P(\alpha) \le (|I| + \delta n^2)/(\alpha n)^2$. Hence, for events $S_h = \left\{\left|X(\sigma) - \bar{X}\right| \ge 2^h \varepsilon n\right\}$ we obtain $\mu(S_h) \le P(2^h\varepsilon) \le 4^{-h}\delta/\varepsilon^2 + O(1/n)$. Therefore,

$$\sup_{S\subset\Omega^{2n}}\left|\sum_{i\in I}\sum_{(\sigma,\tau)\in S}\mu(\sigma)\mathbf{1}\{\sigma_i=\omega\} - \bar{\mu}(\tau)\mathbf{1}\{\tau_i=\omega\}\right| \le \left\langle\left|X-\bar{X}\right|,\mu\right\rangle \le \sum_{h\ge 0}\mu(S_h)\cdot\varepsilon 2^h \le o(1) + \sum_{h\ge 0}2^{-h}\delta/\varepsilon = 2\delta/\varepsilon + o(1). \tag{4.7}$$

Thus, (4.5) follows from (4.7) and the choice of δ. □

4.2. **Continuity.** Recall that for a given $\mu \in \mathfrak{L}$ the pinned $\mu_{\hat{\boldsymbol{\sigma}}^\mu\downarrow n} \in \mathfrak{L}$ is random. Thus, for the pinned laws we consider the $D_\boxtimes$-Wasserstein metric. The aim in this paragraph is to establish the following key statement.

Proposition 4.5. *The operator $\mu \mapsto \mu_{\hat{\boldsymbol{\sigma}}^\mu\downarrow n}$ is $(D_\boxtimes(\cdot,\cdot),\mathscr{D}_\boxtimes(\cdot,\cdot))$-continuous for any $n \ge 1$.*

Toward the proof of Proposition 4.5 we need to consider a slightly generalised version of the pinning operation. Specifically, for a measurable map $\kappa : [0,1]^2 \to [0,1]^\Omega$ and $\tau \in \Omega^n$ let

$$\boldsymbol{z}_\tau(\kappa) = \int_0^1 \prod_{i=1}^n \kappa_{s,\hat{\boldsymbol{x}}_i}(\tau_i)\mathrm{d}s.$$

Thus, $\boldsymbol{z}_\tau$ is a random variable, dependent on the uniformly and independently chosen $\hat{\boldsymbol{x}}_1,\ldots,\hat{\boldsymbol{x}}_n \in [0,1]$. Also let $\boldsymbol{z}(\kappa) = \sum_{\tau\in\Omega^n}\boldsymbol{z}_\tau(\kappa)$. Further, define $\kappa_{\tau\downarrow n} \in \mathscr{K}_1$ as follows. If $\boldsymbol{z}_\tau(\kappa) = 0$, then we let $\kappa_{\tau\downarrow n} = \kappa$. But if $\boldsymbol{z}_\tau(\kappa) > 0$, then we let $\kappa_{\tau\downarrow n}$ be a kernel representation of the probability distribution

$$\int_0^1 \frac{\prod_{i=1}^n \kappa_{s,\hat{\boldsymbol{x}}_i}(\tau_i)}{\boldsymbol{z}_\tau(\kappa)}\delta_{\kappa_s}\mathrm{d}s \in \mathscr{P}(\mathscr{S}_1)$$

Additionally, let $\hat{\boldsymbol{\sigma}}^\kappa \in \Omega^n$ denote a vector drawn from the distribution $(\boldsymbol{z}_\tau(\kappa)/\boldsymbol{z}(\kappa))_{\tau\in\Omega^n}$ if $\boldsymbol{z}(\kappa) > 0$, and let $\hat{\boldsymbol{\sigma}}^\kappa \in \Omega^n$ be uniformly distributed otherwise.

Lemma 4.6. *For any $n \ge 1$, $\varepsilon > 0$ there is $\delta > 0$ such that for all $\kappa \in \mathscr{K}$ and all $\kappa' \in \mathscr{K}_1$ with $D_1(\kappa,\kappa') < \delta$ we have*

$$\mathscr{D}_\boxtimes\left(\kappa_{\hat{\boldsymbol{\sigma}}^\kappa\downarrow n},\kappa'_{\hat{\boldsymbol{\sigma}}^{\kappa'}\downarrow n}\right) < \varepsilon.$$

Toward the proof of Lemma 4.6 we require the following statement.

Lemma 4.7. *For any $n \ge 1$, $\varepsilon > 0$ and $\kappa \in \mathscr{K}$ we have $\mathbb{P}\left[\boldsymbol{z}_{\hat{\boldsymbol{\sigma}}^\kappa}(\kappa) < \varepsilon \mid \hat{\boldsymbol{x}}_1,\ldots,\hat{\boldsymbol{x}}_n\right] < \varepsilon|\Omega|^n$.*

Proof. We have $\mathbb{P}\left[\boldsymbol{z}_{\hat{\boldsymbol{\sigma}}}(\kappa) < \varepsilon \mid \hat{\boldsymbol{x}}_1,\ldots,\hat{\boldsymbol{x}}_n\right] = \sum_{\tau\in\Omega^n}\mathbf{1}\{\boldsymbol{z}_\tau(\kappa) < \varepsilon\}\boldsymbol{z}_\tau(\kappa) < \varepsilon|\Omega|^n$. □

Proof of Lemma 4.6. Given $\varepsilon > 0$ pick small enough $\eta = \eta(\varepsilon,n) > 0$, $\delta = \delta(\eta) > 0$. Consider $\kappa \in \mathscr{K}, \kappa' \in \mathscr{K}_1$ such that $D_1(\kappa,\kappa') < \delta$ and let $\mu = \mu^\kappa$, $\mu' = \mu^{\kappa'}$. Then we see that

$$\mathbb{P}\left[1-\eta < \boldsymbol{z}(\kappa') < 1+\eta\right] > 1-\eta. \tag{4.8}$$

Hence, in the following we may condition on the event that $1-\eta < \boldsymbol{z}(\kappa') < 1+\eta$. Given that this is so, choose $\hat{\boldsymbol{\sigma}},\hat{\boldsymbol{\sigma}}' \in \Omega^n$ from the distributions

$$\mathbb{P}\left[\hat{\boldsymbol{\sigma}} = \sigma \mid \hat{\boldsymbol{x}}_1,\ldots,\hat{\boldsymbol{x}}_n\right] = \boldsymbol{z}_\sigma(\kappa)/\boldsymbol{z}(\kappa) = \boldsymbol{z}_\sigma(\kappa), \qquad \mathbb{P}\left[\hat{\boldsymbol{\sigma}}' = \sigma \mid \hat{\boldsymbol{x}}_1,\ldots,\hat{\boldsymbol{x}}_n\right] = \boldsymbol{z}_\sigma(\kappa')/\boldsymbol{z}(\kappa') \qquad (\sigma\in\Omega^n).$$

Further, define the probability density functions

$$p_\kappa(s) = \frac{1}{\boldsymbol{z}_{\hat{\boldsymbol{\sigma}}}(\kappa)}\prod_{i=1}^n \kappa_{s,\hat{\boldsymbol{x}}_i}(\hat{\boldsymbol{\sigma}}_i), \qquad p_{\kappa'}(s) = \frac{1}{\boldsymbol{z}_{\hat{\boldsymbol{\sigma}}'}(\kappa')}\prod_{i=1}^n \kappa'_{s,\hat{\boldsymbol{x}}_i}(\hat{\boldsymbol{\sigma}}'_i) \qquad \text{and set}$$

$$\hat{p}(s) = p(s)\wedge p'(s), \qquad \hat{p}_\kappa(s) = p_\kappa(s) - \hat{p}(s), \qquad \hat{p}_{\kappa'}(s) = p_{\kappa'}(s) - \hat{p}(s)$$

so that

$$\mu_{\hat{\boldsymbol{\sigma}}\downarrow n} = \int_0^1 p_\kappa(s)\delta_{\kappa_s}\mathrm{d}s, \qquad \mu'_{\hat{\boldsymbol{\sigma}}'\downarrow n} = \int_0^1 p_{\kappa'}(s)\delta_{\kappa'_s}\mathrm{d}s.$$

To couple $\mu_{\hat{\boldsymbol{\sigma}}\downarrow n}, \mu'_{\hat{\boldsymbol{\sigma}}'\downarrow n}$ draw a pair $(\boldsymbol{t}, \boldsymbol{t}') \in [0,1]^2$ from the following distribution: with probability $\int_0^1 \hat{p}(s)\mathrm{d}s$, we draw $\boldsymbol{t} = \boldsymbol{t}'$ from the distribution $(\int_0^1 \hat{p}(s)\mathrm{d}s)^{-1}\hat{p}(s)\mathrm{d}s$, and with probability $1-\int_0^1 \hat{p}(s)\mathrm{d}s$ we draw $\boldsymbol{t}, \boldsymbol{t}'$ independently from the distributions

$$\left(1-\int_0^1 \hat{p}(s)\mathrm{d}s\right)^{-1}\hat{p}_\kappa(s)\mathrm{d}s, \qquad \left(1-\int_0^1 \hat{p}(s)\mathrm{d}s\right)^{-1}\hat{p}'_\kappa(s)\mathrm{d}s,$$

respectively. Then $(\kappa_{\boldsymbol{t}}, \kappa'_{\boldsymbol{t}'})$ provides a coupling of $\mu_{\hat{\boldsymbol{\sigma}}\downarrow n}, \mu'_{\hat{\boldsymbol{\sigma}}'\downarrow n}$. Consequently,

$$\mathscr{D}_{\boxtimes}(\mu_{\hat{\boldsymbol{\sigma}}\downarrow n}, \mu'_{\hat{\boldsymbol{\sigma}}'\downarrow n}) \le D_1(\kappa,\kappa') + \mathbb{P}\left[\boldsymbol{t}\neq\boldsymbol{t}'\right] + \mathbb{P}\left[\boldsymbol{z}(\kappa')\notin(1-\eta,1+\eta)\right] < \delta + \mathbb{P}\left[\boldsymbol{t}\neq\boldsymbol{t}'\right] + \mathbb{P}\left[\boldsymbol{z}(\kappa')\notin(1-\eta,1+\eta)\right]. \tag{4.9}$$

To estimate $\mathbb{P}\left[\boldsymbol{t}\neq\boldsymbol{t}'\right]$ let

$$\mathscr{E} = \left\{\sum_{\tau\in\Omega^n}\int_0^1 \left|\prod_{i=1}^n \kappa_{s,\hat{\boldsymbol{x}}_i}(\tau_i) - \prod_{i=1}^n \kappa'_{s,\hat{\boldsymbol{x}}_i}(\tau_i)\right|\mathrm{d}s < \eta^2\right\}.$$

Picking δ sufficiently small ensures that

$$\mathbb{P}\left[\mathscr{E}\right] > 1-\eta \tag{4.10}$$

and on the event $\mathscr{E}$ we have

$$d_{\mathrm{TV}}\left(\hat{\boldsymbol{\sigma}}, \hat{\boldsymbol{\sigma}}'\right) = \frac{1}{2}\sum_{\sigma\in\Omega^n}\left|\mathbb{P}\left[\hat{\boldsymbol{\sigma}}=\sigma\right] - \mathbb{P}\left[\hat{\boldsymbol{\sigma}}'=\sigma\right]\right| = \sum_{\sigma\in\Omega^n}\left|\boldsymbol{z}_\sigma(\kappa) - \boldsymbol{z}_\sigma(\kappa')/\boldsymbol{z}(\kappa)\right| < \eta.$$

Hence, on $\mathscr{E}$ we can couple $\hat{\boldsymbol{\sigma}}, \hat{\boldsymbol{\sigma}}'$ such that

$$\mathbb{P}[\hat{\boldsymbol{\sigma}}\neq\hat{\boldsymbol{\sigma}}'] < \eta. \tag{4.11}$$

Additionally, let $\mathscr{E}' = \left\{\hat{\boldsymbol{\sigma}}=\hat{\boldsymbol{\sigma}}',\, \boldsymbol{z}_{\hat{\boldsymbol{\sigma}}}(\kappa) \ge \eta^{1/3}\right\}$. Then Lemma 4.7, (4.10) and (4.11) imply that

$$\mathbb{P}\left[\mathscr{E}' \mid \mathscr{E}\right] \ge 1 - 2\eta^{1/3}|\Omega|^n. \tag{4.12}$$

