Structural Similarities of Complex Networks: A Computational Model by Example of Wiki Graphs

Alexander Mehler
Bielefeld University
Computational Linguistics & Text Technology
Alexander.Mehler@uni-bielefeld.de

Abstract

This paper elaborates a framework for representing and classifying large complex networks by example of wiki graphs. By means of this framework we reliably measure the similarity of document, agent and word networks by solely regarding their topology. In doing so the paper departs from classical approaches to complex network theory which focus on topological characteristics in order to check their small world-property. This does not only include characteristics which have been studied in complex network theory, but also some of those which were invented in social network analysis and hypertext theory. We show that network classifications come into reach which go beyond the hypertext structures traditionally analyzed in web mining. The reason is that we focus on networks as a whole as units to be classified — *above the level of websites and their constitutive pages*. As a consequence, we bridge classical approaches to text and web mining on the one hand and complex network theory on the other hand. Last but not least, this approach also provides a framework for quantifying the linguistic notion of intertextuality.

1 Introduction

With the advent of the Web 2.0 the rise of a novel linguistic unit can be observed which because of its size, structure and complexity has been widely unknown so far. This relates to linguistic networks and their collaborative construction by large communities of interacting agents in computer-mediated communication. These networks induce a further level of discourse structure far away from the focus of prevalent approaches to computational linguistics. The latter approaches have been concerned with discourse internal *micro* and *meso level* structures which range from sentences and turns up to the level of texts and dialogues (Manning and Schütze, 1999; Jurafsky and Martin, 2000). With the availability of Web 2.0 corpora (Mehler, 2008a) complex networks of discourse units as a whole and, thus, *macro* level structures come into the fore as a *new* research object hardly — if at all — accounted for in the past (cf. Mehler 2008a for an overview of such approaches).

In a narrow sense one might say that linguistic networks span a further level of logical document structure (Power et al., 2003) above the level of single (web) documents. This, in turn, suggests a renaissance of classical hypertext theory (Landow, 1987) with its focus on hypertext authoring. In fact, this position is misleading for two reasons: firstly, it neglects the interplay of the semantic and

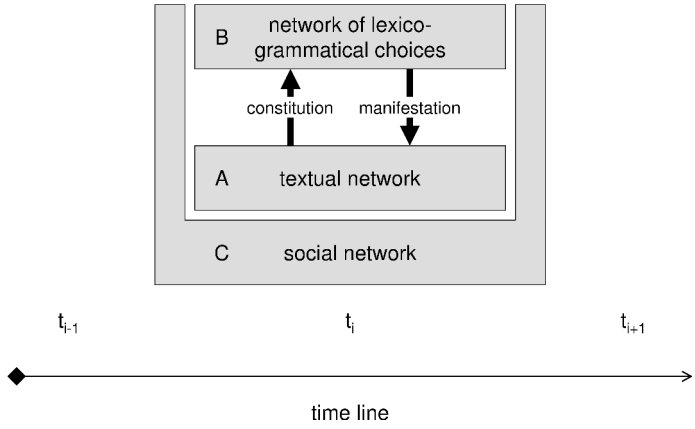


Figure 1: A multilevel or, more specifically, 3-level network of textual nodes (Level A) as manifestations of networked lexico-grammatical choices (Level B) and their embedding into a language community spanning a further network (Level C). The reference to the time line indicates the co-evolvement of these networks.

social dimension of networking as being constitutive of the Web 2.0. Secondly, it underestimates the structural complexity of linguistic networks. Classical approaches start from a notion of text to be augmented/replaced by hypertexts whose central characteristic was seen to be *nonlinearity* (Storrer, 2002). This notion is far away from the kind of networks spanning the Web 2.0 whose characteristic is the *co-evolvement* of social and semantic networking of agents *and* documents cooperatively generated by the former. Networking is a social-semantic phenomenon which demands a unified theory of structure, content and usage mining (Chakrabarti, 2002) in the line of open hypertext systems (Hammwöhner, 1997; Kuhlen, 2004). One reason is that agent networks need to have a critical size and structure so that document networks can evolve which meet certain criteria of information processing. As will become clear subsequently this relates to the notion of small worlds and related network topologies as prerequisites of efficient information processing in terms of space and time complexity (Steyvers and Tenenbaum, 2005).

When modeling networking as a social-semantic phenomenon we do not only leave the narrow stance of hypertext theory and go beyond the discourse level as an upper bound of document structures in computational linguistics. Rather, we also extend the narrow focus of complex network theory on *unipartite* networks. That is, for a decade or so laws of networking have been explored almost exclusively by example of directed or undirected graphs (Newman, 2003) more and more often with weighted edges (Barrat et al., 2004; Serrano et al., 2006a) sometimes partitioned into bipartite structures (Watts, 2003). However, when dealing with linguistic systems we have to distinguish at least three levels of networking as exemplified in Figure 1. It starts from a network of textual nodes (Level A) linked by intertextual relations of referential, typological or some other provenance (Fraas, 1997; Thibault, 1997). Generally speaking, any such textual node manifests a system of lexico-grammatical choices which span a network by their syntagmatic and paradigmatic relations and, thus, the linguistic system of a certain language community (Hjelmslev, 1969; Halliday, 1978; Martin, 1992). Hence, we get a second network perspective of a linguistic system (Level B) whose parts are both manifested as well as changed or newly constituted by discourse units (see Figure 1). However, the choices constitut-

ing the linguistic system are made by members of a corresponding language community interlinked by social relationships as, e.g., common interest or acquaintance. As these relationships constrain which members communicate with each other we get a third network perspective (Level C) which embeds the community's language system and its textual manifestations.

As an example of such tripartite networks think of a special wiki of a certain field of knowledge (as, e.g., an open source software project).[1] In this case, the interlinked wiki pages span a textual network which manifests the corresponding system of terminological or, more general, lexical choices shared and collectively shaped by those "wikilocutors" who participate in collaboratively writing the wiki.[2] This sort of tripartite networks is at the core of the present paper. It introduces an integrative model of (i) the networking of agent communities which collectively author (ii) a network of web documents in order to manifest and further develop (iii) a social ontology (Steels, 2006) of some field(s) of knowledge. The computational model of *multilevel* networks presented in this paper is meant to replace the unipartite graph model of prevalent approaches to social-semiotic networks (Steyvers and Tenenbaum, 2005; Mehler, 2008a). The idea is that multilevel networks open the perspective to a more thorough understanding of social-semiotic networking in web-based communication.

For a concrete example of a multilevel network see Figure 2. It shows the agent, document and word network of the project wiki about the Mozilla Firefox browser. This wiki serves as a platform for the technical documentation of the browser software. More specifically, in the case of Network A vertices denote agents whose collaborations span the edges. That is, an edge is built between any two wikilocutors if they co-author at least one page of the wiki. In Network B vertices denote pages whose hyperlinks span the edges of the graph. Finally, in Network C vertices denote words whose lexical associations (as confirmed by the corpus of all textual nodes of Network B) span the edges.[3]

Our approach to such multilevel networks follows complex network theory (Newman, 2003) which provides overwhelming evidence for the exceptionality of biological, social, semiotic and technological networks in contrast to certain classes of random and regular graphs. The reason is that the former are known for their S*mall* W*orld* (SW) property (Watts and Strogatz, 1998). That is, other than random graphs S*mall* W*orld* N*etworks* (SWN) do not only have short geodesic distances, but also a higher cluster value. Steyvers and Tenenbaum (2005) relate this property with the time and space complexity of linguistic networks where it is seen to guarantee efficient memory storage and retrieval. On the other hand, Blanchard and Krüger (2004) show that social networks obey the cameo principle which says that agents tend to connect to those with rare, exceptional features. Further, Newman (2003) reports on assortativity in social networks where agents with alike connectivity patterns tend to be linked. These are all well established results so that we do not present yet another study of SW networks. Rather, we provide evidence in support of the following hypotheses about tripartite networks:

- *Network Correlation Hypothesis (NCH):* Agent, document and word networks correlate with respect to their small world-property. That is, linguistic networks emerge as small worlds only

[1] Such wikis evolve more and more frequently as documented by numerous wiki directories (as, e.g., `wikiindex.org` or `www.wikimatrix.org`).

