Hindawi
Complexity
Volume 2020, Article ID 4607025, 47 pages
https://doi.org/10.1155/2020/4607025

WILEY | Hindawi

Research Article

From Topic Networks to Distributed Cognitive Maps: Zipfian Topic Universes in the Area of Volunteered Geographic Information

Alexander Mehler,[1] **Rüdiger Gleim**,[1] **Regina Gaitsch**,[2] **Wahed Hemati**,[1] **and Tolga Uslu**[1]

[1]*Goethe-University Frankfurt, Frankfurt, Germany*
[2]*Justus Liebig University Giessen, Giessen, Germany*

Correspondence should be addressed to Alexander Mehler; mehler@em.uni-frankfurt.de

Received 30 August 2018; Revised 7 June 2019; Accepted 2 July 2019; Published 27 April 2020

Guest Editor: Nicole Beckage

Are nearby places (e.g., cities) described by related words? In this article, we transfer this research question in the field of lexical encoding of geographic information onto the level of intertextuality. To this end, we explore *Volunteered Geographic Information* (VGI) to model texts addressing places at the level of cities or regions with the help of so-called topic networks. This is done to examine how language encodes and networks geographic information on the aboutness level of texts. Our hypothesis is that the networked thematizations of places are similar, regardless of their distances and the underlying communities of authors. To investigate this, we introduce *Multiplex Topic Networks* (MTN), which we automatically derive from *Linguistic Multilayer Networks* (LMN) as a novel model, especially of thematic networking in text corpora. Our study shows a Zipfian organization of the thematic universe in which geographical places (especially cities) are located in online communication. We interpret this finding in the context of *cognitive maps*, a notion which we extend by so-called *thematic maps*. According to our interpretation of this finding, the organization of thematic maps as part of cognitive maps results from a tendency of authors to generate shareable content that ensures the continued existence of the underlying media. We test our hypothesis by example of special wikis and extracts of Wikipedia. In this way, we come to the conclusion that geographical places, whether close to each other or not, are located in neighboring semantic places that span similar subnetworks in the topic universe.

1. Introduction

In this article, we explore crowd-sourced resources for automatically characterizing geographical places with the help of so-called *topic networks*. Our goal is to model the thematic structure of corpora of natural language texts that are about certain places seen as thematic frames. This is done in order to automatically compare the thematic structures of corpora of texts about these places, which will be represented as topic networks. In this way, we want to investigate the regularity or systematicity according to which geographical objects (i.e., cities and regions) are dealt with, especially in online communication.

Our work relates to what is described by Crooks et al. [1] as a novel paradigm of modeling "*urban morphologies*." We not only add special wikis such as regional and city wikis as candidates to the resources listed in [1] but also introduce a novel method for modeling their content. This concerns local media of collaborative writing about places (cf. [2]), which contain *everyday place descriptions* [3] authored and networked according to the wiki principle. The corresponding wikis and the subgraphs of Wikipedia that we additionally analyze manifest *Volunteered Geographic Information* (VGI) [4–6] and thus relate to what is called the wikification of *Geographical Information Systems* (GIS) [7]. VGI is "*completing traditional authoritative geographic information*" [8], an information source which is still "*underutilized*" in geography [9] as a source of big textual data [8] making natural language processing an indispensable prerequisite for its analysis. According to Hardy

et al. [6], authoring VGI has a *spatial* component in the sense that people likely write about *local* content though this also holds for Wikipedia for a minor degree [10]. This spatial component can be accompanied by a lack of quality assurance, which makes VGI susceptible to deficiencies and to a distorted resource of still unknown extent [5]. In any event, the biased coverage of VGI is a characteristic of resources like Wikipedia so that the same region can be displayed very differently in its various language editions [11], a sort of biasing which is typical for user-generated content. Nevertheless, Hahmann and Burghardt [12] show that more than 50% of the articles in the German Wikipedia contain georeferenced data (at least indirectly via links to other articles), so that such media can be regarded as rich resources of VGI. Moreover, Goodchild and Li [5] point to the fact that crowd-sourcing or, more precisely, crowd-curation [13], as enabled by wikis, is a means of quality assurance.

We follow this concept and assume that geographic data, as manifested linguistically in online media, are a valuable resource to investigate how communities form a common sense for addressing places of common interest. In line with Clare ([14], 41), we additionally assume that "*[a]s people communicate more about a place, social consensus will create increased similarity between and within people's judgments of it*." However, we also assume that the latter similarity can affect communications of different communities about different places. In this way, we assume a kind of horizontal self-similarity [15] of the thematic structure of online media, which is more or less independent of the underlying theme and the community. That is, our hypothesis on the theming of places is as follows.

Hypothesis 1. Thematizations of different places at a certain level of thematic abstraction tend to be similar among each other (rather than being dissimilar) (1) in the sense that they focus on similar topics and (2) the way these topics are networked and (3) with respect to the skewness of this focus, regardless of whether the underlying media are generated by different communities and whether these communities address related or unrelated places at near or distant spaces.

The intuition behind Hypothesis 1 is that thematizations of places in web-based communication are seemingly somehow thematically redundant: in reporting, for example, on the cities in which people live, they may aim to emphasize the special character of these places. It seems, however, as if a thematic trend is breaking ground that ultimately makes such reports appear thematically very similar. Whether or not this intuition is actually a trend that can be observed specifically in the field of wiki-based media is something this study is intended to clarify. From this point of view, it is obvious that Hypothesis 1 is only a starting point which in itself needs further clarification in order to be testable: similarity, for example, is a highly context-sensitive attribute [17] that needs further definitional specifications in order to be computable. Likewise, the concept of thematization (theme or topic)—a concept which according to Adamzik [18] has so far found comparatively less attention in linguistics—is not yet specified in Hypothesis 1. Thus, an appropriate elaboration and concretization of Hypothesis 1 is one of the main tasks of the present paper. To this end, it is developing a *generic topic network model* in conjunction with a measurement procedure which will specify both the notion of *similarity* (which will be defined in terms of the graph similarity of topic networks) and of the *thematization of places* (which will be defined in terms of topic labeling and topic networking). This topic network model will allow Hypothesis 1 to be reformulated and concretized in the form of variants (i.e., Hypotheses 2–4), which will be presented in Section 3.2.7 and whose formulations presuppose the topic network model that this paper develops in the preceding sections.

The skewness that is mentioned by Hypothesis 1 reminds one of a Zipfian process, according to which a few topics dominate, while the majority of candidate topics are underrepresented or disregarded. Therefore, we speak of *Zipfian thematic universes*, which are spanned by the thematization of the same places in online media such as special wikis of the sort studied here. By the term *topic*, we refer to the notion of aboutness of texts [18, 19]. From a linguistic point of view, the terminology of Hypothesis 1 seems to be confusing when referring to places as *what is given* and with topic to *what is said about these places*. The reason is that linguistics distinguishes between what is given (*theme* or *topic*) and what is said about it (*rheme*, *comment*, or *focus*) in a given piece of text [18, 20–22]: a mention of a city like *Vienna*, for example, can be connected with certain subtopics (e.g., *classical music*), which characterize this place rhematically by providing new information about it. The latter distinction is meant when we relate subtopics in the role of rhemes to places in the role of topics in the linguistic sense. Thus, when talking about topics as part of a computational model, we will use the term *topic* (topic_2), while when talking about places as topics in the linguistic sense (topic_1), we will use the term *theme* and speak about its rhemes as its subtopics modeled by topics (topic_2) as units of our model. This scenario and its relation to Hypothesis 1 are depicted in Figure 1. It shows a generalization of a hypothesis of Louwerse and Zwaan [16] according to which language encodes geographical information: the places p and q, which are understood as conceptual units (i.e., mental models), are described by or expressed in two discourse units (texts, dialogs, etc.) x and y. From the latter units, the topic representations α and β are derived by means of a computational model (e.g., *Latent Dirichlet Allocation* (LDA) [23] or the topic network model introduced in Section 3). While such derived topics are part of the computational model, the underlying discourses belong to the modeled system. We assume that the conceptual unit p (q) is structured into a system of networked rhemes or subtopics p_i (q_m). Ideally, the derived topic α in Figure 1 is a valid model of one of the rhemes of place p (e.g., p_i) and β of one of the rhemes of place q (e.g., q_m). If we assume now that p and q are conceptually related (e.g., similar) to each other, then the linguistic encoding hypothesis implies that this is possibly reflected by a relatedness (e.g., similarity) relation among some rhemes of these places (e.g., by the relatedness of p_i and q_m). From the point of view of modeling, this relation is ideally mapped by the relatedness (e.g., similarity)

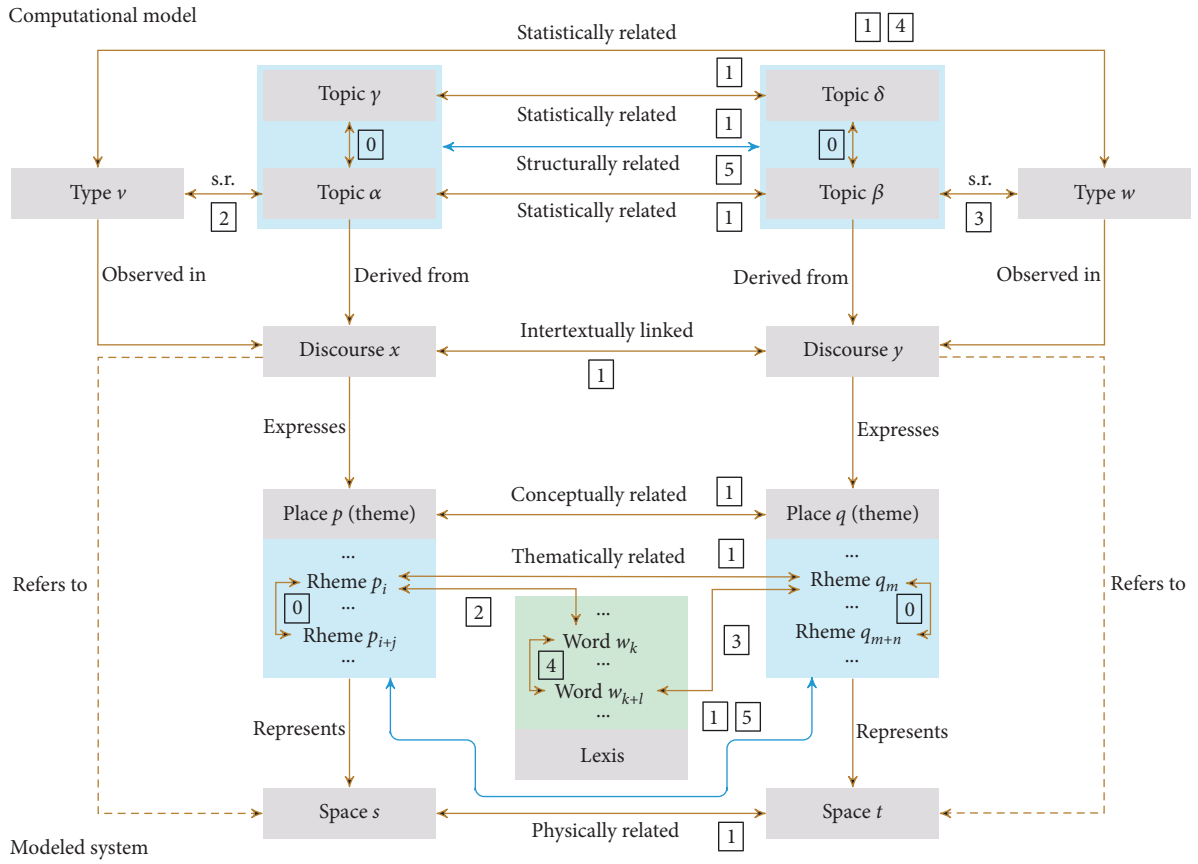


FIGURE 1: Schematic depiction of a generalization of a hypothesis of Louwerse and Zwaan [16] saying that language encodes geographical information: the places p and q are expressed in the discourses x and y, respectively, from which the topic representations α and β are computationally derived. Places are structured into systems of networked rhemes or subtopics. The conceptual relatedness of p and q is grounded in the relatedness of the rhemes p_i and q_m and modeled by the relatedness of the derived topics α and β modeling these rhemes. According to the semiotic triangle, we assume that the relation of signs (here, texts) to their referents (here, spaces) is mediated by sign processes. We use dashed arcs to express the indirect relation of the former to the latter. In lexical variants of this approach, p and q are preferably denoted or described by some words w_k and w_{k+l} of the underlying lexis, which are syntagmatically or paradigmatically associated and modeled by some types v and w. Framed numbers indicate relations that potentially parallelize each other. *s.r.* means *statistically related*.

of the derived topics α and β. We assume that conceptual relations between places can be parallelized by relations of physical proximity or distance between spaces that are mentally modeled by these places. If one additionally assumes that proximity in space correlates with relatedness in conceptual space (the less the distant, the more the similar, for example), one obtains a linguistic variant of Tobler's so-called first law (see Section 2). If we look at the literature (see Section 2), we find that the approaches in this area differ in terms of the linguistic level at which they observe the linguistic encoding of platial [13] relations: for example, at the level of intertextually linked texts, at the level of the topics these texts are about, or at the level of lexical elements used by these and other texts to deal with the latter topics. In lexical variants of this approach, the places p and q, for which we assume that they are conceptually related, are preferably referred to or described by means of lexical items w_k and w_{k+l} (see Figure 1) of the underlying lexis that are syntagmatically or paradigmatically associated. From the point of view of modeling, we have to then assume the two types v and w (as models of the words w_k and w_{k+l}) for which we automatically detect, for example, their (paradigmatic) closeness in semantic space (cf. [24, 25]) or the similarity of their (syntagmatic) co-occurrence statistics (cf. [26]).

From this analysis, we obtain a series of reference points or means for encoding geographical information about conceptual relations (see [1] in Figure 1) of places. This concerns more precisely a series of possible parallelizations of such relations, which may ultimately be parallelized by relations between the spaces designated by these places (for the numbers in brackets, see Figure 1): at the level of the modeled system, this refers to thematically linked rhemes, intertextually linked discourse units (e.g., texts), and syntagmatically or paradigmatically linked words ([1]). From a modeling point of view, we distinguish the statistical relatedness of types or of topics as candidate parallelizations ([1]). Beyond that, we find the parallelization of the relatedness of rhemes and words on the one hand and of types and topics on the other ([2], [3]), as well as that of the relatedness of words on the one hand and of types on the other ([4]). The parallelization of the relatedness of rhemes of the same place ([0]) by the relatedness of the rhemes of

another place concerns the core of our network approach. Such relations among rhemes constitute rhematic networks or networks of rhemes on both sides of the affected places. Our main assumption is now that any such rhematic network, which manifests the thematic structure of a place, can be related *as a whole* to that of another place. In doing so, it is, from a modeling point of view, ideally parallelized *by the structural relatedness (e.g., similarity or complementarity) of topic networks*, which are derived from corpora of texts, each of which describes one of these places ([5]). This type of parallelization affects entire networks of linguistic objects and yet offers a means of encoding the conceptual relationship of places ([1]) or the proximity of spaces, respectively. In the present paper, we explore relations of Type [5] in order to learn about the encoding of geographical information in natural language texts, that is, about relations of Type [1]. To this end, we develop, instantiate, and empirically test a formal model of multiplex topic networks derived from so-called linguistic multilayer networks as a model of relations of Type [5].

From this point of view, Hypothesis 1 means that certain rhemes of places and the structure they span resemble each other, regardless of how far the quantified distances of the spaces represented by these places are and regardless of the fact that the texts in which these rhemes are described are written by different communities. To test this hypothesis, we introduce topic networks to make the networking of topics a research object according to the scenario described in Figure 1, that is, in relation to the hypothesis of linguistic encoding of geographical information. The contributions of this article are of theoretical, methodical, and empirical nature.

(1) Formal modeling: we develop a generic, extensible formalism for the representation of topic networks that cover a wide range of informational sources for spanning and weighting topic links. To this end, we introduce the notion of *multiplex topic networks* derived from so-called *multilayer linguistic networks*. In this way, we enable the same place to be represented by a family of thematic networks that offer different perspectives on the networking of its rhemes. We exemplify this model by means of two perspectives provided by so-called *Text Topic Networks* (TTN) and their corresponding *Author Topic Networks* (ATN).

(2) Procedural modeling: we develop a measurement procedure for instantiating our formal model. To this end, we introduce novel measures of the similarity of labeled graphs that are sensitive to their links and to their nodes.

(3) Experimentation: we further develop the range of baseline statistics in network theory in order to better assess the quality of our measurements. To this end, we test our model by means of a threefold classification experiment that compares a set of TTNs with each other, a set of corresponding ATNs with each other, and the former TTNs with the latter ATNs.

(4) Theory formation: we interpret our findings in the context of cognitive maps, thus building a bridge between our network-theoretical approach and approaches to the cognitive representation of geographical information. We show how to integrate the analysis of entire networks into the research about the linguistic encoding of geographical information (see Figure 1).

This paper is organized as follows: Section 2 discusses related work. Section 3 introduces our formal model of linguistic multilayer networks and the multiplex topic networks derived from them. Section 4 describes our experiments in detail, and Section 5 discusses our findings. Finally, Section 6 concludes and gives an outlook on future work.

2. Related Work

Our work is related to linguistic research on Tobler's [27] first law (TFL) which says that "[. . .] *everything is related to everything else, but near things are more related than distant things*" ([27], p. 236). Due to its underspecification, this so-called law raised many questions about what it means to be *related* or *distant* [28]. Accordingly, a range of approaches exist that make different proposals to interpret relatedness also in terms of *semantic relatedness*. In the context of information visualization, Montello et al. [29] test a variant of TFL called the first law of *cognitive* geography which says that "*people believe closer things to be more similar than distant things*" ([29], p. 317), where spatial distance is referred to for judging the similarity of information objects. This approach is contrasted with a study by Hecht and Moxley [30] who model relations of Wikipedia articles as a function of the probability of being linked in the web graph and find that this probability is related to the geographical distance of toponyms described in the articles. Hecht and Moxley relate their finding to the transitivity of networks by stating that the smaller the geographical distance of nodes, the higher their clustering coefficient ([30], 101). This work is extended by Li et al. [31], who calculate semantic relationships of articles instead of hyperlinks and show that TFL holds independently of the geographical domain up to a certain distance threshold. A lexical variant of TFL is mentioned by Yang et al. [32], according to which geographically close words tend to be clustered into the same geographical topics. This phenomenon has earlier been studied by Louwerse et al. (cf. the review in [26]) who reformulated Firth's famous dictum by saying that "[. . .] *you shall know the physical distance between locations by the lexical company they keep*" ([26], p. 1557). This means that the distance of places correlates with syntagmatic associations between the lexical items used to describe them. That is, language encodes geographical information [16] at least regarding the distances of semantically related places. From this perspective, TFL appears to be reformulated as a candidate for a geolinguistic law that is compatible with the more general *Symbol Interdependency Hypothesis* (SIH) [33]. According to SIH, linguistic information encodes perceptual

information so that the former serves as a shortcut to the latter [33]. Finally, a rather text-linguistic variant of TFL is proposed by Adams and McKenzie [34], which states that near places are each described by texts whose topics are more similar than in the case of texts about distant places.

In contrast to these approaches, we hypothesize that places, no matter how far apart, have similar topic distributions when their descriptions are transmitted by media such as city and region wikis. If we find evidence for this hypothesis, there are various candidates for explaining it: Firstly, such a finding could indicate a trivial meaning of TFL (cf. [28]) in relation to the topics modeled by us, implying that everything, distant or not, is highly related. Secondly, it could indicate the (in)effectiveness of distances and similarities at different scales: at the level of local, specific topics (within the scope of TFL) and at the level of global, more general topics (outside the scope of TFL). Thirdly, such a finding could indicate a hidden similarity of processes of collaboratively writing wikis about different places, even if the wikis are written by different communities (see Hypothesis 1). In order to decide between these alternatives, we need a new topic model that derives networks of thematic structures at different scales from texts in online media about the same places. This should at least include the networking of topics along relations of intertextuality and coauthorship in order to allow for revealing similarities of the underlying processes of collaborative writing. To this end, we will develop multiplex networks that integrate text- and author-driven topic networks.

So far, most approaches to thematic aspects of places use topic modeling based on *Latent Dirichlet Allocation* (LDA) to associate topics and texts about geographical units, where topics are represented as sets of thematically related words. An early approach in this regard is described by Mei et al. [35] who model *spatiotemporal theme patterns* to identify dominant topics in texts that are connected to places. A related approach is proposed by Qiang et al. [36], who aim to detect topics that are "localized" in places. This is done to ground their similarities in relations of their thematic representations—a scenario that is omnipresent in linguistically motivated work in the context of TFL (cf. Figure 1). Likewise, Adams and McKenzie [34] extract topic models from travel blogs to detect topics as groups of semantically related words associated to places, so that relations among places can be identified by shared topics. Another example is proposed by Bahrehdar and Purves [37]: instead of documents written by individual authors, they analyze tagging data extracted from image descriptions in Flickr. A hybrid model of topic modeling comes from Yin et al. [38], in which representations of regions are used instead of documents to link topics to places. A related *region-topic model* that uses regions as topics to map words, sentences, and texts to distributions of regions or to ground them semantically (cf. [39]) is proposed by Speriosu et al. [40]. A promising extension is developed by Gao et al. [41] who aim at detecting higher-level functional regions as semantically coherent areas of interest. To this end, they analyze co-occurrence relations between topics to describe many-to-many relations of locations and urban functions. Another direction is pursued by Lansley and Longley [42], who investigate the location- and time-based distribution of topics in Twitter, setting a number of twenty topics as a target for LDA. See also Jenkins et al. [13] who utilize a list of six high-level topic categories. One of the largest studies in this context is the one of Gao et al. [43] who present an integrative approach to modeling texts from a range of different media such as Wikipedia, Twitter, and Flickr to demarcate cognitive regions [44]. All these approaches start from topic modeling to map natural language texts onto distributions of topics in order to relate the places thematized by these texts (cf. Figure 1).

A prominent precursor of topic models [45] is given by *Latent Semantic Analysis* (LSA) [46]. Consequently, there are studies in the context of TFL based on this predecessor. Davies [24], for example, interprets the associations of place names computed by LSA from place descriptions as a model of the cognitive representation of the corresponding spaces (cf. [47]). This approach opens up a perspective for measuring biased cognitive representations of spatial systems: according to Davies, her approach provides representations of cognitive geographies that are explored by the associations of semantically close place names in accordance or not with the underlying geographical relations, that is, in accordance or not with TFL (cf. [39]). These and related studies produce interesting results about the localization of topics or vice versa about the thematization of places in texts. However, they mostly disregard topic networking, not to mention the networking of topics viewed from different angles. Although it is easy to derive a network approach from binary relations of topic similarity, relationships that cannot be traced back to sharing similar words are hardly mapped by topic models of the sort considered so far. By generating topic distributions per location, for example, we know nothing about the dynamics of the coauthorship of the underlying texts: in the extreme case, one observes (dis) similarities, which result from the activity of a small number of authors or even only one author—in contrast to the assumed collaboration density of online media such as Wikipedia. Therefore, it is our goal to develop a model of topic networks that simultaneously addresses the dynamics of the coauthorship of the underlying texts. A subtask will be to develop a formal model of thematic networking that is generic enough to integrate a wide range of sources of networking—at least theoretically.

While most of the approaches considered so far ignore aspects of networking, a second branch of research tends to follow the paradigm of network theory. Hu et al. [48], for example, measure the semantic relatedness of cities as nodes of a city network [9] depending on the co-occurrences of city names in news articles. This approach is related to Liu et al. [49], who explore co-occurrences of toponyms to induce city networks that can be used to test predictions associated with TFL. Hu et al. [48] further develop this approach to networking cities by reference to topics of articles in which the corresponding toponyms are observed. They use Labeled LDA [50] to learn to extract topics α from texts to finally determine the α-relative similarity of cities based on the co-occurrences of their names in texts about α. Another

approach to city networks using Wikipedia as a data source is proposed by Salvini and Fabrikant [9]: they link cities as a function of the number of articles "co-siting" [51] their Wikipedia articles. A comprehensive perspective on modeling spatial information is developed by Luo et al. [52], who propose a three-part network model that integrates representations of spatial, social, and semantic networks. In this conceptual model, semantics plays the role of interpreting behavior in spatial and social space and thus of bridging them. Although we share this hybridization of the network perspective on spatial information, we strive for a more concrete model that can be empirically tested.

Any such study has to face various aspects of the vagueness [44, 53] or informational uncertainty [5] of concepts of regions [44] and places [13] and especially of the names of such entities [43]. According to Winter and Freksa [54], this includes *semantic ambiguity*, *indeterminacy of spatial extent*, or *boundary vagueness* [43], preference-oriented *re-scaling of extent*, and the *dynamics of salience* affected by various dimensions of contrast. Beyond boundary vagueness, Gao et al. [43] speak of the shape and location vagueness by example of cognitive regions. Furthermore, Jenkins et al. [13] refer to the temporal dynamics of places as evolving concepts as a source of uncertainty. From a methodological point of view, this multifaceted uncertainty has two implications: in relation to the model, which should be flexible enough to map these facets, and in relation to the object itself, which could complicate its modeling by unsystematically distorting it.

In accordance with Hu's study [55], we assume that the thematic perspective complements the spatial and temporal perspective of the study of places. A rheme can be understood as the "content" of a geographical region that expands its dimensionality [44]. This content may be further specified in terms of affordances, functions, or shared conceptual representations associated by members of a community with the corresponding place so that different places can be related by being associated with similar content. This thematic perspective will be at the core of our article. To this end, we follow the approach of Jenkins et al. [13], according to which places are connected with meanings generated by collaborators of crowd-sourcing media such as Wikipedia: their collaboration creates what Jenkins et al. call *platial themes*, namely, themes that are characteristic for certain places. As shared meanings, these platial themes ultimately create a "*collective sense of place*," as it is perceived by the corresponding community. In this context, Jenkins et al. [13] propose to study *politics*, *business*, *education*, *recreation*, *sports*, and *entertainment* as six high-level topics of places. However, by reference to the *Dewey Decimal Classification* (DDC), we will instead deal with more than six hundred hierarchically organized topics, each of which is manifested by a range of Wikipedia articles. In any event, we have to consider that thematic aspects may distort the conceptualization and perception of spatial objects [43]. A central question then concerns the regularity or systematicity of this distortion in the sense of asking to what extent thematic representations of different places show similar aspects of being biased. This question will be at the core of this article.

3. Multiplex Topic Networks: A Novel Approach to Topic Modeling

In order to study relations of thematic preference in VGI as a manifestation of distributed cognition, we introduce *Topic Network*s (TNs) as an alternative to *Topic Model*s (TMs) [23, 58, 59]. TMs are based on the idea that texts manifest probabilistic distributions of topics which are represented as probability distributions over the lexical constituents of these texts, where these distributions may be affected by style, the underlying genre, or any other (syntactic, semantic, or pragmatic) criterion of text production [60–62]. Regardless of its success, this model is unsuitable for modeling TNs as manifestations of distributed cognitive maps because of the following problems:

(P1) Corpus specificity: the corpus specificity of TMs impairs comparability and transferability to ever new corpora, since the topic distributions are learned from the input corpora whose topics are to be modeled. This approach apparently cannot use a transferable topic model as a basis for representing the topics of a large number of different corpora.

(P2) Topic labeling: the corpus-specific derivation of topic labels from the input corpora makes it difficult to compare their topic distributions. As reviewed by Herzog et al. [63], external resources can be used for this task. However, there are hardly any such resources for all possible topic combinations—unless one wants to explore an overarching system such as Wikidata making such a project considerably more difficult due to its size. The labeling problem can be addressed using, for example, Labeled LDA [50], an approach that leads us into the area of supervised classification, which is also followed here.

(P3) Scalability: instead of dealing with corpora of equally large texts, online communication often leads to sparse, tiny texts that sometimes consist of a single sentence, a single phrase, or a single word. Regardless of the size of the text, we need a procedure that determines its topic distributions so that texts of different sizes can be compared using topic models of comparable size. Even if small texts are postprocessed (after topic modeling) in such a way that their topic distributions are derived from their lexical constituents, such an approach would nevertheless mean to exclude text snippets from the training process.

(P4) Rare topics: one reason to prefer training by means of corpora as large as Wikipedia is to allow for detecting topics even if they form a kind of thematic hapax legomenon in the corpora to be analyzed. If we try to identify rare topics directly from these corpora, we will probably not detect them, since by definition these corpora do not provide enough information to identify such topics. In any event, the rarity of evidence about a topic should not be an impediment to identifying its occurrences even at the level of single sentences.

(P5) Methodical closeness: instead of deriving all distributions of all dependent and independent variables

as part of the same topic model, one possibly wants to include different information sources that are computed by different methods based on diverse computational paradigms (e.g., ontological approaches to measuring sentence similarities, approaches to word embeddings based on neural networks, and topic models). In order to enable this, we look for a methodologically open topic model that allows such different resources to be easily integrated.

In a nutshell, we are looking for an approach that (i) allows thematic comparisons of previously unforeseen text corpora using an underlying reference corpus, (ii) offers a generic solution to the problem of topic labeling, (iii) is highly scalable and can therefore map even the smallest text snippets to topic distributions, (iv) simultaneously takes rare topics into account, and (vii) is methodologically open and expandable. Such a topic network model is now developed in two steps: in Section 3.1, we introduce the underlying formal apparatus. This is done by deriving multiplex topic networks from linguistic multilayer networks. Section 3.2 describes a method by which this model is instantiated as a prerequisite for its empirical testing.

3.1. From Linguistic Multilayer Networks to Multiplex Topic Networks. In this section, we introduce multiplex topic networks. This is a type of network that is based on the idea of deriving the networking of topics of textual units by evaluating evidence from different sources of information such as text vocabulary, higher-level text components, distributed authorship or readership, genre, register, or medium. Since these sources of evidence can be explored in different compositions, this can lead to different perspectives on the salience and networking of the topics addressed by the same texts. Topic networks are multiplex precisely in this respect: the different evidence-providing perspectives may lead to different topic networks that allow comparisons to be made through which differences in the linguistic, social, or otherwise contextual embedding of thematizations become visible. This concept of a multiplex topic network is now being generically formalized.

To introduce multiplex topic networks, we start with defining *linguistic multilayer networks* (Definition 1) whose layeredness allows for distinguishing several (non)linguistic information sources of topic networking. We refer to supervised topic classifiers trained by means of large reference corpora to tackle the challenges P1, P2, P3, and P4. Based thereon, we introduce so-called *text topic networks* (Definition 3), which evaluate intra- and intertextual relations for the purpose of topic networking. Then, we introduce *two-level topic networks* (Definition 4) and exemplify them by *author* (Definition 5) and *word topic networks* (Definition 6), which explore relations of (co)authorship and lexical relatedness, respectively, as sources of topic networking. These notions are generalized to arrive at *n-level topic networks* (Definition 7) which are based on $n > 1$ informational sources of topic networking (cf. challenge P5). Finally, *multiplex topic networks* are defined as families of n-level topic networks (Definition 8) representing the networking of the same set of topics from different informational perspectives and thus allowing for mapping the thematic dynamics, for example, of descriptions of the same place.