Moreover, on $\mathscr{E}\cap\mathscr{E}'$ we have

$$\left|\boldsymbol{z}_{\hat{\boldsymbol{\sigma}}}(\kappa') - \boldsymbol{z}_{\hat{\boldsymbol{\sigma}}}(\kappa)\right| \le \eta$$

and consequently

$$\begin{aligned}\mathbb{P}\left[\boldsymbol{t}\neq\boldsymbol{t}' \mid \mathscr{E}\cap\mathscr{E}'\right] &= 1-\int_0^1 \hat{p}(s)\mathrm{d}s = 1-\frac{1}{2}\int_0^1\left(p(s)+p'(s)-\left|p(s)-p'(s)\right|\right)\mathrm{d}s = \frac{1}{2}\int_0^1\left|p(s)-p'(s)\right|\mathrm{d}s\\ &\le \frac{1}{2\boldsymbol{z}_{\hat{\boldsymbol{\sigma}}}(\kappa)}\int_0^1\left|\prod_{i=1}^n\kappa_{s,\hat{\boldsymbol{x}}_i}(\hat{\boldsymbol{\sigma}}_i) - \prod_{i=1}^n\kappa'_{s,\hat{\boldsymbol{x}}_i}(\hat{\boldsymbol{\sigma}}_i)\right|\mathrm{d}s + \frac{\boldsymbol{z}_{\hat{\boldsymbol{\sigma}}}(\kappa)-\boldsymbol{z}_{\hat{\boldsymbol{\sigma}}}(\kappa')}{2\boldsymbol{z}_{\hat{\boldsymbol{\sigma}}}(\kappa)\boldsymbol{z}_{\hat{\boldsymbol{\sigma}}}(\kappa')} \le \sqrt{\eta}.\end{aligned} \tag{4.13}$$

Finally, the assertion follows from (4.9), (4.10), (4.12) and (4.13). □

Lemma 4.8. *For any $\varepsilon>0$, $\ell\ge 1$ there is $\delta>0$ such that for all $\kappa\in\mathscr{K}$ such that $\mu^\kappa\in\mathscr{L}$ is supported on a set of size at most ℓ and all $\iota\in\mathscr{K}_1$ with $D_\square(\kappa,\iota)<\delta$ we have $\mathscr{D}_{\boxtimes}(\kappa_{\hat{\boldsymbol{\sigma}}^\kappa\downarrow n}, \iota_{\hat{\boldsymbol{\sigma}}^\iota\downarrow n}) < \varepsilon$.*

Proof. Pick $\alpha=\alpha(\varepsilon,\ell,n)$, $\beta=\beta(\alpha)$, $\xi=\xi(\beta)$, $\zeta=\zeta(\xi)$, $\eta=\eta(\zeta)>0$ and $\delta=\delta(\eta)>0$ sufficiently small. To summarise,

$$0<\delta\ll\eta\ll\zeta\ll\xi\ll\beta\ll\alpha\ll\varepsilon/(n+\ell). \tag{4.14}$$

We may assume that there is a partition $S_1,\dots,S_\ell$ of $[0,1]$ such that κ is constant on $S_i\times\{x\}$ for all $x\in[0,1]$. Moreover, we may assume without loss that there is $k\in[\ell]$ such that $\lambda(S_i)>\eta$ for all $i\le k$, while $\lambda(S_i)<\eta$ for all $i>k$. Let $t_i:[0,1]\to S_i$ be a measurable bijection that maps the Lebesgue measure on $[0,1]$ to the probability measure $\lambda(S_i)^{-1}\mathrm{d}s$ on S_i for $i\le k$ (see Lemma 2.3). Assuming that δ is small enough, we see that the kernels

$$\kappa^{(i)}_{s,x} = \kappa_{t_i(s),x}, \qquad \iota^{(i)}_{s,x} = \iota_{t_i(s),x}$$

have cut distance

$$D_\square(\kappa^{(i)},\iota^{(i)}) < \zeta \qquad \text{for all } i\le k. \tag{4.15}$$

Combining Proposition 3.8 and (4.15), we conclude that after an n-fold application of the $\oplus'$-operation we have $D_\square(\kappa^{(i)\oplus' n}, \iota^{(i)\oplus' n}) < \xi$. Since for every $x \in [0,1]$ the map $s \mapsto \kappa^{(i)\oplus' n}_{s,x}$ is constant, we therefore find that

$$\sum_{\tau\in\Omega^n} \mathbb{E}\left|\mathbb{E}\left[\prod_{j=1}^n \kappa^{(i)}_{\boldsymbol{s},\hat{\boldsymbol{x}}_1,\ldots,\hat{\boldsymbol{x}}_n}(\tau_j) - \prod_{j=1}^n \iota^{(i)}_{\boldsymbol{s},\hat{\boldsymbol{x}}_1,\ldots,\hat{\boldsymbol{x}}_n}(\tau_j)\,\middle|\,\hat{\boldsymbol{x}}_1,\ldots,\hat{\boldsymbol{x}}_n\right]\right| < \beta \qquad \text{for all } i \le k. \tag{4.16}$$

Because $\lambda(S_i) < \eta$ for all $i > k$ and $\ell\eta < \beta$ for small enough η, (4.16) implies that

$$\sum_{\tau\in\Omega^n} \mathbb{E}\left|\mathbb{E}\left[\prod_{j=1}^n \kappa_{\boldsymbol{s},\hat{\boldsymbol{x}}_1,\ldots,\hat{\boldsymbol{x}}_n}(\tau_j) - \prod_{j=1}^n \iota_{\boldsymbol{s},\hat{\boldsymbol{x}}_1,\ldots,\hat{\boldsymbol{x}}_n}(\tau_j)\,\middle|\,\hat{\boldsymbol{x}}_1,\ldots,\hat{\boldsymbol{x}}_n\right]\right| < 2\beta. \tag{4.17}$$

Combining (4.17) with Markov's inequality, we conclude that

$$\mathbb{P}[\mathscr{E}] > 1-\beta^{1/3}, \qquad \text{where} \qquad \mathscr{E} = \left\{\sum_{\tau\in\Omega^n}\left|\int_0^1 \prod_{j=1}^n \kappa_{s,\hat{\boldsymbol{x}}_1,\ldots,\hat{\boldsymbol{x}}_n}(\tau_j) - \prod_{j=1}^n \iota_{s,\hat{\boldsymbol{x}}_1,\ldots,\hat{\boldsymbol{x}}_n}(\tau_j)\mathrm{d}s\right| < \beta^{1/3}\right\}. \tag{4.18}$$

Consequently, on $\mathscr{E}$ we have

$$\sum_{\tau\in\Omega^n} |\boldsymbol{z}_\kappa(\tau) - \boldsymbol{z}_\iota(\tau)| < \beta^{1/3}. \tag{4.19}$$

In particular, there exists a coupling of the reference configurations $\hat{\boldsymbol{\sigma}}^\kappa, \hat{\boldsymbol{\sigma}}^\iota \in \Omega^n$ such that $\mathbb{P}\left[\hat{\boldsymbol{\sigma}}^\kappa = \hat{\boldsymbol{\sigma}}^\iota\right] \ge 1-\beta^{1/4}$. Hence, Lemma 4.7 implies that the event $\mathscr{E}' = \left\{\hat{\boldsymbol{\sigma}}^\kappa = \hat{\boldsymbol{\sigma}}^\iota,\, \boldsymbol{z}_\kappa(\hat{\boldsymbol{\sigma}}^\kappa) \ge \alpha\right\}$ satisfies

$$\mathbb{P}\left[\mathscr{E}' \mid \mathscr{E}\right] \ge 1 - |\Omega|^n\alpha. \tag{4.20}$$

To complete the proof let

$$p_\kappa(s) = \prod_{i=1}^n \kappa_{s,\hat{\boldsymbol{x}}_i}(\hat{\boldsymbol{\sigma}}^\kappa_i), \qquad p_\iota(s) = \prod_{i=1}^n \iota_{s,\hat{\boldsymbol{x}}_i}(\hat{\boldsymbol{\sigma}}^\iota_i) \qquad \text{and}$$
$$\hat{p}_{\kappa,i} = \int_{S_i} p_\kappa(s)\mathrm{d}s, \qquad \hat{p}_{\iota,i} = \int_{S_i} p_\iota(s)\mathrm{d}s.$$

Further, let $\mathscr{E}'' = \left\{\sum_{i=1}^\ell \left|\hat{p}_{\kappa,i} - \hat{p}_{\iota,i}\right| < \alpha^3\right\}$. Then (4.16), (4.18) and (4.20) imply that

$$\mathbb{P}\left[\mathscr{E}'' \mid \mathscr{E}\cap\mathscr{E}'\right] > 1-\alpha. \tag{4.21}$$

Moreover, since $\boldsymbol{z}_\kappa(\hat{\boldsymbol{\sigma}}^\kappa) \ge \alpha$ and $\hat{\boldsymbol{\sigma}}^\kappa = \hat{\boldsymbol{\sigma}}^\iota$, on $\mathscr{E}\cap\mathscr{E}'\cap\mathscr{E}''$ we have $\boldsymbol{z}_\iota(\hat{\boldsymbol{\sigma}}^\iota) \ge \alpha/2$. Therefore, on $\mathscr{E}\cap\mathscr{E}'\cap\mathscr{E}''$ the probability distributions $(p_{\kappa,i})_{i\in[\ell]}, (p_{\iota,i})_{i\in[\ell]}$ with

$$p_{\kappa,i} = \hat{p}_{\kappa,i}/\boldsymbol{z}_\kappa(\hat{\boldsymbol{\sigma}}^\kappa), \qquad p_{\iota,i} = \hat{p}_{\iota,i}/\boldsymbol{z}_\iota(\hat{\boldsymbol{\sigma}}^\iota)$$

have total variation distance $d_{\mathrm{TV}}((p_{\kappa,i})_{i\in[\ell]}, (p_{\iota,i})_{i\in[\ell]}) < 2\alpha$. Consequently, there exists a coupling of random variables $\boldsymbol{i}_\kappa, \boldsymbol{i}_\iota$ with these distributions such that

$$\mathbb{P}\left[\boldsymbol{i}_\kappa \neq \boldsymbol{i}_\iota \mid \mathscr{E}\cap\mathscr{E}'\cap\mathscr{E}''\right] < 2\alpha. \tag{4.22}$$

We extend this coupling to a coupling γ of $\mu_{\hat{\boldsymbol{\sigma}}^\mu\downarrow n}, \nu_{\hat{\boldsymbol{\sigma}}^\nu\downarrow n}$: given $\boldsymbol{i}_\kappa, \boldsymbol{i}_\iota$, pick any $\boldsymbol{s}_\kappa \in S_{\boldsymbol{i}_\kappa}$ and choose $\boldsymbol{s}_\iota \in S_{\boldsymbol{i}_\iota}$ from the distribution $p_\iota(s)/\hat{p}_{\iota,\boldsymbol{i}_\iota}\mathrm{d}s$. Then $\kappa_{\boldsymbol{s}_\kappa}$, $\iota_{\boldsymbol{s}_\iota}$ have distribution $\mu_{\hat{\boldsymbol{\sigma}}^\mu\downarrow n}, \nu_{\hat{\boldsymbol{\sigma}}^\nu\downarrow n}$, respectively. Further, we claim that on $\mathscr{E}\cap\mathscr{E}'\cap\mathscr{E}''$,

$$\left|\int_B\int_X \sigma_x(\omega) - \tau_x(\omega)\mathrm{d}x\mathrm{d}\gamma(\sigma,\tau)\right| < \varepsilon \qquad \text{for all } B\subset\mathscr{S}\times\mathscr{S}, X\subset[0,1], \omega\in\Omega. \tag{4.23}$$

Indeed, thanks to (4.22), we may condition on the event $\boldsymbol{i}_\iota = \boldsymbol{i}_\kappa \le k$. Hence, to prove (4.23) it suffices to show that for any $S\subset S_{\boldsymbol{i}_\iota}$, $X\subset[0,1]$, $\omega\in\Omega$,

$$\int_X\int_S \frac{p_\iota(s)}{\hat{p}_{\iota,\boldsymbol{i}_\iota}}\left(\iota_{s,x}(\omega) - \kappa_{\boldsymbol{s}_\kappa,x}(\omega)\right)\mathrm{d}s\mathrm{d}x < \varepsilon/|\Omega|. \tag{4.24}$$

Because $\boldsymbol{z}_\iota(\hat{\boldsymbol{\sigma}}^\kappa) \ge \alpha/2$ we may also assume that $\hat{p}_{\iota,\boldsymbol{i}_\iota} \ge \alpha^2/\ell$, and we observe that $p_\iota(s) \le 1$. Now, assume for contradiction that there exist S, X, ω for which (4.24) is violated. Letting

$$S^+ = \left\{s\in S : \int_X \left(\iota_{s,x}(\omega) - \kappa_{\boldsymbol{s}_\kappa,x}(\omega)\right)\mathrm{d}x > \alpha\right\},$$

we conclude that

$$\frac{\varepsilon}{2|\Omega|} \le \int_X \int_{S^+} \frac{p_\iota(s)}{\hat{p}_{\iota,\boldsymbol{i}_\iota}} \big(\iota_{s,x}(\omega) - \kappa_{\boldsymbol{s}_\kappa,x}(\omega)\big)\,\mathrm{d}s\mathrm{d}x = \int_{S^+} \frac{p_\iota(s)}{\hat{p}_{\iota,\boldsymbol{i}_\iota}} \int_X \big(\iota_{s,x}(\omega) - \kappa_{\boldsymbol{s}_\kappa,x}(\omega)\big)\,\mathrm{d}x\mathrm{d}s$$
$$\le \frac{\ell}{\alpha^2} \int_{S^+} \int_X \big(\iota_{s,x}(\omega) - \kappa_{\boldsymbol{s}_\kappa,x}(\omega)\big)\,\mathrm{d}x\mathrm{d}s \le \ell\alpha^{-2} D_\square(\iota^{\boldsymbol{i}_\iota}, \kappa^{\boldsymbol{i}_\iota}) \le \ell\alpha^{-2}\zeta \qquad \text{due to (4.15).} \tag{4.25}$$