[2] We use the portmanteau word *wikilocutor* — as a composition of *wiki* and *interlocutor* — in order to name agents who participate in collaborative writing of a wiki.

[3] The details of building such word networks are explained in Section 3.3.

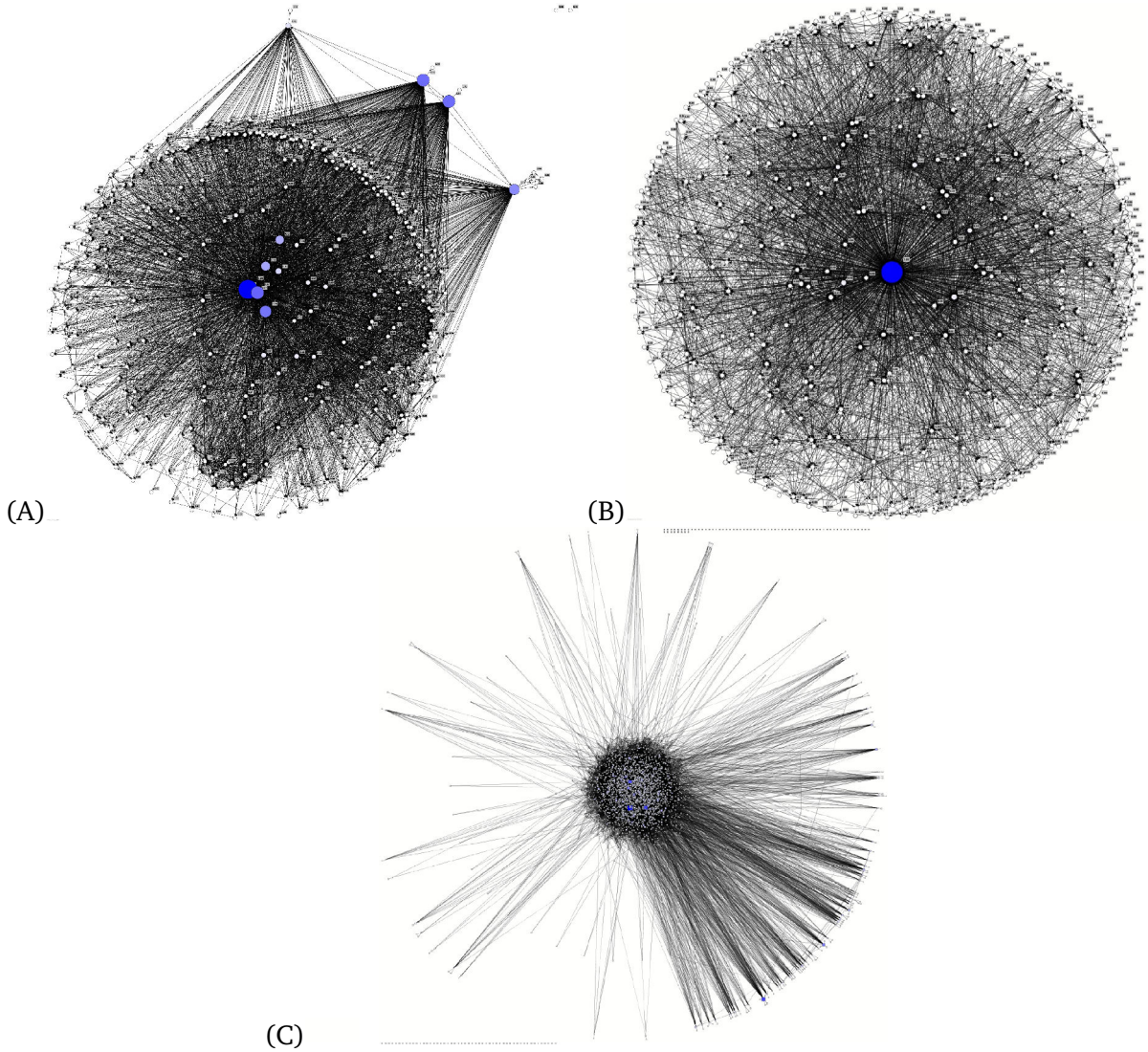


Figure 2: The agent (A), document (B) and word network (C) of the project wiki about the Mozilla Firefox browser (`http://www.firefox-browser.de/wiki`).

if this kind of topology also appears on the level of the underlying agent community. In this sense one might formulate that agents build communities in the form of small worlds in order to generate knowledge networks which themselves are small worlds.

- *Network Separability Hypothesis (NSH):* However, social and linguistic networks can be reliably separated by means of their topological characteristics. According to this hypothesis one might say that agent, document and word networks tend to evolve as small worlds though *very differently*.

Network characteristics in support of these two hypotheses are both informative about the commonalities of SW-like networks as well as about their topological differences. From the point of view of language evolution a law-like co-evolvement of social and linguistic networks as claimed by the

network correlation hypotheses is well justified. The reason is that small worlds enable efficient information processing in terms of space and time complexity (Steyvers and Tenenbaum, 2005). As language learning is a social process (Steels, 2002) any model thereof should be aware of this co-evolvement. Such an awareness might open new perspectives to text and web mining in general as indicated by a combination of social software and machine learning. This paper is about network-theoretical foundations of such a synthesis.

By approaching a unified theory of content, structure and usage mining in terms of multilevel networks we face a methodological challenge. The reason is that we deal with structure formations far away from tree-like structures, multi-rooted trees and tree-related decompositions of networks (Emmert-Streib et al., 2005). That is, the full-fledged apparatus of graph theory has to be used in order to build adequate computational models of the networks in question. As will be shown in this paper this includes layered polypartite graph models of the social-semantic levels of networking. However, in spite of the novelty of this research object the questions induced by it relate to a classical research scenario in web mining, *but with a focus on networking*. This scenario can be summarized as follows:

- Firstly, we have to ask how to reliably segment and classify networks in order to map their constituents and similarity distributions.
- Secondly, we seek a generative model by which the built-up of networks is adequately simulated in order to prospect/retrospect their development.
- Thirdly, we are interested in models of network-based information processing which adequately predict its effort and success rate subject to the corresponding network topology.
- Fourthly, we are concerned with efficient data structures for representing networks when solving these and related tasks.

In this paper, we address the first and the fourth of these research topics with a focus on classification — the second topic is tackled by Mehler (2007a, 2008b). That is, we are concerned with a model of multilevel networks as input to classification algorithms by which networks of various content and provenance can be automatically classified. We describe three building blocks of this task:

1. A graph model expressive enough to map multilevel networks of the kind of wikis (Section 3).
2. A computational model of the similarities of instances of this graph model (Section 4) together with a classification algorithm in terms of Q*uantitative* N*etwork* A*nalysis* (QNA) — as will be seen, QNA makes linguistic networks an input to explorative data analysis.
3. Finally, a model of the distribution of the kind of networking manifested by wikis (Section 5).

This paper presents an in-depth experiment on network classification which evaluates QNA by example of a corpus of wiki networks which — to the best of our knowledge — is the largest corpus of this sort analyzed so far (see Section 6). Our evaluation does not only prove the feasibility of QNA but also provides a procedure for measuring the information value of topological characteristics

content function communication situation	$\approx$	distributional patterns of elementary units distributional patterns and structure of complex units

Table 1: The correlational hypothesis of linguistic classification.

of complex networks in general. However, the paper is not limited to network classification as yet another task in machine learning. Rather, we extend approaches to structure learning by including the level of linguistic networks. Generally speaking, the idea to classify the content, function or situational context of a linguistic unit by the patterns it instantiates comes from quantitative linguistics. Its basic tenet — summarized in Table 1 — is that distributional patterns vary with the semantic/pragmatic status of their linguistic manifestations (Biber, 1995; Tuldava, 1995). In the past, lexical features were mainly explored as a source of such patterns. However, as far as the structure of units varies — though not deterministically — with their semantics/pragmatics they can be seen to be classifiable by their structure (Brinker, 1992). In terms of the weak contextual hypothesis of Miller and Charles (1991) one might say that structural differences reflect semantic or pragmatic ones while similar contents, functions or situations tend to be manifested by structurally similar units. As there is a many-to-many relation of patterns on the one hand and function or content units on the other hand[4], learning semantic or pragmatic classes by exploring the structure of their instances is nontrivial.[5] In this paper we do not classify sentences, paragraphs or texts, nor do we classify turns, dialogue games or dialogues but large complex networks of thousands of nodes. That is, we apply the weak contextual hypothesis on the level of networks. Our expectation is that it is possible to predict the communication area (in terms of related functions and contents) of wiki-based networks by solely examining their topological characteristics while abstracting from the content of their nodes. In this sense, we extend the distributional hypotheses onto the level of topological patterns of networks and, thus, gain networks as a further level of linguistic manifestation.