Definition 1. Let $X = \{x_1, \ldots, x_n\}$ be a corpus of texts and $l \in \mathcal{N}, l > 1$. A *Linguistic Multilayer Network* (LMN) is a tuple (Mehler [57] speaks of *multilevel graphs*; see Boccaletti et al. [64] for a comprehensive overview of related notions whose formalism is used here; see Stella et al. [65] for an example of a multiplex network of lexical systems)

$$\begin{aligned}\mathscr{L}(X, l) &= (\mathbb{L}, \mathbb{C}),\\ \mathbb{L} &= \{L_i = (V_i, A_i, \mu_i, \nu_i, \lambda_i, \kappa_i) \mid i = 1, \ldots, l\},\\ \mathbb{C} &= \{C_{i.j} = (V_{i.j}, A_{i.j}, \mu_{i.j}, \nu_{i.j}, \lambda_{i.j}, \kappa_{i.j}) \mid i, j = 1, \ldots, l : i \neq j\},\end{aligned} \tag{1}$$

of two sets of directed graphs such that the set of *kernel layers* $\mathbb{L}$ consists of a pivotal text layer and several derivative layers, that is, a coauthoring layer, a language-systematic word layer, and possibly several layers modeling the networking of constituents of the pivotal texts:

(1) The pivotal text layer $L_1 = (V_1, A_1, \mu_1, \nu_1, \lambda_1, \kappa_1)$, also called text network, is spanned by texts of the corpus $V_1 = X$ such that A_1 is manifesting intratextual (as in the case of reflexive arcs) or intertextual relations

(2) The author layer $L_2 = (V_2, A_2, \mu_2, \nu_2, \lambda_2, \kappa_2)$, also called agent network, is spanned by the network of agents (co)authoring the texts in V_1 and their social relations

(3) The lexicon layer $L_3 = (V_3, A_3, \mu_3, \nu_3, \lambda_3, \kappa_3)$, also called word network, is spanned by the language-systematic lexical signs (i.e., lexemes and related units) used by agents of V_2 as part of their agent lexica to author the texts in V_1

(4) For $3 < i \leq l' < l$, $L_i = (V_i, A_i, \mu_i, \nu_i, \lambda_i, \kappa_i)$ is called a constituent layer modeling the networking of (e.g., lexical, phrasal, and sentential) constituents of texts $x \in V_1$ such that A_i maps intratextual (e.g., anaphoric) or intertextual (e.g., sentence similarity) relations

(5) For $l' < i \leq l$, $L_i = (V_i, A_i, \mu_i, \nu_i, \lambda_i, \kappa_i)$ is called a contextual layer modeling the networking of units (e.g., media, genres, and registers [66]) of the contextual embedding of texts $x \in V_1$ such that A_i maps, for example, relations of the switching, merging, or embedding [67, 68] of these contextual units

(6) For each $i, j \in \{1, \ldots, l\}, i \neq j, C_{i.j} \in \mathbb{C}, |\mathbb{C}| = l(l-1)$, is called a margin layer where $V_{i.j} = V_i \cup V_j$, $A_{i.j} \subseteq V_i \times V_j$, $\mu_{i.j} = \mu_i \cup \mu_j$, and $\lambda_{i.j} = \lambda_i \cup \lambda_j$.

For $i, j = 1, \ldots, l, i \neq j$, μ_i and $\mu_{i.j}$ are vertex weighting functions, ν_i and $\nu_{i.j}$ are arc weighting functions, λ_i and $\lambda_{i.j}$ are vertex labeling functions, and κ_i and $\kappa_{i.j}$ arc labeling functions. We say that the linguistic multilayer network $\mathscr{L}(X, l)$ is *spanned over the text corpus* X and *layered into* l *layers*.

Example 1. To illustrate our definitions, we construct a minimized example. Suppose a corpus of four texts $V_1 = X = \{x_1, x_2, x_3, x_4\}$, each containing three lexemes

$x_1 = \{w_1, w_2, w_3\}$, $x_2 = \{w_1, w_2, w_4\}$, $x_3 = \{w_5, w_6, w_7\}$, and $x_4 = \{w_4, w_8, w_9\}$ (for reasons of simplicity, we exemplify texts as bag-of-words), that is, $V_3 = \{w_1, \ldots, w_9\}$, $V_{3.1} = \{w_1, \ldots, w_9, x_1, \ldots, x_4\}$, and $A_{3.1} = \{(w_1, x_1), (w_2, x_1), (w_3, x_1), \ldots, (w_4, x_4), (w_8, x_4), (w_9, x_4)\}$. Further, we assume four authors $V_2 = \{a_1, a_2, a_3, a_4\}$ such that a_1 and a_2 coauthored x_1 and x_2, while a_3 and a_4 coauthored x_3 and x_4; that is, $V_{2.1} = \{a_1, \ldots, a_4, x_1, \ldots, x_4\}$ and $A_{2.1} = \{(a_1, x_1), (a_2, x_1), (a_1, x_2), (a_2, x_2), (a_3, x_3), (a_4, x_3), (a_3, x_4), (a_4, x_4)\}$. Further, we assume that the texts x_1 and x_2 are linked by some intertextual coherence relations (e.g., by a rhetorical relation, an argument relation, or some hyperlinks) as are the texts x_3 and x_4 so that $A_1 = \{(x_1, x_2), (x_3, x_4)\}$. Note that additional arcs of the layers L_1, L_2, and L_3 will be generated according to the subsequent definitions. For simplicity reasons, we assume all weighting functions to be limited to the set $\{0, 1\}$ of vertex/arc weights. Since we assume no additional constituent layer, we get $l = 3$. Thus, any linguistic multilayer network $\mathscr{L}(X, 3)$ based on this setting is layered into three layers.

Throughout this paper, we use the following simplifying notation: for any graph $G = (V, A, \lambda)$ of order $|G| = |V|$, arc set $A \subseteq V^2$ of size $|A|$ and vertex labeling function λ, and any vertex $v \in V$, we write $\dot{v} = \lambda(v)$. Thus, for any two graphs G_i and G_j with vertex labeling functions λ_i and λ_j, for which $\lambda_i(v) = \lambda_j(w)$, $v \in V_i$ and $w \in V_j$, we can write $\dot{v} = \dot{w}$. Further, for any function $f : X \times Y \longrightarrow Z$, for which $f(x, y) = z$, we use the following alternative notations:

$$f(x, y) = z \Longleftrightarrow x \longrightarrow_f y = z \Longleftrightarrow x \xrightarrow{f} y = z \Longleftrightarrow f_y(x) = z. \tag{2}$$

Finally, for any function $f : Z^n \longrightarrow Z$, Z being any set, we introduce the following notation based on square brackets:

$$\begin{aligned} &f(\ldots, x \xrightarrow{f} y, y \xrightarrow{g} x \ldots) = z \\ &\Longleftrightarrow f\left[\ldots, x \underset{g}{\overset{f}{\longleftrightarrow}} y \ldots\right] = z \Longleftrightarrow f\left[\ldots, x_g \longleftrightarrow_f y \ldots\right] \\ &\qquad = z. \end{aligned} \tag{3}$$

To leave no room for ambiguity, we assume that expressions of the sort $x \longrightarrow_f y, y \longrightarrow_g x$ are replaced from left to right into expressions of the sort $x_g \longleftrightarrow_f y$. Henceforth, a structure such as $x \longrightarrow_f y$ will be called *information link*. Based on Definition 1, we start now with introducing text topic networks using the following auxiliary notion.

Definition 2. Let $\mathscr{C} = (V_{\mathscr{C}}, A_{\mathscr{C}})$ be a directed *Generalized Tree* (GT) according to Mehler [69, 70] representing a hierarchical topic structure, henceforth called *Reference Classification System* (RCS), that is spanned by kernel arcs which are possibly superimposed by upward, downward, lateral, sequential, external, or reflexive arcs. (See Figure 2 for an example of a GT. This notion is required since we may decide for using, for example, the category system of Wikipedia as an RCS, which spans a GT [70]). That is, vertices $t \in V_{\mathscr{C}}$ represent topics, while kernel arcs $(t, u) \in A_{\mathscr{C}}$ represent subordination relations according to which u is a thematic specialization of t. Let further, θ denote a *hierarchical text classifier* [71] taking values in $V_{\mathscr{C}}$ that has been trained, validated, and tested by means of a reference corpus $\mathscr{R}$. Let now $\mathscr{L}(X, l) = (\mathbb{L}, \mathbb{C})$ be a LMN spanned over the text corpus X and layered into l layers. We call the structure

$$\mathscr{S} = (\mathscr{C}, \theta, \mathscr{L}(X, l)), \tag{4}$$

a *Definitional Setting* for defining topic networks.

Example 2. Given the LMN of Example 1, the *Dewey Decimal Classification* (see Section 3.2), and the topic classifier θ of [72], which uses the DDC as its Reference Classification System $\mathscr{C}$, a definitional setting is exemplified by $(\mathrm{DDC}, \theta, \mathscr{L}(X, 3))$. More specifically, by t_1, t_2, t_3 we will denote three topic labels of the third level of the DDC so that $V_{\mathscr{C}} = \{\ldots, t_1, t_2, t_3, \ldots\}$. Note that by using the DDC as a reference classification, the generalized tree of Definition 2 is reduced to a tree (see Section 3.2 for more details).

Definition 3. Given a definitional setting $\mathscr{S} = (\mathscr{C}, \theta, \mathscr{L}(X, l))$ according to Definition 2, a *Text Topic Network* (TTN) is a vertex- and arc-weighted simple directed graph

$$T(L_1) = T(L_1, \{\}) = (V, A, \mu, \nu, \lambda, \kappa), \tag{5}$$

with vertex set V and arc set $A \subseteq V^2$ which is said to be *derived from* $\mathscr{S}$ and *inferred from* L_1 by means of the optional classifier $\theta^{\leftarrow}$ and the monotonically increasing functions $\alpha, \beta, \gamma, \delta : \mathbb{R}_0^+ \longrightarrow \mathbb{R}_0^+$ if and only if $\forall v \in V$ and $\forall a = (v, w) \in A$:

$$\begin{aligned} \mu(v) &= \alpha\left(\sum_{x \in V_1} \beta\left(\theta(x, \lambda(v)), \theta^{\leftarrow}(\lambda(v), x)\right)\right) \\ &= \alpha\left(\sum_{x \in V_1} \beta\left(\theta_x(\dot{v}), \theta^{\leftarrow}_{\dot{v}}(x)\right)\right) \\ &= \alpha\left(\sum_{x \in V_1} \beta\left(x \xrightarrow{\theta} \dot{v}, \dot{v} \xrightarrow{\theta^{\leftarrow}} x\right)\right) \\ &= \alpha\left(\sum_{x \in V_1} \beta\left(x \underset{\theta^{\leftarrow}}{\overset{\theta}{\longleftrightarrow}} \dot{v}\right)\right) > 0, \end{aligned} \tag{6}$$

$$\begin{aligned} \nu(a) &= \gamma\left(\sum_{x, y \in V_1} \delta\left(\theta(x, \lambda(v)), \theta^{\leftarrow}(\lambda(v), x), \theta(y, \lambda(w)), \right.\right. \\ &\qquad \left.\left. \theta^{\leftarrow}(\lambda(w), y), \nu_1(x, y)\right)\right) \\ &= \gamma\left(\sum_{x, y \in V_1} \delta\left[x \underset{\theta^{\leftarrow}}{\overset{\theta}{\longleftrightarrow}} \dot{v}, y \underset{\theta^{\leftarrow}}{\overset{\theta}{\longleftrightarrow}} \dot{w}, x \xrightarrow{\nu_1} y\right]\right) > 0, \end{aligned} \tag{7}$$

where $\mu : V \longrightarrow \mathbb{R}^+$ is a vertex weighting function, $\nu : A \longrightarrow \mathbb{R}^+$ an arc weighting function, $\lambda : V \longrightarrow V_{\mathscr{C}}$ an injective vertex labeling function, $V_{\mathscr{C}}(V) = \{\lambda(v) \mid v \in V\} \subseteq V_{\mathscr{C}}$, and κ an injective arc labeling function. $T(L_1)$ is called a

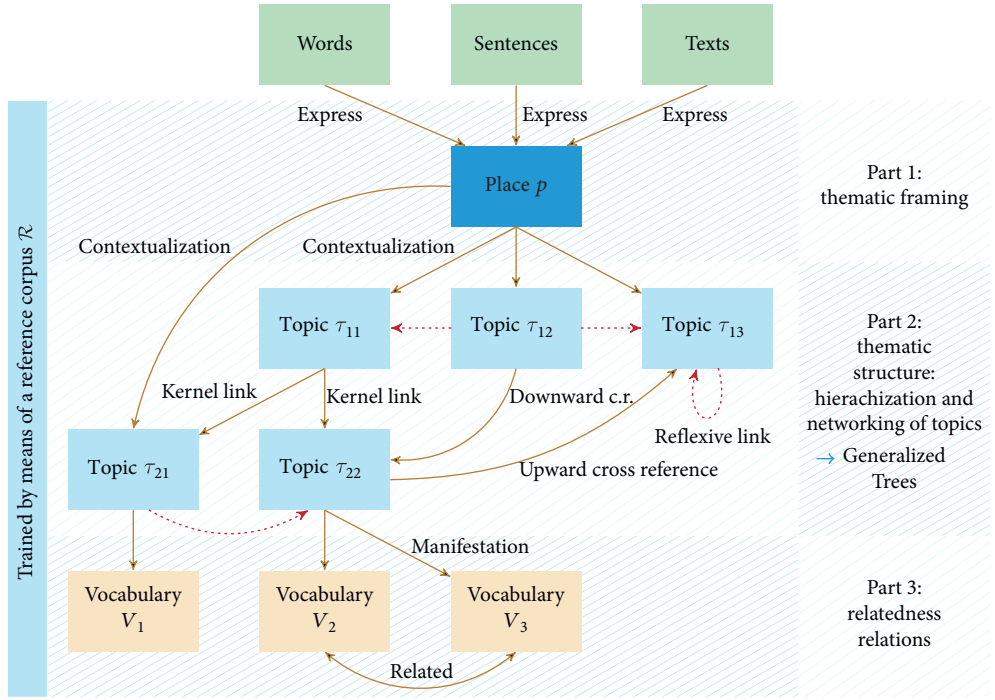


FIGURE 2: Schema of mapping texts onto hierarchically organized topic networks: words, sentences, and texts describing a certain thematic frame (e.g., a place as the central topic of a city wiki) are mapped onto a topic hierarchy as an example of a so-called generalized tree [56, 57]. Based on kernel links of thematic specialization, the topics are organized hierarchically, whereby this organization is superimposed by up- and downward cross-references. Dashed links are inferred as a result of modeling the thematic networking of input words, sentences, or texts. As we assume that the underlying topic model has been trained by means of a reference corpus $\mathcal{R}$ (see Definition 2), each topic is associated with a distribution of lexical elements of $\mathcal{R}$ that are preferably used to *manifest* this topic (see the types ν and w in relation to the topics α and β in Figure 1). This preference relation may be extended to higher-level units such as sentences.

one-layer topic network that is generated by the *generating layer* L_1.

Formulas (6) and (7) require that the weighting values for nodes and arcs are greater than 0: otherwise, the candidate vertices and arcs do not exist in the TTN. $\theta^{\leftarrow}$ is a classifier mapping pairs (t, x) of topics $t \in V_{\mathscr{C}}$ and texts x onto real numbers indicating the extent to which x is a "prototypical" instance of t (obviously, the textual arguments of the functions θ and $\theta^{\leftarrow}$ are not restricted to elements of X.)

Example 3. Given Example 2, we assume that $\lambda(\nu_1) = t_1$, $\lambda(\nu_2) = t_2, \lambda(\nu_3) = t_3$ and $\theta(x_1, t_1) = 1$, $\theta(x_2, t_2) = 1$, $\theta(x_3, t_3) = \theta(x_4, t_3) = 1$ so that $V = \{\nu_1, \nu_2, \nu_3\}$. In our example, we disregard $\theta^{\leftarrow}$. Further, we assume that the functions α, β, γ, and δ are identity functions. Thus, $\mu(\nu_1) = \mu(\nu_2) = 1$ and $\mu(\nu_3) = 2$. Now, we can generate a topic link between ν_1 and ν_2 by exploring the intertextual relation $(x_1, x_2) \in A_1$: To this end, we assume that

$$\delta\left[x \underset{\theta^{\leftarrow}}{\overset{\theta}{\longleftrightarrow}} \dot{\nu}, y \underset{\theta^{\leftarrow}}{\overset{\theta}{\longleftrightarrow}} \dot{w}, x \overset{\nu_1}{\longrightarrow} y\right] \leftarrow \delta\left[x \overset{\nu_1}{\longrightarrow} y\right] \leftarrow \mathrm{id}\left(x \overset{\nu_1}{\longrightarrow} y\right) = x \overset{\nu_1}{\longrightarrow} y, \tag{8}$$

so that $\nu((\nu_1, \nu_2)) = 1$. By analogy to this case, we link topic ν_3 by means of a reflexive link so that $A = \{(\nu_1, \nu_2), (\nu_3, \nu_3)\}$.

Note that these simplifications are made for simplicity's sake only: Section 3.2 will elaborate a realistic weighting scenario. However, the function of the latter illustration is to show that by the intertextual linkage of both texts, we get evidence about the linkage of the topics instantiated by these texts. TTNs always operate according to this premise: they network topics as a function of the networking of an underlying set of texts. Figure 3 gives a schematic depiction of this scenario, which is varied subsequently to illustrate the other types of topic networks developed in this paper.

A concrete example of a TTN that is derived from the articles of the so-called Dresden wiki (see Section 4.1) is depicted in Figure 4. It shows the highest weighted topics addressed by these articles and their (undirected) links. The TTN has been computed by means of the procedural model of Section 3.2. Evidently, the topic *Transportation; ground transportation* is most prominent in this wiki followed by the topic *Central Europe; Germany*. Most topics belong to the areas *transportation* (red), *geography and history* (turquoise), and *architecture* (gray) (for the color code, see Appendix). More examples of TTNs can be found in Figures 5–7.

Arguments of the sort $x \longrightarrow_{\theta} \dot{\nu}$ can be used to quantify evidence about text x as an instance of topic $\dot{\nu}$: the more the evidence of this sort, the higher possibly the impact of x in formula (6) and the higher possibly the final weight of ν. The adverb *possibly* refers to what is licensed by the parameters γ and δ. Arguments of the sort $x \longrightarrow_{\nu_1} y$, where $x \neq y$, can be

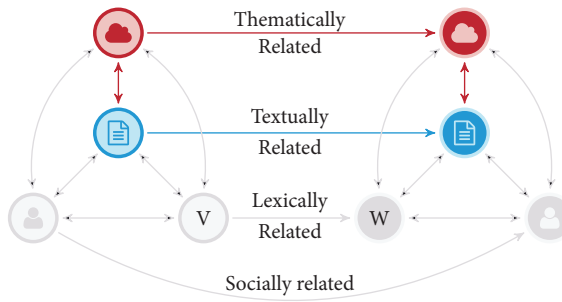


Figure 3: Schematic depiction of the informational sources of linking topics (red vertices) in text topic networks as a function of the textual relatedness of two texts (blue vertices) (belonging to layer L1 of a corresponding LMN—see Definition 1). Bidirectional red arcs denote arcs of the corresponding margin layers: in the present case, this concerns the relation between texts and topics (see below). Relations of thematic relatedness are inferred in this example (see Definition 3). Gray nodes and arcs indicate unconsidered sources of evidence.

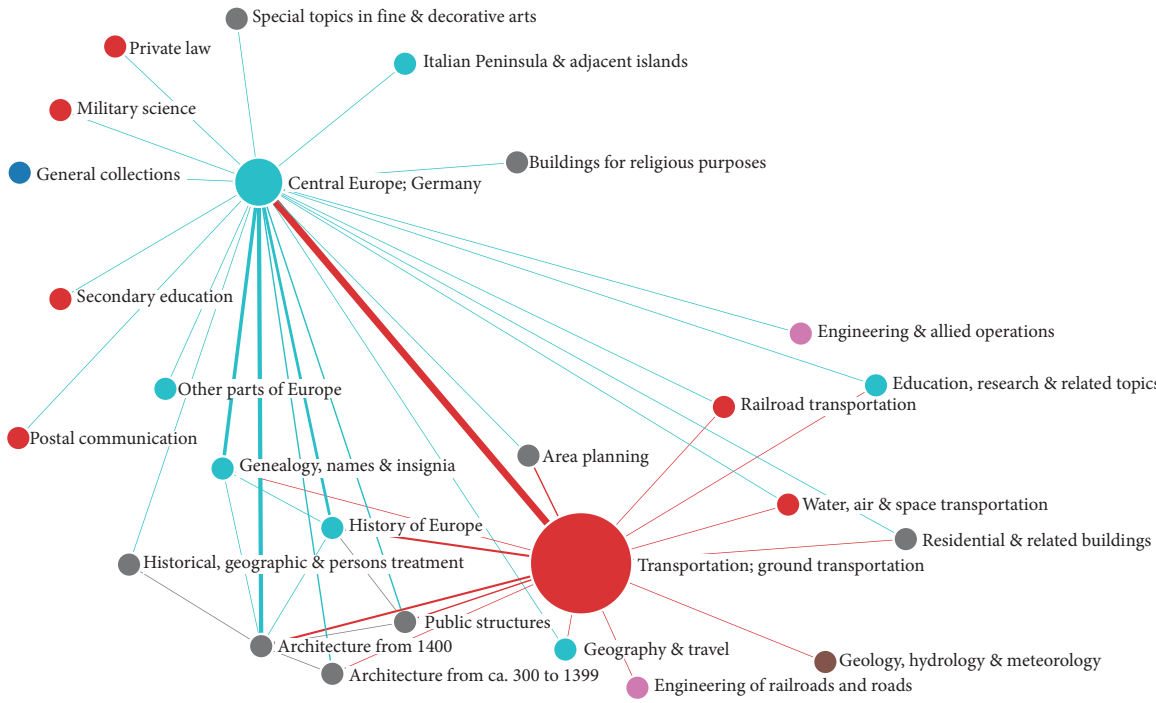


Figure 4: Visualization of a segment of the TTN of the city wiki Dresden (http://www.stadtwikidd.de/wiki/Hauptseite) using the 3rd level of the DDC as the underlying RCS for the definition of topics according to Section 3.2. The segment shows the highest weighted topics and their (undirected) links. Edges have been colored to show the two centers of this graph.

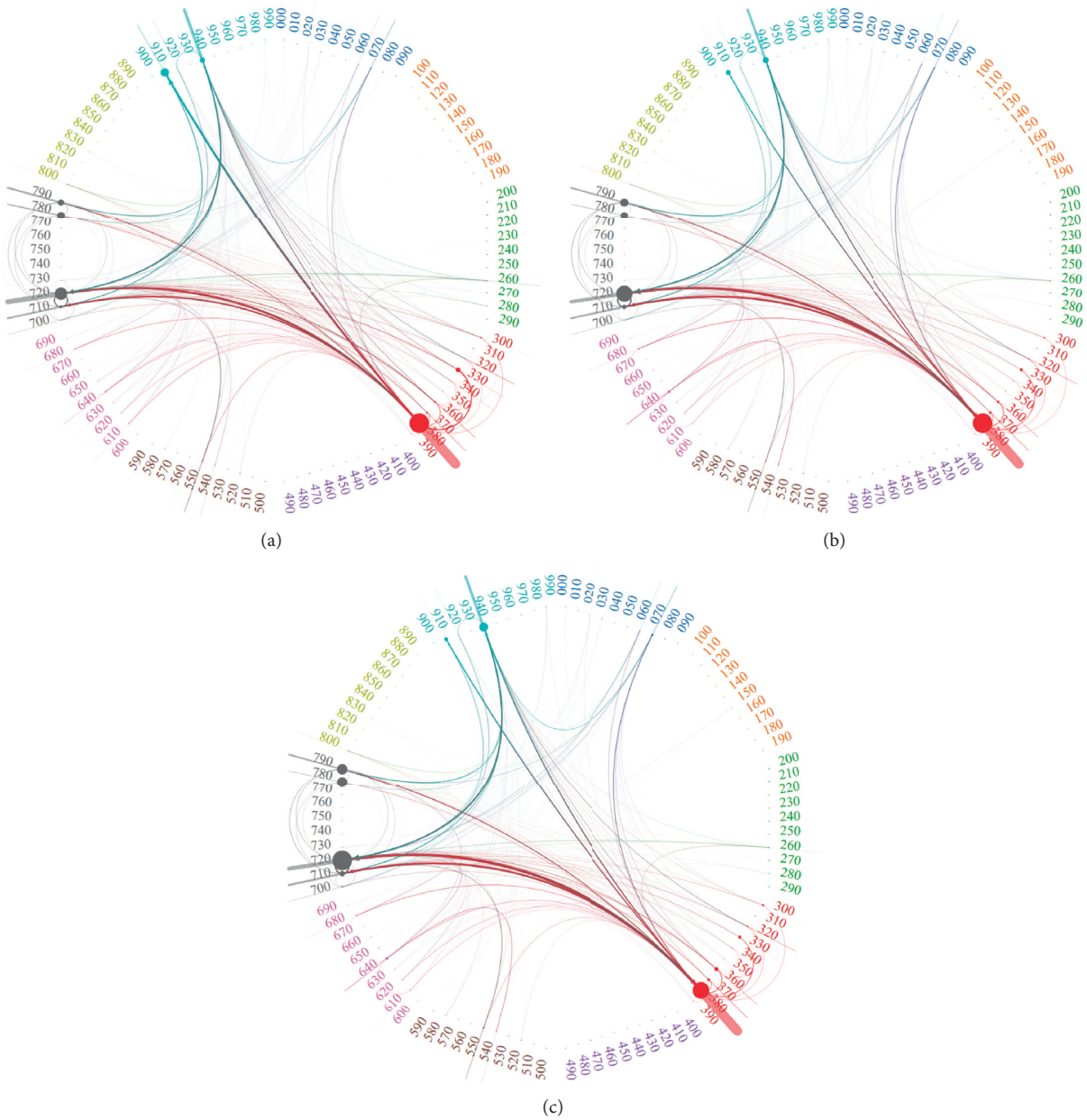
used to quantify evidence that text x is intertextually linked to text y: the more the evidence of this sort, the higher possibly the weight of the link from x to y and the higher possibly the influence of this link onto the weight of the link from topic v to topic w in formula (7) (in cases in which there is no explicit information about intertextual links, one can use functions of aggregated word embeddings of the lexical constituents of texts to calculate their intertextual similarity). In this and related definitions, we do not fully specify the functions $\theta, \theta^{\leftarrow}, \alpha, \beta, \gamma, \delta$ to leave enough space for different instances of topic networks.

Definition 3 relies on the pivotal text layer for deriving topic networks. To integrate further layers into the process of inferring topic networks, we introduce the following generalized schema.

Definition 4. Given a definitional setting $\mathcal{S} = (\mathscr{C}, \theta, \mathscr{L}(X, l))$ according to Definition 2, an $(L_1, \mathbb{L}')$-*Topic Network*, $\mathbb{L}' \in \{\varnothing\} \cup \{\{L_i\} \mid i \in \{2, \ldots, l\}\}$, is a vertex- and arc-weighted simple directed graph

$$T(L_1, \mathbb{L}') = (V, A, \mu, \nu, \lambda, \kappa), \tag{9}$$

which is said to be *derived from* $\mathcal{S}$ and *inferred from* L_1 and the elements of $\mathbb{L}'$ by means of the optional classifiers $\vartheta^{\leftarrow}, \vartheta : V_i \times V_{\mathscr{C}} \longrightarrow \mathbb{R}_0^+, \vartheta^{\leftarrow} : V_{\mathscr{C}} \times V_i \longrightarrow \mathbb{R}_0^+$ and monotonically increasing functions $\alpha, \beta, \gamma, \delta : \mathbb{R}_0^+ \longrightarrow \mathbb{R}_0^+$ iff $\forall v \in V$ and $\forall a = (v, w) \in A$:

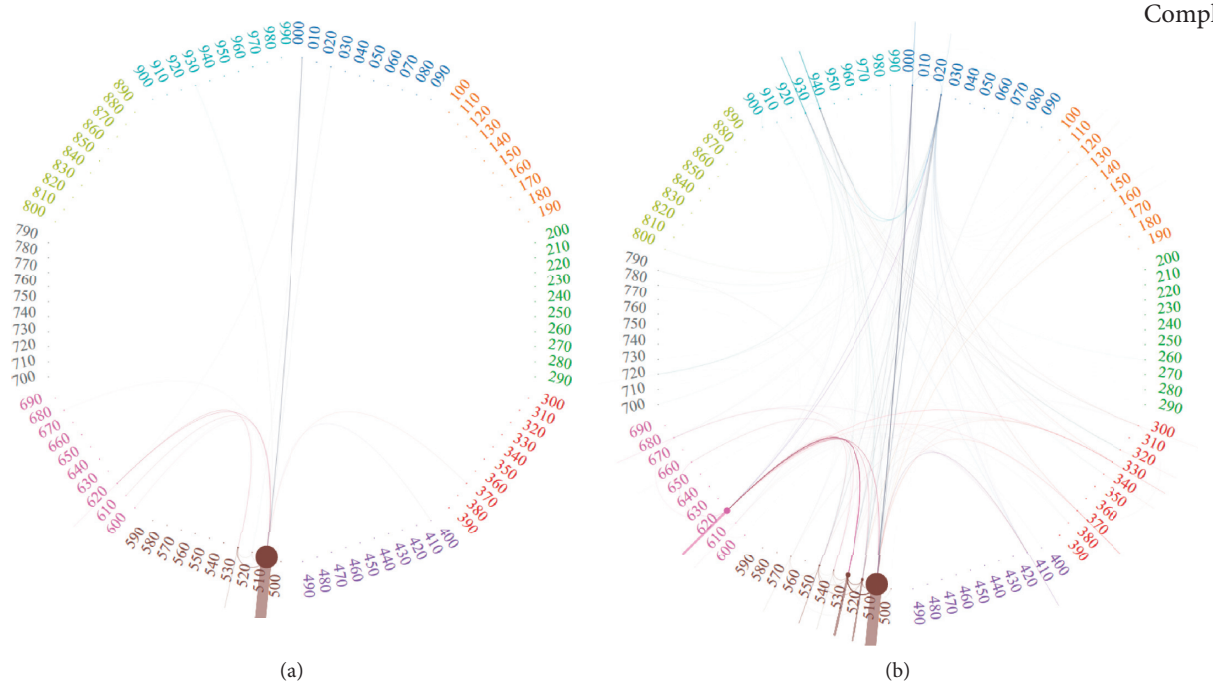


(a) (b) (c)

Figure 5: Visualizations of a TTN (top left) and two corresponding ATNs. The TNs are derived from the city wiki München (https://www.muenchenwiki.de/wiki/Hauptseite) (see Section 4 for statistics about this wiki) using the procedural model of Section 3.2. Top right shows the ATN for which (co)authorship activities are estimated by means of Wikipedia (see Section 3.2.3). The ATN for which these activities are estimated via the wiki itself is displayed below. The visualizations are carried out by means of PolyViz [73] regarding the 2nd level of the DDC: nodes are labeled (with numbers denoting the respective 2nd level class) and colored to encode their membership to one of the top 10 DDC classes (see Appendix). The higher the weight of a topic, the larger the node, and the higher the weight of an arc, the thicker the line. Node and line sizes are defined relative to the maximum vertex and arc weights of the underlying network.

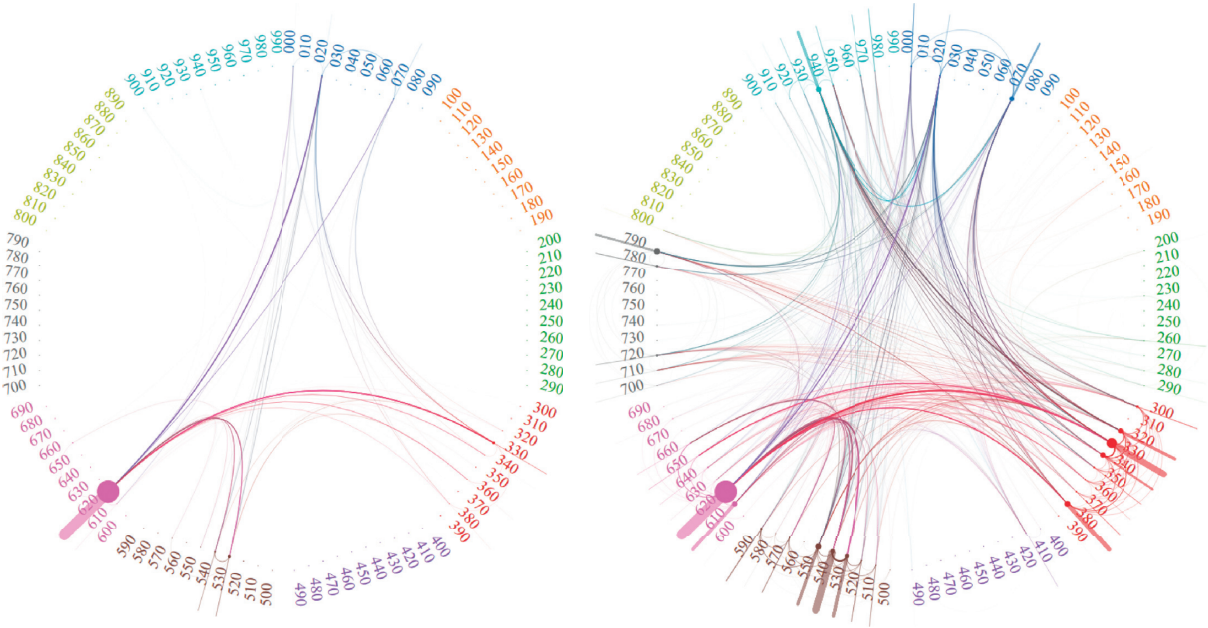
$$\mu(v) = \alpha\left(\sum_{x\in V_1, r\in V_i} \beta\left[x \underset{\theta^{\leftarrow}}{\overset{\theta}{\longleftrightarrow}} \dot{v}, r \underset{\vartheta^{\leftarrow}}{\overset{\vartheta}{\longleftrightarrow}} \dot{v}, r \underset{\nu_{1.i}}{\overset{\nu_{i.1}}{\longleftrightarrow}} x\right]\right) > 0. \tag{10}$$

$$\nu(a) = \gamma\left(\sum_{x,y\in V_1, r,s\in V_i} \delta\left[x \underset{\theta^{\leftarrow}}{\overset{\theta}{\longleftrightarrow}} \dot{v}, y \underset{\theta^{\leftarrow}}{\overset{\theta}{\longleftrightarrow}} \dot{w}, r \underset{\vartheta^{\leftarrow}}{\overset{\vartheta}{\longleftrightarrow}} \dot{v}, s \underset{\vartheta^{\leftarrow}}{\overset{\vartheta}{\longleftrightarrow}} \dot{w}, r \underset{\nu_{1.i}}{\overset{\nu_{i.1}}{\longleftrightarrow}} x, s \underset{\nu_{1.i}}{\overset{\nu_{i.1}}{\longleftrightarrow}} y, r \overset{\nu_i}{\longrightarrow} s, x \overset{\nu_1}{\longrightarrow} y\right]\right) > 0, \tag{11}$$



(a) (b)

Figure 6: Visualization (by means of PolyViz [73]) of the TTN of the 1st orbit (a) and of the 2nd orbit (b) of the German Wikipedia article *Integralrechnung* (*Integral*). The TTNs are derived from the corpora of articles in the 1st and the 2nd orbit (see formula (30)) of this article. Obviously, the most prominent 2nd level DDC class in both TTNs is 510 (*Mathematics*).