But (4.25) contradicts the choice of the parameters from (4.14). Hence, we obtain (4.24) and thus (4.23). Finally, the assertion follows from (4.18), (4.20), (4.21) and (4.23). □

Lemma 4.9. *For every sequence $(k_i)_i$ in $\mathscr{K}_1$ that converges to a kernel $k \in \mathscr{K}_1$ with respect to $D_\square(\cdot,\cdot)$ and for every kernel $k' \in \mathscr{K}_1$ there is a sequence of kernels $(k'_i)_i$, $k'_i \in \mathscr{K}_1$, s.t. $D_\square(k'_i, k') \to 0$ and $D_1(k_i, k'_i) \to D_1(k, k')$.*

Proof. Let $(\kappa^\omega)_\omega, (\kappa'^\omega)_\omega, (\kappa_i^\omega)_\omega, (\kappa_i'^\omega)_\omega$ be the families of bipartite graphons representing $k, k', (k_i)_i, (k'_i)_i$ given by (2.6). From the definition of $D_1(\cdot,\cdot)$ and Lemma 2.5 we get

$$D_\square(k_i, k) = \frac{1}{2}\max_\omega D_\square(\kappa_i^\omega, \kappa^\omega) \qquad \text{and} \qquad D_1(k, k') = \frac{1}{2}\sum_\omega D_1(\kappa^\omega, \kappa'^\omega). \tag{4.26}$$

The lemma follows from (4.26) and [32, Proposition 8.25]. □

The following lemma and the proof of Proposition 4.5 are adaptations of [32, Proof of Lemma 9.16].

Lemma 4.10. *Let $\varepsilon, \delta > 0$ and let $k \in \mathscr{K}$. Let $U_k(\delta,\varepsilon)$ be the set of all $\kappa \in \mathscr{K}$ such that there exists $\kappa' \in \mathscr{K}_1$ with $D_\square(k,\kappa') < \delta$ and $D_1(\kappa',\kappa) < \varepsilon$. Then $U_k(\delta,\varepsilon)$ is $D_\square$-open.*

Proof. Suppose that $\kappa \in U_k(\delta,\varepsilon)$ and that the sequence $(\kappa_i)_{i\ge1}$ satisfies $\lim_{i\to\infty} D_\square(\kappa,\kappa_i) = 0$. It suffices to show that $\kappa_i \in U_k(\delta,\varepsilon)$ for all large enough i. To this end consider $\kappa' \in \mathscr{K}_1$ such that $D_\square(k,\kappa') < \delta$ and $D_1(\kappa',\kappa) < \varepsilon$. By Lemma 4.9 there exists a sequence $\kappa'_i \in \mathscr{K}_1$ such that $\lim_{i\to\infty} D_\square(\kappa',\kappa'_i) = 0$ and $\lim_{i\to\infty} D_1(\kappa'_i,\kappa_i) = D_1(\kappa,\kappa')$. For this sequence we have

$$D_\square(\kappa'_i,\kappa') \to 0, \qquad\qquad D_1(\kappa_i,\kappa'_i) \to D_1(\kappa,\kappa') < \varepsilon.$$

Therefore, for large enough i we have $D_1(\kappa_i,\kappa'_i) < \varepsilon$ and $D_\square(k,\kappa'_i) \le D_\square(k,\kappa') + D_\square(\kappa'_i,\kappa') < \delta$, whence $\kappa'_i \in U_k(\delta,\varepsilon)$. □

Proof of Proposition 4.5. Fix $\varepsilon > 0$. Lemma 4.6 shows that there exists $\delta_0 > 0$ such that for all $\kappa \in \mathscr{K}, \kappa' \in \mathscr{K}_1$,

$$D_1(\kappa,\kappa') < \delta_0 \quad\Rightarrow\quad \mathscr{D}_\boxtimes(\mu^\kappa_{\hat{\boldsymbol{\sigma}}^\kappa\downarrow n}, \mu^{\kappa'}_{\hat{\boldsymbol{\sigma}}^{\kappa'}\downarrow n}) < \varepsilon/2. \tag{4.27}$$

Similarly, by Lemma 4.8 there exists a sequence $(\delta_\ell)_\ell$ such that for all $\mu,\nu \in \mathscr{L}$ with μ supported on at most $\ell \ge 1$ configurations we have

$$D_\boxtimes(\mu,\nu) < \delta_\ell \quad\Rightarrow\quad \mathscr{D}_\boxtimes(\mu_{\hat{\boldsymbol{\sigma}}^\mu\downarrow n}, \nu_{\hat{\boldsymbol{\sigma}}^\mu\downarrow n}) < \varepsilon/2. \tag{4.28}$$

Suppose that $k : [0,1]^2 \to \mathscr{P}(\Omega)$ is a step function that takes $\ell \ge 1$ different values and let $\mathscr{U}_k = \mathscr{U}_k(\delta_\ell,\delta_0)$ be as in Lemma 4.10. Then $\mathscr{U}_k$ is $D_\square$-open and $\bigcup_k \mathscr{U}_k = \mathscr{K}$ because $\mathscr{U}_k$ contains the δ_0-ball around k with respect to the D_1-metric. Further, let $\mathfrak{U}_k \subset \mathfrak{K}$ be the projection of $\mathscr{U}_k$ onto $\mathfrak{K}$. Then $\mathfrak{U}_k$ is open because the canonical map $\mathscr{K} \to \mathfrak{K}$ is open. Moreover, $\bigcup_k \mathfrak{U}_k = \mathfrak{K}$. Hence, a finite number of sets $\mathfrak{U}_k$ cover $\mathfrak{K}$. Thus, the assertion follows from (4.27) and (4.28). □

4.3. **Proof of Theorem 1.12.** Let $\varepsilon > 0$ and pick a small enough $\delta > 0$ and then a large enough $N > 0$. Also let $T = T(\varepsilon) = 64\varepsilon^{-8}\log|\Omega|$. Given $\mu \in \mathfrak{L}$ we apply Theorem 1.9 to obtain a probability distribution $\nu \in \mathscr{L}_N$ such that $D_\boxtimes(\mu,\hat{\nu}) < \delta$. Invoking Theorem 1.1 and Proposition 4.5, we find

$$D_\boxtimes(\mu_{\hat{\boldsymbol{\sigma}}^\mu\downarrow n}, \hat{\nu}_{\hat{\boldsymbol{\sigma}}^{\hat{\nu}}\downarrow n}) < \varepsilon/4 \qquad\qquad \text{for all } n \le T(\varepsilon). \tag{4.29}$$

By construction, for any n the law $\hat{\nu}_{\hat{\boldsymbol{\sigma}}^{\hat{\nu}}\downarrow n}$ obtained by first embedding $\nu \in \mathscr{L}_N$ into $\mathscr{L}$ and then applying the pinning operation coincides with the law obtained by first applying (4.1) to ν and then embedding the resulting $\hat{\nu}$ into $\mathfrak{L}$. Hence, Theorem 4.1 and Lemma 4.4 show that for a uniform $\boldsymbol{\theta} \le T(\varepsilon)$,

$$\mathbb{E}[D_\boxtimes(\overline{\hat{\nu}_{\hat{\boldsymbol{\sigma}}^{\hat{\nu}}\downarrow\boldsymbol{\theta}}}, \hat{\nu}_{\hat{\boldsymbol{\sigma}}^{\hat{\nu}}\downarrow\boldsymbol{\theta}})] < \varepsilon^2/2. \tag{4.30}$$

Further, Theorem 1.2, Theorem 1.11 and (4.29) show that

$$\begin{aligned} D_{\boxtimes}\left(\overline{\mu_{\hat{\boldsymbol{\sigma}}^{\mu}\downarrow\boldsymbol{\theta}}}, \mu_{\hat{\boldsymbol{\sigma}}^{\mu}\downarrow\boldsymbol{\theta}}\right) &\leq D_{\boxtimes}\left(\mu_{\hat{\boldsymbol{\sigma}}^{\mu}\downarrow\boldsymbol{\theta}}, \dot{\nu}_{\hat{\boldsymbol{\sigma}}^{\dot{\nu}}\downarrow\boldsymbol{\theta}}\right) + D_{\boxtimes}\left(\overline{\dot{\nu}_{\hat{\boldsymbol{\sigma}}^{\dot{\nu}}\downarrow\boldsymbol{\theta}}}, \dot{\nu}_{\hat{\boldsymbol{\sigma}}^{\dot{\nu}}\downarrow\boldsymbol{\theta}}\right) + D_{\boxtimes}\left(\overline{\dot{\nu}_{\hat{\boldsymbol{\sigma}}^{\dot{\nu}}\downarrow\boldsymbol{\theta}}}, \overline{\mu_{\hat{\boldsymbol{\sigma}}\downarrow\boldsymbol{\theta}}}\right) \\ &\leq 2D_{\boxtimes}\left(\mu_{\hat{\boldsymbol{\sigma}}\downarrow\boldsymbol{\theta}}, \dot{\nu}_{\hat{\boldsymbol{\sigma}}^{\dot{\nu}}\downarrow\boldsymbol{\theta}}\right) + D_{\boxtimes}\left(\overline{\dot{\nu}_{\hat{\boldsymbol{\sigma}}^{\dot{\nu}}\downarrow\boldsymbol{\theta}}}, \dot{\nu}_{\hat{\boldsymbol{\sigma}}^{\dot{\nu}}\downarrow\boldsymbol{\theta}}\right) < \varepsilon + D_{\boxtimes}\left(\overline{\dot{\nu}_{\hat{\boldsymbol{\sigma}}^{\dot{\nu}}\downarrow\boldsymbol{\theta}}}, \dot{\nu}_{\hat{\boldsymbol{\sigma}}^{\dot{\nu}}\downarrow\boldsymbol{\theta}}\right). \end{aligned} \tag{4.31}$$

Combining (4.30) and (4.31) and applying Markov's inequality, we obtain the first part of Theorem 1.12. The second assertion follows from a similar argument.

4.4. **Proof of Theorem 1.2.** We postponed the proof Theorem 1.2, because it relies on some of the prior results from this section. To finally carry the proof out we adapt the proof strategy from [32], where a statement similar to Theorem 1.2 was established for graphons, to the present setting of probability distributions. We begin with the following simple bound.

Lemma 4.11. *For any $\mu, \nu \in \mathscr{L}_n$ we have $\Delta_{\boxtimes}(\mu,\nu) \leq n^3 D_{\boxtimes}(\dot{\mu},\dot{\nu})$.*

Proof. Let $\psi \in \mathbb{S}$ and let $\gamma \in \Gamma(\dot{\mu},\dot{\nu})$. We are going to show that there exist a coupling $g \in \Gamma(\mu,\nu)$ and a permutation $\phi \in \mathbb{S}_n$ such that

$$\max_{\substack{S\subset\Omega^n\times\Omega^n\\ X\subset[n]\\ \omega\in\Omega}} \left| \sum_{(\sigma,\sigma')\in S, x\in X} g(\sigma,\sigma')\left(\mathbf{1}\{\sigma_x=\omega\} - \mathbf{1}\left\{\sigma'_{\phi(x)}=\omega\right\}\right)\right| \leq n^4 \sup_{\substack{S\subset\mathscr{S}\times\mathscr{S}\\ X\subset[0,1]\\ \omega\in\Omega}} \left|\int_S\int_X \left(\sigma_x(\omega)-\sigma'_{\psi(x)}(\omega)\right)\mathrm{d}x\mathrm{d}s\right|. \tag{4.32}$$

The assertion is immediate from (4.32) and the definitions (1.1), (1.2).

With respect to the coupling g, matters are easy: the construction of $\dot{\mu},\dot{\mu}' \in \mathscr{L}$ ensures that the coupling γ readily induces a coupling g of the original probability distributions μ,ν such that $g(\sigma,\tau) = \gamma(\dot{\sigma},\dot{\tau})$ for all $\sigma,\tau\in\Omega^n$.