In summary, this paper presents a framework for representing and classifying large complex social-semiotic networks. By means of this framework we reliably measure the similarity of document, agent and word networks by solely regarding their topology. In doing so we depart from classical approaches to complex network theory which focus on topological characteristics in order to check their small world-property. Rather, we investigate the separability of various topological features with respect to their classificatory information value. As a result, we distinguish less informative from more informative topological characteristics. This does not only include characteristics which have been studied in complex network theory (Newman, 2003), but also some of those which were invented in social network analysis (Wasserman and Faust, 1999) and hypertext theory (Botafogo et al., 1992). However, our classification algorithm does not explore "universal" characteristics (Barthélemy, 2004). Such features were introduced in order to demarcate general classes of networks irrespective of their

[4]Neither can we deterministically infer a unique function based on observing some text pattern, nor is the same function always manifested by the same pattern — the same holds for content categories.

[5]See, for example, Pustylnikov and Mehler (2007) and Mehler et al. (2007a) for such an approach to text classification.

Communication Area	Example
▽ consumer communication	----------
▷ consumer advisory	www.taxalmanac.org
▷ product & promotion communication	wiki.siemens-enterprise.com
▷ corporate & enterprize communication	wikis.sun.com
▷ health communication	autism.wikia.com
▽ knowledge communication	----------
▶ subject area communication	www.glottopedia.org
▷ bibliographic communication	en.wikisource.org
▷ encyclopedic communication	en.wikipedia.org
▷ metalinguistic communication	en.wiktionary.org
▽ leisure communication	----------
▷ fan communication	memory-alpha.org
▶ game communication	finalfantasy.wikia.com
▷ sport communication	www.kletterwiki.de
▷ travel communication	wikitravel.org/en
▷ news communication	en.wikinews.org
▷ social, political & environmental communication	sustainability.mit.edu
▼ regional communication	allmende.stadtwiki.net
▷ city wikis	daviswiki.org
▷ region wikis	wiki.rhein-neckar.de
▽ scientific & educational communication	----------
▷ educational communication	en.wikiversity.org
▶ scientific networking	aclweb.org/aclwiki
▶ technical communication	wiki.apache.org

Table 2: Classifying wikis by the communication area to which they mainly belong. Communication areas which are taken into account by means of a corpus of wiki networks are marked with ▶ or ▼, respectively.

ontological provenance.[6] Rather, we start from an assumption about the latent separability of network characteristics as a result of their interaction.

Finally, we show that network classifications come into reach which go beyond the hypertext structures traditionally analyzed in web mining. The reason is that we focus on networks as a whole as units to be classified — *above the level of websites and their constitutive pages*. As a consequence, we bridge classical approaches to text and web mining on the one hand and complex network theory on the other hand.

2 Springing up like Mushrooms: Special Wikis in Different Communication Areas

Facing the number of newly released wikis they seem to spring up like mushrooms in ever new content areas with functions beyond knowledge communication as manifested by the most prominent wiki, i.e., the Wikipedia. These newly met functions range from regional & political communication, via scientific, educational & technical communication to community building, broadcasting (in the area of news communication), archiving (in bibliographic communication), promotion (in the area of consumer communication) and entertainment. This increase is confirmed by equally frequently emerging

[6] A good example of such a classification regarding scale-free networks is Amaral et al. (2000).

wiki directories and hosts as, for example, wikiindex.org and www.wikia.com.

Table 2 lists communication areas for which wiki instances are easily identified.[7] This relates to so called special wikis (e.g. city wikis, conference wikis, eduwikis, enterprize wikis, event wikis, game wikis, news wikis, project wikis, product wikis, region wikis, team wikis, translation wikis etc.). Although being mentioned in Table 2 the Wikipedia is not analyzed here — for studies of its topology see Capocci et al. (2006), Mehler (2006) and Zlatic et al. (2006). Instead, we focus on special wikis which have been barely, if ever, investigated before. More specifically, we analyze wikis of the following communication areas (cf. Table 2):

1. *Leisure communication*: Wikis of this area deal with a *single* hobby or leisure activity. Common are wikis on multi-user dungeons or massively multiplayer online role-playing games (e.g. finalfantasy.wikia.com) and related games (e.g. www.cachewiki.de). Other game wikis as, e.g., of computer games are also subsumed under this header. Their thematic core is the documentation of the topic defining game, e.g., by means of background information, manuals, user guides and tutorials. Additionally, they serve agent-related functions such as exchange of experience, community building among (advanced) players, team formation and sub-project building, instruction of beginners and self-portrayal (e.g., by means of high score lists). Typical is also the provision of gaming strategies, tips and tricks. Note that wikilocutors tend to have personal experience with practicing the focal game. Note further, that these wikis are not primarily encyclopedic as they focus on single leisure activities.

2. Wikis of *regional communication* provide portals for regions (e.g. the region wiki wiki.rhein-neckar.de about the metropolitan area Rhine-Neckar) or cities (e.g. the city wiki www.bloomingpedia.org about Bloomington, Indiana). They are information portals about various subtopics (as, e.g., geography, regional institutions and business companies, education, culture, leisure and entertainment) whose thematic priority is their common reference to a certain region or city. Thus, regional wikis are thematically more diversified than game wikis as their wikilocutors have divergent thematic backgrounds and interests to participate (whether, e.g., mainly business or leisure oriented). As a consequence we expect a broader spectrum of participants and this may, in turn, result in different topologies compared to those of wikis of other communication areas.

3. Wikis of technical communication are — similar to leisure wikis — thematically focused. However, they deal with professional communication: their topics range from open source projects (e.g. wiki.mozilla.org) via programming languages (e.g. gcc.gnu.org/wiki), operating systems (e.g. wiki.debian.org) and IDEs (e.g. eclipsewiki.editme.com) to hardware (e.g. www.c64-wiki.de) and other technical systems. Other than the Wikipedia, technical wikis serve as *how-to* manuals and therefore might be called *wikimanuals* (a fusion of *wiki* and *manual* in contrast to *wikipedia* as a portmanteau word of *wiki* and *encyclopedia*). Consequently, wikimanuals serve as extended or supplementary manuals or detailed technical documentations (beyond offline available technical manuals) which are open to the community of corresponding users or developers. Additionally, they serve the agent-related function of exchange of experience with

[7] For a list of URLs of all special wikis analyzed here in conjunction with URLs of wikis of the same or related communication areas see ariadne.coli.uni-bielefeld.de/indogram, *resources*.

respect to using, operating or further developing the focal system while the function of community building is less pronounced than in the case of game wikis. In this sense, wikimanuals have a non web-based predecessor in the form of manuals which they extend, replace or augment by wiki software-based opportunities to cooperate.[8] Note that such professional competitors in the form of on- or offline media do not exist in the area of leisure communication.

4. Wikis of *knowledge communication* are analyzed by example of so called *subject area wikis*. These are wikis which provide encyclopedic knowledge about a single subject field which — though being manifested by a special wiki — might be part of the wikipedia. The reason to focus on subject area wikis while disregarding bibliographic, metalinguistic and encyclopedic wikis is that the latter are thematically far too diversified while the former tend to make untypical use of the wiki principle. Look, for example, at `en.wikiquote.org` as an instance of wiki-based bibliographic communication which is a collection of quotes linked by word co-occurrences. Another instance is `en.wikisource.org` which provides corpora of barely linked source texts. On the other hand, `en.wiktionary.org` functions as a metalinguistic wiki, that is, as a dictionary of highly schematized entries linked by lexicographic relations. Thus, in spite of being thematically focused, subject area wikis function according to wikipedia-related principles. This is explicitly excluded, for example, by many technical wikis while being part of the conception of subject area wikis. We expect that subject area wikis manifest agent-centered functions (e.g. self-portrayal) according to their encyclopedia-related self-conception to a lower degree than leisure wikis. Consequently, a corresponding topology-related impact is expected which distinguishes them from wikis of other communication areas.