(a) (b)

Figure 7: Visualization (by means of PolyViz [73]) of the TTN of the 1st orbit (a) and of the 2nd orbit (b) of the German Wikipedia article *Kernkraftwerk* (*Nuclear power plant*). The TTNs are derived from the corpora of articles in the 1st and 2nd orbits (see formula (30)) of this article. Obviously, the most prominent 2nd level DDC class in both TTNs is 620 (*Engineering*). Compared to the example in Figure 6, the 2nd orbit is now thematically much more diversified.

where $\mathbb{L}' = \{L_i\}$. $\mu : V \longrightarrow \mathbb{R}^+$ is a vertex weighting function, $\nu : A \longrightarrow \mathbb{R}^+$ an arc weighting function, $\lambda : V \longrightarrow V_{\mathscr{C}}$ an injective vertex labeling function, $V_{\mathscr{C}}(V) = \{\lambda(v) \mid v \in V\} \subseteq V_{\mathscr{C}}$, and κ an injective arc labeling function. For $\mathbb{L}' = \{L_i\}$, we say that $T(L_1, \mathbb{L}')$ *is a two-level topic network* that is generated by the *generating layers* L_1 and L_i. If $\mathbb{L}' = \varnothing$, then formula (10) changes to formula (6) and formula (11) to formula (7). By omitting any optional classifier $g \in \{\theta^{\leftarrow}, \vartheta^{\leftarrow}\}$,

expressions of the sort $r_g \longleftrightarrow_f \dot{v}$ change to $r \longrightarrow_f \dot{v}$. ϑ is treated analogously.

To understand formula (10) look at Figure 8: among other things, formula (10) collects the triangle spanned by v, x, and a supposed that the two-level topic network is based on text and authorship links. Obviously, Definition 4 generalizes Definition 3. Now, it should be clear why we speak of the text network of an LMN as its pivotal level: it is the reference layer of any additional layer that is integrated into a two-level topic network according to Definition 4. This role is maintained below when we generalize this definition to capture n layers, $n > 2$. With the help of Definition 4, we can immediately derive so-called *author topic networks*.

Definition 5. An *Author Topic Network* (ATN) is a directed graph

$$T(L_1, \mathbb{L}') = (V, A, \mu, \nu, \lambda, \kappa), \tag{12}$$

according to Definition 4 such that $\mathbb{L}' = \{L_2\}$.

The relational arguments of this definition can be motivated as follows—assuming that they are instantiated appropriately:

(1) $x \xrightarrow{\theta} \dot{v}$ can be used to represent evidence that text x is about topic $\dot{v}$ possibly in relation to other topics of $V_{\mathscr{C}}$.

(2) $\dot{v} \xrightarrow{\theta^{\leftarrow}} x$ can be used to represent evidence that text x is a prototypical instance of topic $\dot{v}$ possibly in relation to other texts in V_1.

(3) $r \xrightarrow{\vartheta} \dot{v}$ can be used to represent the extent to which agent r tends to write about topic $\dot{v}$ possibly in relation to other topics of $V_{\mathscr{C}}$.

(4) $\dot{v} \xrightarrow{\vartheta^{\leftarrow}} r$ represents evidence that agent r is a prototypical author writing about topic $\dot{v}$ possibly in relation to other agents in V_2.

(5) For $x \neq y$, $x \xrightarrow{\nu_1} y$ can be calculated to represent evidence about text x to be intertextually linked to text y (e.g., in the sense of linking contributions of different authors). Otherwise, if $x = y$, $x \longrightarrow_{\nu_1} y$ can be used to quantify evidence about x being intratextually structured.

(6) $r \xrightarrow{\nu_{2.1}} x$ can be used to quantify evidence about the role of agent r as an author of text x possibly in relation to other texts authored by r. Typically, $\nu_{2.1}$ is a function of the number of edit actions performed by r on x [74].

(7) $x \xrightarrow{\nu_{1.2}} r$ can be used to quantify evidence about the role of agent r as a prototypical author of text x possibly in relation to other authors of x. In the simplest case, $\nu_{2.1}$ is symmetric making $\nu_{1.2}$ obsolete.

(8) $r \xrightarrow{\nu_2} s$ represents evidence that agent r is a coauthor of or interacting with s. For instantiating ν_2, the literature knows a wide range of alternatives [74, 75] (which mostly concern symmetric measures of coauthorship). Note that we do not require that $r \neq s$.

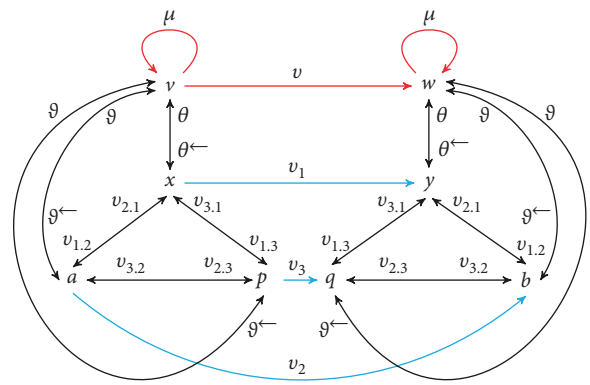


FIGURE 8: A diagrammatic depiction of inferred arcs (red) in topic networks, inferred by means of various arcs (black and blue) of an underlying LMN. Orientation of inferred arcs is provided by three types of input arcs (blue). $x, y \in V_1$ denotes two texts, $a, b \in V_2$ denotes two authors working on x, and y, respectively, $p, q \in V_3$ denotes two lexical units occurring in x and y, respectively. Inferred weights of vertices are denoted by means of (red) reflexive arcs.

Example 4. Starting from Example 3 to exemplify arcs between topics in *author topic networks*, we can now additionally explore the evidence, that text x_1 and x_2 are both coauthored by the agents a_1 and a_2. That is, we can assume a coauthorship link $(a_1, a_2) \in A_2$ (A_2 is the arc set of the author layer in Definition 1) of weight $\nu(a_1, a_2) = 1$. Let us now assume the following simplification of the function δ in Definition 4, for which we assume that it simply multiplies and adds up its argument values in the following way:

$$\begin{aligned}
&\delta\left[x \underset{\theta^{\leftarrow}}{\overset{\theta}{\longleftrightarrow}} \dot{v}, y \underset{\theta^{\leftarrow}}{\overset{\theta}{\longleftrightarrow}} \dot{w}, r \underset{\vartheta^{\leftarrow}}{\overset{\vartheta}{\longleftrightarrow}} \dot{v}, s \underset{\vartheta^{\leftarrow}}{\overset{\vartheta}{\longleftrightarrow}} \dot{w}, r \underset{\nu_{1.2}}{\overset{\nu_{2.1}}{\longleftrightarrow}} x, s \underset{\nu_{1.2}}{\overset{\nu_{2.1}}{\longleftrightarrow}} y, r \xrightarrow{\nu_2} s, x \xrightarrow{\nu_1} y\right] \longleftarrow \\
&\qquad \delta\left[x \xrightarrow{\theta} \dot{v}, y \xrightarrow{\theta} \dot{w}, r \xrightarrow{\nu_{2.1}} x, s \xrightarrow{\nu_{2.1}} y, r \xrightarrow{\nu_2} s, x \xrightarrow{\nu_1} y\right] \longleftarrow \\
&(x \xrightarrow{\theta} \dot{v}) \cdot (y \xrightarrow{\theta} \dot{w}) \cdot \left(r \xrightarrow{\nu_{2.1}} x\right) \cdot \left(s \xrightarrow{\nu_{2.1}} y\right) \cdot \left(r \xrightarrow{\nu_2} s + x \xrightarrow{\nu_1} y\right) = (1 \cdot 1 \cdot 1 \cdot 1)(1 + 1) \\
&\qquad = 2.
\end{aligned} \tag{13}$$

In our example, we get $\dot{v} = t_1 = \lambda(v_1)$, $\dot{w} = t_2 = \lambda(v_2)$, $x = x_1$, $y = x_2$, $r = a_1$, and $s = a_2$. Since there is no other interlinked pair of texts (see Example 1), instantiating the topics v_1 and v_2, we get $\nu((v_1, v_2)) = 2$ as the weight of this topic link in the corresponding ATN. By this simplified example of an ATN, we get the information that the link of topic v_1 to topic v_2 is additionally supported by the coauthorship of agents a_1 and a_2: this information extends the evidence about the topic link as provided by the underlying TTN of Example 3. Likewise, the reflexive link of topic v_3 is augmented by 1 compared to the underlying TTN, while there is no other topic link to be considered in this example of an ATN. By analogy to Figure 3, Figure 9 gives a schematic depiction of this scenario. Note that in our example, the weight of the link between authors a_1 and a_2 (cf. $r \xrightarrow{\nu_2} s$) is a function of their coauthorship: this is only one alternative to weight the social relatedness of both agents, actually one that can be measured by exploring (special) wikis. However, any other social relatedness might be explored to weight the interaction of agents.

By comparing a text topic network $T(L_1) = (V_{l+1}, A_{l+1}, \mu_{l+1}, \nu_{l+1}, \lambda_{l+1}, \kappa_{l+1})$ with an author topic network $T(L_1, \{L_2\}) = (V_{l+2}, A_{l+2}, \mu_{l+2}, \nu_{l+2}, \lambda_{l+2}, \kappa_{l+2})$ derived from the same LMN $\mathscr{L}(X, l)$, we can learn how the topics of $V_{\mathscr{C}}$ are manifested in the texts of corpus X in the form of a concomitance or a disparity of intertextual and coauthorship-based networking. Consider, for example, two vertices $v \in V_{l+1}$ and $w \in V_{l+2}$ such that $\dot{v} = \dot{w}$; let further $\perp$ and $\top$ denote the minimum and maximum that the vertex weighting functions of both graphs can assume. Then, we can distinguish four extremal cases:

(1) Cases of the sort

$$\perp \ll \mu_{l+1}(v) \approx \mu_{l+2}(w) \approx \top, \tag{14}$$

provide information on prominent topics that tend to be addressed by many texts which are coauthored by many authors.

(2) Situations like

$$\top \gg \mu_{l+1}(v) \approx \mu_{l+2}(w) \approx \perp, \tag{15}$$

probably apply to the majority of the topics in $V_{\mathscr{C}}$, which are hardly or even not at all addressed by texts in $V_1 = X$ due to the narrow thematic focus of these texts.

(3) Cases like

$$\top \approx \mu_{l+1}(v) \gg \mu_{l+2}(w) \approx \perp, \tag{16}$$

suggests a Zipfian topic effect, according to which a prominent topic is addressed by a small group of agents or even by a single author.

(4) Finally, situations of the sort

$$\perp \approx \mu_{l+1}(v) \ll \mu_{l+2}(w) \approx \top, \tag{17}$$

refer to rarely manifested topics addressed by a few but highly coauthored texts. In conjunction with many cases of the sort described by formula (16), situations of this kind indicate a Zipfian coauthoring effect, according to which many authors write only a few texts, while many texts are written by a few authors without encountering many (relevant) coauthors.

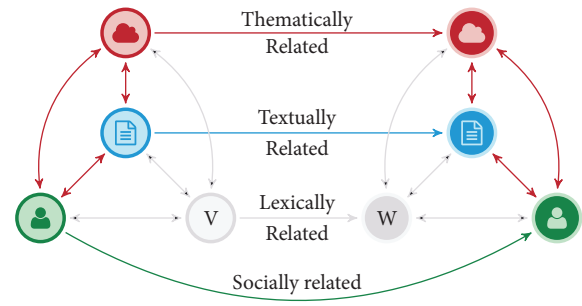


Figure 9: Schematic depiction of the informational sources of linking topics (red vertices) in *author topic networks* as a function of the textual relatedness of two texts (blue vertices) (that belong to layer L_1 of a corresponding LMN—see Definition 1) and the social relatedness of corresponding authors (green vertices) (that belong to layer L_2 of a corresponding LMN). Bidirectional red arcs denote arcs of the corresponding margin layers in Definition 1.

Formulas (14)–(17) compare the node weighting functions of a TTN with those of a related ATN. The same can be done regarding their arc weighting functions. That is, for two arcs $a = (r, s) \in A_{l+1}$ and $b = (v, w) \in A_{l+2}$, for which $\dot{r} = \dot{v} \wedge \dot{s} = \dot{w}$, we distinguish again four cases ($\perp$ and $\top$ now denote the minimum and maximum the arc weighting functions of both graphs can assume):

(1) In the case of

$$\perp \ll \nu_{l+1}(a) \approx \nu_{l+2}(b) \approx \top, \tag{18}$$

topic $\dot{v}$ is intertextually linked more strongly to topic $\dot{w}$ and authors of its text instances tend to cooperate with those of instances of topic $\dot{w}$ likewise to a greater extent.

(2) In the case of

$$\top \gg \nu_{l+1}(a) \approx \nu_{l+2}(b) \approx \perp, \tag{19}$$

topic $\dot{v}$ is intertextually less strongly linked to topic $\dot{w}$ and the few authors of its textual instances tend to cooperate with authors of instances of topic $\dot{w}$ likewise to a lesser extent.

(3) In the case of

$$\top \approx \nu_{l+1}(a) \gg \nu_{l+2}(b) \approx \perp, \tag{20}$$

topic $\dot{v}$ is intertextually more strongly connected with topic $\dot{w}$, while authors of its text instances tend to cooperate with those of instances of topic $\dot{w}$ to a lesser extent, if at all.

(4) Finally, in the case of

$$\perp \approx \nu_{l+1}(a) \ll \nu_{l+2}(b) \approx \top, \tag{21}$$

topic $\dot{v}$ is intertextually less strongly linked to topic $\dot{w}$, while the numerous authors of its text instances tend to cooperate with those of instances of topic $\dot{w}$ to a much greater extent.

Our central question regarding the relationship between TTNs and ATNs *derived from the same LMN* is whether these networks are similar or not. If they are similar, we expect that cases of the sort described by formulas (14), (15), (18), and (19) predominate so that cases matched by formula (14) are parallelized by those considered by formula (18) and where cases according to formula (15) are concurrent to those described by formula (19). An opposite situation would be that two topic nodes in the TTN are highly weighted but weakly linked, while they are weakly weighted but strongly linked in the corresponding ATN. In this case, a few or even only a single author is responsible for the thematic focus of the TTN. Note that this scenario reminds again of a Zipfian effect regarding the relation of TTNs and ATNs. By characterizing TTNs in relation to ATNs along these and related scenarios, we want to investigate laws of the interdependence of both types of networks, which may consist, for example, in the simultaneity of dense or sparse intertextuality-based networking on the one hand and dense or sparse coauthorship-based networking on the other. We may expect, for example, that the more related the two topics, the more likely the authors of their textual instances cooperate. However, not so much is known about such scenarios in the area of VGI especially with regard to Hypothesis 1. Thus, we address this gap at least by introducing a novel theoretical model which may help filling it.

Figure 5 exemplifies two ATNs in relation to a corresponding TTN (T1) which were computed using the apparatus of Section 3.2 to instantiate the formal model of this section. The upper right ATN (A1) is computed by globally weighting coauthorship activities based on Wikipedia (as explained in Section 3.2.3); the ATN (A2) below is calculated by weighting of these activities relative to the city wiki itself. Figure 5 shows that the topic with DDC number 720 (*Architecture*) is weighted higher in A1 than in T1. This is all the more pronounced in A2, where 720 becomes the most prominent topic and consequently displaces the top subject from T1, that is, topic 380 (*Commerce, communications* & *transportation*). That is, although topic 380 is most frequently addressed in this wiki's texts, topic 720 not only is almost as salient but also attracts many more activities among its interacting coauthors. Similar observations concern the switch of the roles of the topics 910 (*Geography* & *travel*) and 940 (*History of Europe*) from T1 to A1 and A2.

Regardless of the answer to this and related questions, we will also ask whether the shape of an ATN can be predicted if one knows the shape of the corresponding TTN and vice versa. To answer this question, we will consider LMNs of different text genres: of city wikis and regional wikis on the one hand and extracts of encyclopedic wikis on the other. We expect that LMNs spanned over corpora of the same genre exhibit a pattern of collaboration- and intertextuality-based networking that makes TTNs and ATNs derived from them mutually recognizable or predictable, whereas for LMNs generated from corpora of different genres this does not apply.

For reasons of formal variety, we now consider an alternative to author topic networks, namely, so-called word topic networks, which in turn are derived from Definition 4.

Definition 6. A *Word Topic Network* (WTN) is a directed graph

$$T\left(L_1, \mathbb{L}'\right) = (V, A, \mu, \nu, \lambda, \kappa), \tag{22}$$

according to Definition 4 such that $L' = \{L_3\}$.

This definition departs by five new relational arguments from Definition 5, which—if being instantiated appropriately—can be motivated as follows:

(1) $a \xrightarrow{\nu_{3.1}} x$ quantifies evidence about the role of word a as a lexical constituent of text x possibly in relation to all other texts in which a occurs. Typically, $\nu_{3.1}$ is implemented by a global term weighting function [76] or by a neural network-based feature selection function.

(2) $x \xrightarrow{\nu_{1.3}} a$ quantifies evidence about the role of the word a as a lexical constituent of the text x possibly in relation to other lexical constituents of x. Typically, $\nu_{1.3}$ is a local term weighting function, such as normalized term frequency [76], or a topic model-based function.

(3) $a \xrightarrow{\vartheta} \dot{v}$ represents evidence about the word a to be associated with the topic $\dot{v}$ possibly in relation to all other topics of $V_{\mathscr{C}}$.

(4) $\dot{v} \xleftarrow{\vartheta^{-}} a$ calculates evidence about the extent to which the topic $\dot{v}$ is prototypically labeled by the word a, possibly in relation to all other words in V_3.

(5) $a \xrightarrow{\nu_3} b$ quantifies evidence about the extent to which the word a associates the word b. Typically, ν_3 is computed by means of word embeddings [77].

Based on this list, we better understand what topic networks offer in contrast to TMs. This concerns the flexibility with which we can include informational resources computed by different methods (e.g., based on neural networks, topic models, and LSA) to generate topic networks (cf. challenge P5). Different relational arguments $X \longrightarrow_Z Y$ can be quantified using different methods, which in turn can belong to a wide range of computational paradigms. Table 2 gives an account of the generality of our approach by hinting at candidate procedures for computing the different relations of Figure 8.

Example 5. Starting from Example 3 to exemplify arcs between topics in *word topic networks*, we have to additionally explore evidence regarding the lexical relatedness of the vocabularies of the texts x_1 and x_2. In Example 1, we assumed that the intersection of both texts (represented as bags-of-words) is given by the set $\{w_1, w_2\}$. By analogy to

Example 4, we assume now the following simplification of the function δ of Definition 4:

$$\delta\Big[x \underset{\theta^{\leftarrow}}{\overset{\theta}{\longleftrightarrow}} \dot{v}, y \underset{\theta^{\leftarrow}}{\overset{\theta}{\longleftrightarrow}} \dot{w}, r \underset{\vartheta^{\leftarrow}}{\overset{\vartheta}{\longleftrightarrow}} \dot{v}, s \underset{\vartheta^{\leftarrow}}{\overset{\vartheta}{\longleftrightarrow}} \dot{w}, r \underset{\nu_{1.2}}{\overset{\nu_{2.1}}{\longleftrightarrow}} x, s \underset{\nu_{1.2}}{\overset{\nu_{2.1}}{\longleftrightarrow}} y, r \xrightarrow{\nu_2} s, x \xrightarrow{\nu_1} y\Big] \longleftarrow (x \xrightarrow{\theta} \dot{v}) \cdot (y \xrightarrow{\theta} \dot{w}) \cdot \left(r \xrightarrow{\nu_{2.1}} x\right) \cdot \left(s \xrightarrow{\nu_{2.1}} y\right) \cdot \left(r \xrightarrow{\nu_2} s + x \xrightarrow{\nu_1} y\right). \tag{23}$$

In this scenario, we have to instantiate Definition 4 as follows: $\dot{v} \longleftarrow t_1 = \lambda(v_1)$, $\dot{w} \longleftarrow t_2 = \lambda(v_2)$, $x = x_1$, $y = x_2$, $r = w_1$, and $s = w_1$ for one summand and—everything else being constant—$r = w_2$ and $s = w_2$ for a second summand (for w_3 (w_4), we do not assume a lexical relatedness w.r.t. the words of text w_4 (w_3)). Note that under this regime, we assume that *relatedness of lexical constituents* only concerns shared usages of identical words—of course, this is a simplifying example. By analogy to the setting of Example 4, we have thus to conclude that $\nu((v_1, v_2)) = 4$ as the weight of the topic link from v_1 to v_2 in the corresponding WTN. For texts x_3 and x_4, we may alternatively assume that lexical relatedness does not only concern shared lexical items but also relatedness that is measured, for example, by means of a terminological ontology [83] or by means of word embeddings [77]. In this way, we may additionally arrive at a topic link between v_2 and v_3. In order to allow for a comparison of a WTN with its corresponding TTN, a more realistic weighting scheme is needed that also reflects above and below average lexical relatednesses of the lexical constituents of interlinked texts—in Section 3.2, we elaborate such a model regarding ATNs in relation to TTNs. Figure 10 gives a schematic depiction of the scenario of WTNs as elaborated so far.

It is worth emphasizing that instead of the (language-systematic) lexicon layer L_3, we may use a constituent layer $L_k, k > 3$, to infer a two-level topic network. For example, we can use the layer spanned by the sentences of the pivotal texts to obtain a sort of *sentence topic network*. In this case, $a \longrightarrow_{\nu_k} b$ may quantify evidence about the extent to which the sentence a entails the sentence b or the extent to which the sentence a is similar to the sentence b, etc., while $x \longrightarrow_{\nu_{1.k}} a$ may quantify evidence about the extent to which the sentence a is thematically central for the text x, etc. In sentence topic networks, topic linkage is a function of sentence linkage: prominent topics emerge from being addressed by many sentences, while prominent topic links arise from the relatedness of many underlying sentences. Another example of inferring two-level topic networks is to link topics as a function of places mentioned (by means of toponyms) within the texts of the underlying corpus X, where geospatial relations of these places can be explored to infer concurrent topic relations: if place p is mentioned in text x about topic $\dot{v}$ and place q in text y about topic $\dot{w}$, where the platial relation $R(p, q)$ relates p and q, this information can be used to link the topic nodes v and w in the corresponding topic network. As a result, we obtain networks manifesting the networking of topics as a function of parallelized geographical relations.

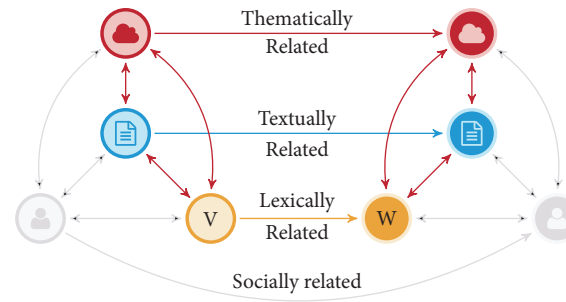


Figure 10: Schematic depiction of the informational sources of linking topics (red vertices) in *word topic networks* as a function of the textual relatedness of two texts (blue vertices) (that belong to layer L_1 of a corresponding LMN—see Definition 1) and the lexical relatedness of corresponding words (orange vertices) (that belong to layer L_3 of a corresponding LMN). Bidirectional red arcs denote arcs of the corresponding margin layers in Definition 1.

Obviously, any other relationship (e.g., entailment among sentences, sentiment polarities shared by linked texts, and co-reference relations) can be investigated to induce such two-level networks. And even more, we can think of n-level networks in which several such relationships are explored at once to generate topic links. We can ask, for example, which locations are linked by which geospatial relations while being addressed in which sentences about which topics where these sentences are related by which sentiment relations. Another example is to ask which authors prefer to write about which topics while tending to use which vocabulary: the higher the number of authors who use the same words more often to write about the same topic, and the higher the number of such words, the higher the weight of that topic. In this case, topic weighting is a function of frequently observed pairs of linguistic (here: lexical) means *and* authors. On the other hand, the higher the degree of coauthorship of two authors contributing to different topics and the higher the degree of association of the words used by these authors to write about these topics, the higher the weight of the link between the topics. This concept of a topic network induced by the text, the coauthorship, and the lexicon layer of an LMN is addressed by the following generalization, which provides a generation scheme for topic networks:

Definition 7. Given a definitional setting $\mathcal{S} = (\mathscr{C}, \theta, \mathscr{L}(X, l))$ according to Definition 2, an $(L_1, \mathbb{L}')$-*Topic Network*, for which

$$\mathbb{L}' = \{L_{i_1}, \ldots, L_{i_n}\} \in 2^{\{L_2, \ldots, L_l\}}, \tag{24}$$

is a vertex- and arc-weighted simple directed graph

$$T(L_1, \mathbb{L}') = (V, A, \mu, \nu, \lambda, \kappa), \tag{25}$$

which is said to be *derived from* $\mathcal{S}$ and *inferred from* L_1 and the elements of $\mathbb{L}'$ by means of the optional classifiers $\theta^{\leftarrow}, \forall i_j \in \{i_1, \ldots, i_n\} : \vartheta_{i_j} : V_{i_j} \times V_{\mathscr{C}} \longrightarrow \mathbb{R}_0^+, \vartheta_{i_j}^{\leftarrow} : V_{\mathscr{C}} \times V_{i_j} \longrightarrow \mathbb{R}_0^+$ and monotonically increasing functions $\alpha, \beta, \gamma, \delta : \mathbb{R}_0^+ \longrightarrow \mathbb{R}_0^+$ iff $\forall v \in V$ and $\forall a = (v, w) \in A$:

$$\mu(v) = \alpha\left(\sum_{\substack{x\in V_1,\\ r_{i_1}\in V_{i_1},\ldots r_{i_n}\in V_{i_n}}} \beta\left[x \overset{\theta}{\underset{\theta^{\leftarrow}}{\longleftrightarrow}} \dot{v}, r_{i_1} \overset{\theta_{i_1}}{\underset{\theta^{\leftarrow}_{i_1}}{\longleftrightarrow}} \dot{v}, \ldots, r_{i_n} \overset{\theta_{i_n}}{\underset{\theta^{\leftarrow}_{i_n}}{\longleftrightarrow}} \dot{v}, r_{i_1} \overset{\nu_{i_1.1}}{\underset{\nu_{1.i_1}}{\longleftrightarrow}} x, \ldots, r_{i_n} \overset{\nu_{i_n.1}}{\underset{\nu_{1.i_n}}{\longleftrightarrow}} x\right]\right) > 0, \tag{26}$$

$$\begin{aligned}\nu(a) = \gamma\Bigg(\sum_{\substack{x,y\in V_1,\\ r_{i_1}\in V_{i_1},\ldots,r_{i_n}\in V_{i_n},\\ s_{i_1}\in V_{i_1},\ldots,s_{i_n}\in V_{i_n}}} \delta\Bigg[& x \overset{\theta}{\underset{\theta^{\leftarrow}}{\longleftrightarrow}} \dot{v},\ y \overset{\theta}{\underset{\theta^{\leftarrow}}{\longleftrightarrow}} \dot{w}, \\ & r_{i_1} \overset{\vartheta_{i_1}}{\underset{\vartheta^{\leftarrow}_{i_1}}{\longleftrightarrow}} \dot{v}, \ldots, r_{i_n} \overset{\vartheta_{i_n}}{\underset{\vartheta^{\leftarrow}_{i_n}}{\longleftrightarrow}} \dot{v}, s_{i_1} \overset{\vartheta_{i_1}}{\underset{\vartheta^{\leftarrow}_{i_1}}{\longleftrightarrow}} \dot{w}, \ldots, s_{i_n} \overset{\vartheta_{i_n}}{\underset{\vartheta^{\leftarrow}_{i_n}}{\longleftrightarrow}} \dot{w}, \\ & r_{i_1} \overset{\nu_{i_1.1}}{\underset{\nu_{1.i_1}}{\longleftrightarrow}} x, \ldots, r_{i_n} \overset{\nu_{i_n.1}}{\underset{\nu_{1.i_n}}{\longleftrightarrow}} x, s_{i_1} \overset{\nu_{i_1.1}}{\underset{\nu_{1.i_1}}{\longleftrightarrow}} y, \ldots, s_{i_n} \overset{\nu_{i_n.1}}{\underset{\nu_{1.i_n}}{\longleftrightarrow}} y, \\ & r_{i_1} \overset{\nu_{i_1}}{\longrightarrow} s_{i_1}, \ldots, r_{i_n} \overset{\nu_{i_n}}{\longrightarrow} s_{i_n}, \\ & r_{i_1} \overset{\nu_{i_1.i_2}}{\underset{\nu_{i_2.i_1}}{\longleftrightarrow}} s_{i_2}, \ldots, r_{i_1} \overset{\nu_{i_1.i_n}}{\underset{\nu_{i_n.i_1}}{\longleftrightarrow}} s_{i_n}, \ldots, r_{i_n} \overset{\nu_{i_n.i_1}}{\underset{\nu_{i_1.i_n}}{\longleftrightarrow}} s_{i_1}, \ldots, r_{i_n} \overset{\nu_{i_n.i_{n-1}}}{\underset{\nu_{i_{n-1}.i_n}}{\longleftrightarrow}} s_{i_{n-1}}, \\ & x \overset{\nu_1}{\longrightarrow} y\Bigg]\Bigg) > 0,\end{aligned} \tag{27}$$

$\mu : V \longrightarrow \mathbb{R}^+$ is a vertex weighting function, $\nu : A \longrightarrow \mathbb{R}^+$ an arc weighting function, $\lambda : V \longrightarrow V_{\mathscr{C}}$ an injective vertex labeling function, $V_{\mathscr{C}}(V) = \{\lambda(v) \mid v \in V\} \subseteq V_{\mathscr{C}}$, and κ an injective arc labeling function. For $|\mathbb{L}'| = n$, we say that $T(L_1, \mathbb{L}')$ *is an m-level,* $m = n + 1$, *topic network* generated by the *generating layers* L_1 and the elements of $\mathbb{L}'$. If $\mathbb{L}' = \varnothing$, formula (26) changes to formula (6) and formula (27) to formula (7). By omitting the optional classifier $g \in \left\{\vartheta^{\leftarrow}_{i_j} \mid j \in \{1, \ldots, n\}\right\}$, expressions of the sort $r_g \longleftrightarrow_f \dot{v}$ change to $r \longrightarrow_f \dot{v}$. θ and ϑ_{i_j} are treated analogously. In order to derive an *undirected* m-level topic network $\overline{T}(L_1, \mathbb{L}') = (V, E, \mu, \overline{\nu}, \lambda, \overline{\kappa})$ from $T(L_1, \mathbb{L}')$, we define $\{v, w\} \in E \longleftrightarrow (v, w) \in A \vee (w, v) \in A$ and

$$\overline{\nu}(\{v, w\}) = \begin{cases} \zeta_1(\nu((v, w)), \nu((w, v))), & (v, w) \in A \wedge (w, v) \in A, \\ \zeta_2(\nu((v, w))), & (v, w) \in A \wedge (w, v) \notin A, \end{cases} \tag{28}$$

and where ζ_1 and ζ_2 are monotonically increasing functions.