We are left to exhibit the permutation ϕ. To this end let $I_j = [(j-1)/n, j/n)$. We construct a bipartite auxiliary graph $\mathscr{G}$ with vertex set $\{v_1,\ldots,v_n\}\cup\{w_1,\ldots,w_n\}$ in which v_i, w_j are adjacent iff $\lambda(I_j\cap\psi(I_i)) \geq n^{-3}$. Then the Hall's theorem implies that $\mathscr{G}$ possesses a perfect matching. Indeed, assume that $\emptyset \neq V \subset \{v_1,\ldots,v_n\}$ satisfies $|\partial V| < |V|$. Then because ψ preserves the Lebesgue measure we obtain

$$\frac{1}{n} \leq \frac{|V|-|\partial V|}{n} = \frac{|V|}{n} - \sum_{v_i\in V,\, w_j\in\partial V} \lambda(I_j\cap\psi(I_i)) = \sum_{v_i\in V,\, w_j\notin\partial V} \lambda(I_j\cap\psi(I_i)) \leq \frac{|V|(n-|V|)}{n^3} < \frac{1}{n},$$

a contradiction. Thus, let ϕ be the permutation of $[n]$ induced by any perfect matching of $\mathscr{G}$.

To complete the proof we claim that g,ϕ satisfy (4.32). Indeed, given a set $S\subset\Omega^n\times\Omega^n$ let $\dot{S} = \{(\dot{\sigma},\dot{\tau}) : (\sigma,\tau)\in S\}$. Further, for $X\subset[n]$ let $\dot{X}\subset[0,1]$ be any measurable set such that $\lambda(\dot{X}\cap I_i) = n^{-3}$ for all $i\in[n]$ and $\dot{X}\cap I_j \subset \psi(I_i)$ if $\phi(i) = j$. Then we obtain (4.32). □

As a second step we will complement the coarse multiplicative bound from Lemma 4.11 with a somewhat more subtle additive bound. To this end, we need an enhanced version of a 'Frieze-Kannan type' regularity lemma for probability distributions. Specifically, let $\mu\in\mathscr{L}_n$ and let $S = \{S_1,\ldots,S_k\}$ and $X = \{X_1,\ldots,X_\ell\}$ be partitions of Ω^n and $[n]$, respectively. We call the partition S *canonical* if there exists a set $\mathscr{I}\subset[n]$ such that

$$S = \left\{\left\{\sigma\in\Omega^n : \forall i\in\mathscr{I} : \sigma_i = \tau_i\right\} : \tau\in\Omega^{\mathscr{I}}\right\}.$$

In words, S partitions the discrete cube Ω^n into the $\Omega^{|\mathscr{I}|}$ sub-cubes defined by the entries on the set $\mathscr{I}$ of coordinates. In this case we define

$$\mu^{S,X}(\sigma) = \sum_{h=1}^{k} \mu(S_h) \prod_{i=1}^{\ell} \prod_{j\in X_i} \sum_{x\in X_i} \frac{\mu_x(\sigma_j|S_h)}{|X_i|} \in \mathscr{L}_n.$$

Thus, $\mu^{S,X}$ is a mixture of product measures, one for each class of the partition S.

Lemma 4.12. *For any Ω there exists $c = c(\Omega) > 0$ such that for every $0<\varepsilon<1/2$, $n>0$ and all $\mu,\nu\in\mathscr{L}_n$ there exist a canonical partition $S_1,\ldots,S_k$ of Ω^n and a partition $X_1,\ldots,X_\ell$ of $[n]$ such that the following statements are satisfied.*

- $k+\ell \leq \exp(\varepsilon^{-c})$.

- *with* $\gamma \in \Gamma(\mu, \mu^{S,X})$ *and* $\gamma' \in \Gamma(\nu, \nu^{S,X})$ *defined by*

$$\gamma(\sigma,\tau) = \sum_{h=1}^{k} \mathbf{1}\{\sigma,\tau \in S_h\}\,\mu(\sigma)\mu^{S,X}(\tau)/\mu(S_h),$$
$$\gamma'(\sigma,\tau) = \sum_{h=1}^{k} \mathbf{1}\{\sigma,\tau \in S_h\}\,\nu(\sigma)\nu^{S,X}(\tau)/\nu(S_h)$$

we have

$$\max_{S\subset\Omega^n\times\Omega^n,\, X\subset[n],\omega\in\Omega} \left| \sum_{(\sigma,\tau)\in S} \sum_{x\in X} \gamma(\sigma,\tau)\left(\mathbf{1}\{\sigma_x=\omega\} - \mathbf{1}\{\tau_x=\omega\}\right)\right| < \varepsilon n, \tag{4.33}$$

$$\max_{S\subset\Omega^n\times\Omega^n,\, X\subset[n],\omega\in\Omega} \left| \sum_{(\sigma,\tau)\in S} \sum_{x\in X} \gamma'(\sigma,\tau)\left(\mathbf{1}\{\sigma_x=\omega\} - \mathbf{1}\{\tau_x=\omega\}\right)\right| < \varepsilon n. \tag{4.34}$$

Hence, $\Delta_{\boxtimes}(\mu,\mu^{S,X}) < \varepsilon$, $\Delta_{\boxtimes}(\nu,\nu^{S,X}) < \varepsilon$.

Proof. Combining Theorem 4.1 and Lemma 4.4, we find a set $\mathscr{F} \subset [n]$ such that the induced canonical partition $S_1,\dots,S_k$ satisfies

$$\sum_{i=1}^{k} \mu(S_i)\Delta_{\square}\left(\mu[\cdot|S_i], \bigotimes_{x=1}^{n} \mu_x[\cdot|S_i]\right) < \varepsilon/8, \qquad \sum_{i=1}^{k} \nu(S_i)\Delta_{\square}\left(\nu[\cdot|S_i], \bigotimes_{x=1}^{n} \nu_x[\cdot|S_i]\right) < \varepsilon/8. \tag{4.35}$$

Moreover, the size k of the partition is bounded by $\exp(\varepsilon^{-c'})$ for some $c' = c'(\Omega)$. Now, for each $i \in [k]$ we can partition the set $[n]$ into at most $32/\varepsilon$ classes $X_{i,1},\dots,X_{i,\ell_i}$ such that for all $x,y \in X_{i,j}$ we have $d_{\mathrm{TV}}(\mu_x[\cdot|S_i],\mu_y[\cdot|S_i]) < \varepsilon/16$. A similar partition $X'_{i,1},\dots,X'_{i,\ell'_i}$ exists for $\nu[\cdot|S_i]$. Hence, the smallest common refinement $X_1,\dots,X_\ell$ of all these partitions $(X_{i,j}),(X'_{i,j})$ has at most $\exp(\varepsilon^{-c})/2$ classes, for some suitable $c = c(\Omega) > 0$. Further, by construction, letting

$$\mu^{(i)}(\sigma) = \prod_{j=1}^{\ell} \prod_{x\in X_j} \frac{1}{|X_j|} \sum_{x\in X_j} \mu_x(\sigma_x|S_i), \qquad \nu^{(i)}(\sigma) = \prod_{j=1}^{\ell} \prod_{x\in X_j} \frac{1}{|X_j|} \sum_{x\in X_j} \nu_x(\sigma_x|S_i),$$

we obtain from (4.35) that

$$\sum_{i=1}^{k} \mu(S_i)\Delta_{\square}\left(\mu[\cdot|S_i],\mu^{(i)}\right) < \varepsilon/4, \qquad \sum_{i=1}^{k} \nu(S_i)\Delta_{\square}\left(\nu[\cdot|S_i],\nu^{(i)}\right) < \varepsilon/4. \tag{4.36}$$

In addition, since $\mu^{(i)},\nu^{(i)}$ are product measures, the couplings $\gamma^{(i)},\gamma^{(i)\prime}$ for which the cut distance in (4.36) attained are trivial, i.e., $\gamma^{(i)} = \mu[\cdot|S_i]\otimes\mu^{(i)}$ and $\gamma^{(i)\prime} = \nu[\cdot|S_i]\otimes\nu^{(i)}$. Therefore, (4.36) implies (4.33)–(4.34). □

Lemma 4.13. *For any* $\mu,\nu \in \mathscr{L}_n$ *we have* $\Delta_{\boxtimes}(\mu,\nu) \le D_{\boxtimes}(\dot\mu,\dot\nu) + o(1)$ *as* $n\to\infty$.

Proof. Let $0 < \varepsilon = \varepsilon(n) = o(1)$ be a sequence that tends to zero sufficiently slowly. By Corollary 4.12 there exist partitions $S_1,\dots,S_k$ of Ω^n and $X_1,\dots,X_\ell$ of $[n]$ such that $\Delta_{\boxtimes}(\mu,\mu^{S,X}) + \Delta_{\boxtimes}(\nu,\nu^{S,X}) < \varepsilon$ and $k+\ell \le \exp(\varepsilon^{-c})$. By the triangle inequality,

$$\begin{aligned} D_{\boxtimes}(\dot\mu^{S,X},\dot\nu^{S,X}) &\le D_{\boxtimes}(\dot\mu,\dot\nu) + D_{\boxtimes}(\dot\mu,\dot\mu^{S,X}) + D_{\boxtimes}(\dot\nu,\dot\nu^{S,X}) \\ &\le D_{\boxtimes}(\dot\mu,\dot\nu) + \Delta_{\boxtimes}(\mu,\mu^{S,X}) + \Delta_{\boxtimes}(\nu,\nu^{S,X}) \le D_{\boxtimes}(\dot\mu,\dot\nu) + 2\varepsilon. \end{aligned}$$

Hence, there exist $\phi \in \mathbb{S}$ and a coupling g of $\mu^{S,X},\nu^{S,X}$ such that the induced coupling $\dot g$ of $\dot\mu^{S,X},\dot\nu^{S,X}$ satisfies

$$\sup_{T\subset\mathscr{S}\times\mathscr{S},\, Y\subset[0,1],\omega\in\Omega} \left| \int_T \int_Y \left(\sigma_y(\omega) - \tau_{\phi(y)}(\omega)\right) \mathrm{d}y\,\mathrm{d}g(\sigma,\tau)\right| < D_{\boxtimes}(\mu,\nu) + 3\varepsilon. \tag{4.37}$$

Because ϕ preserves the Lebesgue measure, there exists a bijection $\varphi:[n]\to[n]$ such that the following is true. For a class $X_i \subset [n]$ let $\dot X_i = \bigcup_{x\in X_i}[(x-1)/n, x/n)$. Then uniformly for all $h,i \in [\ell]$ we have

$$\left|X_h \cap \varphi(X_i)\right| = n\lambda(\dot X_h \cap \phi(\dot X_i)) + O(1). \tag{4.38}$$

Further, we construct a coupling $G \in \Gamma(\mu,\nu)$ by letting

$$G(\sigma,\tau) = \sum_{\substack{\sigma'\in\Omega^n:\mu^{S,X}(\sigma')>0\\ \tau'\in\Omega^n:\nu^{S,X}(\tau')>0}} \frac{\gamma(\sigma,\sigma')g(\sigma',\tau')\gamma'(\tau,\tau')}{\mu^{S,X}(\sigma')\nu^{S,X}(\tau')}$$

and we claim that

$$\frac{1}{n}\max_{\substack{T\subset\Omega^n\times\Omega^n\\ Y\subset[n]\\ \omega\in\Omega}}\left|\sum_{(\sigma,\tau)\in T} G(\sigma,\tau)(\sigma_Y(\omega)-\tau_{\varphi(Y)}(\omega))\right| < D_{\boxtimes}(\mu,\nu)+6\varepsilon, \text{ where } \sigma_Y(\omega)=\sum_{y\in Y}\mathbf{1}\{\sigma_y=\omega\}. \tag{4.39}$$

Clearly, (4.39) readily implies the assertion.