5. Finally, we analyze wikis of scientific communication: focusing on a scientific topic they provide platforms for the build-up and organization of scientific communities from the level of project groups and departments up to the level of conference participants and other large (international) communities. This includes so called department and team wikis as well as project and conference wikis. Wikis of this sort do not primarily describe a certain subject area in terms of an encyclopedia, but enable scientists to share information on their topic-related projects so that they can network with each other.

In the following sections, networks of these five communication areas will be numerically classified by their topological characteristics. This will be done by example of interrelated agent, document and terminological networks. In the next section we present our formal model as a prerequisite of this classification experiment.

3 Extracting Wiki Graphs from Wiki Networks

A prerequisite of any similarity model of network structures is an underlying data model. Such a model can be built with the help of graph theory (Diestel, 2005). For decades it has been used to represent

[8] A prototypical example is `http://www.typolight.org/wiki/german:documentation` which makes the user manual downloadable as an alternative to the wiki offered as an online manual.

linguistic structures up to the level of hypertext graphs.[9] In this paper we utilize graph theory to map the structure of wikis. This is done by the notion of a *wiki graph* which introduces a polypartite model of networking on the level (i) of documents, (ii) of the agents who co-authored them and (iii) of the linguistic means used to generate the documents. Starting from the level of document networking we distinguish three reference points of graph-like structure formation (Mehler, 2007b):

- On the *micro*-level wiki page-internal structures (e.g. of single articles or talks) can be distinguished by analogy to web pages.
- On the *meso*-level those aspects of wiki-based structure formation come into focus which correspond to websites as thematically and functionally closed units of web-based communication. In this sense, we distinguish, for example, portals which consist of at least one wiki page.
- Finally, on the *macro*-level the topology of the corresponding wiki document network can be studied as a whole.

Leaving out the micro-level of wiki document structures we start with the meso-level (in Section 3.1) before we present our macro-level model of networks (in Section 3.3) which are finally input to network classification (in Section 6). A note on terminology: throughout the paper we use the terms *node* and *link* when speaking about empirical networks, while *vertex* and *edge* are used to speak about *graphs* as their formal models (cf. Mehler, 2008a).

3.1 Setting the Scene: Meso-Level Structures in Wiki Networks

In this section we study the structural complexity of meso-level discourse units of wiki systems. The reason is to demarcate elementary building blocks of network classification. However, this is not meant as a structuralist reduction of the semantics and pragmatics of collaborative writing by means of wikis. Rather, the aim is to identify structural manifestations by which networking in different communication areas and for different purposes can be distinguished. Thus, this section provides a first step towards instantiating the correlational hypothesis of Table 1 in the area of document networks.

Generally speaking, a wiki is more than just a network of articles. There is a wide range of page types according to the wiki namespace.[10] Articles, for example, are the smallest self-contained, non-recursively organized content unit in wiki media. They are non-recursively organized in the sense that they do not include subpages[11] of the namespace *article*. On the other hand, portals are complex content units — henceforth called *wiki documents* — which may recursively consist of other portals down to the level of articles. Moreover, articles, portals and other wiki pages are complemented by talk and history pages which support their collaborative generation. In this sense, a wiki document network is best conceived as a non-uniquely typed or labeled graph in which the *wiki page* is the

[9] See Dehmer (2005a) for an overview of graph models in web mining.

[10] In this article we refer to the namespace of the Wikipedia as a source of typing pages.

[11] A subpage is a lower level web page extending its name by a slash-separated name suffix (see `http://en.wikipedia.org/wiki/Wikipedia:Subpages`). In the case of a talk, e.g., subpages are used to archive parts of it. Note that the subpages of a wiki page span trees.

reference unit of typing or namespace mapping, respectively.[12] Even in the case of composite units as, e.g., portals their entry pages are the primary object of typing. Thus, *wiki pages* can be defined as the smallest self-contained building block of wikis in general and of wiki documents (which may consist of *subpages* or link to subordinate units by means of inclusion links as done in portals) in particular. Note that wiki pages are generally seen as elementary meso-level units of a wiki whether mapped onto a namespace or not.[13]

A basic guideline of our meso-level model of wikis to be presented now is the distinction between wiki *documents* — which are usually omitted in the research on wiki networks — and wiki *pages* which are usually referred to as vertices inducing units of network analysis (Capocci et al., 2006; Mehler, 2006; Zlatic et al., 2006). As the literature lacks this distinction it disregards an important level of structure formation with a yet unknown impact on wiki topologies. In contrast to this, we need to decide, firstly, which types of pages are considered at all and, secondly, whether wiki documents or pages are explored as vertices. If the decision is 'wiki documents', their internal arcs do not count when computing, e.g., cluster coefficients. In this case two agents can be said to collaborate even if they author different pages of the same document — this would be impossible if the decision is 'wiki page' instead.[14] The literature on wiki networks obscures these decisions so that wikis have been modeled as purely page-based networks (disregarding, for example, the differences of portals, articles, their talks etc.). As far as the wiki type system is neglected, too, pages of whatever type are taken into account. This approach disregards the varying impact of knowledge, collaboration and network management types on networking. Resolving redirects, for example, raises the probability of a higher cluster coefficient and, thus, affects the small world model, e.g., of Watts and Strogatz (1998). In summary, previous approaches model wikis as unlabeled undirected graphs and therefore disregard page typing as well as the formation of wiki documents. The present paper is a way out of this structural ignorance.

In order to take typing according to the wiki namespace into account Table 3 introduces a terminological ontology based on the following distinctions:[15]

- *Knowledge management types* classify wiki pages as proper content units in the sense of encyclopedias, knowledge portals or reference books.

[12] We use the terms typing and namespace mapping interchangeably.

[13] History pages, for example, are not mapped onto a namespace in Wikimedia systems.

[14] Such a leveled model of elementary pages and composite documents based thereon has been elaborated by Björneborn (2004) in the area of the WWW.

[15] Table 3 presents an is-a-hierarchy although there is plenty room for classifying wiki pages in terms of acyclic graphs. A talk, e.g., might be seen as an additional resource to explore the content of the corresponding article. Nevertheless, Table 3 subsumes talks among those types whose instances help to keep wiki-based collaboration ongoing. The reason is that they primarily focus on user collaboration, but do not serve as *the* reference for the topic of the corresponding article.

[15] `Article` is known as the main category in the English Wikipedia. That is, wiki pages without an explicit mapping to a namespace are — with the exception of history pages — known to be articles.

[16] `ArticleTalk` is known as `Talk` in the namespace of the English Wikipedia.

[17] `Site` is also not part of the namespace of the Wikipedia but a generalized representation of those namespaces which denote the corresponding wiki as, e.g., `Wikipedia` or `Ontowiki.org` in the corresponding wikis. A wiki page typed as `Site` contains general information about that specific wiki. Analogously, `SiteTalk` is the type of talks about corresponding wiki pages typed as `Site`.

Type	Composite	Basic Type	Added/Adapted
▼ GENERIC DOCUMENT TYPE (⊤)	– – – – – – – – – – – – – – – – – –		
▼ KNOWLEDGE MANAGEMENT TYPE	– – – – – – – – – – – – – – – – – –		
▶ Article[15]	–	+	+
▼ CATEGORIZATION TYPE	– – – – – – – – – – – – – – – – – –		
▼ Category	+	+	–
▶ Root Category	+	–	+
▶ Disambiguation	–	–	+
▶ Image	+	+	–
▶ Portal	+	+	–
▼ COLLABORATION MANAGEMENT TYPE	– – – – – – – – – – – – – – – – – –		
▼ TALK TYPE	– – – – – – – – – – – – – – – – – –		
▶ ArticleTalk[16]	+	+	+
▶ CategoryTalk	+	+	–
▶ HelpTalk	+	+	–
▶ ImageTalk	+	+	–
▶ MediaWikiTalk	+	+	–
▶ PortalTalk	+	+	–
▶ SiteTalk[16]	+	+	–
▶ TemplateTalk	+	+	–
▶ UserTalk	+	+	–
▶ Help	+	+	–
▶ History	+	–	+
▶ Site[17]	+	+	+
▶ User	+	+	–
▼ NETWORK MANAGEMENT TYPE	– – – – – – – – – – – – – – – – – –		
▶ MediaWiki	+	+	–
▶ Redirect	–	–	+
▶ Template	+	+	–

Table 3: Types of wiki pages and the number of their instances studied within the present paper.