Evidently, Definition 7 is a generalization of Definition 3 by considering higher numbers of generating layers. A schematic depiction of the scenario addressed by this definition is shown in Figure 11 by example of a 3-level topic network that explores evidence about topic linking starting from the text, the author, and the lexicon layer of Definition 1. Likewise, Figure 12 depicts an n-level topic network, $n > 3$, in which additional resources are explored beyond the word, author, and text level. Figure 8 illustrates more formally the inference process underlying Definition 7, and in particular of the arguments used. It illustrates the inference of an arc that connects two topics by exploring the links of the text, author, and lexicon layers of an underlying LMN. In this example, the blue and black arcs are evaluated to determine the weights of red arcs connecting the focal topic nodes. Blue arcs are used to orientate inferred arcs. We will not develop this apparatus further, nor will we empirically examine $n + 1$-layer topic networks for $n > 2$. Rather, the apparatus developed so far serves to demonstrate the generality, flexibility, and extensibility of our formal model.

In the above, we explained that one of the reasons for introducing a flexible and extensible formalism of topic networks is to compare topic networks derived from different layers (e.g., from the text layer on the one hand and the author layer on the other). In order to systematize this approach, we finally introduce the concept of a *multiplex topic network*, which is derived from the same or from different linguistic multilayer networks:

Definition 8. Given a definitional setting $\mathcal{S} = (\mathscr{C}, \theta, \mathscr{L}(X, l))$ according to Definition 2, a *Multiplex Topic Network* (MTN) is a k-layer network

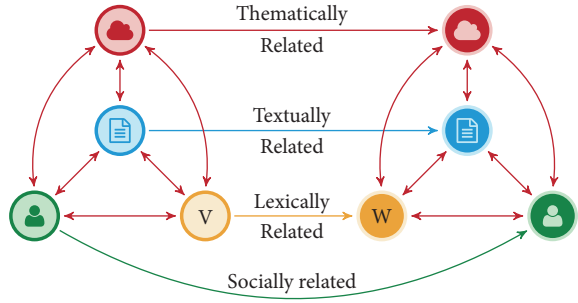


Figure 11: Schematic depiction of informational sources explored to link topics (red vertices) in a *3-level topic network* as a function of the textual relatedness of texts (blue vertices) (belonging to layer L_1 of Definition 1), the social relatedness of corresponding authors (green vertices) (belonging to layer L_2 of Definition 1), and the lexical relatedness of corresponding words (orange vertices) (belonging to layer L_3 of Definition 1). In this scenario, thematic relatedness is the information to be inferred, while textual, lexical, and social relations concern given information or evidence. Bidirectional red arcs denote arcs of corresponding margin layers of Definition 1.

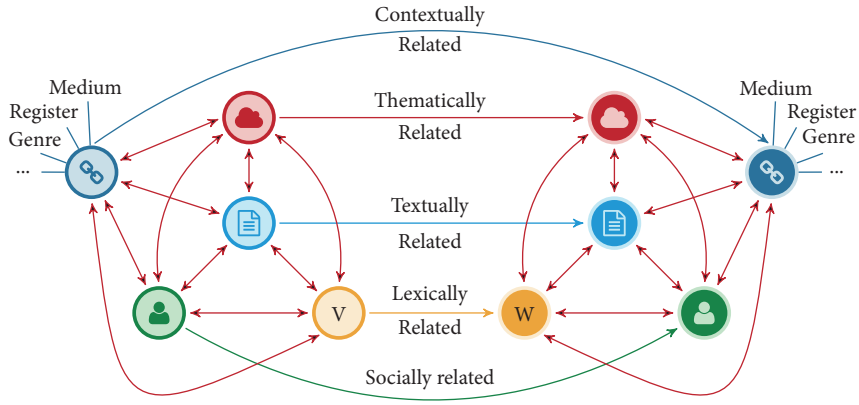


Figure 12: Schematic depiction of informational sources explored to link topics (red vertices) in an *n-level topic network*, $n > 3$, as a function of the textual relatedness of texts (blue vertices) (belonging to layer L_1 of Definition 1), the social relatedness of corresponding authors (green vertices) (belonging to layer L_2 of Definition 1), the lexical relatedness of corresponding words (orange vertices) (belonging to layer L_3 of Definition 1), and additional layers of contextual patterns concerning, for example, the underlying medium, genre, or register instantiated by the texts under consideration.

$$\begin{aligned}\mathscr{M}(X,k) &= (\mathbb{M},\mathbb{D}),\\ \mathbb{M} &= \{M_i = (V_i, A_i, \mu_i, \nu_i, \lambda_i, \kappa_i) \mid i = 1,\ldots,k\},\\ \mathbb{D} &= \{D_{i.j} = (V_{i.j}, A_{i.j}, \mu_{i.j}, \nu_{i.j}, \lambda_{i.j}, \kappa_{i.j}) \mid i,j = 1,\ldots,k : i \neq j\},\end{aligned} \tag{29}$$

such that each M_i, $i \in \{1,\ldots,k\}$, is an $(L_1, \mathbb{L}_i')$-Topic Network derived from $\mathcal{S}$ according to Definition 7 and for each $i,j \in \{1,\ldots,l\}$, $i \neq j$, $D_{i.j} \in \mathbb{D}$, $|\mathbb{D}| = k(k-1)$, is called a *margin layer* fulfilling the following requirements: $V_{i.j} = V_i \cup V_j$, $A_{i.j} = \{(v,w) \in V_i \times V_j \mid \dot{v} = \dot{w}\}$, $\mu_{i.j} = \mu_i \cup \mu_j$, and $\lambda_{i.j} = \lambda_i \cup \lambda_j$.

See Figure 13 for a schematic depiction of the comparison of two MTNs. Note that because of Definition 7, it does not necessarily hold that $V_{\mathscr{C}}(V_i) = V_{\mathscr{C}}(V_j)$, but it always holds that $V_{\mathscr{C}}(V_i) \subseteq V_{\mathscr{C}} \supseteq V_{\mathscr{C}}(V_j)$. In this respect, we depart from [64], which instead require more strongly that $V_i = V_j$. In the case of topic networks, this would be too restrictive, as different topic networks derived from the same definitional setting can focus on different subsets of topics, while ignoring the rest of the topics in the co-domain $V_{\mathscr{C}}$ of θ. (A way to extend Definition 8 is to include the RCS $\mathscr{C} = (V_{\mathscr{C}}, A_{\mathscr{C}})$ of Definition 2 as an additional layer. This would allow for directly relating its constituent topic networks with the hierarchical classification system $\mathscr{C}$.)

In this paper, we quantify similarities of the different layers of MTNs to shed light on Hypothesis 1. More specifically, we generate an LMN for each corpus of a set of different text corpora in order to derive a separate two-layer MTN for each of these LMNs, each consisting of a TTN and an associated ATN. Then, among other things, we conduct a triadic classification experiment: firstly with respect to the subset of all TTNs derived from our corpus, secondly with respect to the subset of all corresponding ATNs, and thirdly with respect to the subset of all TTNs in relation to the subset of the corresponding ATNs. In the next section, we explain the measurement procedure for carrying out this triadic classification experiment.

3.2. A Procedural Model of Topic Network Analysis. In order to instantiate topic networks as manifestations of the rhematic networking of places, we employ the procedure depicted in Figure 14. It combines nine modules for the induction, comparison, and classification of topic networks.

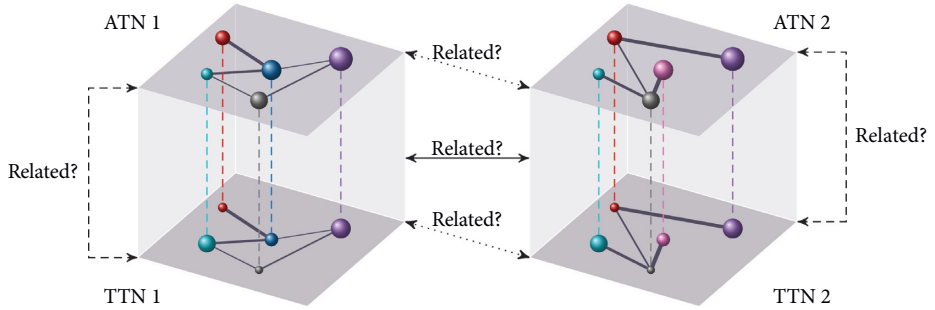


Figure 13: 3D depiction of two MTNs (left and right) each consisting of two layers (including a TTN at the bottom and an ATN at the top of the respective cube). Shared colors of nodes and dashed vertical lines indicate identically labeled vertices. The depiction disregards the orientation of the arcs. In this example, all four layers span topic networks over the same set of topics (vertices). Any such two-layer MTN can be used to represent the intertextuality- and coauthorship-based networking of the topics derived from the same corpus of texts about the same place. In this way, we gain several perspectives for the analysis of such multiplex networks: by comparing the TTNs or the ATNs of different MTNs (dotted arcs), by comparing the TTNs of different networks with their corresponding ATNs (dashed arcs), or by comparing the different MTNs as a whole with each other (solid arc).

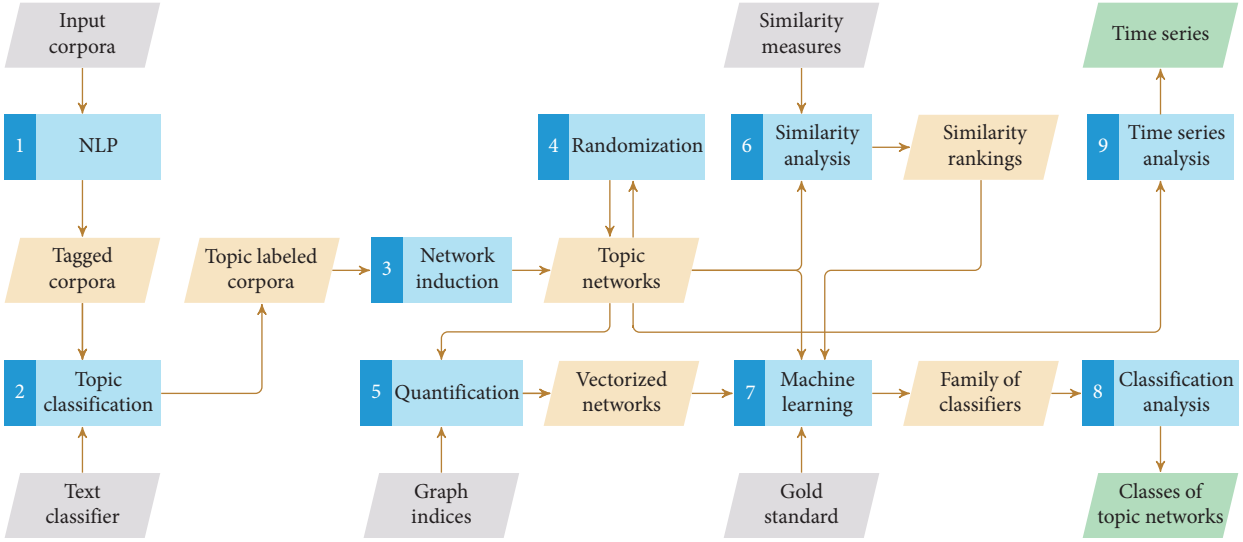


Figure 14: A procedural model of investigating LMNs and MTNs: generating, randomizing, and quantifying topic networks in 9 steps including *Natural Language Processing* (NLP) (1), topic classification using a classifier θ according to Definition 2 (2), topic network induction according to Definition 8 (3), network randomization according to Section 3.2.4 (4), network quantification (5), and network similarity analysis (6) both based on Section 3.2.6, machine learning of network classifiers (7) and classification analysis (8) both based on Section 3.2.7, and finally, time series analysis of topic networks (which will not be performed here) (9).

3.2.1. Module 1: Natural Language Processing. Preparatory for all modules is the natural language processing of the input text corpora. To this end, we utilize the NLP tool chain of *TextImager* [84] to carry out tokenization, sentence splitting, part of speech tagging, lemmatization, morphological tagging, named entity recognition, dependency parsing [85], and automatic disambiguation—the latter by means of fastSense [86]. For more details on these submodules, see [86, 87]. As a result of Module 1, the topic classification can be fed with texts whose lexical components are disambiguated at the sense level. As a sense model, we use the disambiguation pages of Wikipedia, currently the largest available model of lexical ambiguity.

3.2.2. Module 2: Topic Classification. According to Definition 2, the derivation of TNs from LMNs requires the specification of a *Reference Classification System* (RCS) $\mathscr{C} = (V_{\mathscr{C}}, A_{\mathscr{C}})$. For this purpose, we utilize the *Dewey Decimal Classification* (DDC), a system that is well established in the area of (digital) libraries. As a result, the generalized tree $\mathscr{C}$ from Definition 2 degenerates into an ordinary tree since the DDC has no arcs superimposing its kernel hierarchy (see Figure 15 for a subtree of the DDC). As a classifier θ, which addresses the DDC, we use $\theta :=$ text2ddc [72], a topic classifier based on neural networks, which has been trained for a variety of languages [88] (see https://textimager.hucompute.org/DDC/).

Starting from the output of Module 1 (NLP), we use text2ddc to map each input text x to the distribution of the 5 top-ranked DDC classes that best match the content of x as predicted by text2ddc. Since text2ddc reflects the three-level topic hierarchy of the DDC, this classifier can output a subset of 98 classes of the 2nd (two classes of this level are unspecified) and a subset of 641 classes of the 3rd DDC level for each input text. (We did not have training for all 3rd level classes (which are partly unspecified). See [72] and the appendix for details.) Thus, each topic network of each input corpus is represented on two levels of increasing thematic resolution. Note that text2ddc classifies input texts of any size (from single words to entire texts in order to meet challenge P3) and works as a multilabel classifier for processing thematically ambiguous input texts. By using an RCS, text2ddc meets challenge P2 simply by referring to the labels of the topic classes of the DDC. Furthermore, since text2ddc is trained with the help of a reference corpus, it can detect topics, even if they occur only once in a text (this is needed to meet challenge P4) and guarantees comparability for different input corpora (challenge P1). text2ddc is based on fastText whose time complexity is $O(h\log_2(k))$, where "k is the number of classes and h the dimension of the text representation" (2, [89]) (making this classifier competitive compared to TMs).

Figures 4–7 show examples of TTNs and ATNs generated by means of text2ddc by addressing the second level of the DDC. Each of these topic networks was generated for a subset of articles of the German Wikipedia that are at most 2 clicks away from the respective start article x (for the statistics of the corpora underlying these topic networks, see Section 4.1). Formally speaking, let $G = (V, A)$ be a directed graph and $v \in V$; the nth orbit induced by v is the subgraph,

$$\begin{aligned} G_v^n &= \left(V_v^n, A_v^n\right), \\ V_v^n &= \{w \in V \mid \delta(v,w) \leq n\}, \\ A_v^n &= \{(r,s) \in A \mid r, s \in V_v^n\}, \end{aligned} \tag{30}$$

that is induced by the subset of vertices whose geodetic distance $\delta(v,w)$ from v is at most n (cf. [90]). We compute the first orbit and the second orbit of a set of Wikipedia articles (so that G denotes Wikipedia's web graph). This is done to obtain a basis for comparison for the evaluation of topic networks derived from special wikis. Since Wikipedia is probably more strongly regulated than these special wikis, we expect higher disparities between networks of different groups (Wikipedia vs. special wiki) and smaller differences for networks of the same group.

3.2.3. Module 3: Network Induction. Network induction is done according to the formal model of Section 3.1. It starts with inducing an LMN $\mathscr{L}(X,2)$ for each input corpus X. That is, for each corpus X, we generate a text network L_1 and an agent network L_2 according to Definition 1:

(1) In this paper, X always denotes the set of texts (web documents) of a corresponding wiki W so that the text layer $L_1 = (V_1, A_1, \mu_1, \nu_1, \lambda_1, \kappa_1)$ of the LMN $\mathscr{L}(X,2)$, in which L_2 is an agent network defined below, can be used to represent the web graph [91] of this wiki. Thus, for any two texts x and y that are linked in W, we generate an arc $a = (v,w) \in A_1$, where $\nu_1(a) = 1$ and $\kappa_1(a) =$ hyperlink. Further, for $\forall x \in V_1 : \mu_1(x) = 1 \wedge \lambda_1(x) = x$.

(2) The author layer $L_2 = (V_2, A_2, \mu_2, \nu_2, \lambda_2, \kappa_2)$ of the LMN $\mathscr{L}(X,2)$ corresponding to L_1 (see Definition 1) is generated as follows: V_2 is the set of all registered authors or TCP/IP addresses of anonymous users working on texts in X so that $\forall v \in V_2 : \lambda_2(v)$ maps to this name or IP address, respectively. Let $✎(r,x)$ be the sum of all additions made by the author $r \in V_2$ to any revision of the edit history of the text x; we use $✎(r,x)$ to approximate the more difficult to measure concept of authorship as introduced by Brandes et al. [74]. Then, we define: $\forall r \in V_2 : \mu_2(r) = \sum_{x \in V_1} ✎(r,x)$. Further, A_2 is the set of all arcs (r,s) between users $r, s \in V_2$, for which there is at least one text x to which both contribute so that $✎(r,x)$, $✎(s,x) > 0$. Then, we define (cf. [92]):

$$\nu_2(r,s) = \sum_{x \in V_1} 2\frac{\min(✎(r,x), ✎(s,x))}{\sum_{u \in V_2} ✎(u,x)} \in (0,1]. \tag{31}$$

Finally, $\kappa_2(a) =$ coauthorship. Obviously, L_2 is symmetric.

Now, given the definitional setting $(\mathscr{C}, \theta, \mathscr{L}(X,2))$, where $\mathscr{C}, \theta$ are instantiated in terms of Section 3.2.2, we induce a TTN $T(L_1) = (V_{L_1}, A_{L_1}, \mu_{L_1}, \nu_{L_1}, \lambda_{L_1}, \kappa_{L_1})$ according to Definition 3 by means of appropriately defined monotonically increasing functions α_1, β_1, γ_1, and δ_1. To this end, we utilize the set

$$\theta_x^{V_{\mathscr{C}}} = \{\theta_x(\dot{v}) > \theta_{\min} \mid \dot{v} \in V_{\mathscr{C}}\}, \tag{32}$$

of the membership values of text $x \in V_1$ to the topics in $V_{\mathscr{C}}$, where the parameter $\theta_{\min}$ denotes a lower bound of an acceptable degree of aboutness. We set $\theta_{\min} := 0$. Further, by

$$\overline{\theta} = \frac{1}{|\mathbb{Y}|} \sum_{y \in \mathbb{Y}} y, \tag{33}$$

we denote the mean value of the set $\mathbb{Y} = \cup_{x \in V_1} \theta_x^{V_{\mathscr{C}}}$ of selected topic membership values and by $\max(\mathbb{X}, m)$ we denote the $m \in \{1, \ldots, |\mathbb{X}|\}$ largest value of the arbitrary set $\mathbb{X}$. Finally, we select a number $0 < m_\perp < |V_{\mathscr{C}}|$ and define $\forall v \in V, \forall x \in V_1$, thereby instantiating the parameters α, β, γ, and δ of formulas (6) and (7) of Definition 3:

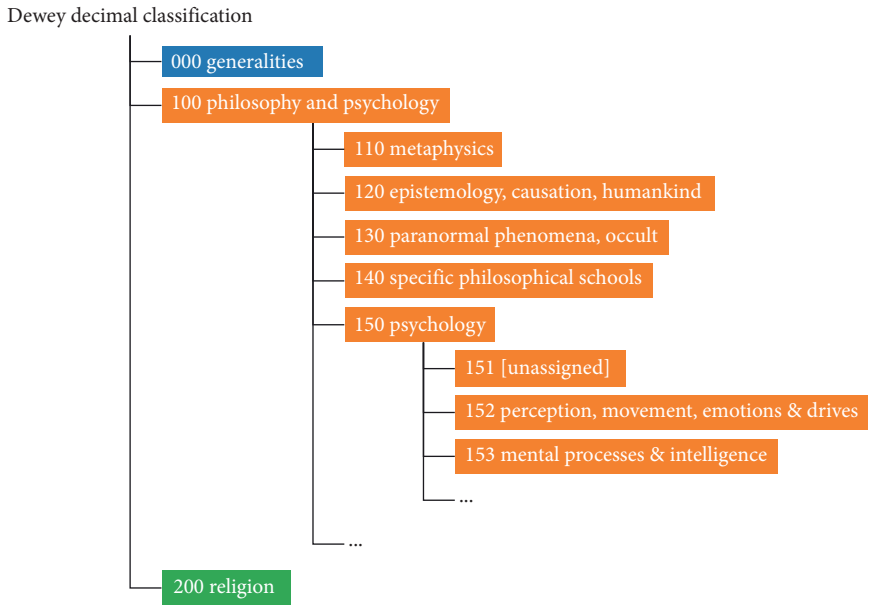


Figure 15: A subtree of the DDC displaying a snapshot of the second class (100) on the first three levels.

$$\alpha \coloneqq \alpha_1 = \mathrm{id}, \tag{34}$$

$$\begin{aligned}\beta\left(x \underset{\theta^{\leftarrow}}{\overset{\theta}{\longleftrightarrow}} \dot{v}\right) &\coloneqq \beta_1\left(x \xrightarrow{\theta} \dot{v}\right) = \beta_1\left(\theta\left(x, \lambda(v)\right)\right) = \beta_1\left(\theta_x(\dot{v})\right) \\ &= \begin{cases} \theta_x(\dot{v}), & \theta_x(\dot{v}) \in \left\{r \in \theta_x^{V_{\mathscr{C}}} \mid \exists m \leq m_{\perp} : r = \max\left(\theta_x^{V_{\mathscr{C}}}, m\right) \geq \overline{\theta}\right\}, \\ 0, & \text{else}, \end{cases}\end{aligned} \tag{35}$$

$$\begin{aligned}\gamma &\coloneqq \gamma_1 = \mathrm{id}, \\ \delta\left[x \underset{\theta^{\leftarrow}}{\overset{\theta}{\longleftrightarrow}} \dot{v}, y \underset{\theta^{\leftarrow}}{\overset{\theta}{\longleftrightarrow}} \dot{w}, x \xrightarrow{\nu_1} y\right] &\coloneqq \delta_1\left(x \xrightarrow{\theta} \dot{v}, y \xrightarrow{\theta} \dot{w}, x \xrightarrow{\nu_1} y\right) \\ &= \begin{cases} \beta_1\left(\theta_x(\dot{v})\right)\beta_1\left(\theta_y(\dot{w})\right), & (x, y) \in A_1, \\ 0, & \text{else}. \end{cases}\end{aligned} \tag{36}$$

According to formula (35), $\beta_1\left(x_{\theta^{\leftarrow}} \longleftrightarrow_{\theta} \dot{v}\right) = \theta_x(\dot{v})$ iff $\theta_x(\dot{v})$ is one of the $m_{\perp}$ highest membership values of x to the topics in $V_{\mathscr{C}}$, supposed that $\theta_x(\dot{v}) > \overline{\theta}$. Otherwise, $\beta_1\left(x_{\theta^{\leftarrow}} \longleftrightarrow_{\theta} \dot{v}\right) = 0$. In this paper, we experiment with $m_{\perp} = 5$. The higher the value of $m_{\perp}$, the more sensitive the generation of $T(L_1)$ to the thematic ambiguity of the underlying texts. However, since θ creates a membership value for each pair of texts and topics, we use $\overline{\theta}$ as a lower bound of aboutness (in the sense of addressing a topic known by θ) so that irrelevant classifications $\theta_x(\dot{v})$ do not affect $\mu_{L_1}(v)$.

Regarding the ATN $T(L_1, \{L_2\}) = (V_{L_2}, A_{L_2}, \mu_{L_2}, \nu_{L_2}, \lambda_{L_2}, \kappa_{L_2})$ corresponding to the TTN $T(L_1)$, we have to define monotonically increasing functions α_2, β_2, γ_2, and δ_2. To this end, we use several auxiliary functions:

(i) By $\overline{\not\in(\cdot,\cdot)}$, we denote the mean activity per author per Wikipedia article.

(ii) By $\overline{|\not\in(\cdot,\cdot)|}$, we denote the average number of active authors per Wikipedia article.

The corresponding estimators are found in Table 4. Now, consider the set $V_2(x)$ of all active authors of the text x and the set $\theta_v(V_1)$ of all texts that potentially contribute to $\mu_{L_2}(v)$ and thus to the weight of the vertex $v \in V_{L_2}$:

$$\begin{aligned}V_2(x) &= \{r \in V_2 \mid \not\in(r, x) > 0\}, \\ \theta_v(V_1) &= \left\{x \in V_1 \mid \beta_1\left(\theta_x(\dot{v})\right) > 0\right\}.\end{aligned} \tag{37}$$

Then, we define the following functions and ratios:

$$\mathrm{scale} = \begin{cases} (0,1]^2 \longrightarrow (0,2], \\ \mathrm{scale}(a,b) \mapsto 2\dfrac{a}{a+b}, \end{cases} \tag{38}$$

$$\omega_x = \mathrm{scale}\left(\left|V_2(x)\right|, \overline{|\not\in(\cdot,\cdot)|}\right), \quad \in (0,2], \tag{39}$$

$$\omega_v = \frac{1}{\left|\theta_v(V_1)\right|} \sum_{x \in \theta_v(V_1)} \omega_x, \quad \in (0,2], \tag{40}$$

where scale is a function which is used to rescale below or above average values (see formula (39)). Formula (40) defines the mean of the rescaled numbers of active users per article in $\theta_v(V_1)$. Based on these preliminaries and regarding the vertex weighting function μ_{L_2}, we define $\forall v \in V$ and $\forall r \in V_2$, thereby instantiating the functions α and β of formula (10) of Definition 4:

$$\alpha \coloneqq \alpha_2 \wedge \forall z \in \mathbb{R} : \alpha_2(z) = \omega_v \cdot z, \tag{41}$$

$$\beta\left[x \underset{\theta^{\leftarrow}}{\overset{\theta}{\longleftrightarrow}} \dot{v}, r \underset{\vartheta^{\leftarrow}}{\overset{\vartheta}{\longleftrightarrow}} \dot{v}, r \underset{\nu_{1,i}}{\overset{\nu_{i,1}}{\longleftrightarrow}} x\right] \coloneqq \beta_2\left(x \overset{\theta}{\longrightarrow} \dot{v}, r \overset{\nu_{2,1}}{\longrightarrow} x\right)$$
$$= \beta_1\left(\theta_x(\dot{v})\right) \cdot \begin{cases} \frac{1}{p} \frac{✎(r,x)}{\sum_{s \in V_2} ✎(s,x)}, & ✎(r,x) < \overline{✎(\cdot,\cdot)}, \\ \frac{✎(r,x)}{\sum_{s \in V_2} ✎(s,x)}, & ✎(r,x) = \overline{✎(\cdot,\cdot)}, \\ p \frac{✎(r,x)}{\sum_{s \in V_2} ✎(s,x)}, & ✎(r,x) > \overline{✎(\cdot,\cdot)}. \end{cases} \tag{42}$$

In the present paper, we experiment with $p = 2$. To understand this definition, we have to run through the cases of formula (42):

(1) The case $✎(r,x) = \overline{✎(\cdot,\cdot)}$: suppose that, for each $x \in \theta_v(V_1)$, the following condition holds: $\forall r, s \in V_2(x) : ✎(r,x) = ✎(s,x) = \overline{✎(\cdot,\cdot)}$. In this case, we obtain for each $x \in \theta_v(V_1)$, the following result:

$$\sum_{r \in V_2} \beta_1\left(\theta_x(\dot{v})\right) \frac{✎(r,x)}{\sum_{s \in V_2} ✎(s,x)} = \beta_1\left(\theta_x(\dot{v})\right) \sum_{r \in V_2} \frac{✎(r,x)}{\sum_{s \in V_2} ✎(s,x)}$$
$$= \beta_1\left(\theta_x(\dot{v})\right). \tag{43}$$

In other words, if all authors of all texts contributing to the weight of a topic contribute to these texts according to the average activity, the weight of this topic in the ATN corresponds to that of the corresponding TTN. In this case, the average activity does not bias the weight of a topic in the ATN compared to the same topic in the corresponding TTN. Obviously, this scenario gives us a *neutral point* or, more specifically, a *calibration point* for the comparison of ATNs and TTNs. Such a calibration point allows us to interpret any down- or upward deviation of the topic weights in both networks, since no deviation means average activity and average number of active users. However, this consideration presupposes that $\omega_v = 1$ so that $\alpha_2 = \alpha_1 = \mathrm{id}$. If $\omega_v > 1$, then the number of authors of texts contributing to the weight of v is on average higher than expected on the basis of Wikipedia, so that the weight of the topic v in the ATN is "biased upwards" compared to the weight of the same topic in the corresponding TTN. Conversely, if $\omega_v < 1$, then the number of authors of texts contributing to the weight of v is on average smaller than expected, so that v's weight in the ATN is "biased downwards" compared to the weight of the same topic in the corresponding TTN. This scenario teaches us the different roles of α_2 and β_2 with respect to the weighting of the β_1 values: while β_2 operates as a function of the activities of authors, α_2 considers their number.

(2) The case $✎(r,x) < \overline{✎(\cdot,\cdot)}$: suppose for each $s \neq r$ that $✎(s,x) = \overline{✎(\cdot,\cdot)}$ while $✎(r,x) < \overline{✎(\cdot,\cdot)}$. Then, we conclude the following:

$$\beta_1\left(\theta_x(\dot{v})\right)\left(\sum_{t \in V_2 \setminus \{r\}} \frac{✎(t,x)}{\sum_{s \in V_2} ✎(s,x)} + \frac{1}{p} \frac{✎(r,x)}{\sum_{s \in V_2} ✎(s,x)}\right) < \beta_1\left(\theta_x(\dot{v})\right) \Longleftrightarrow$$
$$\sum_{t \in V_2 \setminus \{r\}} \frac{✎(t,x)}{\sum_{s \in V_2} ✎(s,x)} + \frac{1}{p} \frac{✎(r,x)}{\sum_{s \in V_2} ✎(s,x)} < 1 \Longleftrightarrow \tag{44}$$
$$\frac{1}{p} \frac{✎(r,x)}{\sum_{s \in V_2} ✎(s,x)} < \frac{✎(r,x)}{\sum_{s \in V_2} ✎(s,x)} \Longleftrightarrow 1 < p.$$

Thus, for $p > 1$, we penalize the contribution of a below-average active author of a text to the weight of the topic to which this text contributes. The different effects of $\omega_v \lessgtr 1$ have already been discussed.

(3) The case $\overline{\not\in}(r,x) \geq \overline{\not\in}(\cdot,\cdot)$: if we suppose now that $\forall s \neq r : \not\in(s,x) = \overline{\not\in}(\cdot,\cdot)$ while $\not\in(r,x) > \overline{\not\in}(\cdot,\cdot)$, we conclude that for $p > 1$, we reward the contribution of an above-average active author of a text to the weight of the topic to which this text contributes.

In a nutshell, α_2 and β_2 implement the following proportionality assumptions:

(i) By α_2 we penalize or reward under- or above-average coauthorships: the higher the above-average number of authors contributing to the texts of a topic, the higher the reward effect and the higher the weight of the topic. And vice versa, the lower the below-average number of authors contributing to the texts of a topic, the higher the penalty effect and the lower the weight of the topic.

(ii) By β_2 we penalize or reward under- or above-average activities of single authors: the higher the above-average activity of a single author contributing to a text of a topic, the higher the reward effect and the higher the contribution of this author-text pair to the weight of the topic. And vice versa, the lower the below-average activity of a single author contributing to a text of a topic, the higher the penalty effect and the lower the contribution of this author-text pair to the weight of the topic.