To verify (4.39) we observe that, due to symmetry and the triangle inequality, it suffices to show that

$$\left|\sum_{(\sigma,\tau)\in T}\sum_{\sigma',\tau'}\frac{\gamma(\sigma,\sigma')g(\sigma',\tau')\gamma'(\tau,\tau')}{\mu^{S,X}(\sigma')\nu^{S,X}(\tau')}\left(\sigma_Y(\omega)-\sigma'_Y(\omega)\right)\right| < \varepsilon n, \tag{4.40}$$

$$\left|\sum_{(\sigma,\tau)\in T}\sum_{\sigma',\tau'}\frac{\gamma(\sigma,\sigma')g(\sigma',\tau')\gamma'(\tau,\tau')}{\mu^{S,X}(\sigma')\nu^{S,X}(\tau')}\left(\sigma'_Y(\omega)-\tau'_{\varphi(Y)}(\omega)\right)\right| < D_{\boxtimes}(\mu,\nu)n+4\varepsilon n. \tag{4.41}$$

for all T, Y, ω. Now, invoking Lemma 4.12, we obtain

$$\sum_{(\sigma,\tau)\in T}\sum_{\sigma',\tau'}\frac{\gamma(\sigma,\sigma')g(\sigma',\tau')\gamma'(\tau,\tau')}{\mu^{S,X}(\sigma')\nu^{S,X}(\tau')}\left(\sigma_Y(\omega)-\sigma'_Y(\omega)\right)_+ \le \sum_{\sigma,\sigma'}\gamma(\sigma,\sigma')\left(\sigma_Y(\omega)-\sigma'_Y(\omega)\right)_+ < \varepsilon n.$$

As the same bound holds for the negative part $\left(\sigma_Y(\omega)-\sigma'_Y(\omega)\right)_-$, we obtain (4.40). Similarly, due to Corollary 4.12, (4.37) and (4.38),

$$\begin{aligned}\sum_{(\sigma,\tau)\in T}\sum_{\sigma',\tau'}\frac{\gamma(\sigma,\sigma')g(\sigma',\tau')\gamma'(\tau,\tau')}{\mu^{S,X}(\sigma')\nu^{S,X}(\tau')}\left(\sigma'_Y(\omega)-\tau'_{\varphi(Y)}(\omega)\right)_+ &\le \sum_{\sigma',\tau'} g(\sigma',\tau')\left(\sigma'_Y(\omega)-\tau'_{\varphi(Y)}(\omega)\right)_+\\ &< nD_{\boxtimes}(\mu,\nu)+3\varepsilon n+O(k\ell) \le nD_{\boxtimes}(\mu,\nu)+4\varepsilon n,\end{aligned}$$

whence (4.41) follows. □

Proof of Theorem 1.2. The theorem follows by combining Lemmas 4.11 and 4.13. □

Acknowledgment. We thank Viresh Patel for bringing [49] to our attention, an anonymous reviewer for their careful reading, which has led to numerous corrections, and a second anonymous reviewer for pointing out several further references.

References

[1] D. Aldous: Representations for partially exchangeable arrays of random variables. J. Multivariate Anal. **11** (1981) 581–598.
[2] N. Alon, W. Fernandez de la Vega, R. Kannan, M. Karpinski: Random sampling and approximation of MAX-CSPs. J. Comput. System Sci. **67** (2003) 212–243
[3] T. Austin: On exchangeable random variables and the statistics of large graphs and hypergraphs. Probab. Surveys **5** (2008) 80–145.
[4] T. Austin: Exchangeable random measures. Annales de l'institut Henri Poincaré, Probabilités et Statistiques **51** (2015) 842–861.
[5] V. Bapst, A. Coja-Oghlan: Harnessing the Bethe free energy. Random Structures and Algorithms **49** (2016) 694–741.
[6] C. Borgs, J. Chayes, L. Lovász, V. Sós, K. Vesztergombi: Convergent sequences of dense graphs I: subgraph frequencies, metric properties and testing. Adv. Math. **219** (2008), 1801–1851.
[7] C. Borgs, J. Chayes, L. Lovász, V. Sós, K. Vesztergombi: Convergent sequences of dense graphs II: multiway cuts and statistical physics. Ann. Math. **176** (2012) 151–219.
[8] C. Borgs, J. Chayes, H. Cohn, N. Holden: Sparse exchangeable graphs and their limits via graphon processes. Journal of Machine Learning Research. **18** (2018) 1 – 71.
[9] C. Borgs, J. Chayes, H. Cohn, Y. Zhao: An Lp theory of sparse graph convergence II: LD convergence, quotients and right convergence. Ann. Probab. **46** (2018) 337–396.
[10] T. Bühler: Functional Analysis. American Mathematical Society (2018).
[11] A. Coja-Oghlan, F. Krzakala, W. Perkins, L. Zdeborová: Information-theoretic thresholds from the cavity method. Advances in Mathematics **333** (2018) 694–795.
[12] A. Coja-Oghlan, W. Perkins: Spin systems on Bethe lattices. Communications in Mathematical Physics **372** (2018) 441–523.
[13] A. Coja-Oghlan, W. Perkins: Bethe states of random factor graphs. Communications in Mathematical Physics **366** (2019) 173–201.
[14] A. Coja-Oghlan, W. Perkins, K. Skubch: Limits of discrete distributions and Gibbs measures on random graphs. European Journal of Combinatorics **66** (2017) 37–59.

[15] D. Cai, T. Campbell, T. Broderick: Edge-exchangeable graphs and sparsity. Advances in Neural Information Processing Systems **29** (2016) 4249–4257.
[16] L. Coregliano, A. Razborov: Semantic Limits of Dense Combinatorial Objects. Uspekhi Matematicheskikh Nauk **75** (2020) 45–152.
[17] H. Crane, W. Dempsey: Edge Exchangeable Models for Interaction Networks, Journal of the American Statistical Association, **113:523** 1311-1326 (2018).
[18] D. Conlon, J. Fox: Bounds for graph regularity and removal lemmas. Geometric and Functional Analysis **22** (2012) 1191–1256.
[19] P. Diaconis, S. Janson: Graph limits and exchangeable random graphs. Rend. Mat. Appl. **28** (2008) 33–61.
[20] R. Eldan: Taming correlations through entropy-efficient measure decompositions with applications to mean-field approximation. arXiv:1811.11530 (2018).
[21] A. Frieze, R. Kannan: Quick approximation to matrices and applications. Combinatoria **19** (1999) 175–220.
[22] A. Galanis, D. Stefankovic, E. Vigoda: Inapproximability for antiferromagnetic spin systems in the tree nonuniqueness region. J. ACM **62** (2015) 50
[23] H.-O. Georgii: Gibbs measures and phase transitions. 2nd edition. De Gruyter (2011).
[24] D. G. Hartig: The Riesz representation theorem revisited. American Mathematical Monthly **90** (1983) 277–280.
[25] D. Hoover: Relations on probability spaces and arrays of random variables. Preprint, Institute of Advanced Studies, Princeton, 1979.
[26] C. Hoppen, Y. Kohayakawa, C. Moreira, B. Rath, R. Sampaio: Limits of permutation sequences. Journal of Combinatorial Theory Series B. **103** (2011) 10.1016/j.jctb.2012.09.003.
[27] S. Janson: Poset limits and and exchangeable random posets. Combinatorica **31** 529–563 (2011).
[28] S. Janson: Graphons, cut norm and distance, couplings and rearrangements. NYJM Monographs **4** (2013).
[29] O. Kallenberg: Probabilistic symmetries and invariance principles. Springer, New York, 2005.
[30] A. S. Kechris: Classical descriptive set theory. Springer (1995).
[31] F. Krzakala, A. Montanari, F. Ricci-Tersenghi, G. Semerjian, L. Zdeborová: Gibbs states and the set of solutions of random constraint satisfaction problems. Proc. National Academy of Sciences **104** (2007) 10318–10323.
[32] L. Lovász: Large Networks and Graph Limits. American Mathematical Society 2012.
[33] L. Lovász, B. Szegedy: Limits of compact decorated graphs. ArXiV 1010.5155 (2010).
[34] L. Lovász, B. Szegedy: Limits of dense graph sequences. J. Combin. Theory Ser. B **96** (2006) 933–957.
[35] L. Lovász, B. Szegedy: Szemerédi's lemma for the analyst. Geom. Funct. Anal. **17** (2007) 252–270.
[36] L. Lovász, B. Szegedy: Regularity partitions and the topology of graphons. In: I. Bárány, J. Solymosi, G. Sági: An Irregular Mind. Bolyai Society Mathematical Studies **21** (2010).
[37] G. W. Mackey: Borel structure in groups and their duals. Trans. Amer. Math. Soc. **85** (1957) 134–165.
[38] E. Marinari, G. Parisi, F. Ricci-Tersenghi, J. Ruiz-Lorenzo, F. Zuliani: Replica symmetry breaking in short-range spin glasses: theoretical foundations and numerical evidences. J. Stat. Phys. **98** (2000) 973
[39] M. Mézard, A. Montanari: Information, physics and computation. Oxford University Press 2009.
[40] A. Montanari: Estimating random variables from random sparse observations. European Transactions on Telecommunications **19** (2008) 385–403.
[41] J. Nešetřil, P. Ossona de Mendez: Existence of modeling limits for sequences of sparse structures. The Journal of Symbolic Logic **84** (2019) 452–472.
[42] S. Nicolay, L. Simons: Building Cantor's Bijection. arXiv 1409.1755 (2014).
[43] D. Panchenko: The Sherrington-Kirkpatrick Model. Springer Monographs in Mathematics (2013).
[44] D. Panchenko: Spin glass models from the point of view of spin distributions. Annals of Probability **41** (2013) 1315–1361.
[45] P. Raghavendra, N. Tan: Approximating CSPs with global cardinality constraints using SDP hierarchies. Proc. 23rd SODA (2012) 373–387.
[46] A. Sly: Computational transition at the uniqueness threshold. Proc. 51st FOCS (2010) 287–296.
[47] A. Sly, N. Sun: The computational hardness of counting in two-spin models on d-regular graphs. Proc. 53rd FOCS (2012) 361–369.
[48] E. Szemerédi: On sets of integers containing no k elements in arithmetic progression. Acta Arithmetica **27** (1975) 199–245.
[49] T. Tao: Szemeredi's regularity lemma via the correspondence principle. Blog entry. https://terrytao.wordpress.com/2009/05/08/szemeredis-regularity-lemma-via-the-correspondence-principle/
[50] V. Veitch, D. Roy: The Class of Random Graphs Arising from Exchangeable Random Measures. arXiv 1512.03099 (2015).
[51] C. Villani: Optimal Transport. Springer (2009).

AMIN COJA-OGHLAN, acoghlan@math.uni-frankfurt.de, GOETHE UNIVERSITY, MATHEMATICS INSTITUTE, 10 ROBERT MAYER ST, FRANKFURT 60325, GERMANY.

MAX HAHN-KLIMROTH, hahnklim@math.uni-frankfurt.de, GOETHE UNIVERSITY, MATHEMATICS INSTITUTE, 10 ROBERT MAYER ST, FRANKFURT 60325, GERMANY.

G. Random perturbation of sparse graphs

RANDOM PERTURBATION OF SPARSE GRAPHS

MAX HAHN-KLIMROTH, GIULIA S. MAESAKA, YANNICK MOGGE, SAMUEL MOHR, AND OLAF PARCZYK

ABSTRACT. In the model of randomly perturbed graphs we consider the union of a deterministic graph $\mathcal{G}_\alpha$ with minimum degree αn and the binomial random graph $\mathbb{G}(n,p)$. This model was introduced by Bohman, Frieze, and Martin and for Hamilton cycles their result bridges the gap between Dirac's theorem and the results by Posá and Koršunov on the threshold in $\mathbb{G}(n,p)$. In this note we extend this result in $\mathcal{G}_\alpha \cup \mathbb{G}(n,p)$ to sparser graphs with $\alpha = o(1)$. More precisely, for any $\varepsilon > 0$ and $\alpha\colon \mathbb{N} \mapsto (0,1)$ we show that a.a.s. $\mathcal{G}_\alpha \cup \mathbb{G}(n,\beta/n)$ is Hamiltonian, where $\beta = -(6+\varepsilon)\log(\alpha)$. If $\alpha > 0$ is a fixed constant this gives the aforementioned result by Bohman, Frieze, and Martin and if $\alpha = O(1/n)$ the random part $\mathbb{G}(n,p)$ is sufficient for a Hamilton cycle. We also discuss embeddings of bounded degree trees and other spanning structures in this model, which lead to interesting questions on almost spanning embeddings into $\mathbb{G}(n,p)$.

1. INTRODUCTION AND RESULTS

For $\alpha \in (0,1)$ we let $\mathcal{G}_\alpha$ be an n-vertex graph with minimum degree $\delta(\mathcal{G}_\alpha) \geq \alpha n$. A famous result by Dirac [15] says that if $\alpha \geq 1/2$ and $n \geq 3$, then $\mathcal{G}_\alpha$ contains a Hamilton cycle, i.e. a spanning cycle through all vertices of $\mathcal{G}_\alpha$. This motivated the more general questions of determining the smallest α such that $\mathcal{G}_\alpha$ contains a given spanning structure. For example, there are results for trees [29], factors [22], powers of Hamilton cycles [26, 28], and general bounded degree graphs [12]. This is a problem for deterministic graphs that belongs to the area of extremal graph theory.

We can consider similar questions for random graphs, in particular, for the binomial random graph model $\mathbb{G}(n,p)$, which is the probability space over n-vertex graphs with each edge being present with probability p independent of all the others. Analogous to the smallest α we are looking for a function $\hat{p} = \hat{p}(n)\colon \mathbb{N} \mapsto (0,1)$ such that if $p = \omega(\hat{p})$ the probability that $\mathbb{G}(n,p)$ contains some spanning subgraph tends to 1 as n tends to infinity and for $p = o(\hat{p})$ it tends to 0. We call this $\hat{p}$ the threshold function for the respective property (an easy sufficient criteria for its existence can be found in [8]) and if the first/second statement holds we say that $\mathbb{G}(n,p)$ has/does not have this property asymptotically almost surely (a.a.s.). One often says that $\mathbb{G}(n,p)$ undergoes a *phase transition* at $\hat{p}$. For the Hamilton cycle problem Posá [39] and Koršunov [31] proved independently that $\hat{p} = \log n/n$ gives the threshold. Similar as above there was a tremendous amount of research on determining the thresholds for various spanning structures, e.g. for matchings [17], trees [32, 36], factors [24], powers of Hamilton cycles [35, 37], and general bounded degree graphs [1, 18, 19, 40]. An extensive survey by Böttcher can be found in [9].