- *Collaboration management types* relate instead to those wiki pages which serve as self-contained content units the maintenance or enhancement of wiki-based collaborative writing.
- Finally, *network management types* refer to wiki pages which are not generated as self-contained content units, but support content generation or processing.

In defining this terminological ontology we take a conservative position by starting from the namespace of the Wikipedia together with some extensions (see below). This also means that we do not directly consider those namespace extensions which have been introduced in special wikis.[18]

Table 3 contains three columns in need of an explanation: The column entitled *composite* distinguishes wiki pages (e.g. portals, talks, images) that may have subpages from those that may not (e.g. articles) or cannot (e.g. redirects). Generally speaking, subpages induce tree-like document structures which span generalized trees by means of hyperlinks (see below). The same holds, e.g., for portals which include subordinate pages via hyperlinks henceforth called *inclusion links*. That is, portals may recursively consist of portals down to the level of articles which altogether span inclusion hierarchies. As in the case of subpages, these hierarchies may be superimposed by graph-inducing up, down and across links so that generalized trees emerge (see below). In order to take this kind of structure for-

[18]The ontoworld.org wiki (`http://ontoworld.org/wiki`), for example, knows the type `Type` as part of its namespace which classifies wiki pages about data types — `Type` is, so to speak, a subtype of the type `Article`.

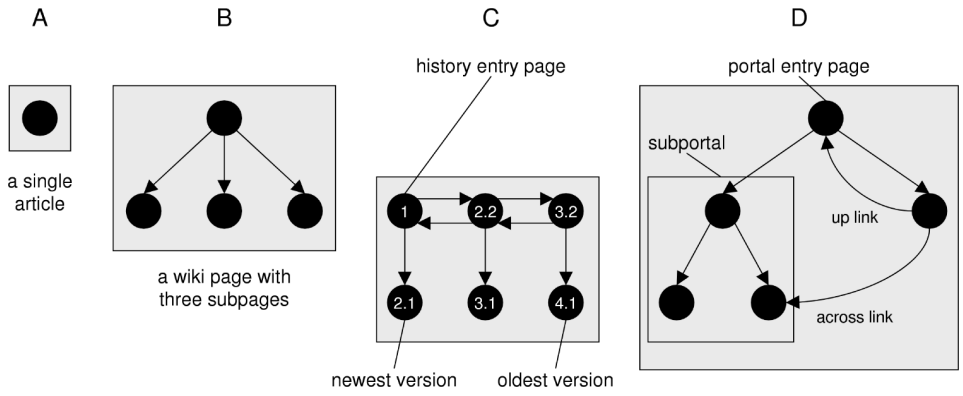


Figure 3: Variants of document structure of wiki documents.

	elementary	composite
constituency structure	wiki page	wiki document
constituency relation	inclusion link	inclusion hierarchy
dependency relation	association link	association path

Table 4: Two-mode structure formation in wiki document networks: starting from a network of wiki pages we distinguish links used to build complex constituents (inclusion links) from those which span the dependency structure of that network (i.e. association links). Based thereon we further distinguish inclusion hierarchies and association paths as complex units of structure formation in wiki networks.

mation into account (which wiki systems solely manifest by namespaces and URIs), we introduce the notion of a *wiki document*: A wiki document either consists of a single wiki page (without subpages or inclusion links to subordinate pages) or of several pages (including subpages or inclusion links to subordinated documents or pages). This is demonstrated in Figure 3: An article is the most simple, non-composite wiki document (Case A).[19] In contrast to this a talk may have several subpages which are archiving parts of it (Case B). Further, a history document consists of one or more history list pages (Case C) together with the corresponding version pages. Next, Case D demonstrates a portal which consists of the corresponding entry page and a subordinate portal and article. Note that articles together with their talk and history documents define the classical cases of composite documents in wikis.[20] Note also that talks, images and many other types of pages may have subpages and, thus, span composite wiki documents (see Table 3).

As wiki documents may contain across, up and down links that induce graphs (cf. Figure 3, Case D) a graph model expressive enough to handle this kind of structure formation as an add-on to the kernel hierarchical structure of a wiki document is needed. Such a graph model is given by generalized trees (Dehmer, 2005b; Dehmer et al., 2007; Mehler, 2007b). We redefine it as follows:

Definition 3.1. (Generalized Tree) Let $T(r) = (V, E, r, \mathcal{O})$ be a directed ordered rooted tree with vertex set V, arc set E, root r and order relation $\mathcal{O} \subset V^2$ which for each vertex $v \in V$ linearly orders

[19] Documents which consist of a single wiki page have, so to speak, a status by analogy to one-word-sentences.

[20] At least in media wiki systems. This medium-related constraint hints at the demand of distinguishing layout from functionally and topically defined units.

	Digital Library	WWW	Wiki	Graph Model
atomar level:	e-text document	web page	wiki page	vertex
level of complex units:	citation cluster	website	wiki document	generalized tree
networking level:	citation network	(sub-)domain	wiki website	document network graph

Table 5: Building blocks of wiki document graphs in comparison with digital libraries and the WWW.

the set $\{w \,|\, e \in E \wedge \mathsf{in}(e) = v \wedge \mathsf{out}(e) = w\}$ of vertices to which v is adjacent. Let further $P_{rv} = (v_{i_0}, e_{j_1}, v_{i_1}, \ldots, v_{i_{n-1}}, e_{j_n}, v_{i_n}), v_{i_0} = r, v_{i_n} = v, e_{j_k} \in E, \mathsf{in}(e_{j_k}) = v_{i_{k-1}}, \mathsf{out}(e_{j_k}) = v_{i_k}, k \in \{1, \ldots, n\}$, be the unique path in $T(r)$ from r to $v \in V$ and $V(P_{rv}) = \{v_{i_0}, \ldots, v_{i_n}\}$ be the set of all vertices of P_{rv}. A *generalized tree*

$$GT(r) = (V, E_{[1]}, E_{[2]}, E_{[3]}, E_{[4]}, E_{[5]}, E_{[6]}, E_{[7]}, r)$$

induced by $T(r)$ is a graph with possibly multiple and parallel arcs whose partitioned arc set is incrementally defined:

$$\begin{aligned}
\textit{kernel arcs:} \quad & E_{[1]} &=\ & E \\
\textit{up arcs:} \quad & E_{[2]} &\subseteq\ & E_{\mathrm{u}} = \{e \,|\, \mathsf{in}(e) = v \in V \ \wedge\ \mathsf{out}(e) = w \in V(P_{rv}) \setminus \{v\}\} \\
\textit{down arcs:} \quad & E_{[3]} &\subseteq\ & E_{\mathrm{d}} = \{e \,|\, \mathsf{in}(e) = w \in V \ \wedge\ \mathsf{out}(e) = v \in V(P_{rw}) \setminus \{w\}\} \\
\textit{reflexive arcs:} \quad & E_{[4]} &\subseteq\ & E_{\mathrm{r}} = \{e \,|\, \mathsf{in}(e) = \mathsf{out}(e) = v \in V\} \\
\textit{across arcs:} \quad & E_{[5]} &\subseteq\ & V^2 \setminus (E \cup E_{\mathrm{u}} \cup E_{\mathrm{d}} \cup E_{\mathrm{r}}) \\
\textit{sequential arcs:} \quad & E_{[6]} &=\ & \mathcal{O} \\
\textit{external arcs:} \quad & E_{[7]} &=\ & \emptyset
\end{aligned}$$

As a naming convention we say that generalized trees are *generalized by their up, down, reflexive, across and sequential arcs*. We abbreviate $GT(r) = (V, E_{[1..7]}, r)$ where $e \in E_{[1..7]}$ iff $e \in \cup_{i=1}^{7} E_{[i]}$. A generalized tree $GT(r) = (V, E_{[1..7]}, r)$ induces a directed tree $\mathrm{Tree}(GT(r)) = T(r)$ called *kernel hierarchy* of $GT(r)$.