Finally, we define the functions γ_2 and δ_2 to get instantiations of the functions γ and δ of formula (11) of Definition 4 (or, in the generalized case, of formula (27) of Definition 7). This is done by means of the following auxiliary function:

$$\widetilde{\nu}_2(r,s) = \text{scale}\left(\nu_2(r,s), \overline{\nu}_2\right) \in \mathbb{R}^+, \tag{45}$$

where $\overline{\nu}_2$ estimates the average degree of coauthorship in Wikipedia according to formula (31). (We estimate $\overline{\nu}_2$ by means of 10,000 randomly selected Wikipedia articles so that $\overline{\nu}_2 \coloneqq 0.002,756.$) $\widetilde{\nu}_2(r,s)$ is a readjustment of $\nu_2(r,s)$ in relation to the mean value $\overline{\nu}_2$: the higher the above-average coauthorship, the higher the value of $\widetilde{\nu}_2$, and the lower the below-average coauthorship, the lower the value of $\widetilde{\nu}_2$. Then, we define

$$\begin{aligned}
\gamma &\coloneqq \gamma_2 = \text{id},\\
\delta\left[x \underset{\theta^{\leftarrow}}{\overset{\theta}{\longleftrightarrow}} \dot{v}, y \underset{\theta^{\leftarrow}}{\overset{\theta}{\longleftrightarrow}} \dot{w}, r \underset{\vartheta^{\leftarrow}}{\overset{\vartheta}{\longleftrightarrow}} \dot{v}, s \underset{\vartheta^{\leftarrow}}{\overset{\vartheta}{\longleftrightarrow}} \dot{w}, r \underset{\nu_{1.i}}{\overset{\nu_{i.1}}{\longleftrightarrow}} x, s \underset{\nu_{1.i}}{\overset{\nu_{i.1}}{\longleftrightarrow}} y, r \overset{\nu_i}{\longrightarrow} s, x \overset{\nu_1}{\longrightarrow} y\right] &\coloneqq\\
\delta_2\left(x \overset{\theta}{\longrightarrow} \dot{v}, y \overset{\theta}{\longrightarrow} \dot{w}, r \overset{\nu_{2.1}}{\longrightarrow} x, s \overset{\nu_{2.1}}{\longrightarrow} y, r \overset{\nu_2}{\longrightarrow} s, x \overset{\nu_1}{\longrightarrow} y\right) &=\\
\begin{cases}
\widetilde{\nu}_2(r,s) \cdot \beta_2\left(\theta_x(\dot{v})\right) \cdot \beta_2\left(\theta_y(\dot{w})\right), & (x,y) \in A_1 \wedge (r,s) \in A_2,\\
0, & \text{else.}
\end{cases}
\end{aligned} \tag{46}$$

In this definition, $\beta_2(\theta_x(\dot{v}))$ quantifies the link $x \overset{\theta}{\longrightarrow} \dot{v}$ and the link $r \overset{\nu_{2.1}}{\longrightarrow} x$ (cf. formula (11)), the product $\beta_2(\theta_x(\dot{v}))\beta_2(\theta_y(\dot{w}))$ quantifies the link $x \longrightarrow_{\nu_1} y$, and $\widetilde{\nu}_2(r,s)$ quantifies the link $r \longrightarrow_{\nu_2} s$. The calibration point of arc weighting is now reached under the conditions of the following scenario (for the first two conditions, see above):

$$\beta_2\left(\theta_x(\dot{v})\right) = \beta_1\left(\theta_x(\dot{v})\right), \tag{47}$$

$$\beta_2\left(\theta_y(\dot{w})\right) = \beta_1\left(\theta_y(\dot{w})\right), \tag{48}$$

$$\widetilde{\nu}_2(r,s) = 1. \tag{49}$$

Under these conditions, the authors r and s contribute to the texts x and y at an average level while interacting at an average level of coauthorship. In this case, the (co)authorship of both authors does not influence the strength of the corresponding arc in the ATN: in terms of neither reducing nor increasing $\nu_2(v,w)$. Note that the size of an ATN (i.e., the number of its arcs) is always less than or equal to that of the corresponding TTN, since the arcs present in a TTN are merely re-weighted in the corresponding ATN: no new arcs are added. The same holds for the order of the ATN since there is no node in a TTN for which there is no author authoring it.

Our instantiation of multiplex text and author topic networks has shown two points: firstly, we demonstrated a single-parameter setting as an element of a huge parameter space spanned by parameters such as p, $\overline{\nu}_2$, $\overline{\not\in}(\cdot,\cdot)$, $|\overline{\not\in}(\cdot,\cdot)|$, θ, α_1, α_2, β_1, β_2, γ_1, γ_2, δ_1, and δ_2. (In the latter eight cases, various information links are included as candidate parameters. Formula (42) shows, for example, that out of the six possible information links, only two are evaluated to instantiate β_2. Obviously, numerous alternatives exist to instantiate this function.) Secondly, anyone who complains about the apparently inherent parameter explosion in our approach should consider the hyperparameter spaces of neuronal networks as an object of parameter optimizations. Regardless of the heuristic character of our approach, compared to the black box character of neural networks, its settings are extensible on the basis of the schematic framework provided by Definition 8 of MTNs and the definitions it is based upon. At the same time, this approach guarantees interpretability as long as the different ingredients entering our model via formulas of the sort as formulas (26) and (26) fulfill this condition—in order to meet challenge P5.

3.2.4. Module 4: Network Randomization. Randomization is conducted to assess the significance of our findings. This is necessary because there is currently no related classification in the area examined here that can serve this role. To fill this gap, we compute the following randomizations:

(1) Baseline B1: a lower bound of a baseline is obtained by randomly assigning the object networks onto the gold standard (target) classes. This can be done by informing the assignment about the true cardinality of these classes (B11) or not (B12). We opt for B11 since this variant yields a higher *F*-score, making it more difficult to surpass. Of course, any serious network representation and classification model should go beyond this baseline. B1 will be averaged over 100,000 iterations.

(2) Baseline B2: an alternative is to randomize the input networks and to derive vector representations (according to Section 3.2.3), which ultimately undergo the same classification process as the original networks. That is, the input networks are randomly rewired to generate Erdős-Rényi (ER) graphs, for which we ask whether they are separable by the same classification model. (An alternative, not considered here, would be to randomize the topic classification of the underlying texts.) If this is successful in terms of high *F*-scores (the *F*-score is a measure of the accuracy of a classification, that is, the harmonic mean of its precision and recall), then we conclude that the network representation model or the operative classifier is not informative enough regarding the hypothetical class memberships of the input networks. Conversely, the lower the average *F*-scores obtained by classifying the randomized networks compared to the classification of the original ones, the more informative the representation model or the classification procedure regarding the underlying hypotheses. By keeping the model constant while varying the classifier, we can ultimately attribute this (non)informativity to the underlying representation model. Conversely, by keeping the classifier constant while varying the model, we can attribute this informativity to the classification model. B2 will be repeated 100 times.

(3) Baseline B3: a third baseline results from randomizing the matrices that form the input of the target classifiers. This means that instead of calculating graph invariants or similarity values to feed the classifiers, we use matrices whose dimensions are chosen uniformly at random from the domain of the corresponding invariants or (dis)similarity measures. (We require that the main diagonal of the random matrix is 1 and that it is symmetric.) If the classification based on the original networks does not exceed this baseline, we are again informed about a deficit of our representation model. Evidently, we are looking for models that significantly exceed this baseline; otherwise, we would have to accept that the same classifiers perform better on random values than on our feature model. B3 will be repeated 100 times.

(4) Baseline B4: finally, we start from randomly reorganizing the set of observations into random classes while using the same representation model to separate the resulting random gold standard. (Obviously, we have to prevent that the gold standard is ever part of the set of these randomizations.) We choose the variant of using randomized cardinalities of the random classes rather than keeping the sizes of the gold standard. Tests have shown that this approach tends to generate higher *F*-scores than the latter. If our network representation and classification model do not outperform this baseline, we learn that the underlying invariants used to characterize the networks are not *specific* enough; rather, they can be related to random classifications of the same objects using the same feature space. Obviously, we are looking for a model characterizing the gold standard (*tendency to specificity*) and not a random counterpart of it (*tendency to non-specificity*). B4 is averaged over 100 repetitions.

B1 is a lower bound: models that fall under this bound are obsolete. B2 concerns the evaluation of the network representation or classification model. B3 focuses on evaluating the classification model, and B4 aims at evaluating the specificity of the operative feature model.

3.2.5. Module 5: Network Quantification. Module 5 is a preparatory step for a subset of network similarity measures. This relates to so-called topology-based approaches to graph similarity [57, 93–96]. The idea behind this approach is to map input networks onto vectors of graph indices or invariants to compare them with each other. That is, graph similarity is traced back to similarity in vector space: the higher the number of indices for which two graphs resemble each other, the more similar the graphs. The apparatus that we employ in this context is described next.

3.2.6. Module 6: Graph Similarity Analysis. Our hypothesis about thematic networks on geographical places says that these networks are similar in terms of the skewness of their thematic focus and their network structure, regardless of whether the underlying texts are written by different communities and regardless of the framing theme. To test this hypothesis, we apply the framework of graph similarity measurement which allows for mapping the second of these three reference points by exploring the structure of topic networks as well as features of their nodes. Since graph similarity measurement is generally known to be computational complex, we take profit from the fact of dealing with *labeled* graphs. By using alignments of the labels of the nodes of the graphs to be compared, we reduce the time complexity of these approaches enormously.

The literature knows a number of approaches for graph similarity measurement. Among other things, this includes

the following approaches (see Emmert-Streib et al. [97] for an overview (cf. [98, 99]); the paper does not aim at a comprehensive study of them but focuses on a selected subset):

(1) *Graph Edit Distance-* (GED-) based approaches [100–102] and their relatives (e.g., the *Vertex and Edge* Overlap (VEO) [103])

(2) Spherical [90] or neighborhood-related approaches (cf. [99])

(3) Network topology-related approaches [57, 93–96, 103]

We will develop and test candidates of each of these classes.

GED-based methods are well studied in the area of web mining [104]. Since we are dealing with labeled graphs, we can compute the GED directly from the vertex and edge sets of the input graphs [99, 100]. Let $G_1 = (V_1, A_1, \mu_1, \nu_1, \lambda_1, \kappa_1)$ and $G_2 = (V_2, A_2, \mu_2, \nu_2, \lambda_2, \kappa_2)$ be two TNs, then their GED is computed as follows:

$$\begin{aligned}\mathrm{GED}(G_1, G_2) = &\left|V_1\right| + \left|V_2\right| - 2\left|V_{\mathscr{C}}(V_1) \cap V_{\mathscr{C}}(V_2)\right| \\ &+ \left|A_1\right| + \left|A_2\right| - 2\left|V_{\mathscr{C}}(A_1) \cap V_{\mathscr{C}}(A_2)\right| \in \mathbb{R}_0^+,\end{aligned} \tag{50}$$

where $V_{\mathscr{C}}(A_i) = \{(\dot{v}, \dot{w}) \mid (v, w) \in A_i\}, i = 1, \ldots, 2$. Since we are targeting graph similarities, we consider GES instead of GED, where overlaps of vertex and arc sets are equally weighted:

$$\mathrm{GES}(G_1, G_2) = 1 - \frac{1}{2}\left(\frac{\left|V_1\right| + \left|V_2\right| - 2\left|V_{\mathscr{C}}(V_1) \cap V_{\mathscr{C}}(V_2)\right|}{\left|V_1\right| + \left|V_2\right|} + \frac{\left|A_1\right| + \left|A_2\right| - 2\left|V_{\mathscr{C}}(A_1) \cap V_{\mathscr{C}}(A_2)\right|}{\left|A_1\right| + \left|A_2\right|}\right) \in [0, 1]. \tag{51}$$

The same is done in the case of Wallis' approach to graph distance [102], which is adapted as follows to get a similarity measure:

$$\mathrm{WAL}(G_1, G_2) = \frac{\left|V_{\mathscr{C}}(V_1) \cap V_{\mathscr{C}}(V_2)\right| + \left|V_{\mathscr{C}}(A_1) \cap V_{\mathscr{C}}(A_2)\right|}{\left|V_1\right| + \left|V_2\right| + \left|A_1\right| + \left|A_2\right| - \left|V_{\mathscr{C}}(V_1) \cap V_{\mathscr{C}}(V_2)\right| - \left|V_{\mathscr{C}}(A_1) \cap V_{\mathscr{C}}(A_2)\right|} \in [0, 1]. \tag{52}$$

A relative of GES is the *Vertex/Edge Overlap* (VEO) graph similarity measure [103]:

$$\mathrm{VEO}(G_1, G_2) = 2\frac{\left|V_{\mathscr{C}}(V_1) \cap V_{\mathscr{C}}(V_2)\right| + \left|V_{\mathscr{C}}(A_1) \cap V_{\mathscr{C}}(A_2)\right|}{\left|V_1\right| + \left|V_2\right| + \left|A_1\right| + \left|A_2\right|} \tag{53}$$

$$= 1 - \frac{\mathrm{GED}(G_1, G_2)}{\left|V_1\right| + \left|V_2\right| + \left|A_1\right| + \left|A_2\right|}, \quad \in [0, 1]. \tag{54}$$

Since node and arc weights are not taken into account by these measures, we compute the following variant of GES to close this gap:

$$\forall x, y \in \mathbb{R}_0^+ : \delta(x, y) = \frac{|x - y|}{\max(x, y)}, \quad \in [0, 1], \tag{55}$$

$$\forall v \in V_1 \forall w \in V_2 : \mathrm{wges}(v, w) = \begin{cases} \delta(\mu_1(v), \mu_2(w)), & \dot{v} = \dot{w}, \\ 0, & \text{else}, \end{cases} \in [0, 1], \tag{56}$$

$$\forall a = (v, w) \in A_1, \forall b = (x, y) \in A_2 :$$

$$\mathrm{wges}(a, b) = \begin{cases} \delta(\nu_1(a), \nu_2(b)), & \dot{v} = \dot{x} \wedge \dot{w} = \dot{y}, \\ 0, & \text{else}, \end{cases} \in [0, 1], \tag{57}$$

$$\mathrm{wges}(V_1, V_2) = \frac{|V_1| + |V_2| - 2\sum_{v \in V_1, w \in V_2} \delta(\mu_1(v), \mu_2(w))}{|V_1| + |V_2|}, \quad \in [0,1], \tag{58}$$

$$\mathrm{wges}(A_1, A_2) = \frac{|A_1| + |A_2| - 2\sum_{a \in A_1, b \in A_2} \delta(\nu_1(a), \nu_2(b))}{|A_1| + |A_2|}, \quad \in [0,1], \tag{59}$$

$$\mathrm{wges}(G_1, G_2) = \frac{\mathrm{wges}(V_1, V_2) + \mathrm{wges}(A_1, A_2)}{2}, \quad \in [0,1], \tag{60}$$

wges is sensitive to arc [99] *and* to vertex weights of TNs, the latter measuring the membership degree of the underlying texts to the topic represented by the corresponding vertex. We say that such measures are *dual weight-dependent*. These measures are of high interest since they cover more information on the underlying networks than single weight-dependent or even weight-independent measures (cf. the axiom of edge weight sensitivity of Koutra et al. [99]).

GED and its relatives share a view of similarity, according to which graphs are considered to be more similar the more (equally weighted) vertices and arcs they share. This notion of similarity is contrasted by spherical approaches (see above) as exemplified by DeltaCon [99]. Roughly speaking, according to DeltaCon, the more similar two graphs resemble each other from the perspective of their vertices, the more similar they are. Since DeltaCon is not dual weight-dependent, we consider a dual weight-dependent relative of it. To this end, we compute the cosine of the vectors of geodetic distances for each pair of equally labeled vertices. Since topic networks can differ in their order, we first have to align their node sets to make them comparable—this is also needed because we aim for a dual weight-dependent measurement. The required alignment is addressed by means of the following auxiliary graphs G_{12} and G_{21}:

$$\begin{aligned}
&\forall i, j \in \{1,2\}, i \neq j : G_{ij} = \left(V_{ij}, A_i, \mu_{ij}, \nu_i, l_{ij}\right),\\
&V_{ij} = V_i \cup \left\{w \in V_j \mid \nexists v \in V_i : \dot{v} = \dot{w}\right\},\\
&\forall v \in V_{ij} : \mu_{ij}(v) = \begin{cases} \mu_i(v), & v \in V_i,\\ 0, & \text{else,} \end{cases}\\
&\forall v \in V_{ij} : l_{ij}(v) = \begin{cases} l_i(v), & v \in V_i,\\ l_j(v), & \text{else.} \end{cases}
\end{aligned} \tag{61}$$

G_{12} and G_{21} are needed to make G_1 and G_2 comparable whose symmetric difference $V_1 \Delta V_2$ can be nonempty while their vertex labeling functions share the same co-domain (since G_1 and G_2 belong to the same multiplex topic network according to Definition 8). Obviously, $|G_{12}| = |G_{21}|$ so that for each $v \in V_i, w \in V_{ij} \backslash V_i; i, j \in \{1,2\}, i \neq j$, there is no path from v to w in G_{ij}. Cases in which no such path exists are denoted by $v \nsim w$; otherwise, if such a path exists, we denote by $\mathrm{ged}_{ij}(v, w)$ the length of the shortest path, that is, the geodetic distance between v and w in G_{ij}. As we deal with graph similarities, we first transform the distance values into similarity values:

$$\begin{aligned}
&\forall v, w \in V_{ij} : \mathrm{gep}_{ij}^{[\omega,\iota]}(v, w)\\
&= \begin{cases} 1 - \left(\dfrac{\mathrm{ged}_{ij}^{[\omega,\iota]}(v, w)}{|V_{ij}|}\right), & v, w \in V_i,\\ 0, & \text{else,} \end{cases} \quad \in [0,1],
\end{aligned} \tag{62}$$

gep is short for geodetic *proximity*. With the denominator $|V_{ij}|$, we penalize situations in which there is no path between v and w, that is, $v \nsim w$. The parameter $\omega \in \{\mathrm{w}, \neg\mathrm{w}\}$ specifies, whether the geodetic distance $\mathrm{ged}_{ij}^{[\omega,\iota]}$ and the geodetic proximity $\mathrm{gep}_{ij}^{[\omega,\iota]}$ are computed for the weighted (w) or unweighted ($\neg$w) variant of G_{ij}. If $\omega = \mathrm{w}$, we assume that each arc weighting value is normalized by means of the nonzero maximum value assumed by the arc weighting function for this network (this means that a graph G_2, which is obtained from a graph G_1 by multiplying the weights of all arcs of G_1 by a factor $c > 0$, will be equal to G_1 in terms of the graph similarity measure to be introduced now (insensitivity to certain scalings)). $\iota \in \mathbb{R}_0^+$ specifies the maximum geodetic distance to be considered: beyond this value, nodes w are considered to be of maximum geodetic distance $|V_{ij}|$ to v—irrespective of their real distance. For $\iota \geq |V_{ij}|$, we have to compute all geodetic distances. For values of $\iota \ll |V_{ij}|$ (e.g., $\iota = 2$), we arrive at variants of gep_{ij} that are less time complex. We consider the variant $\iota = \infty$ so that we take all path-related information into account. Now, we calculate the dual weight-dependent cosine of G_1 and G_2 as follows:

$$\forall v \in V_{12} \forall w \in V_{21} : \cos[\omega, \iota](v, w) = \frac{\sum_{x \in V_{12}, y \in V_{21}, \dot{x} = \dot{y}} \text{gep}_{ij}^{[\omega,\iota]}(v, x) \text{gep}_{ij}^{[\omega,\iota]}(w, y)}{\sqrt{\sum_{u \in V_{12}} \text{gep}_{ij}^{[\omega,\iota]}(v, u)^2} \sqrt{\sum_{u \in V_{21}} \text{gep}_{ij}^{[\omega,\iota]}(w, u)^2}}, \quad \in [0, 1], \tag{63}$$

$$\cos_{\mathscr{A}}[\omega, \iota, \phi, \mathbb{L}](G_1, G_2) = \frac{\sum_{v \in V_{12}, w \in V_{21}, \dot{v} = \dot{w} \in \mathbb{L}} \phi(v, w) \cos[\omega, \iota](v, w)}{\sum_{v \in V_{12}, w \in V_{21}, \dot{v} = \dot{w} \in \mathbb{L}} \phi(v, w)}, \quad \in [0, 1], \tag{64}$$

$$\cos_{\mathscr{V}}(G_1, G_2) = \frac{\sum_{v \in V_{12}, w \in V_{21}, \dot{v} = \dot{w}} \mu_{12}(v) \mu_{21}(w)}{\sqrt{\sum_{v \in V_{12}} \mu_{12}(v)^2} \sqrt{\sum_{w \in V_{21}} \mu_{21}(w)^2}}, \quad \in [0, 1], \tag{65}$$

$$\cos_{\mathscr{A}\mathscr{V}}[\omega, \iota, \phi, \mathbb{L}](G_1, G_2) = \frac{\cos_{\mathscr{V}}(G_1, G_2) + \cos_{\mathscr{A}}[\omega, \iota, \phi](G_1, G_2)}{2}, \quad \in [0, 1], \tag{66}$$

$\cos[\omega, \iota, \phi, \mathbb{L}](G_1, G_2)$ is the weighted cosine of the vectors of geodetic proximities of the same-named vertices in G_{12} and G_{21}. In this article, we consider two instantiations of parameter ϕ:

$$\forall v \in V_{12}, w \in V_{21}, \quad \dot{v} = \dot{w} : \phi_1(v, w) = 1, \tag{67}$$

$$\forall v \in V_{12}, w \in V_{21}, \quad \dot{v} = \dot{w} : \phi_2(v, w) = \max(d(v), d(w)), \tag{68}$$

where ϕ_1 implements an arithmetic mean. ϕ_2 is a function of the degree centrality [105] of its arguments: the more linked a topic in a network, the higher its impact onto the similarity of the input networks. The similarity view behind this approach is that while $\cos_X[\omega, \iota, \phi_1, \mathbb{L}]$, $X \in \{\mathscr{A}, \mathscr{A}\mathscr{V}\}$, treats all—peripheral or central—nodes equally, $\cos_X[\omega, \iota, \phi_2, \mathbb{L}]$ gives central nodes more influence. Take the example of two city networks [106]: it is plausible to say that if city networks look similar from the point of view of their central places, this should have more impact on the general similarity assessment than similarities from the point of view of peripheral locations. An extension would be to use more informative node weighting measures (e.g., closeness centrality). Finally, parameter $\mathbb{L}$ limits the number of vertices for which cosine values are computed. In the unlimited case, $\mathbb{L} \coloneqq \mathbb{L}_{12} = \{l_{12}(v) \mid v \in V_{12}\}$. It is easy to see that formulas (64)–(66) are similarity measures. For $X \in \{\mathscr{A}, \mathscr{V}, \mathscr{A}\mathscr{V}\}$, this can be shown as follows:

(1) *Symmetry*:
$\cos_X[\omega, \iota, \phi, \mathbb{L}](G_1, G_2) = \cos_X[\omega, \iota, \phi, \mathbb{L}](G_2, G_1)$
since formulas (63)–(66) are all symmetric.

(2) *Positivity:* since we are considering only positive arc weights, it always holds that

$$\cos_X[\omega, \iota, \phi, \mathbb{L}](G_1, G_1) \geq 0, \tag{69}$$

for any ω, ι, ϕ and $\mathbb{L} \neq \varnothing$.

(3) *Upper bound*: $\cos[\omega, \iota, \phi, \mathbb{L}](G_1, G_1) = 1$ for any ω, ι, ϕ and $\mathbb{L} \neq \varnothing$ and thus

$$\forall G_2 \neq G_1 : \cos[\omega, \iota, \phi, \mathbb{L}](G_1, G_1) \geq \cos[\omega, \iota, \phi, \mathbb{L}](G_1, G_2). \tag{70}$$

It is worth noticing that the range of values of formulas (63) and (65) is limited to $[0, 1]$, since the values of *gep* are always positive and we only consider positive membership values of texts to topic nodes.

So far we looked at measures that mostly processed the arc set A of TNs. This is contrasted by measures operating on topological indices of graphs. An example is NetSimile [107], which is based on the idea of characterizing networks by vectors of graph indices, which mostly draw on theories of social networks or egonets. Starting from seven local, node-related structural features (e.g., node degree, node clustering, or size of a node's egonet (see Berlingerio et al. [107] for the details of this approach)), it computes the mean and the first four moments of the corresponding distributions to generate 35-dimensional feature vectors per network where the Canberra Distance is used to compute their distances: let $\vec{x}, \vec{y} \in \mathbb{R}^k$ be two vectors, then their Canberra Distance is defined as

$$d_{\text{Can}}(\vec{x}, \vec{y}) = \sum_{i=1}^{k} \frac{\left|\vec{x}_i - \vec{y}_i\right|}{\left|\vec{x}_i + \vec{y}_i\right|}. \tag{71}$$

Soundarajan et al. [108] show that NetSimile is consistently close to the consensus among all measures studied by them, showing that it approximates the results of more complex competitors. This finding makes NetSimile a first choice in any comparative study of graph similarities.

Following on from this success, we introduce a topology-related approach to graph similarity, which draws on the hierarchical classification of the texts underlying the topic networks by reference to the *Dewey Decimal Classification* (DDC) (see Section 3.2.2). Starting from a pretest which essentially showed that graph invariants of complex network theory [109] do not sufficiently distinguish networks from their random counterparts, we decided to calculate a series of graph indices that evaluate the assignment of topics to the second level of the DDC. More specifically, we compute three node type-sensitive variants of the four cluster coefficient C_{ws} [110], C_{br} [111], C_{bbpv} [112], and C_{zh} [113] (cf. [114]). This variation can be exemplified by means of C_{ws}: to derive the desired variants from C_{ws}, we use the

following scheme, where mode $\in$ {intra, inter, heter} serves as a parameter to distinguish these alternatives (d_i is the degree of $v_i \in V$):

$$C_{\mathrm{ws}_{\mathrm{mode}}} = \frac{1}{n}\sum_{i=1}^{n} 2\frac{\mathrm{adj}_{\mathrm{mode}}(v_i)}{d_i^2 - d_i}, \quad \in [0,1], \tag{72}$$

where $\mathrm{adj}_{\mathrm{intra}}(v_i)$ is the number of adjacent neighbors of $v_i \in V$ sharing their 2nd level topic classification with v_i, $\mathrm{adj}_{\mathrm{inter}}(v_i)$ is the number of adjacent neighbors of v_i whose identical classification differs from that of v_i, and $\mathrm{adj}_{\mathrm{heter}}(v_i)$ is the number of adjacent neighbors of v_i whose classification differs among each other and from that of v_i (a 4th case is that v_i shares with a single neighbor its 2nd level topic while differing from the topics of all other neighbors). In this way, we compute for each of the cluster values C_{ws} (unweighted), C_{br} (unweighted), C_{bbpv} (weighted), and C_{zh} (weighted) three variants considering intra- and interrelational as well as heterogeneous type-sensitive clustering so that topic networks are finally represented by 12-dimensional feature vectors which are compared using the cosine measure. We call this approach *ToSi* (as short for *topological similarity*).

As a result of this candidate show of graph similarity measures, we consider the set of measures displayed in Table 5 for measuring the similarities of topic networks in order to shed light on Hypothesis 1, part (2).

3.2.7. Modules 7 and 8: Machine Learning and Classification Analysis. We conduct experiments in supervised learning with the aim of training classifiers to detect the layer (TTN or ATN) to which a topic network of a MTN belongs and the genre of the corpus from which the underlying LMN is derived. That is, our machine learning starts from a set of n genres $\mathcal{G}_i$, $i = 1, \ldots, n$, each of which is represented by a set $\mathbb{C}_i = \{C_{ij} \mid j = 1, \ldots, n_i\}$ of text corpora C_{ij} (see Figure 16). The set $\{\mathbb{C}_i \mid i = 1, \ldots, n\}$ defines a gold standard for which we assume that $\forall i, j = 1, \ldots, n,\ i \neq j : \mathbb{C}_i \cap \mathbb{C}_j = \varnothing$. Next, for each corpus C_{ij} of each genre $\mathcal{G}_i$, we span an LMN $\mathscr{L}(C_{ij}, 2)$ that in turn is used to derive a two-layer MTN $\mathscr{M}(C_{ij}, 2) = (\mathbb{M}_{ij}, \mathbb{D}_{ij})C_{ij}$ such that $\mathbb{M}_{ij} = \{M_{ij}, N_{ij}\}$ consists of exactly two topic networks: a TTN M_{ij} and an ATN N_{ij} both derived from $\mathscr{L}(C_{ij}, 2)$. In this way, we obtain the set $\mathbb{M}_{\mathrm{ttn}}$ and the set $\mathbb{M}_{\mathrm{atn}}$ of all TTNs and ATNs, respectively, both derived from $\mathscr{L}(C_{ij}, 2)$ according to Section 3.2.3. Next, each of the sets $\mathbb{M}_{\mathrm{ttn}}$ and $\mathbb{M}_{\mathrm{atn}}$ is randomized according to the procedure described in Section 3.2.4 (Baseline B2). In this way, we obtain the sets $\mathbb{M}'_{\mathrm{ttn}}$ and $\mathbb{M}'_{atn}$ as the randomized counterparts of $\mathbb{M}_{ttn}$ and $\mathbb{M}_{atn}$. As a result, we distinguish a range of classification experiments (1–14) only a subset of which will be conducted in Section 4 to tackle Hypothesis 1. We start with distinguishing TTNs from ATNs. The underlying classification hypothesis is as follows.

Hypothesis 2. Topic networks of the same layer (also called mode) (i.e., TTN or ATN) are more similar than networks of different modes (this concerns Scenario 1 (observed data) and Scenario 6 (randomized data) in Figure 16).

The similarity of TNs will be quantified by means of the apparatus of Section 3.2.6. Regardless of which genre (*urban* vs. *regional* vs. *encyclopedic communication*) the underlying corpus belongs to, Hypothesis 2 assumes that one can always distinguish TTNs from ATNs by their structure, while TTNs and ATNs are less distinguishable among themselves. This scenario is depicted in Figure 14 by Arrow 1. If we falsify the alternative to this hypothesis, we can assume that (poor, rich, or moderate) thematic intertextuality, as manifested by TTNs, is different from coauthorship-based networking of topics in ATNs. Collaboration- and intertextuality-based networking would then differ in a way that characterizes their layer. In order to test genre sensitivity as disregarded by Hypothesis 2, we carry out two experiments: one in which we classify TTNs (ATNs) by genre and one in which we combine both classifications by simultaneously classifying by genre and layer. When classifying by genre, we distinguish TNs derived from city wikis (*urban communication*), regional wikis (*regional communication*), and subnetworks of Wikipedia (*knowledge communication*) (see Section 3.2.2). Finally, we generate two control classes of wikis and Wikipedia-based networks outside of these three genres. The corresponding wikis are sampled in a way that their members are rather dissimilar. Our similarity measurement should therefore not work with them. In a nutshell, the underlying classification hypothesis is as follows.

Hypothesis 3. Topic networks of the same genre are more similar than those of different genres (this concerns Scenarios 2–4 (observed) and Scenarios 7–9 (random data) in Figure 16).

As we consider the genre-sensitive classification in the context of the layer-sensitive one, we get different classification scenarios:

(1) Scenario 2 in Figure 16 denotes the task of training a classifier that detects TTNs of the same genre while distinguishing TTNs of different ones. If this is successful, we can assume that the TTNs analyzed here are genre-sensitive or that the communication functions that we hypothetically associate with these genres influence the structure of these TTNs.

(2) Scenario 3 from Figure 16 regards the analog experiment for the genre-sensitive classification of ATNs.

(3) Scenario 4 concerns the alternative in which the modal difference of TTNs and ATNs is ignored in order to classify topic networks independently of their modal difference according to their underlying genre.

(4) This scenario is contrasted with Scenario 5, which considers classifiers for simultaneously detecting the genre and layer of TNs. The underlying classification hypothesis is as follows.

Hypothesis 4. Topic networks of the same layer and genre are more similar than networks of different layers or genres (this concerns Scenario 5 (observed data) and Scenario 10 (random data) in Figure 16).

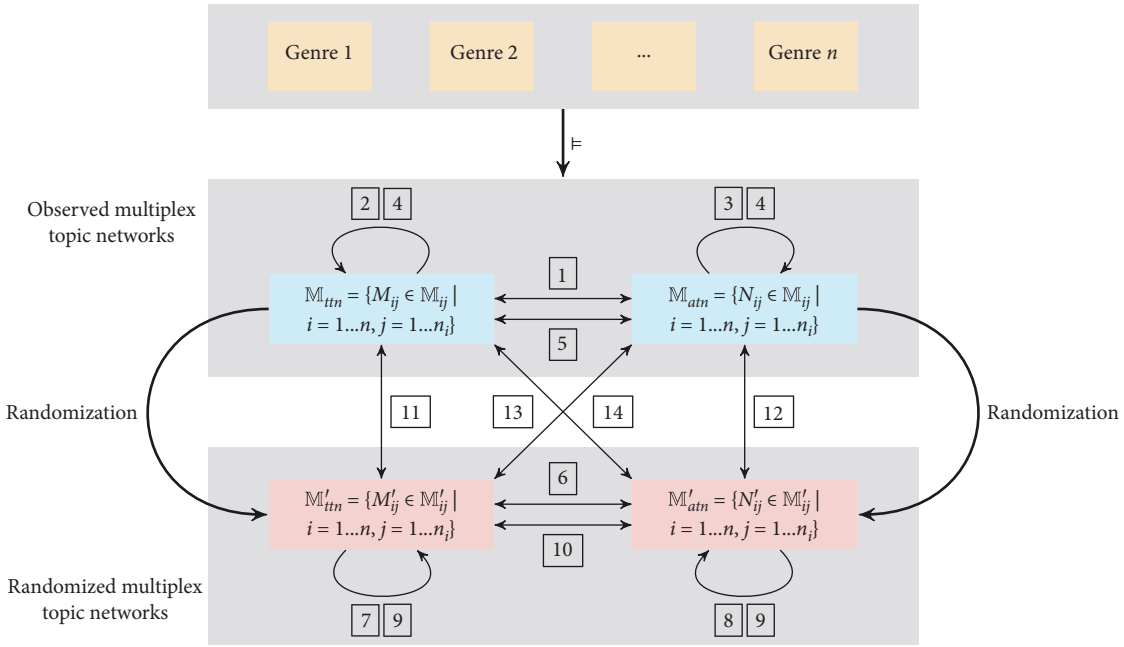


FIGURE 16: From sets of corpora of different genres to multiplex topic networks and their randomizations: corpora of different genres are the starting point for spanning LMNs which are then used to derive two-layer multiplex topic networks (⊨). In a second step, randomized counterparts according to Section 3.2.4 are derived from these MTNs to obtain a further basis for evaluating their significance. In this way, we arrive at fourteen candidate scenarios for classifying topic networks.