Motivated by the smoothed analysis of algorithms [41], both these worlds were combined by Bohman, Frieze, and Martin [7]. For any fixed $\alpha > 0$, they defined the model of randomly perturbed graphs as the union $\mathcal{G}_\alpha \cup \mathbb{G}(n,p)$. They showed that $1/n$ is the threshold for a Hamilton cycle, meaning that there is a graph $\mathcal{G}_\alpha$ such that with $p = o(1/n)$ there a.a.s. is no Hamilton cycle in $\mathcal{G}_\alpha \cup \mathbb{G}(n,p)$ and for any $\mathcal{G}_\alpha$ and $p = \omega(1/n)$ there a.a.s. is a Hamilton cycle in $\mathcal{G}_\alpha \cup \mathbb{G}(n,p)$. It is important to note that in $\mathbb{G}(n,p)$, $p = 1/n$ is also the threshold for an almost spanning cycle, this is for any $\varepsilon > 0$ a cycle on at least $(1-\varepsilon)n$ vertices. It should be further remarked that if $p = o(\log n/n)$ there are a.a.s. isolated vertices in $\mathbb{G}(n,p)$ and the purpose of $\mathcal{G}_\alpha$ is to compensate for this and to help in turning the almost spanning cycle into a Hamilton cycle.

This first result on randomly perturbed graphs [7] sparked a lot of subsequent research on the thresholds of spanning structures in this randomly perturbed graphs model, e.g. trees [10, 25, 34], factors [4], powers

The research on this project was initiated during a workshop in Cuxhaven. We would like to thank the Hamburg University of Technology for their support. OP was supported by Technische Universität Ilmenau, the Carls Zeiss Foundation, and DFG Grant PA 3513/1-1. MHK was supported by Stiftung Polytechnische Gesellschaft. SM was supported by DFG Grant 327533333. GSM is supported by the European Research Council (Consolidator Grant PEPCo 724903).

of Hamilton cycles [5, 11], and general bounded degree graphs [11]. As for a Hamilton cycle there is often a log-factor difference to the thresholds in $\mathbb{G}(n,p)$ alone, which is there for local reasons similar to isolated vertices. In most of these cases a $\mathcal{G}_\alpha$, that is responsible for the lower bound, is the complete imbalanced bipartite graph $K_{\alpha n,(1-\alpha)n}$. In this model there are also results with lower bounds on α [6, 16, 23, 38] and for Ramsey-type problems [13, 14].

1.1. **Hamiltonicity in randomly perturbed sparse graphs.** The aim of this note is to investigate a new direction. Instead of fixing an $\alpha \in (0,1)$ in advance we allow α to tend to zero with n. This extends the range of $\mathcal{G}_\alpha$ to sparse graphs and we want to determine the threshold probability in $\mathcal{G}_\alpha \cup \mathbb{G}(n,p)$. For example, with $\alpha = 1/\log n$ we have a sparse deterministic graph $\mathcal{G}_\alpha$ with minimum degree $n/\log n$. Then $p = \omega(1/n)$ does not suffice in general, but it is sufficient to take $\mathcal{G}_\alpha \cup \mathbb{G}(n,\Theta(\log\log n)/n)$ to a.a.s. guarantee a Hamilton cycle. More generally, we can prove the following.

Theorem 1.1. *Let $\alpha = \alpha(n) : \mathbb{N} \mapsto (0,1)$ and $\beta = \beta(\alpha) = -(6+o(1))\log(\alpha)$. Then a.a.s. $\mathcal{G}_\alpha \cup \mathbb{G}(n,\beta/n)$ is Hamiltonian.*

This extends the result of Bohman, Frieze, and Martin [7] for constant $\alpha > 0$. For even n a direct consequence of this theorem is the existence of a perfect matching in the same graph. To prove Theorem 1.1 we use a result by Frieze [20] to find a very long path in $\mathbb{G}(n,p)$ alone and then use the switching technique developed in [11] to turn this into a Hamilton cycle. As it turns out, our method allows to prove the existence of a perfect matching with a slightly lower edge probability.

Theorem 1.2. *Let $\alpha = \alpha(n) : \mathbb{N} \mapsto (0,1)$ and $\beta = \beta(\alpha) = -(4+o(1))\log(\alpha)$. Then a.a.s. $\mathcal{G}_\alpha \cup \mathbb{G}(n,\beta/n)$ contains a perfect matching.*

To see that in both theorems β is optimal up to the constant factor, consider $\mathcal{G}_\alpha = K_{\alpha n,(1-\alpha)n}$ and note that there cannot be a perfect matching, if we have more than αn isolated vertices on the $(1-\alpha)n$ side. The number of isolated vertices in $\mathbb{G}(n,\beta/n)$ roughly is $n(1-\beta/n)^{n-1} \cong n\exp(-\beta)$, which is larger than αn if $\beta = o(-\log(\alpha))$.

For proving results in the randomly perturbed graphs model good almost spanning results are essential. Typically, by almost spanning one means that for any $\varepsilon > 0$ we can embed the respective structure on at least $(1-\varepsilon)n$ vertices. For paths and cycles in $\mathbb{G}(n,C/n)$ this can, for example, be done using expansion properties and the DFS-algorithm [33]. These almost spanning results are much easier than the spanning counterpart, because there is always a linear size set of available vertices. But for the proof of Theorem 1.1 this is not sufficient, because if $\alpha = o(1)$ we will not be able to take care of a linear sized leftover. Instead we exploit that we have $\mathbb{G}(n,\beta/n)$ and use the following result showing that we can find a long cycle consisting of all but sublinearly many vertices.

Lemma 1.3 (Frieze [20])**.** *Let $0 < \beta = \beta(n) \le \log n$. Then $\mathbb{G}(n,\beta/n)$ a.a.s. contains a cycle of length at least*

$$\big(1-(1-o(1))\,\beta\exp\big(-\beta\big)\big)\,n.$$

This is optimal, because this is asymptotically the size of the 2-core (maximal subgraph with minimum degree 2) of $\mathbb{G}(n,p)$ [21, Lemma 2.16]. A similar result holds for large matchings.

Lemma 1.4 (Frieze [20])**.** *Let $0 < \beta = \beta(n) \le \log n$. Then $\mathbb{G}(n,\beta/n)$ a.a.s. contains a matching consisting of at least $\big(1-(1-o(1))\exp\big(-\beta\big)\big)\,n$ vertices.*

Again this is optimal, because the number of isolated vertices is a.a.s. $(1+o(1))e^{-\beta}n$ [21, Theorem 3.1]. Observe, that also a bipartite variant of this lemma holds, which can be proved by removing small degree vertices and employing Halls theorem.

Lemma 1.5. *Let $0 < \beta = \beta(n) \le \log n$. Then the bipartite binomial random graph $\mathbb{G}(n,n,\beta/n)$ a.a.s. contains a matching consisting of at least $\big(1-(1-o(1))\exp\big(-\beta\big)\big)\,n$ edges.*

1.2. **Bounded degree trees in randomly perturbed sparse graphs.** After Hamilton cycles and perfect matchings, the next natural candidates are n-vertex trees with maximum degree bounded by a constant Δ. In $\mathbb{G}(n,p)$ the threshold $\log n/n$ was determined in a breakthrough result by Montgomery [36], in $\mathcal{G}_\alpha$ it is enough to have a fixed $\alpha > 1/2$ [27], and in $\mathcal{G}_\alpha \cup \mathbb{G}(n,p)$ with constant $\alpha > 0$ the threshold is $1/n$ [34]. To obtain a result similar to Theorem 1.1 for bounded degree trees using our approach we need an almost spanning result similar to Lemma 1.3. With a similar approach as for Theorem 1.1 and 1.2 we obtain the following modular statement.

Theorem 1.6. *Let $\Delta \geq 2$ be an integer and suppose that $\alpha, \beta, \varepsilon \colon \mathbb{N} \mapsto [0,1]$ are such that $4(\Delta+1)\varepsilon < \alpha^{\Delta+1}$ and a.a.s. $\mathbb{G}(n, \beta/n)$ contains a given tree with maximum degree Δ on $(1-\varepsilon)n$ vertices. Then any tree with maximum degree Δ on n vertices is a.a.s. contained in the union $\mathcal{G}_\alpha \cup \mathbb{G}(n,\beta/n)$.*

Next we discuss the almost spanning results that we can obtain in the relevant regime. Improving on a result of Alon, Krivelevich, and Sudakov [2], Balogh, Csaba, Pei, and Samotij [3] proved that for $\Delta \geq 2$ there exists a $C > 0$ such that for $\varepsilon > 0$ a.a.s. $\mathbb{G}(n, \beta/n)$ contains any tree with maximum degree Δ on at most $(1-\varepsilon)n$ vertices provided that $\beta \geq \frac{C}{\varepsilon}\log\frac{1}{\varepsilon}$. For the proof they only require that the graph satisfies certain expander properties. This can be extended to the range where $\varepsilon \to 0$ and $\omega(1) = \beta \leq \log n$ and following along the lines of their argument we get the following.

Lemma 1.7. *For $\Delta \geq 2$ there exists a $C > 0$ such that for any $0 < \beta = \beta(n) \leq \log n$ and $\varepsilon = \varepsilon(n) > 0$ with $\beta \geq \frac{C}{\varepsilon}\log\frac{1}{\varepsilon}$ the following holds. $\mathbb{G}(n,\beta/n)$ a.a.s. contains any bounded degree tree on at most $(1-\varepsilon)\,n$ vertices.*

Then together with Theorem 1.6 we obtain the following.

Corollary 1.8. *For $\Delta \geq 2$ there exists a $C > 0$ such that for $\alpha = \alpha(n) : \mathbb{N} \mapsto (0,1)$ and $\beta = \beta(\alpha) = C\alpha^{-(\Delta+1)}\log\frac{1}{\alpha}$ the following holds. Any n-vertex tree T with maximum degree Δ is a.a.s. contained in $\mathcal{G}_\alpha \cup \mathbb{G}(n,\beta/n)$.*

The proof for the dense case in [34] uses regularity and it is unlikely to give anything better in the sparse regime. As remarked in [2] the condition on the almost spanning embedding in $\mathbb{G}(n,\beta/n)$ could possibly be improved to $\beta > \log\frac{C}{\varepsilon}$, then covering almost all non-isolated vertices. More precisely this asks for the following.

Question 1.9. *For every integer Δ there exists $C > 0$ such that with $0 < \beta = \beta(n) \leq \log n$ the following holds. Is any given tree with maximum degree Δ on*

$$(1 - C\exp(-\beta))\,n$$

vertices a.a.s. contained in $\mathbb{G}(n,\beta/n)$?

With Theorem 1.6 this would then give that already $\beta = -(\Delta+1)\log(C\alpha)$ suffices, which would be optimal up to the constant factors. We want to briefly argue why it is possible to answer this question for large families of trees and what the difficulties are. For simplicity we only discuss the case $\beta = \log\log n$ and note that by Lemma 1.7 above we can embed trees on roughly $(1 - 1/\log\log n)n$ vertices. A very helpful result for handling trees by Krivelevich [32] states that for any integer $n, k > 2$, a tree on n vertices either has at least $n/4k$ leaves or a collection of at least $n/4k$ bare paths (internal vertices of the path have degree 2 in the tree) of length k. If there are at least $n/(4\log\log n)$ leaves, we can embed the tree obtained after removing the leaves. Then we can use a fresh random graph and Lemma 1.5 to find a matching for all the leaves, completing the embedding of the tree.

On the other hand, if there are at least $n\log\log n/(4\log n)$ bare paths of length $\log n/\log\log n$, it is possible to embed all but $n/\log n$ of these paths, which are all but $n/\log\log n$ vertices. Then one has to connect the remaining paths, again using ideas from [36]. In between both cases it is not clear what should be done, because we might have $n/\log n$ leaves and $n/(4\log\log n)$ bare paths of length $\log\log n$. The length of the paths are too short to connect them and the leaves are too few for the above argument. Answering this questions and thereby improving the result of Alon, Krivelevich, and Sudakov [2] is a challenging open problem.