Remark. The arc set of a connected directed graph can always be partitioned into that of a generalized tree. We just have to select a kernel tree and to type the remaining arcs as up, down or across links. Hence the notion of a generalized tree seems to be hardly informative. But this is the wrong way of thinking about this notion. Look, for example, at directed in relation to undirected graphs: Starting from an undirected graph there are the more possibilities to orient its edges the higher the number of these edges. However, it is out of place to question the notion of a directed graph simply because there is always at least one such orientation. Rather, which candidate orientation is empirically justified is a question which goes beyond purely formal considerations. Likewise, the question which arcs have to be typed as kernel, up, down or across links has to be answered on empirical grounds. Thus, the notion of a generalized tree is best understood by starting from an empirically well-founded tree-like skeleton which in the next step is augmented by graph-inducing links as enumerated above.

By reference to the kernel hierarchy of a generalized tree, Definition 3.1 defines various types of arcs in purely structural terms. Thus, Definition 3.1 does not include a labeling function in order to distinguish types of arcs.

Note that kernel arcs model inclusion links while arcs of other types are used to model association links (cf. Table 4). Note also that wiki pages and documents are modeled as vertices and generalized trees, respectively — this is summarized in Table 5 and explained at length in Section 3.2. Finally, for an example of generalized trees with non-empty sets of external arcs see Mehler (2007b).

The next step of introducing the meso level model of wiki document networks is to relate generalized trees with the ontology of Table 3. This is done as follows:

Definition 3.2. (Labeled Typed Generalized Tree) A *labeled typed generalized tree* $GT(r) = (V, E_{[1..7]}, r, \mathcal{L}_0, \mathcal{L}_1)$ is a generalized tree together with two labeling functions $\mathcal{L}_0 : V \rightarrow \mathrm{L}_0$ and $\mathcal{L}_1 : V \rightarrow \mathrm{L}_1$ where L_0 is a set of so called *vertex labels* and L_1 is a set of so called *vertex types*.

Labeled generalized trees are the preferred means to model wiki documents. They capture the typing of elementary wiki pages according to Table 3 and the typing of hyperlinks according to Definition 3.1. This is done by equating T_V with $\mathrm{T}_{\text{wiki}} = \{$`Article`, `ArticleTalk`, `Category`, `CategoryTalk`, `Image`, `ImageTalk`, `Help`, `HelpTalk`, `History`, `MediaWiki`, `MediaWikiTalk`, `Portal`, `PortalTalk`, `RootCategory`, `Site`, `SiteTalk`, `Template`, `TemplateTalk`, `User`, `UserTalk`$\}$, $|\mathrm{T}_{\text{wiki}}| = 20$. The following example shows how this helps to map the Cases A-D of Figure 3.

Example. *Case A of Figure 3 is a trivial case of a generalized tree. Further, as trees are always generalized trees, Case B is straightforwardly mapped by Definition 3.1. A history document which distributes the list of all modifications of the corresponding article among sequentially linked list pages in which each list item links to the corresponding version of that article induces a generalized tree as follows: the history entry page serves as the root of the generalized tree while all successive list pages and version pages are subordinated as demonstrated in Figure 3, Case C, by its numbering. Note that in this case, back links between list pages are treated as up links. Case D is a typical example of a generalized tree which in the present case contains only one across and one up link.*

The second and third column of Table 3 in need of an explanation are entitled *Basic Type* and *Added/Adapted*, respectively. *History*, for example, is a type not included in the wiki namespace — it is not a basic type —, but introduced as an additional label. The attribute '*added to the wiki namespace*' holds for the type *root category* which uniquely identifies the artificial root added to the wiki category graph in order to guarantee a rooted kernel tree as its taxonomic skeleton. Further, *disambiguation* is a type which classifies non-article pages that contain links to pages disambiguating the reading of an ambiguous lemma — note that in Wikimedia-based wikis disambiguation pages are marked by a suffix[21]. As mentioned above pages typed as *Site* refer to the corresponding basic type which for reasons of homogeneous naming is adapted here. Finally, redirect pages may circumvent duplicate entries and, thus, help to manage the wiki network structure. They are not basic types but added to the type system in Table 3.

The following section utilizes the notion of a generalized tree in order to span wiki document networks which consist of wiki pages and documents based thereon.

Object Area	Graph-Theoretical Correspondent	Definition
node / wiki page	vertex	
link / hyperlink	edge	
wiki document	labeled generalized tree / hyperedge	(3.2)
category system	multi-rooted generalized tree (not defined here)	
wiki page network	type graph	(3.3)
wiki document network	type hypergraph	(3.8)
wiki network	multilevel graph	(3.10)

Table 6: Naming conventions for wiki networks and their graph-theoretical correspondents.

3.2 Weaving the Graph: A Model of Wiki Document Networks

What graph model is best suited to model wiki document networks? There are at least two candidate models to be considered: either we start from wiki pages as vertex inducing units or from wiki documents whose structure is modeled as encapsulated generalized trees. Both approaches retain pages and documents as reference units of graph modeling. However, in this paper we realize the first variant and, thus, stress the accentuated role of wiki pages while we map the special role of wiki documents by means of subgraphs in the form of generalized trees. The reason behind this approach is that it stays in the framework of classical graph theory (while the second variant utilizes the notion of a hierarchical graph[22]). This is done by means of the notion of a polypartite directed multigraph with labeled vertices and arcs (cf. Mehler et al., 2007b):

Definition 3.3. (Type Graph) A *type graph* is a directed multigraph

$$TG = (V, E, \{\mathcal{V}_i\}_{i \in I \cup \{0\}}, \{\mathcal{E}_j\}_{j \in J}, \omega)$$

with the set of vertices V, the set of arcs E, the arc weighing function $\omega : E \rightarrow \mathbb{R}$, a set of labeling functions $\mathcal{V}_i : V \rightarrow \mathrm{V}_i$, $i \in I$, for a set V_i of vertex types and a set of labeling functions $\mathcal{E}_j : E \rightarrow \mathrm{E}_j$, $j \in J$, for a set E_j of arc types.

A special vertex labeling is introduced by the function $\mathcal{V}_0 : V \rightarrow \mathrm{V}_0$ which labels vertices (though not necessarily uniquely) by their names (i.e. lemmata) $l \in \mathrm{V}_0$. As TG is directed, E comes together with two functions in: $E \rightarrow V$ and out: $E \rightarrow V$ such that for all arcs $e \in E$, $\mathsf{in}(e) \in V$ is the *initial* and $\mathsf{out}(e) \in V$ the *end vertex* of e. Further, TG is multiple in the sense that there are possibly multiple arcs $e_i, e_j \in E$, $e_i \neq e_j$, for which $\{\mathsf{in}(e_i), \mathsf{out}(e_i)\} = \{\mathsf{in}(e_j), \mathsf{out}(e_j)\}$.

Definition 3.3 provides a generalized notion of a directed graph with labeled vertices and arcs. By including multiple arcs this graph model maps symmetric as well as asymmetric vertex relations. This is needed in order to distinguish links which are parallelized in the opposite direction from those which are not.[23] Starting from Table 3 and the set T_{wiki} we now introduce k-partite type networks whose k subsets are induced by the types in T_{wiki}. Thereby, we assume that T_{wiki} induces a partition of the set of wiki pages. This special role of T_{wiki} is formalized as follows:

[21] However, this is not the only way to manifest disambiguation pages in mediawikis.

[22] For a graph model in the line of this variant see Mehler (2007b).

[23] This also means that we do not start from back links in order to induce multiple arcs for any hyperlink by analogy with undirected edges.