Falsifying the alternative to part (2) of Hypothesis 1 implies that TNs derived from corpora written by different communities by addressing different thematic frames (e.g., cities) appear nevertheless similar in their gestalt. Such a finding is very unlikely in cases in which the underlying corpora serve very different communication functions: Hypothesis 1 is not saying that everything is similar irrespective of the heterogeneity of the underlying function or the thematic orientation. Thus, a genre-oriented classification that shows that TNs of the same genre (serving a certain communication function and having a certain thematic orientation) are more similar than those belonging to different genres would rather correspond to such a finding. From this point of view, Hypotheses 3 and 4 are of interest: to deal with them experimentally could pave the way for testing the second part (2) of Hypothesis 1.

As explained in Section 3.2.4, we randomize input networks so that we obtain five additional classification scenarios labeled 6–10 in Figure 16. The experiments corresponding to these scenarios will be conducted here, as far as they concern the baseline scenario B2 of Section 3.2.4. Furthermore, scenarios are to be enumerated which attempt to distinguish observed networks directly from their randomized counterparts. In this context, Scenario 11 aims at distinguishing TTNs from their randomized counterparts by means of the classifiers trained to detect TTNs. Analogously, Scenario 12 considers ATNs in relation to their randomized counterparts, while Scenario 13 aims to separate observed topic networks (whether ATNs or TTNs) from randomized ones. Finally, Scenario 14 extends the latter scenario by trying to additionally account for the modal difference of ATNs and TTNs. These scenarios are only listed for theoretical reasons.

4. Experimentation

To test Hypothesis 1 and its relatives (i.e., Hypotheses 2–4), we conduct several experiments using two resources: a corpus of special wikis, called the *Frankfurt Regional Wiki Corpus*, and a corpus of subnetworks of Wikipedia that mostly contain information about cities and regions.

4.1. Tools and Resources. The *Frankfurt Regional Wiki Corpus* (FRWC) contains 43 wikis collected from online wiki lists (e.g., https://de.wikipedia.org/wiki/Regiowiki). Table 1 shows the statistics of this corpus, which is divided into three genres: CITIES relates to wikis describing certain cities, REGIONS includes wikis focusing on a specific region, while the residual class OTHERS collects wikis that are not off-topic w.r.t. regional communication but are unusual in their structure or the described rhemes. We consider only articles that are not redirects. Wiki authors use redirect pages to lead readers of articles with outdated, incorrect, or alternative spelling titles to the desired target page. We remove all such redirects and rewire all affected links accordingly. As a result, the number of processed articles is smaller than their overall number (see Table 1). In addition to the FRWC, we extracted a corpus of Wikipedia subgraphs (see Section 3.2.2 for the formal definition of these graphs and Table 3 for the corpus statistics). Subsequently, we denote the two variants in this Wikipedia corpus WP-REGIO-1 and WP-REGIO-2. We choose 25 articles

about cities or regions matching the titles of the wikis in the FRWC and additionally include the subgraphs of six off-topic articles to build two additional corpora, called WP-Others-1 and WP-Others-2, for purposes of comparison.

We process the content, link structure, and metadata (e.g., authorship-related information) of all articles in our corpora. This includes their history, that is, the chains of revisions which led to their current state. We do not consider past states of link structure and content itself but incorporate the authorship and the amount of content being added or removed per revision (see Section 3.2.3). The wikis considered here are based on MediaWiki. The structure of their articles varies from wiki to wiki so that HTML-based extractions are error-prone. To circumvent this problem, we use WikiDragon [115], a Java-based framework for importing and processing wikis offline.

For our experiments we used, adapted, and newly developed several tools including the so-called *GeneticClassifierWorkbench* (GCW), a Python library for performing feature selections and sensitivity analyses in classification experiments. Since our experiments are based on feature vectors with a size of sometimes more than 100 features, a complete sensitivity analysis of all feature combinations was not possible. Therefore, we conducted a genetic search for the best-performing subset of features due to maximizing the *F*-score. That is, a population of p features is evaluated and mutated over a number of t rounds. Instances which score best are saved unchanged for the next round and partly added in a slightly mutated form. The worst-performing instances are removed and replaced by random feature combinations. The Workbench is based on the Python library *scikit-learn* [116], allowing us to abstract from the underlying machine learning paradigm so that the same genetic search can be applied to optimize different classifiers. We experimented with neural networks which produced similar results on our test data but took too much time to be used for genetic searches and random baseline computations. Therefore, we decided for *Support Vector Machine*s (SVM) as the embedded method of supervised learning using the *Radial Basis Function* (RBF) as a kernel. Our source code is open source on GitHub (https://github.com/texttechnologylab/GeneticClassifierWorkbench).

4.2. Classification Experiments. We investigate the similarities of our seven corpora of regional wikis (Cities, Regions, and Others) and of Wikipedia-based subgraphs (WP-Regio-1, WP-Regio-2, WP-Others-1, and WP-Others-2) (each defining a corpus of texts) in order to test Hypothesis 1 and its derivatives, that is, Hypotheses 2–4. Thus, we distinguish up to seven target classes in our experiments. For reasons of simplicity, we call each element of these corpora *wiki* and each of the seven classes *genre*. Unless otherwise stated, the experiments are performed on all of them. In the case of WP-Regio-2 and WP-Others-2, we did not induce the corresponding ATNs, as some of these would have included several million edit events. Thus, in this case, we have at most five target classes. Each experiment includes three consecutive steps:

(1) *The all variant:* the first step, denoted by *all*, is a hyperplane parameter optimization and evaluation using the entire feature set. The optimized parameters of the respective classifier are then used in subsequent steps. Ideally, the parameters are optimized independently for each step, but this would have slowed down the genetic search.

(2) *The opt variant:* in the 2nd step, denoted by *opt*, genetic searches for optimal feature subsets are performed using a population of 20 feature vector instances and 50 rounds, trying to maximize the *F*-score of the classification. Note that these searches may only reach a local maximum.

(3) *The ext variant:* for experiments which are not conducted on random baseline data, we perform an extended genetic search for optimal feature subsets based on 20 instances and 500 rounds. In an additional step, a bit-wise genetic optimization attempts to further minimize the number of used features while keeping or even improving the *F*-score, using 20 instances and 500 rounds.

4.2.1. Graph-Similarity-Based Classification. Using the apparatus of Section 3.2.6, each TN (ATN or TTN) of each MTN is represented by a vector of values indicating its similarities to the wikis of the underlying experiment. Any such vector is separately computed for each of the 11 similarity measures of Table 5. Thus, if $\mathbb{T}$ is the set of all TNs of whatever mode (ATN or TTN) and genre (Cities, Regions, etc.) and if $\mathbb{T}' \subseteq \mathbb{T}$ is a subset of these TNs used in a classification experiment concerning the genres (target classes) $\text{Genre}\,i_1, \ldots, \text{Genre}\,i_j$ (cf. Figure 16), then each topic network $T \in \mathbb{T}'$ is represented for each similarity measure by a $|\mathbb{T}'|$-dimensional feature vector which is processed by the three-step algorithm described above. If for a given similarity measure the topic networks derived from wikis of the same genre are mapped to neighboring similarity vectors, then they belong to overlapping neighborhoods in vector space: *related networks are similar in their similarity and dissimilarity relations.* In this way, TNs of the same genre should become as recognizable as TNs of different genres. Now we see why a genetic search for optimal subsets of features is necessary: the reason is that otherwise we would assume that all dimensions of our feature vectors are equally informative—an assumption that is probably wrong.

Relating to Hypothesis 3, Tables 6 and 7 summarize our findings regarding the genre-sensitive classification of TTNs and ATNs, respectively. Cosine-based measures always perform best. Especially in the case of ATNs we see that accounting for arcs *and* for nodes secures better performance: dual weight-dependent measures (see Section 3.2.6) outperform single weight-dependent or weight-insensitive measures. However, in the case of TTNs, we also see that as long as we do not perform an extended optimization (ext), the measure $\cos_{\mathcal{AV}}[\neg w, \infty, \phi_1, L_{12}]$, which disregards arc weights, is a best performer. Of special interest is

Table 1: Statistics of the FRWC showing the number of articles with (#art. 1) and without (#art. 2) redirects, the number of revisions (#rev.), and the number of distinct authors (#authors).

Wiki	#art.1	#art.2	#rev.	#authors
Baden-Baden	999	844	3,576	138
Boppard	24	23	107	17
Cuxhaven	2,884	2,722	28,284	619
Dresden	11,479	9,796	76,776	2,702
Erfurt	2,275	2,267	30,314	129
Esslingen	252	219	2,646	353
Fürth	9,686	8,055	109,467	2,546
Görlitz	1,897	1,735	11,412	555
Hamm	16,602	14,439	99,307	1,353
Karlsruhe	38,870	25,575	306,143	11,002
Köln	3,925	3,184	13,394	400
Linz	6,776	4,250	28,923	343
Lüneburg	105	96	422	108
Lustenau	812	553	3,185	241
München	20,344	15,829	111,681	8,016
Münster	4,096	3,703	24,226	984
Olsberg	376	360	2,403	140
Reutlingen	583	545	3,122	368
Schiltach	505	489	560	14
Schorndorf	1,035	1,005	4,778	73
Strausberg	3,906	3,668	12,860	111
Stuttgart	1,260	1,076	6,784	228
Tübingen	4,749	4,211	38,540	1,513
Weißenburg	436	393	5,436	63
Wulfen	746	722	23,218	767
Würzburg	22,432	17,661	283,773	2,726

Wiki	#art.1	#art.2	#rev.	#authors
Ahrweiler	24,194	22,814	149,345	690
Attersee/Attergau	922	813	17,944	53
Dithmarschen	2,155	1,712	29,981	185
Ennstal	12,774	11,936	76,721	135
Franken	5,511	4,510	78,371	887
Göttingen	8,695	7,755	36,393	488
Niederbayern	33,751	20,504	196,525	1,392
Pforzheim-Enz	14,763	12,821	67,604	3,213
Rhein-Main	5,276	2,801	17,290	40
Rhein-Neckar	12,241	10,413	62,830	2,807
Sachenanhalt	4,644	4,173	36,264	1,153
Waldviertel	266	264	1,906	124

Wiki	#art.1	#art.2	#rev.	#authors
Graz	10,226	9,436	35,490	32
RegioWikiAT	12,085	8,551	113,436	3,221
Wallis	3,174	3,149	18,054	86
Wetzikon	1,737	1,302	23,999	446
Wien-Geschichte	45,473	43,919	296,467	402

Note: the last three columns disregard redirecting articles. Left table: genre CITIES; upper right: genre REGIONS; lower right: genre OTHERS. The German, Austrian, and Swiss wikis were downloaded in early 2018.

$\cos_{\mathscr{AV}}[w, \infty, \phi_2, L_{12}]$, the best performer regarding the classification of ATNs (Table 7), which is not only arc and node sensitive but also weights nodes as a function of their degree centrality and therefore covers the highest amount of structural information among all candidates considered here. This measure is also a robust candidate working at a high level in both experiments (it is the 2nd best performer in the case of TTNs if being optimized by an extended genetic search). Thus, we conclude that spherical measures clearly outperform GED-related approaches and especially network-topology-based approaches (ToSi and NetSimile) which perform worst: the kind of information we seek is apparently ignored or "abstracted away" by the latter measures. However, NetSimile has at least a high optimization potential (see the column ext in Table 6)—a potential which is missing in the case of ToSi. In any event, none of the measures considered here is outperformed by our baselines. But in Table 6, we also see that B3 (opt) approaches ToSi (all); in Table 7, we make analog observations also by example of other measures. A serious problem concerns NetSimile in relation to Baseline B2 regarding the classification of ATNs (Table 7): the baseline surpasses the topology-related measure whether being optimized (opt) or not (all). The graph indices collected by NetSimile have obviously difficulties in making observed networks distinguishable from their random counterparts—at least in some of the cases considered

Table 2: Building blocks of topic networks (texts, topics, words, agents, etc.), their relations according to Figure 8, and candidate procedures for weighting the corresponding arcs (last column).

Source	Relation	Target	Candidate procedure
Text	$\xrightarrow{\theta}$	Topic	text2ddc [72]
Topic	$\xrightarrow{\theta^{-}}$	Text	text2ddc^{-1}
Text	$\xrightarrow{\nu_1}$	Text	Measures of sentence/text similarity, text embeddings [78]
Agent	$\xrightarrow{\vartheta}$	Topic	Topic models [59]
Topic	$\xrightarrow{\vartheta^{-}}$	Agent	Topic models [59]
Agent	$\xrightarrow{\nu_{2,1}}$	Text	Edit networks [74]
Text	$\xrightarrow{\nu_{1,2}}$	Agent	Edit networks [74]
Agent	$\xrightarrow{\nu_2}$	Agent	Coauthorship [74, 79]
Word	$\xrightarrow{\vartheta}$	Topic	text2ddc [72], topic models [59]
Topic	$\xrightarrow{\vartheta^{-}}$	Word	text2ddc^{-1}, topic models [59]
Word	$\xrightarrow{\nu_{3,1}}$	Text	fastText, topic models [59]
Text	$\xrightarrow{\nu_{1,3}}$	Word	fastText, topic models [59]
Word	$\xrightarrow{\nu_3}$	Word	Word embeddings [77, 80–82]
—	—	—	—

Table 3: Wikipedia-based corpora: number of content articles (#articles n), revisions (#revisions n), and authors (#authors n) of non-redirecting articles in WP-Regio-1 ($n = 1$) and WP-Regio-2 ($n = 2$) of the German Wikipedia dump from 201807-01 (subgraphs 1–25); the variable n codes the nth orbit (see formula (30)). Subgraphs 26–31 are used to generate the corpora WP-Others-1 and WP-Others-2.

	Seed article	#articles 1	#revisions 1	#authors 1	#articles 2	#revisions 2	#authors 2
1	Ahrweiler	90	66,217	16,772	11,413	5,602,327	930,621
2	Dithmarschen	210	156,862	38,180	30,386	10,006,785	1,506,634
3	Dresden	1,615	1,180,743	239,747	127,675	27,746,644	3,566,957
4	Erfurt	943	850,786	179,282	100,052	23,644,822	3,158,299
5	Fürth	504	598,687	130,445	77,663	19,481,686	2,657,440
6	Görlitz	790	468,641	99,606	62,896	17,305,177	2,431,331
7	Göttingen	922	786,663	170,082	93,726	22,448,816	2,995,497
8	Hamm	764	697,437	150,502	82,099	20,436,567	2,799,384
9	Karlsruhe	1,021	842,723	180,652	97,984	23,178,185	3,103,192
10	KöIn	1,485	1,090,676	223,801	122,446	26,851,098	3,483,785
11	Linz	816	602,346	130,520	79,376	20,188,052	2,792,374
12	Metropolregion Rhein-Neckar	296	157,356	37,960	23,250	8,608,771	1,388,939
13	München	1,421	1,077,626	216,774	120,725	26,727,317	3,472,725
14	Munster	1,139	894,916	193,090	103,436	24,330,809	3,251,427
15	Niederbayern	239	142,392	33,551	22,466	7,796,961	1,222,744
16	Rhein-Main-Gebiet	390	297,276	65,804	42,238	12,750,028	1,870,354
17	Sachsen-Anhalt	603	459,933	96,116	59,565	16,392,237	2,291,304
18	Schorndorf	362	226,153	51,264	32,562	11,738,799	1,746,169
19	Steirisches Ennstal	43	19,702	6,322	4,400	2,101,467	386,487
20	Strausberg	265	215,854	49,617	30,284	10,602,198	1,579,390
21	Stuttgart	1,317	1,089,313	215,788	123,906	26,648,581	3,403,376
22	Tübingen	623	385,288	85,266	54,525	15,884,637	2,265,358
23	Wetzikon	204	145,207	33,914	20,607	8,044,399	1,306,780
24	Wien	1,380	874,419	170,952	102,792	23,357,095	3,087,254
25	Würzburg	959	885,109	185,495	106,381	24,484,274	3216,674
26	Hydraulik	121	59,874	19,400	8,287	3,600,636	700,341
27	Integralrechnung	194	75,082	21,787	6,708	2,663,563	508,606
28	Kernkraftwerk	287	196,202	49,279	20,773	8,195,232	1,387,491
29	Neuronales Netze	85	27,878	9,750	3,739	1,488,680	332,714
30	Schlacht bei Waterloo	200	97,290	25,614	18,674	6,990,403	1,097,749
31	Zecken	112	58,582	16,350	7,500	3,896,913	734,269

here. B3 is also of interest with regard to the classification of ATNs, which achieves F-scores of up to 40% and thus makes representation models based on measures such as NetSimile, ToSi, and *wges* problematic candidates. The values of B4 opt are also remarkably high and can therefore be regarded as a challenge for the measures.

Figure 17 shows that the baselines B1, B3, and B4 are outperformed by the results obtained for TTNs. However, it

Table 4: Estimates of the average number of active authors per Wikipedia article ($\overline{|✎(\cdot,\cdot)|}$) and the average activity of authors per article ($\overline{✎(\cdot,\cdot)}$) differentiated for the complete set of articles in the German Wikipedia (downloaded at 2018-07-01) with and without redirect articles (numbers of articles in parentheses).

Corpus of articles	$\overline{\lvert✎(\cdot,\cdot)\rvert}$	$\overline{✎(\cdot,\cdot)}$
Without redirects (2,195,812)	27.34	234.52
With redirects (3,657,483)	17.07	226.61

also shows that feature optimization affects the random baselines. This is particularly evident in the case of B3, which is based on random matrices. This gain in F-score can be explained by random numbers that allow the target classes to be separated—at least to some extent. These features are then selected by the genetic feature selection. The baseline results for ATNs show a similar picture (see Figure 17(b)). Regarding B2, we make the following observations in Figure 17(b) (for reasons of complexity, we did not consider all measures to compute B2): although the best B2 candidates are better than the average F-scores calculated on the basis of real data, B2 is clearly surpassed on average. Thus, we come to the conclusion that we found effective measures for comparing networks—this concerns in particular the spherical approach based on the cosine measure. From these experiments, we conclude the following:

(1) Hypothesis 3 is not falsified: we know the genre of a topic network by its structure. Note that this only concerns Scenarios 2 and 3 of Figure 16—Scenario 4 is not computed here. Similarly, by calculating our baselines, this also involves Scenarios 7 and 8 while ignoring Scenario 9. The classification benefits especially from information that is explored by dual weight-dependent measures. This holds regardless of the mode (ATN or TTN).

(2) Spherical measures should be preferred to GED-based measures and these in turn to topology-based measures:

$$\text{spherical} \succ \text{GED} \succ \text{topological}. \tag{73}$$

The boxplots in Figure 18 give another perspective on the classification results by summarizing the distributions of precision and recall values generated by the graph similarity measures. Except for the results on ATN using all features, the average precision is higher than the average recall. The figure also demonstrates the strong effect of feature selection.

So far, we considered classifications as a whole and thus abstracted from the scores obtained for individual genres. The boxplots in Figure 19 give insights into these genre-related scores regarding the classification of TTNs by means of the extended feature optimization (ext). The members of the genre Cities are well identified: in terms of recall and precision. The genre Regions is far less separable and causes many classification errors (low recall). Apparently, this class contains more heterogeneous TTNs. In any event, the Wikipedia-based genres WP-Regio-1 and WP-Regio-2 are very well separated. By contrast, instances of the category Others are extremely difficult to detect (as predicted in Section 3.2.7). Similarly, elements of the classes WP-Others-1 and WP-Others-2 are difficult to identify—albeit to a minor degree. Thus, we conclude that the upper bound of separability concerns Wikipedia-based regional wikis. The corresponding subgraphs are very similar. This upper bound is approached by city wikis. Region wikis are less homogeneous, making the corresponding class Regions rather blurred and therefore question its status as a genre. Figure 20 shows the corresponding results of classifying ATNs. The general picture is quite similar to that of the TTNs.

We take another perspective on the results to examine classification errors. The best results on TTNs using all features is achieved by $\cos_{\mathscr{MV}}[\neg w, \infty, \phi_1, \mathbb{L}_{12}]$. Figure 21 shows to what degree wikis of a target class are wrongly classified using this measure. The labels show the proportion of the categories according to the gold standard (top) and the classification result (bottom). The picture is diverse, but some details become clear: wikis of the classes Regions and Others are often falsely categorized as Cities. City wikis on the other hand are wrongly classified as WP-Others-1 or WP-Regio-1.

Genetic feature selection has proven to increase F-score significantly. In the extended optimization (ext), the last step is to minimize the number of features used. Since our features stand for similarities to networks, we have to ask whether some of the wikis underlying these networks are more relevant for the differentiation of the target classes than others—possibly because of their prototypical status. If all wikis were equally important, an equal distribution of the frequencies with which these features are selected by the genetic optimization would be expected. Figure 22 shows the corresponding rank frequency distribution: it shows that we are far from evenly distributed features. From this, we conclude that the selection of features is indispensable and that the underlying wikis are very different in their roles in our classification experiments.

Next, we try to distinguish TTNs from ATNs thereby addressing Hypothesis 2 (or more specifically, Scenario 1 of Figure 16). The error analysis in Figure 23 shows that networks of these two modes are not separable using our approach. Table 8 differentiates this outcome by reporting the results obtained for different measures. It shows that this classification scenario is far exceeded by Baseline B1 and is therefore irrelevant. From this result, we conclude that ATNs are so similar to their corresponding TTNs that they cannot be distinguished by our measures, or alternatively, our similarity measures are not suitable to distinguish them. This is not surprising, as the order and the size of an ATN always correspond to the order and the size of the TTN from which it was derived, so that they can only differ by the weighting of their nodes and arcs. By concerning Hypothesis 4 and thus by distinguishing twelve target classes (in the case of WP-Others-2 and WP-Regio-2, we do not induce ATNs), Table 8 shows a somehow different scenario: though the F-scores are still rather low, Baseline B1 is clearly outperformed when using a cosine measure for graph similarity measurement. From this observation, we conclude that while Hypothesis 2 is

Table 5: The list of measures of graph similarity used for computing the similarities of topic networks.

	Measure	Approach	Formula	Reference
1	GES	Graph edit similarity	(51)	[100]
2	WAL	Graph edit similarity	(52)	[102]
3	VEO	Vertex and edge overlap	(54)	[103]
4	wges	Weighted graph edit similarity	(60)	
5	$\cos_{\mathscr{A}}[w, \infty, \phi_1, L_{12}]$	Cosine graph similarity	(64)	
6	$\cos_{\mathscr{A}\mathscr{V}}[w, \infty, \phi_1, L_{12}]$	Cosine graph similarity	(66)	
7	$\cos_{\mathscr{A}\mathscr{V}}[w, \infty, \phi_2, L_{12}]$	Cosine graph similarity	(66)	
8	$\cos_{\mathscr{A}}[\neg w, \infty, \phi_1, L_{12}]$	Cosine graph similarity	(64)	
9	$\cos_{\mathscr{A}\mathscr{V}}[\neg w, \infty, \phi_1, L_{12}]$	Cosine graph similarity	(66)	
10	NetSimile	Topological similarity	(70)	[107]
11	ToSi	Topological similarity	(72)	

Table 6: F-scores of classifying TTNs into seven target classes (Cities, Regions, Others, WP-Regio-1, WP-Regio-2, WP-Others-1, and WP-Others-2) by means of SVMs using RBF kernels.

	Measure	all	opt	ext	B1	B3 all	B3 opt	B4 all	B4 opt
1	GES	0.653	0.753	0.798	0.143	0.130	0.286	0.121	0.213
2	WAL	0.649	0.751	0.788	0.143	0.130	0.286	0.109	0.216
3	VEO	0.677	0.773	0.816	0.143	0.130	0.286	0.120	0.221
4	*wges*	0.559	0.620	0.650	0.143	0.130	0.286	0.120	0.199
5	$\cos_{\mathscr{A}}[w, \infty, \phi_1, L_{12}]$	0.638	0.722	0.764	0.143	0.130	0.286	0.119	0.211
6	$\cos_{\mathscr{A}\mathscr{V}}[w, \infty, \phi_1, L_{12}]$	0.729	0.768	**0.853**	0.143	0.130	0.286	0.125	0.223
7	$\cos_{\mathscr{A}\mathscr{V}}[w, \infty, \phi_2, L_{12}]$	0.694	0.766	0.832	0.143	0.130	0.286	0.127	0.229
8	$\cos_{\mathscr{A}}[\neg w, \infty, \phi_2, L_{12}]$	0.642	0.681	0.717	0.143	0.130	0.286	0.122	0.212
9	$\cos_{\mathscr{A}\mathscr{V}}[\neg w, \infty, \phi_2, L_{12}]$	**0.742**	**0.773**	0.790	0.143	0.130	0.286	0.102	0.156
10	NetSimile	0.479	0.629	0.722	0.143	0.130	0.286	0.127	0.229
11	ToSi	0.390	0.433	0.465	0.143	0.130	0.286	0.108	0.229

Column "all": F-scores, if all features are used by the similarity measure (row). Column "opt": F-scores, if a subset of features selected by the genetic search is used. Column "ext": F-scores, if a subset of features selected by the extended genetic search is used. The last five columns display the F-scores of the random baselines B1, B3, and B4, in the case of B3 and B4 differentiated for the variants *all* and *opt*.

Table 7: F-scores of classifying ATNs into five classes (Cities, Regions, Others, WP-Regio-1, and WP-Others-1) by means of SVMs using RBF kernels.

	Measure	all	opt	ext	B1	B2 all	B2 opt	B3 all	B3 opt	B4 all	B4 opt
1	GES	0.598	0.649	0.752	0.200	0.226	0.325	0.182	0.397	0.176	0.294
2	WAL	0.610	0.635	0.707	0.200	0.168	0.222	0.182	0.397	0.158	0.289
3	VEO	0.636	0.706	0.783	0.200	0.213	0.306	0.182	0.397	0.170	0.308
4	*wges*	0.458	0.576	0.618	0.200	0.311	0.348	0.182	0.397	0.173	0.281
5	$\cos_{\mathscr{A}}[w, \infty, \phi_1, L_{12}]$	0.567	0.673	0.737	0.200	—	—	0.182	0.397	0.173	0.300
6	$\cos_{\mathscr{A}\mathscr{V}}[w, \infty, \phi_1, L_{12}]$	**0.740**	0.777	0.854	0.200	0.242	0.440	0.182	0.397	0.181	0.320
7	$\cos_{\mathscr{A}\mathscr{V}}[w, \infty, \phi_2, L_{12}]$	0.612	**0.816**	**0.875**	0.200	—	—	0.182	0.397	0.187	0.340
8	$\cos_{\mathscr{A}}[\neg w, \infty, \phi_2, L_{12}]$	0.559	0.600	0.652	0.200	—	—	0.182	0.397	0.182	0.307
9	$\cos_{\mathscr{A}\mathscr{V}}[\neg w, \infty, \phi_2, L_{12}]$	0.721	0.811	0.865	0.200	0.240	0.464	0.182	0.397	0.182	0.317
10	NetSimile	0.467	0.507	0.610	0.200	0.494	0.602	0.182	0.397	0.173	0.272
11	ToSi	0.431	0.567	0.585	0.200	-		0.182	0.397	0.179	0.254

Column "all": F-scores using all features in terms of the respective similarity measure. Column "opt": using a subset of features detected according to a genetic search. Column "ext": subset selection according to extended genetic optimization. Additionally, F-scores of random baselines B1, B2, B3, and B4 are displayed, in the latter three cases differentiated for the variants *all* and *opt*.

falsified, there is at least a potential regarding the simultaneous distinction of genre and mode: ATNs do not uniformly resemble their corresponding TTNs.

So far we considered part (2) of Hypothesis 1 by showing that TTNs (and also ATNs) with similar functions resemble each other, while differing from networks of other genres. It remains to be shown that these networks are also thematically focused—in a highly skewed manner. To test this, we fit power laws to the distributions of node weights in TTNs. Remember that these weights result from detecting textual instances of the topic represented by the respective node so that the more such instances are detected, the more salient the topic in the network. Fitting a power law to such a distribution means that there is a minority of topics or just one topic that surpasses all other topics in its importance, while the

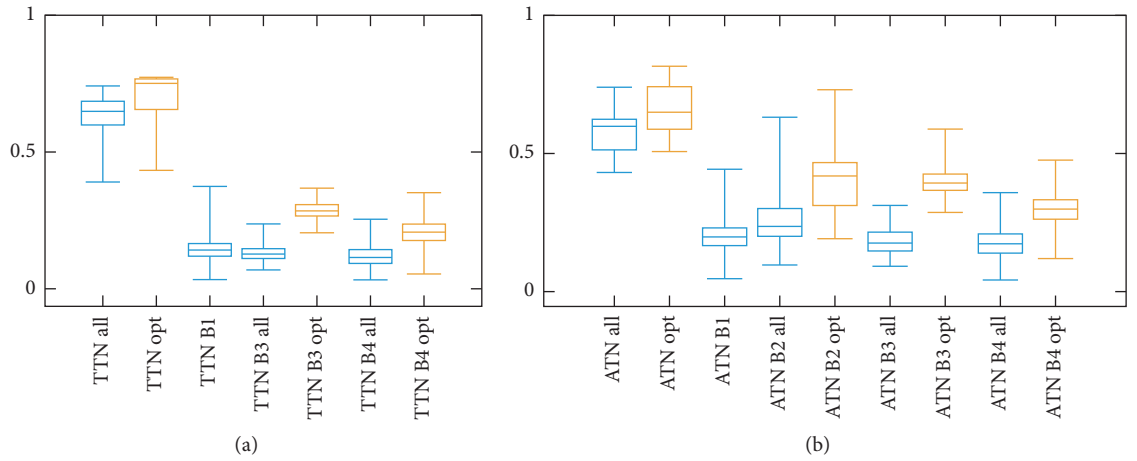


Figure 17: (a) Boxplots of F-scores obtained for classifying TTNs contrasted by the baselines B1, B3, and B4. (b) Boxplots of F-scores obtained for classifying ATNs contrasted by the baselines B1, B2, B3, and B4.

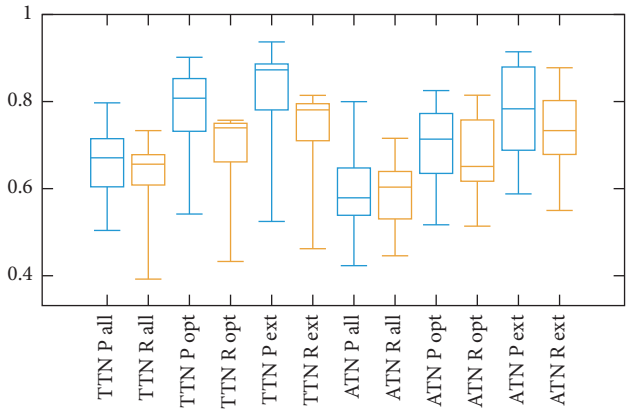


Figure 18: Boxplots of precision P and recall R values (y-axis) induced by the measures of Table 5 and underlying the F-scores of Table 6 (first six columns) and Table 7 (last six columns). Distributions are distinguished by considering all features (all) or subsets of them generated by the genetic optimization *opt* or *ext*.