1.3. **Other spanning structures.** As mentioned above, embeddings of spanning structures in $\mathcal{G}_\alpha$, $\mathbb{G}(n,p)$, and $\mathcal{G}_\alpha \cup \mathbb{G}(n,p)$ for fixed $\alpha > 0$ have also been studied for other graphs such as powers of Hamilton cycles, factors, and general bounded degree graphs. In most of these cases almost spanning embeddings (e.g. Ferber, Luh, and Nguyen [18]) can be generalised such that previous proofs can be extended to the regime $\alpha = o(1)$ with $\beta = \alpha^{-1/C}$, similar to what we do in Corollary 1.8. Further improvements seem to be hard, because better almost spanning results are similar in difficulty to spanning results in $\mathbb{G}(n,p)$ alone. We want to discuss this on one basic example, the triangle factor, which is the disjoint union of $n/3$ triangles.

In $\mathcal{G}_\alpha$ we need $\alpha \geq 2/3$, in $\mathbb{G}(n,p)$ the threshold is $n^{-2/3}\log^{1/3} n$, and in $\mathcal{G}_\alpha \cup \mathbb{G}(n,p)$ with a fixed $\alpha > 0$ it is $n^{-2/3}$. Note that the log-term in $\mathbb{G}(n,p)$ is needed to ensure that every vertex is contained in a triangle, which is essential for a triangle factor. Using Janson's inequality [21, Theorem 21.12] it is not hard to prove the almost spanning result for a triangle factor on at least $(1-\varepsilon)n$ vertices with $p = \omega(n^{-2/3})$. This can be generalised to $\mathbb{G}(n,\beta n^{-2/3})$ giving a.a.s. a triangle factor on at least $(1-C/\beta)n$ vertices. Again, this can only give something with $\beta = \alpha^{-1/C}$ in $\mathcal{G}_\alpha \cup \mathbb{G}(n,\beta n^{-2/3})$ and to improve this we ask the following.

Question 1.10. *Let $0 < \beta = \beta(n) \leq \log^{1/3} n$. Does $\mathbb{G}(n,\beta n^{-2/3})$ a.a.s. contain a triangle factor on at least*

$$\big(1-(1-o(1))\exp(-\beta^3)\big)\,n$$

vertices?

Observe, that this is a.a.s. the number of vertices of $\mathbb{G}(n,\beta n^{-2/3})$ that are not contained in a triangle. Similar questions for other factors or more general structures would be of interest. It took a long time until Johannson, Kahn, and Vu [24] determined the threshold for the triangle factor. This conjecture seems to be of similar difficulty, whereas for our purposes it would already be great to obtain a triangle factor on at least $(1-C\exp(-\beta^3))n$ vertices for some $C > 1$.

For the remainder of this note we prove Theorem 1.1 and 1.6 in Section 2 and 3 respectively.

2. Hamiltonicity

We will prove the following proposition that will be sufficient to prove the theorem together with known results on Hamilton cycles in $\mathbb{G}(n,p)$.

Proposition 2.1. *Let $\alpha = \alpha(n) : \mathbb{N} \mapsto (0,1)$ such that $\alpha = \omega(n^{-1/6})$, and let $\beta = \beta(\alpha) = -(6+o(1))\log(\alpha)$. Then a.a.s. $\mathcal{G}_\alpha \cup \mathbb{G}(n,\beta/n)$ is Hamiltonian.*

Proof of Theorem 1.1. Let $\alpha, \beta > 0$ such that $\beta = -(6+o(1))\log(\alpha)$. If $\alpha = O(n^{-1/6})$, we have $\beta \geq (1+o(1))\log n$ and we can infer that a.a.s. there is a Hamilton cycle in $G(n,\beta/n)$ (this follows from an improvement on the result concerning the threshold for Hamiltonicity [30]). On the other hand, if $\alpha = \omega(n^{-1/6})$, then we apply Proposition 2.1 to a.a.s. get the Hamilton cycle. □

Proof of Proposition 2.1. To prove the proposition we apply the following strategy. We first find a long path in $\mathbb{G}(n,p)$ alone. Then, by considering the union with $\mathcal{G}_\alpha$, we obtain a reservoir structure for each vertex that allows us to extend the length of the path iteratively. Finally, we will also be able to close this path to a cycle on all vertices. W.l.o.g. we can assume that $\alpha < 1/10$.

Finding a long path. Let $P = p_1, \dots, p_\ell$ be the longest path that we can find in $\mathcal{G}_1 = \mathbb{G}(n,(\beta-1)/n)$ and let $V' = \{v_1, \dots, v_k\} = V(\mathcal{G}_1) \setminus \{p_1, ..., p_\ell\}$ be the left-over. Then, by Lemma 1.3, we get a.a.s. that

$$k = \big|V'\big| = n - \ell \leq (1-o(1))\,\beta\exp\big(1-\beta\big)\,n. \tag{2.1}$$

Next, let P' be a collection of vertices of P, where we take every other vertex, excluding the last, that is

$$P' = \big\{p_i : i \equiv 0 \pmod 2\big\} \setminus \{p_\ell\} \tag{2.2}$$

In the following, we will work on P' instead of all of P, ensuring that certain absorbing structures do not overlap.

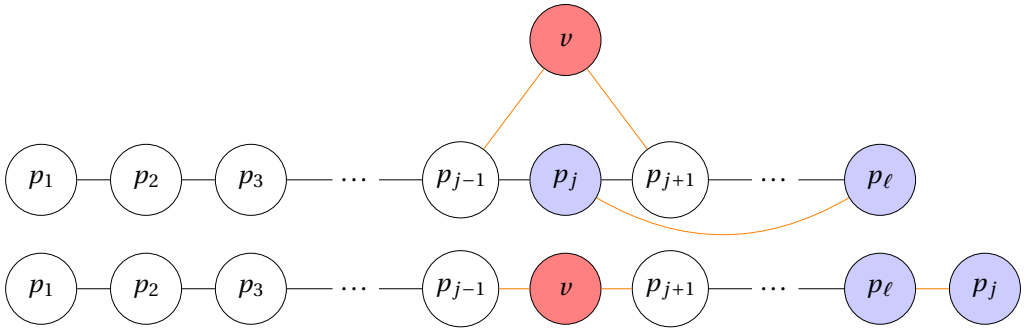

FIGURE 1. The top shows a path $P = p_1, \dots, p_\ell$ and the left-over vertex v. Black edges belong to the random graph, orange edges can be found in $\mathcal{G}_\alpha$. The bottom shows the situation after absorbing v using that $p_j \in \boldsymbol{B}(p_\ell, v)$.

Absorbing the left-over. We now consider the union $\mathcal{G}_\alpha \cup \mathcal{G}_1$. The following absorbing structure is the key to the argument.

Definition 2.2. *For any vertices $u, v \in V(\mathcal{G}_\alpha \cup \mathcal{G}_1)$ let*

$$\boldsymbol{B}(u,v) = \left\{x \in N_{\mathcal{G}_\alpha}(u) \cap P' \mid N_P(x) \subseteq N_{\mathcal{G}_\alpha}(v)\right\}. \tag{2.3}$$

If for some $v \in V'$ there is an $p_j \in \boldsymbol{B}(p_\ell, v)$ we can proceed as follows (see Figure 1). By definition we have $p_{j-1}, p_{j+1} \in N_{\mathcal{G}_\alpha(v)}$ and $p_j \in N_{\mathcal{G}_\alpha}(p_\ell) \cap P$. Then p_j can be replaced by v in the path P and can now be appended to the path P at p_ℓ. So we get the path $\tilde{P} = p_1, \dots, p_{j-1}, v, p_{j+1}, \dots, p_\ell, p_j$, where $\tilde{P} \subset P \cup \mathcal{G}_\alpha$.

To iterate this argument we show that a.a.s. for any pair of vertices u and v, the set $\boldsymbol{B}(u,v)$ is large enough.

Claim 2.3. *We have a.a.s. $|\boldsymbol{B}(u,v)| \geq \alpha^3 n/4$ for any $u, v \in V(\mathcal{G}_\alpha \cup \mathcal{G}_1)$.*

Proof. Let u, v be arbitrary vertices in $V = V(\mathcal{G}_\alpha \cup \mathcal{G}_1)$. The set $\boldsymbol{B}(u,v)$ is uniformly distributed over P', because $\mathbb{G}(n, (\beta-1)/n)$ is sampled independently of the deterministic graph $\mathcal{G}_\alpha$. Then by definition

$$\mathbb{E}\left[|\boldsymbol{B}(u,v)|\right] \geq \frac{9}{10}\alpha^3 \left|P'\right| \geq \frac{2}{5}\alpha^3 \left(1 - (1-o(1))\,\beta \exp\left(1-\beta\right)\right) n \geq \alpha^3 n/3. \tag{2.4}$$

An immediate consequence of $\boldsymbol{B}(u,v)$ being uniformly settled over $\mathbb{G}(n, (\beta-1)/n)$ is that $|\boldsymbol{B}(u,v)| \sim \mathrm{Bin}(\left|P'\right|, \alpha^3)$. It follows from (2.4) and the Chernoff bound that there is a sufficiently small, but constant, $\delta > 0$ s.t.

$$\mathbb{P}\left(|\boldsymbol{B}(u,v)| < \alpha^3 n/4\right) \leq \mathbb{P}\left(|\boldsymbol{B}(u,v)| < (1-\delta)\mathbb{E}\left[|\boldsymbol{B}(u,v)|\right]\right) \leq \exp\left(-\delta^2/8\alpha^3 n\right) < \exp\left(-\sqrt{n}\right). \tag{2.5}$$

The lemma follows from a union bound over all $\binom{n}{2}$ choices for u, v and (2.5). □

We now have everything at hand to absorb all but two of the left-over vertices $v \in V'$ onto a path of length $n-2$. We do this inductively using Algorithm 1.

Let $\tilde{P}, B_i(\cdot,\cdot)$ be defined as in Algorithm 1. In order to see that the algorithm terminates with $\tilde{P} = P_k$ it suffices to prove, that $B_i(u,v)$ is not empty for any $u, v \in V$ and $i = 1 \dots k$. By definition of P' in (2.2) we have $|\boldsymbol{B}(u,v) \setminus B_i(u,v)| \leq i$ and using Claim 2.3 and (2.1) we get

$$|B_i(u,v)| \geq \alpha^3 n/8, \tag{2.6}$$

whenever $\beta \exp\left(1-\beta\right) < \alpha^3/8$. As this holds by definition of $\beta = -(6+o(1))\log(\alpha)$ and with $\alpha < 1/10$, we get that (2.6) holds for all $u, v \in V$ and any $i = 1, \dots, k$.

Closing the cycle. We have found a path $\tilde{P} = p_1, \dots, p_{n-2}$ and we are left with two vertices v_{k-1}, v_k that are not on the path. It is possible to close the Hamilton cycle by absorbing v_{k-1} and v_k if there is an edge between $A := B_k(p_1, v_{k-1})$ and $B := B_k(p_{n-2}, v_k)$. Indeed, we then have w.l.o.g. $i < j$ such that $p_i \in A$, $p_j \in B$, and there is an edge $p_i p_j$. By definition of A and B we can then obtain the Hamilton cycle

$$p_i, p_1, \dots, p_{i-1}, v_{k-1}, p_{i+1}, \dots, p_{j-1}, v_k, p_{j+1}, \dots, p_{n-2}, p_j.$$

Algorithm 1: Absorbs all but two vertices of the left-over set V' onto a path.

Input : Path $P = p_1 \dots p_\ell$, set of left-over vertices $V' = \{v_1, \dots, v_k\}$.
Output: Path $\tilde{P}$ in $P \cup \mathcal{G}_\alpha$ on $n-2$ vertices.

Define $\ell_1 = \ell$, $P_1 = P$ with $P_1 = u_1^1 \dots u_{\ell_1}^1$;
Define for any u, v the set $B_1(u,v) = \boldsymbol{B}(u,v)$;
Define $V_1' = V'$;
for $i = 1$ **to** $k-2$ **do**
  Choose $u_j^i \in B_i(u_{\ell_i}^i, v_i)$ and absorb v_i onto P_i;
  Denote by $P_{i+1} = u_1^i \dots u_{j-1}^i v_i u_{j+1}^i \cdots u_{\ell_i}^i u_j^i = u_1^{i+1} \dots u_{\ell_i+1}^{i+1}$ the resulting path;
  Update $\ell_{i+1} = \ell_i + 1$, $V_{i+1}' = V_i' \setminus \{v_i\}$;
  Set $B_{i+1}(u,v) = B_i(u,v) \setminus \left\{u_j^i\right\}$ for any u, v;
end
$\tilde{P} = P_k$;

It remains to prove that we have an edge between A and B. For this we reveal $\mathcal{G}_2 = \mathbb{G}(n, 1/n)$. As $|A|, |B| \geq \alpha^3 n/8$ by (2.6) we get

$$\mathbb{E}\left[e_{\mathcal{G}_2}(A,B)\right] \geq \frac{1}{n} \cdot \left(\frac{\alpha^3 n}{16}\right)^2 = \omega(1), \tag{2.7}$$

as $\alpha = \omega(n^{-1/6})$. Together with Chernoff's inequality this implies that a.a.s $e_{\mathcal{G}_2}(A,B) > 0$. As the union of $\mathcal{G}_1$ and $\mathcal{G}_2$ can be coupled as a subgraph of $\mathbb{G}(n, \beta/n)$ this implies that a.a.s. there is a Hamilton cycle in $\mathcal{G}_\alpha \cup \mathbb{G}(n,p)$ and finishes the proof of Proposition 2.1. □

Observe, that when running the same proof for Theorem 1.2 we can obtain the better constant by adapting the definition of the $\boldsymbol{B}(u,v)$ to the setup of perfect matchings and then proving that a.a.s. $|\boldsymbol{B}(u,v)| \geq \alpha^2 n/4$. We spare the details here.