Definition 3.4. (k-Partite Type Graph) Let $TG = (V, E, \{\mathcal{V}_i\}_{i\in I\cup\{0\}}, \{\mathcal{E}_j\}_{j\in J}, \omega)$ be a type graph. For the vertex labeling function $\mathcal{V}_1 : V \to \mathrm{V}_1$, $|\mathrm{V}_1| = k$, a subset $E' \subset E$ of arcs and a partition $\mathbb{P}_{\mathcal{V}_1}(V) = \{V_1, \ldots, V_k\}$ of V such that

1. the function $\mathcal{V}_0 \times \mathcal{V}_1 \colon V \to \mathrm{V}_0 \times \mathrm{V}_1$ with $\mathcal{V}_0 \times \mathcal{V}_1(v) = (\mathcal{V}_0(v), \mathcal{V}_1(v))$ for all $v \in V$ is injective,
2. $\forall V_i, V_j \in \mathbb{P} \forall v, w \in V_i, x, y \in V_j \colon \mathcal{V}_1(v) = \mathcal{V}_1(w) \wedge \mathcal{V}_1(x) = \mathcal{V}_1(y) \wedge \mathcal{V}_1(v) \neq \mathcal{V}_1(x)$ and
3. $\forall e \in E' \exists! i, j \in \{1, \ldots, k\}, i \neq j \colon \mathsf{in}(e) \in V_i \wedge \mathsf{out}(e) \in V_j \;\leftrightarrow\; \mathsf{in}(e) \in V_j \wedge \mathsf{out}(e) \in V_i$

we call $\mathcal{V}_1(TG) = (V, E', \{\mathcal{V}_i\}_{i\in I\cup\{0,1\}}, \{\mathcal{E}'_j\}_{j\in J}, \omega')$ a *k-partite type graph* where $\mathcal{E}'_j$, $j \in J$, is the restriction of $\mathcal{E}_j$ to E' and ω' the one of ω to E'. Further, $\mathbb{P}_{\mathcal{V}_1}(V)$ is called k-partition of TG induced by $\mathcal{V}_1$. Finally, a *constraint l-partite type graph* $\mathcal{V}'_1(TG) = (V'', E'', \{\mathcal{V}''_i\}_{i\in I\cup\{0,1\}}, \{\mathcal{E}''_j\}_{j\in J}, \omega'')$ induced by $\mathcal{V}''_1 \subseteq \mathcal{V}_1$, $\mathcal{V}''_1 \colon V'' \to \mathrm{V}''_1 \subseteq \mathrm{V}_1$, $|\mathrm{V}''_1| = l \leq k$, is an l-partite type graph with $V'' \subseteq V$, $\forall e \in E'' \colon \{\mathsf{in}(e), \mathsf{out}(e)\} \subseteq V''$ and where $\mathcal{V}''_i, \mathcal{E}''_j$ and ω'' are restrictions in the usual sense. In order to stress the role of $\mathcal{V}_1$ we call this labeling function a *vertex typing*.

Remark. Condition 1 implies that the label and type of a vertex uniquely identify it among all vertices of V. This is done by equating V_1 with T_{wiki}. That is, while $\mathcal{V}_0$ models the lemma-related labeling of wiki pages, their namespace mapping is modeled by $\mathcal{V}_1$. Condition 2 demands that $\mathcal{V}_1$ induces a partitioning of V while Condition 3 guarantees the build-up of a k-partite graph. Finally, l-partite graphs are generated by selecting a subset of vertex types which induce a subgraph spanned by their instances. This approach is needed when selecting, for example, the article graph together with the corresponding category graph of a wiki network while disregarding all other nodes and their links.

Definition 3.4 is instantiated by setting $\mathrm{V}_1 = \mathrm{T}_{\text{wiki}}$ to get a polypartite graph TG as a model of a wiki document network. On the one hand, this allows to extract subgraphs for a certain task in wiki mining by constraining T_{wiki} to a subset of types. On the other hand, extending T_{wiki} allows to consider more fine-grained vertex classifications. The only precondition is that T_{wiki} induces a partition over the vertex set V of TG while the same is demanded for a corresponding subset of V in the case of restricting T_{wiki}. In a nutshell, the parameter k is not fixed but has to be fit to the given task in wiki mining. This role of k is normally disregarded in the literature. Rather, previous approaches to complex network theory by example of wikis underestimate the role of vertex and arc types. Now we are in a position to give a graph-theoretical account of wiki documents. As this definition makes use of the notion of a hypergraph we first repeat two central definitions of this area of graph theory — for more details on hypergraphs see (Berge, 1989; Gallo et al., 1993; Melnikov et al., 1994):

Definition 3.5. (Hypergraph) Let V be a finite nonempty set and $\hat{E}$ be a finite collection of not necessary distinct subsets of V. $H = (V, \hat{E})$ is called *(undirected) hypergraph* with the set of vertices V and the set of *(hyper-)edges* $\hat{E}$. Identical edges of $\hat{E}$ are called multiple. Further, a vertex $v \in V$ and a hyperedge $\hat{e} \in \hat{E}$ are *incident* if $v \in \hat{e}$. $\hat{E}(v)$ is the set of all hyperedges incident with v. $|\hat{E}(v)| \geq 0$ is the *degree of vertex* v. If $|\hat{E}(v)| = 0$, then v is called *isolated*. $|\hat{e}| \geq 0$ is the *degree of the hyperedge* $\hat{e} \in \hat{E}$. Two vertices $v, w \in V$ are *adjacent* if there is a hyperedge $\hat{e} \in \hat{E}$ such that $\{v, w\} \subseteq \hat{e}$. Two hyperedges $\hat{e}, \hat{f} \in \hat{E}$ are *adjacent* if $\hat{e} \cap \hat{f} \neq \emptyset$.

A *directed hypergraph* $H = (V, \hat{A})$ consists of a set of vertices V and a set $\hat{A}$ of hyperarcs $\hat{a} = (X, Y)$, that is, tuples of not necessary distinct subsets $X, Y \subset V$. $\mathsf{orig} \colon \hat{A} \to 2^V$ and $\mathsf{dest} \colon \hat{A} \to 2^V$ are two functions such that for all arcs $\hat{a} \in \hat{A}$, $\mathsf{orig}(\hat{a}) \subseteq V$ is the *origin* and $\mathsf{dest}(\hat{a}) \subseteq V$ is the *destination* of $\hat{a}$.

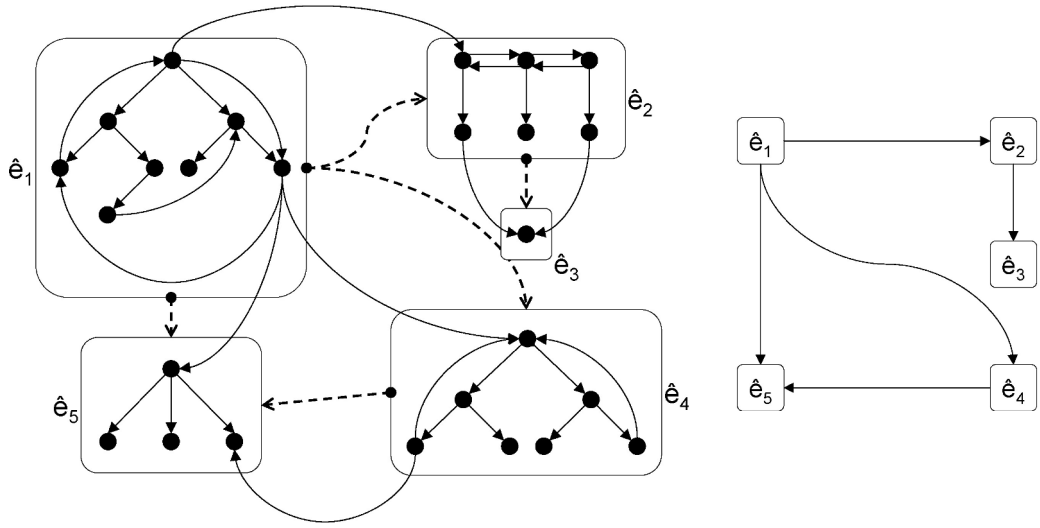


Figure 4: A type hypergraph as a model of a wiki document network (left) and its realization (right).