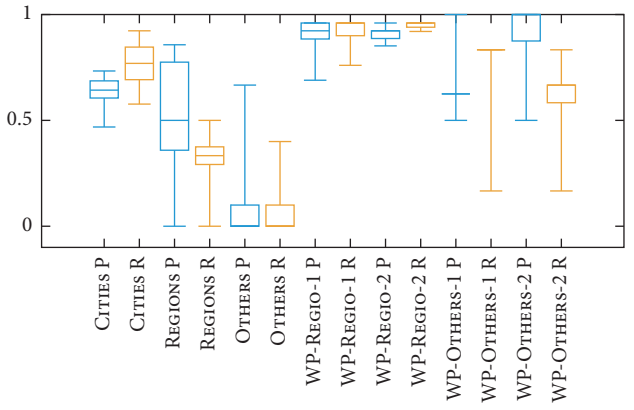


Figure 19: Boxplots of precision (P) and recall (R) values (y-axis) induced by the measures of Table 5 underlying the F-scores of Table 6. Distributions are distinguished by the respective target class of the classification.

majority of topics are of little or no importance. The boxplots in Figure 24(a) show the distribution of the exponents of the power laws fitted to these distributions, differentiated by the genres considered here. To assess the goodness of the fittings, we compute the adjusted R-squares and display the value distributions in Figure 24(b). Obviously, the fits are very good (the adjusted R-squares are on average above 95%) while the averages of the exponents range between 0.5 and 1.5: from this analysis, we conclude that the underlying wikis are all thematically focused and skewed by dealing with a minority of topics in depth. The five most detected DDC labels per genre are shown in Table 9. It shows that *Transportation; ground transportation* is by far the most dominant topic in city wikis and in region wikis. *Obviously, these wikis are thematically focused in a highly skewed manner.*

It remains to be shown that our findings about urban wikis neither depend on the distances of the

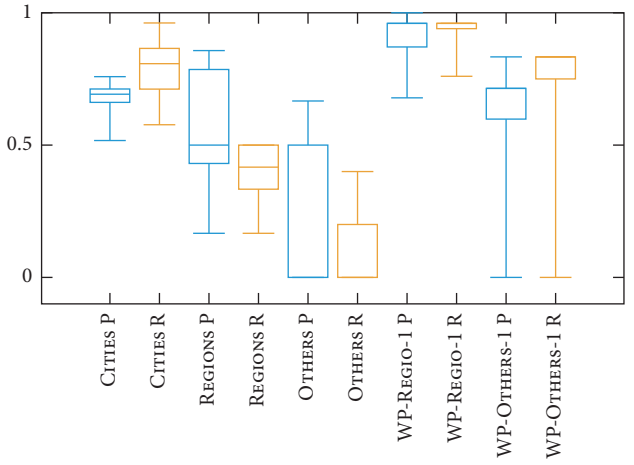


Figure 20: Boxplots of precision P and recall R values (y-axis) induced by the measures of Table 5 underlying the F-scores of Table 7. Distributions are distinguished by the respective target class of the classification.

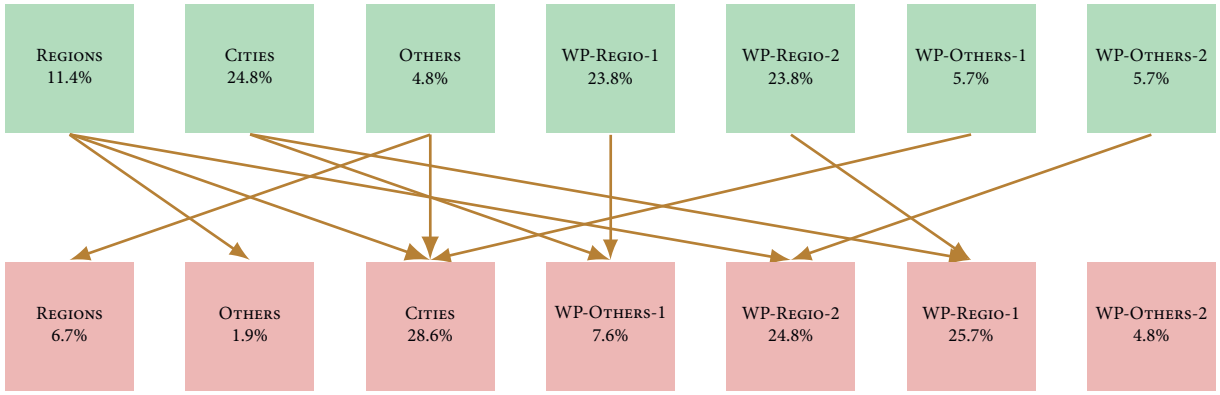


Figure 21: Error analysis regarding the classification of TTNs by means of $\cos_{\mathscr{AV}}[\neg w, \infty, \phi_1, \mathbb{L}_{12}]$.

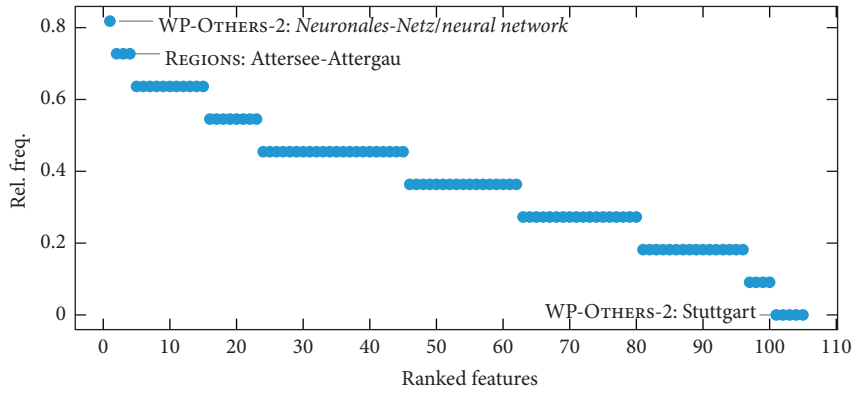


Figure 22: Ranking of the relative frequencies of features as a result of being selected by the extended genetic feature optimization in the classification of TTNs.

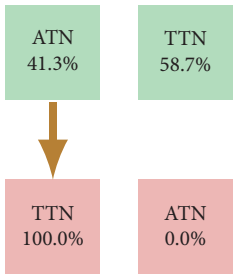


Figure 23: Error analysis regarding the classification of TTNs vs. ATNs by means of $\cos_{\mathscr{AV}}[\neg w, \infty, \phi_1, \mathbb{L}_{12}]$.

corresponding places nor on the communities writing these wikis. Figure 25 shows that the similarities detected by us do hardly correlate with the underlying distances of the places. In the heatmap in Figure 25(a), a connection between two city wikis is the greener, the closer, and the more similar they are to each other, while a pair of wikis is the more red, the less similar, and the more distant they are. Similarity is measured by $\cos[w, \infty, \phi_1, \mathbb{L}_{12}]$ while distance is converted into closeness and normalized to the unit interval (the values of the heatmap scale to $[-1, 1]$ by calculating $-1 +$ closeness + similarity). Figure 25(b) shows that there is hardly a tendency to being more similar when being more close to each other. The lower similarity values are mostly induced by the rather unusually small wikis such as Boppard (see Table 1). Figure 26 shows the Fuzzy Jaccard of the communities underlying the wikis, that is, the overlap of these communities weighted by the activities of their authors: the lower the number of shared authors of two wikis and the less active these authors, the lower the fuzzy overlap of these wikis. The Fuzzy Jaccard is computed as follows (cf. [117]): let $\text{authors}(\mathbb{W})$ be the set of all registered users contributing to any of the wikis in $\mathbb{W} = \text{Cities} \cup \text{Regions}$, $\text{Others} \cup \text{WP-Regio-1} \cup \text{WP-Others-1}$ and let $\text{texts}(W)$ be the set of all (nonredirect) articles of wiki $W \in \mathbb{W}$, then we compute

$$\forall A, B \in \mathbb{W} : J_\mu(A, B) = \frac{\sum_{r \in \text{authors}(\mathbb{W})} \mu_{A \cap B}(r)}{\sum_{r \in \text{authors}(\mathbb{W})} \mu_{A \cup B}(r)} \in [0, 1], \tag{74}$$

where

Table 8: Left: *F*-scores obtained for different measures and optimizations by classifying ATNs vs. TTNs according to Scenario 1 of Table 16—two target classes are considered. B1 considers Scenario 6 of Figure 16. Right: *F*-scores obtained for different measures and optimizations by classifying simultaneously for mode and genre according to Scenario 5—twelve target classes are considered. B1 considers Scenario 10.

	Measure	all	opt	ext	B1
1	GES	0.370	0.370	0.370	0.500
2	VEO	0.370	0.370	0.370	0.500
3	$\cos[w, \infty, \phi 1, L_{12}]$	0.370	0.370	0.370	0.500
4	$\cos[\neg w, \infty, \phi_1, L_{12}]$	0.370	0.370	0.370	0.500

	Measure	all	opt	ext	B1
1	GES	0.152	0.178	0.194	0.082
2	VEO	0.181	0.228	0.259	0.082
3	$\cos[w, \infty, \phi_1, L_{12}]$	0.315	0.363	0.407	0.082
4	$\cos[\neg w, \infty, \phi_1, L_{12}]$	0.284	0.339	0.409	0.082

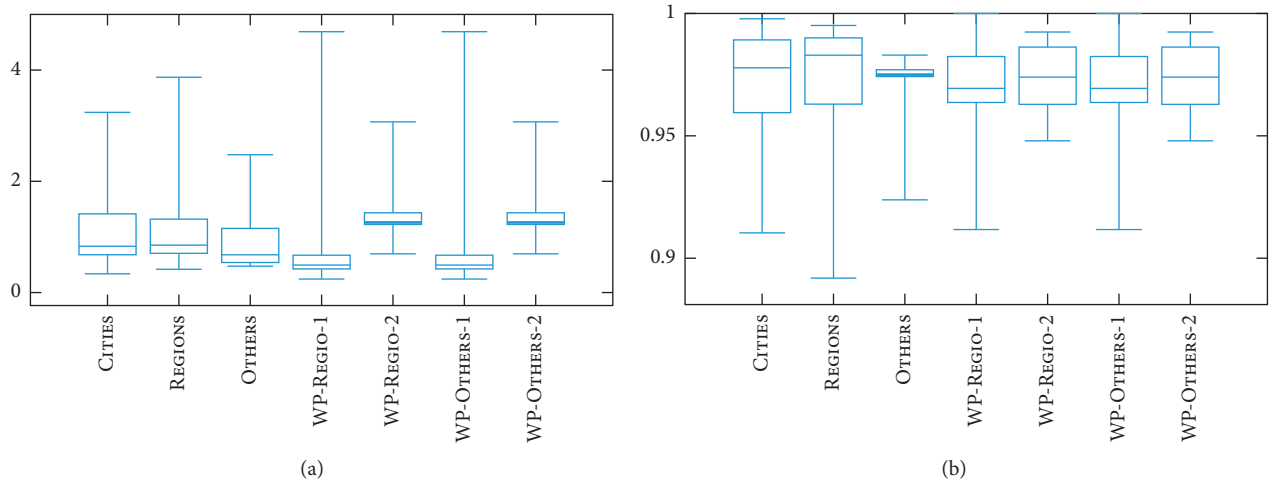
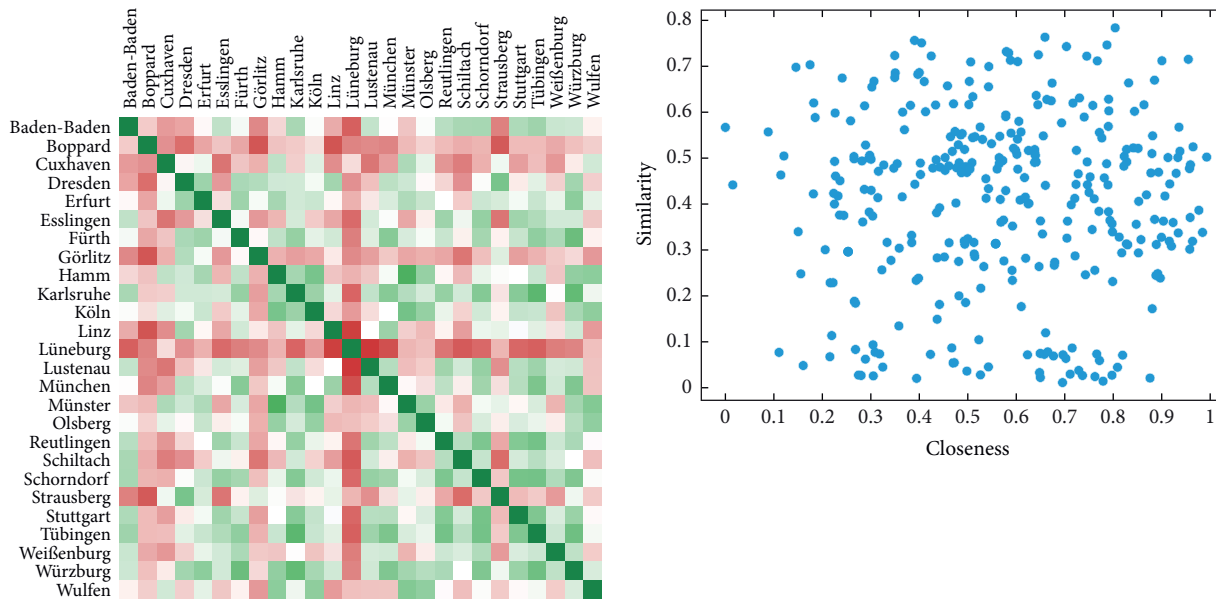


Figure 24: (a) Boxplots of the distribution of the exponents of the power laws fitted to the weight distributions of the nodes in the TTNs differentiated by the target classes. (b) The boxplots of the corresponding fitting accuracies computed by means of the adjusted *R*-squared.

Table 9: The five most detected DDC labels for the genres Cities, Regions, and Others.

Rank	Genre	Node weight sum	Avg weight	DDC	Description
1	City	10,325.830	397.147	388	Transportation; ground transportation
2	City	2,404.631	92.486	943	Central Europe; Germany
3	City	1,570.010	60.385	726	Buildings for religious purposes
4	City	1,512.536	58.174	725	Public structures
5	City	964.262	37.087	711	Area planning
1	Region	5,127.546	427.296	388	Transportation; ground transportation
2	Region	1,692.267	141.022	943	Central Europe; Germany
3	Region	1,385.013	115.418	726	Buildings for religious purposes
4	Region	1,289.722	107.477	551	Geology, hydrology & meteorology
5	Region	1,171.656	97.638	796	Athletic & outdoor sports & games
1	Other	5,335.555	1,067.111	929	Genealogy, names & insignia
2	Other	1,640.042	328.008	726	Buildings for religious purposes
3	Other	715.084	143.017	723	Architecture from ca. 300 to 1399
4	Other	701.298	140.260	725	Public structures
5	Other	680.309	136.062	720	Architecture



(a) (b)

FIGURE 25: (a) The heatmap of thematic similarity and spatial closeness among city wikis. Red means that the wikis are thematically dissimilar and distant in space; green means that they are thematically similar and close in space. (b): the distribution of the similarities (y-axis) as a function of the closenesses (x-axis) of the different pairs of city wikis.

$$\mu_{A\cap B}(r) = \min\left(\frac{\sum_{x\in \text{texts}(A)} \measuredangle(r,x)}{\sum_{s\in \text{authors}(A)}\sum_{x\in \text{texts}(A)} \measuredangle(s,x)}, \frac{\sum_{x\in \text{texts}(B)} \measuredangle(r,x)}{\sum_{s\in \text{authors}(B)}\sum_{x\in \text{texts}(B)} \measuredangle(s,x)}\right),$$
$$\mu_{A\cup B}(r) = \max\left(\frac{\sum_{x\in \text{texts}(A)} \measuredangle(r,x)}{\sum_{s\in \text{authors}(A)}\sum_{x\in \text{texts}(A)} \measuredangle(s,x)}, \frac{\sum_{x\in \text{texts}(B)} \measuredangle(r,x)}{\sum_{s\in \text{authors}(B)}\sum_{x\in \text{texts}(B)} \measuredangle(s,x)}\right). \quad (75)$$

Figure 25 shows that while among the Wikipedia-based extractions the overlap is remarkably high, it does nearly not exist between any of the city or region wikis: these wikis are written by mostly completely different communities. The picture is not different if one considers all authors—registered and unregistered.

5. Discussion

Section 4 has shown that topic networks, whether TTNs or ATNs, are similar if they belong to the same genre, while they are characterized by a high degree of thematic focusing. In order to operationalize this notion of network similarity, we tested further or newly developed 11 different measures of network similarity by relying on four different paradigms of measuring the similarity of graphs (see Table 5 and the discussion of graph/network similarity measures in Section 3.2.6) as instantiated by the complex networks studied here. All these measures and paradigms come along with a different notion of network similarity. We have shown that a subclass of them, especially cosine-based measures of network similarity, allow for detecting similarities of topic networks in line with Hypotheses 3 and 4. At the same time, the concept of network similarity underlying this class of dual weight-dependent measures seems to be the most promising from a research point of view, as it is based on node and arc weights and instantiates a very intuitive concept of network similarity: The more similar the two networks are from the perspective of the more of their nodes, the more similar they are. Thus, at the level of thematic abstraction examined here, there seems to be a hidden tendency to write about very prominent topics when it comes to thematizing places and linking the underlying texts in such a way that the resulting networks become almost indistinguishable.

Starting from this kind of thematic distortion of VGI as conveyed by online media, we now ask for a more general explanation of our findings. The candidate we are considering for this purpose is given by *Cognitive Maps* (CM), which were introduced as models of the cognitive representation and processing of spatial information to explain a number of different cognitive biases. Because of bridging the gap between geographical information and its biased representation, CMs promise to be a candidate for our task. At the same time, this notion allows for the connection of cognitive geography on the one hand and our generalized model of linguistic encoding of geographical information on the other (see Figure 1). The reason is that as mental representations, CMs are seen to integrate a wide range of representations of spatial objects, their relations, and

Figure 26: The Fuzzy Jaccard overlap of the communities of registered authors of the wikis in the corpora Cities, Regions, Others, WP-Regio-1, and WP-Others-1 weighted by means of the writing activities of the authors: the greener the link, the higher the fuzzy overlap.

thematic units (see below). We may argue now that we developed a method to represent and analyze a particular type of thematic information which can be subsumed under the latter list. If this is true, then the thematic distortion observed by us could be seen as a result of the biased processing of geographic information by a community of agents dealing with the same place to generate a common cognitive map, thereby manifesting a particular type of distributed cognition. When creating such a common CM of the same place, agents tend to focus on a highly selected set of rhemes (see Figure 1), even if there is no explicit agreement among these agents about this selection and even if there is little or no direct communication between them and also irrespective of the focal place. It seems that the agents participate in processes of distributed cognition in such a way that their own thematically distorted maps flow into the formation of a shared, stable but likewise distorted "thematic map." These maps then appear as the result of a sort of swarm behavior regarding the formation of a particular distribution of the preference and salience of certain place-related rhemes. From this perspective, topic networks serve as models of these thematic maps which in turn are parts of CMs. To underpin this interpretation, we briefly summarize the research on CMs and, above all, ask about distortions that are distinguished by the research in this area.

Understood as mental representations of spatial knowledge, CMs have been subject of scientific work for

decades. Starting from different disciplinary perspectives, this research provides insights into how people perceive their environment, think about it, and how this influences their spatial behavior. The interdisciplinary research on CMs has led to a multitude of notions, research designs, and outcomes, the integration of which is still pending. Over the years, researchers worked, for example, with different terms for the mental representations in question such as *cognitive maps* [118], *environmental images* [119], *mental maps* [120], *mental sketch maps* [121], *narrative space maps* [122], or *internal representations* [123], where the constituent *map* is most common. However, there has been a discussion as to whether the term *map* is generally misleading. In this context, Kitchin ([124], 3 pp.) distinguishes approaches that understand CMs as

(1) Three-dimensional maps

(2) An analogy to maps (because of their map-like characteristics)

(3) A metaphor for maps (because they function as if they were maps), or

(4) A hypothetical construct used to explain spatial behavior

While we refer to cognitive maps as an auxiliary notion, we adhere to the fourth of these variants. Regardless of this discussion, there is a greater consensus on some characteristics of CMs as mental representations: CMs are understood as complexes of mental images and concepts that humans have in mind when thinking about places, their *location* (in terms of distance and direction), *accessibility* (regarding questions like how to get there), and the *meanings* associated with them. They serve as a means of understanding spatial circumstances and as a frame of reference for the interpretation, preference, and prediction of spatial structures, their relations, and events in which they participate (see [125], 100 pp, 313), ([120], 3), and ([119], 5p.). Beyond that, they also serve as a basis for decision-making regarding spatial behavior (e.g., in route planning). In a nutshell, humans activate, generate, and utilize CMs in spatial thinking and spatial behavior (cf. [126], 233). CMs are distinguished according to the entities they model. Kitchin and Blades ([127], 5p) distinguish CMs of *object spaces* (e.g., rooms and cars), *environmental spaces* (e.g., buildings, streets, neighborhoods, and cities), *geographical spaces* (e.g., regions and countries), *panoramic spaces*, and *map spaces* (including models) (cf. [128]). In this way, they cover existing as well as imagined places, where facts about the former can be mixed with imaginations of the latter [129]. This list includes the kind of places that are central to our study, especially cities.

To build a bridge between the notion of CMs and our analysis, we need to look more closely at their content and the principles by which they are created. Generally speaking, CMs are seen to cover at least two types of information (see [124] 1p. and [129] 314p.):

(1) Regarding *spatial cognition*, this concerns information about where entities are located in the environment of a person (location, distance, and direction in relation to her location or to reference points like landmarks)

(2) Regarding *environmental cognition*, this concerns information about the kind of these entities, their attributes, meanings, valuations, and attitudes that the person associates with them—individually, socially, or culturally mediated ([126], 224, 235)

Our study focuses on the second part of this distinction: it is related to the rhemes that are associated with places as framing themes (see Section 1). In any event, CMs are systematically characterized by distortions ([129], 315) concerning judgments about locations, distances, and directions as well as the formation of preferences which affect spatial or environmental cognition. One example is the *localization effect* [120] according to which people can discriminate nearby places better and have stronger preferences for them, see also [126]. This relates to errors in distance judgments depending on the perspective from which they are made: more differences are seen between closer areas than between more distant ones, so that shorter distances are exaggerated, while longer distances are underestimated [130], 133). Furthermore, spatial knowledge can be organized by reference to landmarks which "distort" places in their "neighborhood" so that buildings, for example, are judged to be closer to them than vice versa [130], 134). Tversky ([130], 135pp.] describes additional modes of distortion: to remember the position and orientation of objects, humans isolate them from their background and organize them by referring to a general frame of reference (rotation) or to other figures (alignment). While these examples primarily concern spatial cognition, the following bias focuses more on environmental cognition. This concerns the hierarchical organization of conceptual systems according to which places of the same category are supposed to be closer in distance than places of different categories, while the direction of a category (with a direction slot) determines the one of its members ([130], 132p). Last but not least, Golledge and Stimson [126] describe distortions of the representation of urban spaces. They observe that interactions influence the perception of a city in the sense that spatial information accumulates along the representations of the paths used to carry out these interactions. Likewise, structural properties of cities which are more salient than others are likely to become anchor points in CMs. In such maps, areas between used paths and anchor points may appear to be "folded" or "wrapped" so that preferred visited places are represented closer to each other. As a result, positional and relational errors can occur in perception (see ([126], 254) and ([131], 7).

To interpret our findings in the light of this research, we need to link the formation of CMs with linguistic processes. The idea that this formation is substantially influenced by human language processing, so that geographical information is nontrivially encoded in linguistic structure, goes back to the work of Louwerse and Benesh (cf. [26]) (see Section 1; see also Montello and Freundschuh ([132], 171)

for an earlier hint on "*obtain[ing] spatial knowledge through language*"). In this context, Golledge and Stimson ([126], 235) distinguish shared components of CMs from personalized ones by stating that "*The common elements facilitate communication with others about the characteristics of an environment; the idiosyncratic elements provide the basis of the personalized responses to such situations*." Our hypothesis is now that at the level of thematic abstraction as modeled here, the organization of platial rhemes shared by the members of a community is influenced by the general law of preferential order, which is most prominently instantiated by Zipf's first law [133]. Such an organization makes the anticipation of a place rather expectable among the members of a community so that communication about this place is facilitated as predicted by Golledge and Stimson [126].

This Zipfian organization allows for relating our findings to the well-known power-law-like degree distributions found in many natural, social, semiotic, or technical networks (see [109, 134] and especially [135] for overviews of this and related research) and also by example of many linguistic systems—especially on the text level [136–138]. Because of this commonality, one might assume that we just detected a well-known *text* or *network* characteristic. Characteristic for our findings, however, is that we developed a measurement procedure that detects a *text (corpus)-related semantic, thematic trend*—with the help of network theory: instead of counting directly observable arcs, for example, in ontological networks or co-occurrences in texts and instead of relying on monoplex networks [70, 93, 139–143], we generated and analyzed a range of different networks in relation to each other in order to determine the corresponding thematic trend by means of multiplex networks. This is not to say that we first discovered a Zipfian process in the organization of linguistic networks, but rather that we observe such a process in a very specific area, in which it has not been observed before and which requires an appropriate explanation as elaborated so far. Indeed, if thematic salience is skewed, and if skewed topic distributions derived from different corpora are similar not only topologically but also regarding the ranking of the majority of salient topics, such an observation requires explanation subject to the fact that the underlying text networks are constituted by different, distributed communities of authors. It is the answer to this question that the paper was about.

At this point, one might further object that we made a rather expectable observation in the sense that descriptions of cities, for example, are very likely related to rhemes like traffic, trade, culture, and history. However, this would mean underestimating our results: (i) the thematic distortions observed by us are extremely skewed, (ii) they seem to emerge rather earlier in the development of a wiki (this is not shown here but is the result of a pretest in which we looked at the life cycles of three different wikis; in future work, we will analyze the underlying time series of multiplex topic networks in detail), and (iii) they make both members of the same genre similar while allowing for distinguishing members of different genres. To phrase it as a question: *If the number of rhemes under which places are thematized is limited, why then should always a tiny subset of them dominate the discourse about a place and why then should the networking of these rhemes make discourses of the same genre identifiable?* From this point of view, we argue that we discovered an additional form of the distortion of CMs, which means that the underlying place is always conceptualized from the point of view of a few but extremely preferred rhemes. When organizing their distributed processes of coauthorship, communities of authors seem to strive to a kind of thematic unification that makes different wikis serving alike functions looking structurally similar—with respect to the preference order of themes and their networking. It seems that people participate in processes of collaborative writing with a tendency to organize their thematic contributions and references in such a way that they remain shareable [144] and communicable among members of the same community. Ensuring shareability means securing the continued existence of the underlying wiki, which could otherwise collapse because of too many personalized or individualized fragmentations. At this point, we can speculate that people unconsciously prefer such thematic contributions that make their social roles and participations expectable and acceptable, whereby this selection behavior produces the described similarity of thematic maps as components of CMs. In other words, the participants anticipate social roles and neglect their personal view of cities and regions, whose documentation would fragment the corresponding media thematically. Instead, they ignore the reproduction of their idiosyncratic, personalized views of places. To say it in terms of the distinction made by Golledge and Stimson [126] between shared and personalized components of CMs: participants overweight the former to the disadvantage of the latter to guarantee the shareability [144, 145] of CMs as a result of distributed cognition.

Note that in our study we did not simply map a frequency effect by our measurements: although we counted frequencies of topic assignments, they were determined by means of an inference process that went through a process of (machine) learning. To support such an interpretation, however, a deeper analysis with a larger corpus of wikis and related media providing different functions is required. This also requires experiments with other and above all much finer classification systems than the DDC to find out how much the use of the DDC has influenced our measurements. And it requires a deeper analysis of the social roles of authors in online media, their interactions, and the regulatory systems under which they interact. But this already concerns future work.

6. Conclusion

We developed a novel model of topic networks in order to investigate the networking of rhemes addressing the same places in underlying corpora of natural language texts. We developed our network model in a way that it enables thematic comparisons of previously unforeseen text corpora using an underlying reference corpus, offers a generic solution to the problem of topic labeling, is highly scalable and can therefore map even the smallest text snippets to topic

000	Computers, internet & systems	100	Philosophy	200	Religion	300	Social sciences, sociology & anthropology
010	Bibliographies	110	Metaphysics	210	Philosophy & theory of religion	310	Statistics
020	Library & information science	120	Epistemology	220	The Bible	320	Political science
030	Encyclopedias & books of facts	130	Astrology, parapsychology & the occult	230	Christianity & christian theology	330	Economics
040	Unassigned	140	Philosophical schools of thought	240	Christian practice & observance	340	Law
050	Magazines, journals & serials	150	Psychology	250	Christian pastoral practice & religious orders	350	Public administration & military science
060	Associations, organizations & museums	160	Logic	260	Church organization, social work & worship	360	Social problems & social services
070	Journalism, publishing & news media	170	Ethics	270	History of christianity	370	Education
080	Quotations	180	Ancient, medieval & Eastern philosophy	280	Christian denominations	380	Commerce, communications & transportation
090	Manuscripts & rare books	190	Modern western philosophy	290	Other religions	390	Customs, etiquette & folklore
400	Language	500	Science	600	Technology	700	Arts
410	Linguistics	510	Mathematics	610	Medicine	710	Landscaping & area planning
420	English & old English languages	520	Astronomy	620	Engineering	720	Architecture
430	German & related languages	530	Physics	630	Agriculture	730	Sculpture, ceramics & metalwork
440	French & related languages	540	Chemistry	640	Home & family management	740	Drawing & decorative arts
450	Italian, Romanian & related languages	550	Earth sciences & geology	650	Management & public relations	750	Painting
460	Spanish & Portuguese languages	560	Fossils & prehistoric life	660	Chemical engineering	760	Graphic arts
470	Latin & Italic languages	570	Biology & life sciences	670	Manufacturing	770	Photography
480	Classical & modern Greek languages	580	Plants (Botany)	680	Manufacturing specific products	780	Music
490	Other languages	590	Animals (Zoology)	690	Building & construction	790	Sports, games & entertainment
800	Literature, rhetoric & criticism	900	History				
810	American literature in English	910	Geography & travel				
820	English & old English literatures	920	Biography & genealogy				
830	German & related literatures	930	History of the ancient world (to ca. 499 A.D.)				
840	French & related literatures	940	History of Europe (ca. 500 A.D. -)				
850	Italian, Romanian & related literatures	950	History of Asia				
860	Spanish & Portuguese literatures	960	History of Africa				
870	Latin & Italic literatures	970	History of North America				
880	Classical & modern Greek literatures	980	History of South America				
890	Other literatures	990	History of other regions				

Figure 27: Color codes of the classes of the 2nd level of the DDC.

distributions, simultaneously takes rare topics into account, and is methodologically open and expandable. Moreover, our model allows for comparatively investigating the networking of thematic units from different angles. In this way, it is open and expandable as it allows for integrating different analytical perspectives into the study of the same semantic networks. We exemplified our model by means of corpora of special wikis and extracts from Wikipedia in order to investigate how textual information encodes geographical information on the aboutness level of texts. Our experiments show that the thematizations of different places on a certain level of abstraction are similar to each other in that they focus on a few themes in a highly distorted manner while networking them in similar ways. This happens regardless of whether the underlying media are generated by different communities and whether these communities address related or unrelated places in nearby or distant places. We interpreted our findings in the context of the notion of cognitive maps. To this end, we proposed to extend this notion in terms of thematic maps and argued that participants or interlocutors of online communication tend to organize their contributions in a way that makes them sharable. This means that the contributions are abstracted and depersonalized at the aboutness level in such a way that the social roles of these participants become expectable and acceptable, while their personal views of places are reduced whose documentation would fragment the corresponding media thematically. Ensuring shareability means securing the continued existence of the wiki, which could otherwise collapse in the face of too many personalized or individualized fragmentations. Future work concerns several tasks: we want to conduct deeper analyses based on larger corpora that manifest a greater variety of communication functions in order to shed more light on the genre sensitivity discovered in our study. Beyond the DDC, we strive for the use of finer structured, higher resolution classification systems in order to model the contents of texts much more precisely. Ideally, this should be carried out with the help of systems like the category system of Wikipedia or even Wikidata, both of which develop as open topic universes [146]. Last but not least, a deeper analysis of the social roles of authors in online media and their coauthorship is required to gain a deeper understanding of the processes of linguistic encoding of geographical information. This will be the task of future work.

Appendix

A. text2ddc

text2ddc is trained by means of corpora that are derived by integrating information from Wikidata, Wikipedia, and the *Integrated Authority File* (*Gemeinsame Normdatei*—GND) of the German National Library: we explore the links of Wikipedia articles to entries in Wikidata containing the property attribute https://www.wikidata.org/wiki/Property:P1036 that directly links to the DDC or to a GND page containing a DDC tag. An example is the article about the *Pythagorean theorem* (https://en.wikipedia.org/wiki/Pythagorean_theorem), which is linked to the GND

page 4176546-1 (https://d-nb.info/gnd/4176546-1) referring to the DDC tag 516 (*geometry*). Using such information, we obtain a corpus for a subset of 98 classes of the 2nd and for a subset of 641 classes of the 3rd DDC level. Since Wikipedia exists for many languages, such corpora can be created for each of them. For preprocessing the input data of text2ddc, we use TextImager [86] and fastSense [88] for disambiguating these data on the sense level. The resulting information is used to train a neural network for classifying any piece of text (down to the word level) into DDC classes (see https://textimager.hucompute.org/DDC/). To this end, text2ddc uses a very efficient classifier, that is, fastText [91], a bag-of-words model to train a neural network with a single hidden layer. To optimize fastText, we optimize the following hyperparameters: learning rate: 0; update rate: 150; minimal number of word occurrences: 5; number of epochs: 10,000. In this way, we increase the *F*-score to 87% for the 2nd level and to 78% for the 3rd level of the DDC.