3. Bounded degree trees

Theorem 1.6 is modular, which turns almost spanning embeddings in the random graph into spanning embeddings in the union $\mathcal{G}_\alpha \cup \mathbb{G}(n, \beta/n)$. The proof is very similar to the proof for Hamilton cycles and we will spare some details.

Proof of Theorem 1.6. Let $\mathcal{G}_\alpha$ be given and $\mathcal{G} = \mathbb{G}(n, \beta/n)$. Let $\mathcal{T}$ be an arbitrary tree on n vertices with maximum degree Δ. Denote by $\mathcal{T}_\varepsilon$ the tree obtained from $\mathcal{T}$ by the following construction.

(1) Set $\mathcal{T}_0 = \mathcal{T}$.
(2) In every step i, check whether $\mathcal{T}_i$ has at most $(1-\varepsilon)n$ vertices.
  - If this is the case, set $\mathcal{T}_\varepsilon = \mathcal{T}_i$ and finish the process.
  - Otherwise, create $\mathcal{T}_{i+1}$ by deleting one leaf of $\mathcal{T}_i$.

We denote by L the left-over, that are the vertices removed during construction of $\mathcal{T}_\varepsilon$. Then

$$|V(\mathcal{T}_\varepsilon)| \leq (1-\varepsilon)n, \qquad |L| \leq \varepsilon n + 1, \qquad \text{and} \qquad V(\mathcal{T}) = V(\mathcal{T}_\varepsilon) \cup L.$$

Next we let T be an independent subset of the vertices of $\mathcal{T}_\varepsilon$ such that the vertices in T do not have neighbours outside of $\mathcal{T}_\varepsilon$ with respect to $\mathcal{T}$. Observe, that there exists such a T such that $|T| \geq \frac{(1-\Delta\varepsilon)n}{\Delta+1}$.

By assumption we a.a.s. have an embedding $\mathcal{T}_\varepsilon'$ of $\mathcal{T}_\varepsilon$ into $\mathcal{G}$ and we denote by T' the image of T under this embedding. We adapt Definition 2.2 and define for any two vertices u, v

$$\boldsymbol{B}(u,v) = \left\{x \in N_{\mathcal{G}_\alpha}(u) \cap T' \mid N_{\mathcal{T}_\varepsilon'}(x) \subset N_{\mathcal{G}_\alpha}(v)\right\}.$$

As before, if we want to embed a vertex w that is a neighbour of an already embedded vertex u in $\mathcal{T}_\varepsilon$ and v is an available vertex we can do it if $\boldsymbol{B}(u,v)$ is non-empty. More precisely, with $x \in \boldsymbol{B}(u,v)$, we can embed

the vertex embedded onto x to v, embed w to x, and obtain a valid embedding of $\mathscr{T}_\varepsilon$ with an additional neighbour of u. Analogous to Claim 2.3 we get the following.

Claim 3.1. *We have a.a.s.* $|\boldsymbol{B}(u,v)| \geq \frac{\alpha^{\Delta+1}n}{4(\Delta+1)}$ *for any* $u, v \in V(\mathscr{G}_\alpha \cup \mathscr{G})$.

Therefore, similar to Algorithm 1, we can iteratively append leaves to $\mathscr{T}_\varepsilon$ to obtain an embedding of $\mathscr{T}$ into $\mathscr{G}_\alpha \cup \mathscr{G}$. As in every step we lose at most one vertex from each $\boldsymbol{B}(u,v)$ this works as long as

$$|L| \leq \varepsilon n + 1 < |\boldsymbol{B}(u,v)|,$$

which holds by Claim 3.1 and the assumption on ε and α. □

REFERENCES

[1] N. Alon and Z. FÃijredi, Spanning subgraphs of random graphs, *Graphs and Combinatorics* 8.**1** (1992), pages 91–94.

[2] N. Alon, M. Krivelevich, and B. Sudakov, Embedding nearly-spanning bounded degree trees, *Combinatorica* 27.**6** (2007), pages 629–644.

[3] J. Balogh, B. Csaba, M. Pei, and W. Samotij, Large bounded degree trees in expanding graphs, *the electronic journal of combinatorics* 17.**1** (2010), page 6.

[4] J. Balogh, A. Treglown, and A. Z. Wagner, Tilings in randomly perturbed dense graphs, *Combinatorics, Probability and Computing* 28.**2** (2019), pages 159–176.

[5] W. Bedenknecht, J. Han, Y. Kohayakawa, and G. O. Mota, Powers of tight Hamilton cycles in randomly perturbed hypergraphs, *arXiv preprint arXiv:1802.08900* (2018).

[6] P. Bennett, A. Dudek, and A. Frieze, Adding random edges to create the square of a Hamilton cycle, *arXiv preprint arXiv:1710.02716* (2017).

[7] T. Bohman, A. Frieze, and R. Martin, How many random edges make a dense graph Hamiltonian?, *Random Structures & Algorithms* 22.**1** (2003), pages 33–42.

[8] B. BollobÃąs and A. G. Thomason, Threshold functions, *Combinatorica* 7.**1** (1987), pages 35–38.

[9] J. BÃűttcher, Large-scale structures in random graphs, *Surveys in Combinatorics* 440 (2017), page 87.

[10] J. BÃűttcher, J. Han, Y. Kohayakawa, R. Montgomery, O. Parczyk, and Y. Person, Universality for bounded degree spanning trees in randomly perturbed graphs, *Random Structures & Algorithms* (2019).

[11] J. BÃűttcher, R. Montgomery, O. Parczyk, and Y. Person, Embedding spanning bounded degree subgraphs in randomly perturbed graphs, *Mathematika* (2019), pages 1–25.

[12] J. BÃűttcher, M. Schacht, and A. Taraz, Proof of the bandwidth conjecture of BollobÃąs and KomlÃşs, *Mathematische Annalen* 343.**1** (2009), pages 175–205.

[13] S. Das, P. Morris, and A. Treglown, Vertex Ramsey properties of randomly perturbed graphs, *arXiv preprint arXiv:1910.00136* (2019).

[14] S. Das and A. Treglown, Ramsey properties of randomly perturbed graphs: cliques and cycles, *arXiv preprint arXiv:1901.01684* (2019).

[15] G. A. Dirac, Some theorems on abstract graphs, *Proceedings of the London Mathematical Society* 3.**1** (1952), pages 69–81.

[16] A. Dudek, C. Reiher, A. RuciÅĎski, and M. Schacht, Powers of Hamiltonian cycles in randomly augmented graphs, *Random Structures & Algorithms* (2018).

[17] P. ErdÅŚs and A. RÃľnyi, On the existence of a factor of degree one of a connected random graph, *Acta Mathematica Hungarica* 17.**3-4** (1966), pages 359–368.

[18] A. Ferber, K. Luh, and O. Nguyen, Embedding large graphs into a random graph, *Bulletin of the London Mathematical Society* 49.**5** (2017), pages 784–797.

[19] A. Ferber and R. Nenadov, Spanning universality in random graphs, *Random Structures & Algorithms* 53.**4** (2018), pages 604–637.

[20] A. M. Frieze, On large matchings and cycles in sparse random graphs, *Discrete Mathematics* 59.**3** (1986), pages 243–256.

[21] A. Frieze and M. KaroÅĎski, *Introduction to random graphs*, Cambridge University Press, 2016.

[22] A. Hajnal and E. SzemerÃľdi, Proof of a conjecture of P. ErdÅŚs, *Combinatorial theory and its applications* 2 (1970), pages 601–623.

[23] J. Han, P. Morris, and A. Treglown, Tilings in randomly perturbed graphs: bridging the gap between Hajnal-SzemerÃľdi and Johansson-Kahn-Vu, *arXiv preprint arXiv:1904.09930* (2019).

[24] A. Johansson, J. Kahn, and V. Vu, Factors in random graphs, *Random Structures & Algorithms* 33.**1** (2008), pages 1–28.

[25] F. Joos and J. Kim, Spanning trees in randomly perturbed graphs, *arXiv preprint arXiv:1803.04958* (2018).

[26] J. KomlÃşs, G. N. SÃąrkÃűzy, and E. SzemerÃľdi, On the PÃşsa-Seymour conjecture, *Journal of Graph Theory* 29.**3** (1998), pages 167–176.

[27] J. KomlÃşs, G. N. SÃąrkÃűzy, and E. SzemerÃľdi, Proof of a packing conjecture of BollobÃąs, *Combinatorics, Probability and Computing* 4.**3** (1995), pages 241–255.

[28] J. KomlÃşs, G. N. SÃąrkÃűzy, and E. SzemerÃl'di, Proof of the Seymour conjecture for large graphs, *Annals of Combinatorics* 2.**1** (1998), pages 43–60.
[29] J. KomlÃşs, G. N. SÃąrkÃűzy, and E. SzemerÃl'di, Spanning trees in dense graphs, *Combinatorics, Probability and Computing* 10.**5** (2001), pages 397–416.
[30] J. KomlÃşs and E. SzemerÃl'di, Limit distribution for the existence of Hamiltonian cycles in a random graph, *Discrete Mathematics* 43.**1** (1983), pages 55–63.
[31] A. D. Korshunov, Solution of a problem of ErdÅŚs and Renyi on Hamiltonian cycles in nonoriented graphs, *Doklady Akademii Nauk*, volume 228, 3, Russian Academy of Sciences, 1976, pages 529–532.
[32] M. Krivelevich, Embedding spanning trees in random graphs, *SIAM Journal on Discrete Mathematics* 24.**4** (2010), pages 1495–1500.
[33] M. Krivelevich, Long paths and Hamiltonicity in random graphs, *Random Graphs, Geometry and Asymptotic Structure* 84 (2016), page 1.
[34] M. Krivelevich, M. Kwan, and B. Sudakov, Bounded-degree spanning trees in randomly perturbed graphs, *SIAM Journal on Discrete Mathematics* 31.**1** (2017), pages 155–171.
[35] D. KÃijhn and D. Osthus, On PÃşsa's conjecture for random graphs, *SIAM Journal on Discrete Mathematics* 26.**3** (2012), pages 1440–1457.
[36] R. Montgomery, Spanning trees in random graphs, *Advances in Mathematics* 356 (2019), page 106793.
[37] R. Nenadov and N. ÅăkoriÄĞ, Powers of Hamilton cycles in random graphs and tight Hamilton cycles in random hypergraphs, *Random Structures & Algorithms* 54.**1** (2019), pages 187–208.
[38] R. Nenadov and M. TrujiÄĞ, Sprinkling a few random edges doubles the power, *arXiv preprint arXiv:1811.09209* (2018).
[39] L. PÃşsa, Hamiltonian circuits in random graphs, *Discrete Mathematics* 14.**4** (1976), pages 359–364.
[40] O. Riordan, Spanning subgraphs of random graphs, *Combinatorics, Probability and Computing* 9.**2** (2000), pages 125–148.
[41] D. A. Spielman and S.-H. Teng, Smoothed analysis of algorithms: Why the simplex algorithm usually takes polynomial time, *Journal of the ACM* 51.**3** (2004), pages 385–463.

MAX HAHN-KLIMROTH, hahnklim@math.uni-frankfurt.de, GOETHE UNIVERSITY, MATHEMATICS INSTITUTE, 10 ROBERT MAYER ST, FRANKFURT 60325, GERMANY.

GIULIA SATIKO MAESAKA, giulia.maesaka@uni-hamburg.de, UNIVERSITÄT HAMBURG, FACHBEREICH MATHEMATIK, 55 BUNDESSTR., HAMBURG 20146, GERMANY.

YANNICK MOGGE, yannick.mogge@tuhh.de, HAMBURG UNIVERSITY OF TECHNOLOGY, MATHEMATICS INSTITUTE, 3 AM SCHWARZENBERG-CAMPUS, HAMBURG 21073, GERMANY.

SAMUEL MOHR, samuel.mohr@tu-ilmenau.de, ILMENAU UNIVERSITY OF TECHNOLOGY, MATHEMATICS INSTITUTE, 25 WEIMARER ST, ILMENAU 98693, GERMANY.

OLAF PARCZYK, o.parczyk@lse.ac.uk, LONDON SCHOOL OF ECONOMICS, DEPARTMENT OF MATHEMATICS, HOUGHTON ST, LONDON, WC2A 2AE, UK.