B. Color Codes and 2nd Class Members of the DDC

Figure 27 shows the colors and labels of the classes of the 2nd level of the DDC.

Data Availability

Parts of the programs that underlie our work are available via GitHub (https://github.com/texttechnologylab/GeneticClassifierWorkbench).

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

Financial support by the Federal Ministry of Education and Research (BMBF) via the *Centre for the Digital Foundation of Research in the Humanities, Social, and Educational Sciences* CEDIFOR is gratefully acknowledged.

References

[1] A. Crooks, D. Pfoser, A. Jenkins et al., "Crowdsourcing urban form and function," *International Journal of Geographical Information Science*, vol. 29, no. 5, pp. 720–741, 2015.

[2] M. Crang and S. Graham, "SENTIENT CITIES ambient intelligence and the politics of urban space," *Information, Communication & Society*, vol. 10, no. 6, pp. 789–817, 2007.

[3] H. Chen, M. Vasardani, S. Winter, and M. Tomko, "A graph database model for knowledge extracted from place descriptions," *ISPRS International Journal of Geo-Information*, vol. 7, no. 6, p. 221, 2018.

[4] M. F. Goodchild, "Citizens as sensors: the world of volunteered geography," *GeoJournal*, vol. 69, no. 4, pp. 211–221, 2007.

[5] M. F. Goodchild and L. Li, "Assuring the quality of volunteered geographic information," *Spatial Statistics*, vol. 1, pp. 110–120, 2012.

[6] D. Hardy, J. Frew, M. F. Goodchild, and Goodchild, "Volunteered geographic information production as a spatial process," *International Journal of Geographical Information Science*, vol. 26, no. 7, pp. 1191–1212, 2012.

[7] D. Z. Sui, "The wikification of GIS and its consequences: or Angelina Jolie's new tattoo and the future of GIS," *Computers, Environment and Urban Systems*, vol. 32, no. 1, pp. 1–5, 2008.

[8] B. Jiang and J.-C. Thill, "Volunteered geographic information: towards the establishment of a new paradigm," *Computers, Environment and Urban Systems*, vol. 53, pp. 1–3, 2015.

[9] M. M. Salvini and S. I. Fabrikant, "Spatialization of user-generated content to uncover the multirelational world city network," *Environment and Planning B: Planning and Design*, vol. 43, no. 1, pp. 228–248, 2016.

[10] B. J. Hecht and D. Gergle, "On the "localness" of user-generated content," in *Proceedings of the 2010 ACM Conference on Computer Supported Cooperative Work, CSCW '10*, pp. 229–232, ACM, New York, NY, USA, February 2010.

[11] M. Graham, B. Hogan, R. K. Straumann, and A. Medhat, "Uneven geographies of user-generated information: patterns of increasing informational poverty," *Annals of the Association of American Geographers*, vol. 104, no. 4, pp. 746–764, 2014.

[12] S. Hahmann and D. Burghardt, "How much information is geospatially referenced? Networks and cognition," *International Journal of Geographical Information Science*, vol. 27, no. 6, pp. 1171–1189, 2013.

[13] A. Jenkins, A. Croitoru, A. T. Crooks, and S. Anthony, "Crowdsourcing a collective sense of place," *PLoS One*, vol. 11, no. 4, Article ID e0152932, 2016.

[14] D. Clare, "Are places concepts? familarity and expertise effects in neighborhood cognition," in *Kathleen Stewart Hornsby, Christophe Claramunt, Michel Denis, and Gérard Ligozat, Spatial Information Theory*, pp. 36–50, Springer, Berlin, Heidelberg, 2009.

[15] A. Mehler, R. . Gleim, A. Lücking, T. Uslu, and C. Stegbauer, "On the self-similarity of Wikipedia talks: a combined discourse-analytical and quantitative approach," *Glottometrics*, vol. 40, pp. 1–45, 2018.

[16] M. M. Louwerse and R. A. Zwaan, "Language encodes geographical information," *Cognitive Science*, vol. 33, no. 1, pp. 51–73, 2009.

[17] D. L. Medin, R. L. Goldstone, and D. Gentner, "Respects for similarity," *Psychological Review*, vol. 100, no. 2, p. 254, 1993.

[18] K. Adamzik, *Textlinguistik: Grundlagen, Kontroversen, Perspektiven*, De Gruyter, Berlin, Germany, 2016.

[19] S. Yablo, *Aboutness*, Princeton University Press, Princeton, NJ, USA, 2014.

[20] K. Brinker, *Linguistische Textanalyse: Eine Einführung in Grundbegriffe und Methoden*, Erich Schmidt, Berlin, Germany, 1992.

[21] F. Daneš, "The paragraph—A central unit of the thematic and compositional build-up of texts," in *Organization in Discourse*, B. Wårvik, S.-K. Tanskanen, and R. Hiltunen, Eds., pp. 29–40, University of Turku, Turku, Finland, 1995.

[22] L. Hoffmann, "Thema, themenentfaltung, makrostruktur," in *Text- und Gesprächslinguistik/Linguistics of Text and Conversation—Ein internationales Handbuch zeitgenössischer Forschung*, K. Brinker, G. Antos, W. Heinemann, and

S. F. Sager, Eds., vol. 1, pp. 344–355, De Gruyter, Berlin, Germany, 2000.

[23] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *Journal of Machine Learning Research*, vol. 3, pp. 993–1022, 2003.

[24] D. Clare, "Reading geography between the lines: extracting local place knowledge from text," in *Spatial Information Theory*, T. Tenbrink, John Stell, A. Galton, and Z. Wood, Eds., pp. 320–337, Springer International Publishing, Cham, Switzerland, 2013.

[25] B. B. Rieger, "Semiotic cognitive information processing: learning to understand discourse. A systemic model of meaning constitution," in *Adaptivity and Learning. An Interdisciplinary Debate*, R. Kühn, R. Menzel, W. Menzel, U. Ratsch, M. M. Richter, and I. O. Stamatescu, Eds., pp. 347–403, Springer, Berlin, Germany, 2003.

[26] M. M. Louwerse and N. Benesh, "Representing spatial structure through maps and language: lord of the rings encodes the spatial structure of middle earth," *Cognitive Science*, vol. 36, no. 8, pp. 1556–1569, 2012.

[27] W. R. Tobler, "A computer movie simulating urban Growth in the Detroit region," *Economic Geography*, vol. 46, pp. 234–240, 1970.

[28] H. J. Miller, "Tobler's first law and spatial analysis," *Annals of the Association of American Geographers*, vol. 94, no. 2, pp. 284–289, 2004.

[29] D. R. Montello, S. I. Fabrikant, M. Ruocco, and R. S. Middleton, "Testing the first law of cognitive geography on point-display spatializations," in *Spatial Information Theory. Foundations of Geographic Information Science*, W. Kuhn, M. F. Worboys, and S. Timpf, Eds., pp. 316–331, Springer, Berlin, Germany, 2003.

[30] B. Hecht and E. Moxley, "Terabytes of Tobler: evaluating the first law in a massive, domain-neutral representation of world knowledge," in *Spatial Information Theory*, K. S. Hornsby, C. Claramunt, M. Denis, and G. Ligozat, Eds., pp. 88–105, Springer Berlin Heidelberg, Berlin, Germany, 2009.

[31] T. J.-J. Li, S. Sen, and B. Hecht, "Leveraging advances in natural language processing to better understand Tobler's first law of geography," in *Proceedings of the 22Nd ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems, SIGSPATIAL '14*, pp. 513–516, ACM, Fort Worth, TX, USA, November 2014.

[32] H. Yang, S. Chen, M. R. Lyu, and I. King, "Location-based topic evolution," in *Proceedings of the 1st International Workshop on Mobile Location-Based Service (MLBS'11)*, pp. 89–98, ACM, Beijing, China, September 2011.

[33] M. M. Louwerse, "Symbol interdependency in symbolic and embodied cognition," *Topics in Cognitive Science*, vol. 3, no. 2, pp. 273–302, 2011.

[34] B. Adams and G. McKenzie, "Inferring thematic places from spatially referenced natural language descriptions," in *Crowdsourcing Geographic Knowledge: Volunteered Geographic Information (VGI) in Theory and Practice*, D. Sui, S. Elwood, and M. Goodchild, Eds., pp. 201–221, Springer, Dordrecht, Netherlands, 2013.

[35] Q. Mei, C. Liu, H. Su, and C.X. Zhai, "A probabilistic approach to spatiotemporal theme pattern mining on weblogs," in *Proceedings of the 15th International Conference on World Wide Web, WWW '06*, pp. 533–542, ACM, Edinburgh, UK, May 2006.

[36] H. Qiang, R. Cai, C. Wang et al., "Equip tourists with knowledge mined from travelogues," in *Proceedings of the 19th International Conference on World Wide Web, WWW '10*, pp. 401–410, ACM, Raleigh, NC, USA, April 2010.

[37] A. R. Bahrehdar and R. S. Purves, "Description and characterization of place properties using topic modeling on georeferenced tags," *Geo-Spatial Information Science*, vol. 23, no. 3, pp. 173–184, 2018.

[38] Z. Yin, L. Cao, J. Han, C. Zhai, and T. Huang, "Geographical topic discovery and comparison," in *Proceedings of the 20th International Conference on World Wide Web (WWW'11)*, pp. 247–256, ACM, Hyderabad, India, March-April 2011.

[39] R. Gabriel and M. Max, "Louwerse. Grounding the ungrounded: estimating locations of unknown place names from linguistic associations and grounded representations," in *Proceedings of the 36th Annual Meeting of the Cognitive Science Society, CogSci 2014*, Quebec City, Canada, July 2014.

[40] M. Speriosu, T. Brown, T. Moon, J. Baldridge, and K. Erk, "Connecting language and geography with region-topic models," in *Proceedings of the Workshop on Computational Models of Spatial Language Interpretation (COSLI)*, vol. 46, Portland, OR, USA, August 2010.

[41] S. Gao, K. Janowicz, and H. Couclelis, "Extracting urban functional regions from points of interest and human activities on location-based social networks," *Transactions in GIS*, vol. 21, no. 3, pp. 446–467, 2017.

[42] G. Lansley and P. A. Longley, "The geography of Twitter topics in London," *Computers, Environment and Urban Systems*, vol. 58, pp. 85–96, 2016.

[43] S. Gao, K. Janowicz, D. R. Montello et al., "A data-synthesis-driven method for detecting and extracting vague cognitive regions," *International Journal of Geographical Information Science*, vol. 31, no. 6, pp. 1245–1271, 2017.

[44] D. R. Montello, "Regions in geography: process and content," *Foundations of Geographic Information Science*, pp. 173–189, 2003.

[45] E. Leopold, "Models of semantic spaces," in *Aspects of Automatic Text Analysis, Volume 209 of Studies in Fuzziness and Soft Computing*, A. Mehler and R. Köhler, Eds., pp. 117–137, Springer, Berlin, Germany, 2007.

[46] T. K. Landauer and S. T. Dumais, "A solution to Plato's problem: the latent semantic analysis theory of acquisition, induction, and representation of knowledge," *Psychological Review*, vol. 104, no. 2, pp. 211–240, 1997.

[47] C. Davies and T. Tenbrink, "Place as location categories: learning from language," in *Proceedings of Workshops and Posters at the 13th International Conference on Spatial Information Theory (COSIT 2017)*, P. Fogliaroni, A. Ballatore, and E. Clementini, Eds., pp. 217–225, Springer International Publishing, Cham,Switzerland, 2018.

[48] Y. Hu, X. Ye, and S.-L. Shaw, "Extracting and analyzing semantic relatedness between cities using news articles," *International Journal of Geographical Information Science*, vol. 31, no. 12, pp. 2427–2451, 2017.

[49] Y. Liu, F. Wang, C. Kang, Y. Gao, and Y. Lu, "Analyzing relatedness by toponym Co-occurrences on web pages," *Transactions in GIS*, vol. 18, no. 1, pp. 89–107, 2014.

[50] D. Ramage, D. Hall, N. Ramesh, D. Christopher, and Manning, "Labeled LDA: a supervised topic model for credit attribution in multi-labeled corpora," in *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing EMNLP '09*, vol. 1, pp. 248–256, Association for Computational Linguistics, Singapore, August 2009.

[51] L. Björneborn, *Small-World link structures across an academic web space: a library and information science approach*, Ph.D. thesis, Royal School of Library and Information

Science, Department of Information Studies, Aalborg, Denmark, 2004.

[52] W. Luo, Y. Wang, X. Liu, and S. Gao, "Cities as spatial and social networks: towards a spatio-socio-semantic analysis framework," in *Cities as Spatial and Social Networks*, X. Ye and X. Liu, Eds., pp. 21–37, Springer International Publishing, Cham, Switzerland, 2019.

[53] P. Agarwal, "Operationalising 'sense of place' as a cognitive operator for semantics in place-based ontologies," in *Spatial Information Theory*, A. G. Cohn and D. M. Mark, Eds., pp. 96–114, Springer, Berlin, Heidelberg, 2005.

[54] S. Winter and C. Freksa, "Approaching the notion of place by contrast," *Journal of Spatial Information Science*, vol. 2012, no. 5, pp. 31–50, 2012.

[55] Y. Hu, "Geospatial semantics," 2017, https://arxiv.org/abs/1707.03550.

[56] M. Dehmer, A. Mehler, and Frank Emmert-Streib, "Graph-theoretical characterizations of generalized trees," in *Proceedings of the 2007 International Conference on Machine Learning: Models, Technologies & Applications (MLMTA'07)*, pp. 113–117, Las Vegas, NV, USA, June 2007.

[57] A. Mehler, "Structural similarities of complex networks: a computational model by example of wiki graphs," *Applied Artificial Intelligence*, vol. 22, no. 7-8, pp. 619–683, 2008.

[58] D. M. Blei, "Probabilistic topic models," *Communications of the ACM*, vol. 55, no. 4, pp. 77–84, 2012.

[59] M. Steyvers, T. Griffiths, T. K. Landauer, D. S. McNamara, S. Dennis, and K. Walter, "Probabilistic topic models," in *Handbook of Latent Semantic Analysis*, pp. 427–448, Lawrence Erlbaum Associates, New York, NY, USA, 2007.

[60] G. Heinrich, *A generic approach to topic models and its application to virtual communities*, Ph.D thesis, University of Leipzig, Leipzig, Germany, 2012.

[61] B.-J. (Paul) Hsu and J. Glass, "Style & topic language model adaptation using HMM-LDA," in *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing, EMNLP' 06*, pp. 373–381, Association for Computational Linguistics, Sydney, Australia, July 2006.

[62] M. Rosen-Zvi, C. Chemudugunta, T. Griffiths, P. Smyth, and M. Steyvers, "Learning author-topic models from text corpora," *ACM Transactions on Information Systems*, vol. 28, no. 1, pp. 1–38, 2010.

[63] A. Herzog, P. John, and S. J. Mikhaylov, "Transfer topic labeling with domain-specific knowledge base: an analysis of UK house of commons speeches 1935-2014," 2018, https://arxiv.org/abs/1806.00793.

[64] S. Boccaletti, G. Bianconi, R. Criado et al., "The structure and dynamics of multilayer networks," *Physics Reports*, vol. 544, no. 1, pp. 1–122, 2014.

[65] M. Stella, N. M. Beckage, M. Brede, and M. De Domenico, "Multiplex model of mental lexicon reveals explosive learning in humans," *Scientific Reports*, vol. 8, no. 1, p. 2259, 2018.

[66] M. A. K. Halliday and R. Hasan, *Language, Context, and Text: Aspects of Language in a Socialsemiotic Perspective*, Oxford University Press, Oxford, UK, 1989.

[67] R. Clarke, "The persistence of systems in organisations," in *Signs of Work. Semiosis and Information Processing in Organisations*, B. Holmqvist, P. B. Andersen, H. Klein, and R. Posner, Eds., pp. 59–91, De Gruyter, Berlin, Germany, 1996.

[68] E. Ventola, *The Structure of Social Interaction: A Systemic Approach to the Semiotics of Service Encounters*, Pinter, London, UK, 1987.

[69] A. Mehler, "Generalized shortest paths trees: a novel graph class applied to semiotic networks," in *Analysis of Complex Networks: From Biology to Linguistics*, M. Dehmer and F. Emmert-Streib, Eds., pp. 175–220, Wiley-VCH, Weinheim, Germany, 2009.

[70] A. Mehler, "Social ontologies as generalized nearly acyclic directed graphs: a quantitative graph model of social ontologies by example of Wikipedia," in *Towards an Information Theory of Complex Networks: Statistical Methods and Applications*, M. Dehmer, F. Emmert-Streib, and A. Mehler, Eds., pp. 259–319, Birkhäuser, Basel, Switzerland, 2011.

[71] F. Sebastiani, "Machine learning in automated text categorization," *ACM Computing Surveys*, vol. 34, no. 1, pp. 1–47, 2002.

[72] T. Uslu, A. Mehler, A. Niekler, and D. Baumartz, "Towards a DDC-based topic network model of wikipedia," in *Proceedings of 2nd International Workshop on Modeling, Analysis, and Management of Social Networks and their Applications (SOCNET 2018)*, February 2018.

[73] T. Uslu and A. Mehler, "PolyViz: a visualization system for a special kind of multipartite graphs," in *Proceedings of the IEEE VIS 2018 (IEEE VIS 2018)*, Berlin, Germany, October 2018.

[74] U. Brandes, P. Kenis, J. . Lerner, and D. van Raaij, "Network analysis of collaboration structure in Wikipedia," in *Proceedings of the 18th International Conference on World Wide Web (WWW'09)*, pp. 731–740, ACM, Madrid, Spain, April 2009.

[75] M. E. J. Newman, "Who is the best connected scientist? A study of scientific coauthorship networks," in *Complex Networks*, E. Ben-Naim, H. Frauenfelder, and Z. Toroczkai, Eds., pp. 337–370, Springer, Berlin Germany, 2004.

[76] G. Salton and C. Buckley, "Term-weighting approaches in automatic text retrieval," *Information Processing & Management*, vol. 24, no. 5, pp. 513–523, 1988.

[77] T. Mikolov, W.-tau Yih, and G. Zweig, "Linguistic regularities in continuous space word representations," in *Proceedings of the NAACL 2013*, pp. 746–751, Redmond, WA, USA, January 2013.

[78] S. Harispe, S. Ranwez, S. Janaqi, and J. Montmain, "Semantic similarity from natural language and ontology analysis," *Synthesis Lectures on Human Language Technologies*, vol. 8, no. 1, pp. 1–254, 2015.

[79] M. E. J. Newman, "Coauthorship networks and patterns of scientific collaboration," *Proceedings of the National Academy of Sciences*, vol. 101, no. 1, pp. 5200–5205, 2004.

[80] A. Komninos and S. Manandhar, "Dependency based embeddings for sentence classification tasks," in *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 1490–1500, San Diego, CA, USA, June 2016.

[81] O. Levy and Y. Goldberg, "Dependency-based word embeddings," in *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, vol. 2, pp. 302–308, Baltimore, MD, USA, June 2014.

[82] L. Wang, C. Dyer, A. Black, and I. Trancoso, "Two/too simple adaptations of word2vec for syntax problems," in *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Association for Computational Linguistics, Denver, CO, USA, June 2015.

[83] A. Budanitsky and G. Hirst, "Evaluating WordNet-based measures of lexical semantic relatedness," *Computational Linguistics*, vol. 32, no. 1, pp. 13–47, 2006.

[84] W. Hemati, T. Uslu, and A. Mehler, "TextImager: a distributed UIMA-based system for NLP," in *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: System Demonstrations*, pp. 59–63, Osaka, Japan, December 2016.

[85] B. Bohnet, J. Nivre, I. Boguslavsky, R. Farkas, F. Ginter, and J. Hajič, "Joint morphological and syntactic analysis for richly inflected languages," *Transactions of the Association for Computational Linguistics*, vol. 1, pp. 415–428, 2013.

[86] T. Uslu, A. Mehler, D. Baumartz, H. Alexander, and W. Hemati, "Fastsense: an efficient word sense disambiguation classifier," in *Proceedings of the 11th Edition of the Language Resources and Evaluation Conference (LREC 2018)*, Miyazaki, Japan, May 2018.

[87] S. Eger, R. . Gleim, and A. Mehler, "Lemmatization and morphological tagging in German and Latin: a comparison and a survey of the state-of-the-art," in *Proceedings of the 10th International Conference on Language Resources and Evaluation, LREC 2016*, Portorož, Slovenia, May 2016.

[88] D. Baumartz, T. Uslu, and A. Mehler, "LTV: labeled topic vector," in *Proceedings of the COLING 2018, the 26th International Conference on Computational Linguistics: System Demonstrations*, The COLING 2018 Organizing Committee, Santa Fe, New Mexico, USA, August 2018.

[89] J. Armand, E. Grave, P. Bojanowski, and T. Mikolov, "Bag of tricks for efficient text classification," 2016, https://arxiv.org/abs/1607.01759.

[90] M. Dehmer, "Information processing in complex networks: graph entropy and information functionals," *Applied Mathematics and Computation*, vol. 201, no. 1-2, pp. 82–94, 2008.

[91] P. Baldi, P. Frasconi, and P. Smyth, *Modeling the Internet and the Web*, Wiley, Chichester, UK, 2003.

[92] A. Mehler, R. . Gleim, W. Hemati, and T. Uslu, "Skalenfreie online soziale Lexika am Beispiel von Wiktionary," in *Proceedings of the 53rd Annual Conference of the Institut für Deutsche Sprache (IDS)*, S. Engelberg, H. Lobin, K. Steyer, and S. Wolfer, Eds., pp. 269–291, De Gruyter, Mannheim, Germany, March 2017.

[93] O. Abramov and A. Mehler, "Automatic language classification by means of syntactic dependency networks," *Journal of Quantitative Linguistics*, vol. 18, no. 4, pp. 291–336, 2011.

[94] L. Geng, M. Semerci, B. . Yener, J. Mohammed, and Zaki, "Graph classification via topological and label attributes," in *Proceedings of the 9th International Workshop on Mining and Learning with Graphs (MLG)*, San Diego, CA, USA, July 2011.

[95] T. Li, D. Han, Y. Shi, and M. Dehmer, "A comparative analysis of new graph distance measures and graph edit distance," vol. 403-404, , pp. 15–21, Information Sciences, 2017.

[96] M. Owen and W. Richards, "Graph comparison using fine structure analysis," in *Proceedings of the 2010 IEEE Second International Conference on Social Computing, SOCIALCOM '10*, pp. 193–200, IEEE Computer Society, Minneapolis, Minnesota, USA, August 2010.

[97] F. Emmert-Streib, M. Dehmer, and Y. Shi, "Fifty years of graph matching, network alignment and network comparison," *Information Sciences*, vol. 346-347, pp. 180–197, 2016.

[98] D. Koutra, A. Parikh, A. Ramdas, and J. Xiang, "Algorithms for graph similarity and subgraph matching," 2011, https://www.cs.cmu.edu/jingx/docs/DBreport.pdf.

[99] D. Koutra, N. Shah, and T. Joshua, "Vogelstein, brian Gallagher, and Christos Faloutsos. DeltaCon: principled massive-graph similarity function with attribution," *ACM Trans. Knowl. Discov. Data*, vol. 10, no. 3, pp. 1–43, 2016.

[100] H. Bunke, P. J. Dickinson, M. Kraetzl, and W. D. Wallis, *A Graph-Theoretic Approach to Enterprise Network Dynamics (Progress in Computer Science and Applied Logic (PCS))*, Birkhäuser, Boston, MA, USA, 2006.

[101] R. Ibragimov, M. Malek, J. Guo, and B. Jan, "GEDEVO: an evolutionary graph edit distance algorithm for Biological network alignment," in *Volume 34 of OpenAccess Series in Informatics (OASIcs)*, T. Beißbarth, M. Kollmar, A. Leha et al., Eds., pp. 68–79, Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik, Dagstuhl, Germany, 2013.

[102] W. D. Wallis, P. Shoubridge, M. Kraetz, and D. Ray, "Graph distances using graph union," *Pattern Recognition Letters*, vol. 22, no. 6-7, pp. 701–704, 2001.

[103] P. Papadimitriou, D. Ali, and H. Garcia-Molina, "Web graph similarity for anomaly detection," in *Proceedings of the 17th International Conference on World Wide Web, WWW '08*, pp. 1167-1168, ACM, Beijing, China, April 2008.

[104] S. Adam, H. Bunke, M. Last, and K. Abraham, *Graph-Theoretic Techniques for Web Content Mining*, World Scientific, Singapore, 2005.

[105] L. C. Freeman, "Centrality in social networks conceptual clarification," *Social Networks*, vol. 1, no. 3, pp. 215–239, 1978.

[106] P. Blanchard and D. Volchenkov, *Mathematical Analysis of Urban Spatial Networks*, Springer, Berlin, Germany, 2009.

[107] M. Berlingerio, D. Koutra, T. Eliassi-Rad, and C. Faloutsos, "Network similarity via multiple social theories," in *Proceedings of the IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, pp. 1439-1440, IEEE, Niagara, Ontario, Canada, August 2013.

[108] S. Soundarajan, T. Eliassi-Rad, and B. Gallagher, "A guide to selecting a network similarity method," in *Proceedings of the 2014 SIAM International Conference on Data Mining*, pp. 1037–1045, SIAM, Philadelphia, PA, USA, April 2014.

[109] M. E. J. Newman, "The structure and function of complex networks," *SIAM Review*, vol. 45, no. 2, pp. 167–256, 2003.

[110] D. J. Watts and S. H. Strogatz, "Collective dynamics of 'small-world' networks," *Nature*, vol. 393, no. 6684, pp. 440–442, 1998.

[111] B. Bollobás and O. M. Riordan, "Mathematical results on scale-free random graphs," in *Handbook of Graphs and Networks. From the Genome to the Internet*, S. Bornholdt and H. G. Schuster, Eds., pp. 1–34, Wiley-VCH, Weinheim, Germany, 2003.

[112] A. Barrat, M. Barthelemy, R. Pastor-Satorras, and A. Vespignani, "The architecture of complex weighted networks," *Proceedings of the National Academy of Sciences*, vol. 101, no. 11, pp. 3747–3752, 2004.

[113] B. Zhang and S. Horvath, "A general framework for weighted gene co-expression network analysis," *Statistical Applications in Genetics and Molecular Biology*, vol. 4, no. 1, 2005.

[114] G. Kalna, J. Desmond, and Higham, "Clustering coefficients for weighted networks," in *Proceedings of the Symposium on network analysis in natural sciences and engineering*, vol. 45, September 2006.

[115] R. Gleim, A. Mehler, Y. Sung, and S. WikiDragon, "A Java framework for diachronic content and network analysis of mediawikis," in *Proceedings of the 11th Edition of the Language Resources and Evaluation Conference*, Miyazaki, Japan, May 2018.

[116] F. Pedregosa, G. Varoquaux, A. Gramfort et al., "Scikit-learn: machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[117] N. Ramli and D. Mohamad, "On the Jaccard index similarity measure in ranking fuzzy numbers," *Matematika*, vol. 25, pp. 157–165, 2009.

[118] E. C. Tolman, "Cognitive maps in rats and men," *Psychological Review*, vol. 55, no. 4, pp. 189–208, 1948.

[119] K. Lynch, *The Image of the City*, Vol. 11, MIT press, Cambridge, MA, USA, 1960.

[120] P. Gould and R. White, *Mental Maps*, Routledge, Abingdon, UK, 1986.

[121] J.J. Gieseking, "Where we go from here: the mental sketch mapping method and its analytic components," *Qualitative Inquiry*, vol. 19, no. 9, pp. 712–724, 2013.

[122] C. Helferich, "Mental maps und narrative raumkarten," in *Methoden der Kulturanthropologie*, K. Oehme-Jüngling and W. L. Christine Bischoff, Eds., pp. 241–256, Haupt Verlag, Bern, Switzerland, 2014.

[123] J. Portugali, "The construction of cognitive maps: an introduction," in *The Construction of Cognitive Maps*, pp. 1–7, Springer, Berlin, Germany, 1996.

[124] R. M. Kitchin, "Cognitive maps: what are they and why study them?," *Journal of Environmental Psychology*, vol. 14, no. 1, pp. 1–19, 1994.

[125] R. M. Downs and D. Stea, *Kognitive Karten. Die Welt in Unseren Köpfen*, Harper & Row, New York, NY, USA, 1982.

[126] R. G. Golledge and R. J. Stimson, *Spatial Behavior: A Geographic Perspective*, Guilford Press, New York, NY, USA, 1996.

[127] R. Kitchin and M. Blades, *The cognition of geographic space*, Vol. 4, Ib Tauris, London, UK, 2002.

[128] S. M. Freundschuh and M. J. Egenhofer, "Human conceptions of spaces: implications for GIS," *Transactions in GIS*, vol. 2, no. 4, pp. 361–375, 1997.

[129] R. M. Downs and David Stea, "Cognitive maps and spatial behaviour: process and product," in *The Map Reader: Theories of Mapping Practice and Cartographic Representation*, D. Martin, R. Kitchin, and C. Perkins, Eds., pp. 312–317, John Wiley & Sons, Hoboken, NJ, USA, 2011.

[130] B. Tversky, "Distortions in cognitive maps," *Geoforum*, vol. 23, no. 2, pp. 131–138, 1992.

[131] R. G. Golledge and T. GÃ¤rling, *Spatial Behavior in Transportation Modeling and Planning*, University of California Transportation Center, Berkeley, CA, USA, 2001.

[132] D. R. Montello and S. M. Freundschuh, "Sources of spatial knowledge and their implications for GIS: an introduction," *Geographical Systems*, vol. 2, no. 1, pp. 169–176, 1995.

[133] G. K. Zipf, *Human Behavior and the Principle of Least Effort: An Introduction to Human Ecology*, Hafner Publishing Company, New York, NY, USA, 1972.

[134] M. E. J. Newman, *Networks: An Introduction*, Oxford University Press, Oxford UK, 2010.

[135] M. Newman, "Power laws, Pareto distributions and Zipf's law," *Contemporary Physics*, vol. 46, no. 5, pp. 323–351, 2005.

[136] S. Naranan and V. K. Balasubrahmanyan, "Models for power law relations in linguistics and information science," *Journal of Quantitative Linguistics*, vol. 5, no. 1-2, pp. 35–61, 1998.

[137] A. Rapoport, "Zipf's law re-visited," in *Studies on Zipf's Law*, H. Guiter and M. V. Arapov, Eds., pp. 1–28, Brockmeyer, Halle, Germany, 1982.

[138] J. Tuldava, *Methods in Quantitative Linguistics*, Wissenschaftlicher Verlag, Trier, Germany, 1995.

[139] D. Raphael Amancio, O. N. Oliveira Jr., and L. F. Costa, "Identification of literary movements using complex networks to represent texts," *New Journal of Physics*, vol. 14, p. 43029, 2012.

[140] A. Baronchelli, R. Ferrer-i-Cancho, R. Pastor-Satorras, N. Chater, and M. H. Christiansen, "Networks in cognitive science," *Trends in Cognitive Sciences*, vol. 17, no. 7, pp. 348–360, 2013.

[141] C. Cattuto, A. Barrat, A. Baldassarri, G. Schehr, and V. Loreto, "Collective dynamics of social annotation," *Proceedings of the National Academy of Sciences*, vol. 106, no. 26, pp. 10511–10515, 2009.

[142] R. F. i Cancho, A. Mehler, O. Pustylnikov, and A. Díaz-Guilera, "Correlations in the organization of large-scale syntactic dependency networks," in *Proceedings of the TextGraphs-2 at NAACL-HLT'07*, Rochester, New York, April 2007.

[143] R. F. i Cancho and R. V. Solé, "The small-world of human language," *Proceedings of the Royal Society of London. Series B, Biological Sciences*, vol. 268, no. 1482, pp. 2261–2265, 2001.

[144] J. Freyd, "Shareability: the social psychology of epistemology," *Cognitive Science*, vol. 7, no. 3, pp. 191–210, 1983.

[145] J. J. Freyd, "What is shareability?," 2005, http://dynamic.uoregon.edu/jjf/defineshareability.html.

[146] A. Mehler and U. Waltinger, "Enhancing document modeling by means of open topic models," *Library Hi Tech*, vol. 27, no. 4, 2009.