Definition 3.6. (Realization of a Hypergraph) An undirected graph $G = (V, E)$ is called *realization of a hypergraph* $H = (V, \hat{E})$ if it meets the following conditions:

1. $\forall e \in E \exists \hat{e} \in \hat{E} : e \subseteq \hat{e}$.
2. For any hyperedge $\hat{e} \in \hat{E}$ the subgraph $G(\hat{e}) = (\hat{e}, \{e \in E \mid e \subseteq \hat{e}\})$ of G induced by $\hat{e}$ is connected.

A *realization of a hyperedge* $e \in \hat{E}$ of H in G is a connected subgraph $G_e = (e, E)$ of G. An *oriented realization of* e is an orientation of G_e, that is, a directed subgraph of an orientation of G.

Definition 3.7. (Directed Graph Induced by a Directed Hypergraph) Let $H = (V, \hat{A})$ be a directed hypergraph. H induces a directed graph $\mathsf{digraph}(H) = (V', A')$ such that $V' = \{\mathsf{orig}(\hat{a}), \mathsf{dest}(\hat{a}) \mid \hat{a} \in \hat{A}\} \setminus \{\emptyset\}$ and $A' = \{a' \mid \exists \hat{a} \in \hat{A} \colon \mathsf{orig}(\hat{a}) = \mathsf{in}(a') = X \in V' \wedge \mathsf{dest}(\hat{a}) = \mathsf{out}(a') = Y \in V'\}$.

Now we are in a position to model wiki document networks which simultaneously manifest the linkage of wiki pages by hyperlinks (modeled as edges of type graphs) and of documents (modeled as hyperedges) whose internal structure is mapped by means of the notion of a generalized tree. That is, wiki document networks are defined as a synthesis of graphs and hypergraphs spanned over them:

Definition 3.8. (Type Hypergraph) Let $G = (V, E, \{\mathcal{V}_i\}_{i \in I \cup \{0,1\}}, \{\mathcal{E}_j\}_{j \in J}, \omega)$ be a type graph with the k-partition $\mathbb{P}_{\mathcal{V}_1}(V) = \{V_1, \ldots, V_k\}$ induced by $\mathcal{V}_1$. A *type hypergraph* $H = (V, E, \hat{E}, \hat{A}, \{\mathcal{V}_i\}_{i \in I \cup \{0,1\}}, \{\mathcal{E}_j\}_{j \in J}, \omega)$ is a type graph together with a collection $\hat{E}$ of hyperedges such that for any $\hat{e} \in \hat{E}$ there exists a generalized tree-like realization $\mathsf{Rel}(\hat{e})$ in G where no two hyperedges $\hat{e}, \hat{f} \in \hat{E}$, $\hat{e} \neq \hat{f}$, have the same realization $\mathsf{Rel}(\hat{e}) = \mathsf{Rel}(\hat{f})$. A *generalized tree-like realization* $\mathsf{Rel}(\hat{e}) = (V', E')$ of a hyperedge $\hat{e} \in \hat{E}$ is a directed subgraph of G with $V' = \hat{e}$ and $E' = \cup_{i=1}^{7} E'_{[i]}$ such that $GT(r) = (V', E'_{[1..7]}, r, \mathcal{V}_0, \mathcal{V}_1)$ is a labeled typed generalized tree with root r. $GT(r)$ is called generalized tree induced by $\hat{e}$ and denoted by $GT(\hat{e})$. Further, $\hat{A}$ is a collection of hyperarcs such that for any $\hat{a} \in \hat{A} \colon \mathsf{orig}(\hat{a}), \mathsf{dest}(\hat{a}) \in \hat{E}$.

Remark. This rather difficult definition is exemplified in Figure 4: its left part shows a structural configuration as being typical of wiki document networks. The undirected hyperedges $\hat{e}_1, \ldots, \hat{e}_5 \in \hat{E}$ model wiki documents, e.g., portals, talks with subpages and unstructured articles (see, e.g., hyperedge $\hat{e}_3$). Any of these hyperedges has a separate realization in the form of a generalized tree induced by its elementary wiki pages (filled circles) and their hyperlinks (directed arcs). Further, the hyperedges $\hat{e}_1, \ldots, \hat{e}_5$ serve as the origins and destinations of hyperarcs (whose visualizations start with a bullet and end with an arrow) of a directed hypergraph. Finally, this directed hypergraph induces a directed graph according to Definition 3.7 which is exemplified in the right part of Figure 4.

Figure 4 demonstrates that given a wiki document network we can always think of two closely related graphs: the directed graph of the elementary wiki pages and the graph induced by the directed hypergraph spanned over it. Note once more that this kind of possibly recursive structure formation has been left out in previous approaches to wiki networks. Note also that our approach also accounts for situations in which the same page belongs to different documents (modeled by different hyperedges).

After having specified how to treat wiki documents in relation to pages we finally need to decide what to do with additional type systems. This includes, for example, the types *disambiguation* and *redirect* (see Table 3): a wiki page typed as an article may additionally serve as a disambiguation page while a wiki page formerly typed as an article may be attributed as a redirect node so that it is no longer directly accessible. Thus, we cannot add these and related types to the set T_{wiki} because this would prevent to induce partitions of the vertex set V of type graphs. An alternative is to map such supplements by additional labeling functions $\mathcal{V}_i$ for some $i > 1$ in order to retain the special role of $\mathcal{V}_0$ (i.e. node labeling) and $\mathcal{V}_1$ (i.e. namespace mapping).

So far we did not specify a system of arc types. However, it is easily seen that this is a further source of clarification with a tremendous impact on the topology of the final graph input to network analysis. Take, for example, pages typed as talks. If we take them into account we have to decide how they are linked with their corresponding articles as there are barely explicit links from articles to their talks. Rather, it is an artefact of the wiki management system how to interrelate articles and talks, whether, for example by some hyperlink or by tabbed windows. What seems to be a trivial question at first glance is actually of importance when considering the cluster value of a network as talk pages are automatically created so that they may induce a higher number of connected triples in relation to the number of triangles in the network. Actually, there are many related questions to be handled when dealing with wikis which all need to be carefully answered. Thus, any decision on what types of wiki pages are considered has to be carefully justified.

3.3 More than the Sum of its Parts: Wiki Networks as Multilevel Graphs

So far, we have modeled the vertex inducing textual part of wiki networks. In this section we complement two models: that of a network of agents who have authored the wiki and that of a network of lexical means used by the agents to write the wiki. Our starting point is the model of a multilevel network as exemplified in Section 1, Figure 1. It exemplifies a threelevel network of interlinked agents, documents and words as prototypically exemplified by wikis. This multilevel network is captured by the following three definitions:

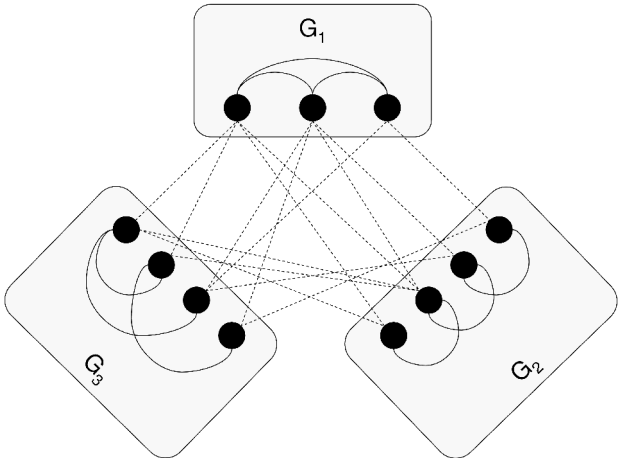


Figure 5: A multilevel graph stratified into three component graphs (edges between vertices of different component graphs are denoted by dashed lines).

Definition 3.9. (Multilevel Graph) Let $G = (V, E, \mathcal{V})$ be a k-partite graph such that the vertex labeling function $\mathcal{V} : V \to \{1, \ldots, k\}$ maps each vertex $v \in V$ onto its unique level $1 \leq \mathcal{V}(v) \leq k$. Starting from G we build a *k-level graph* (or *multilevel graph of order k*) $M = (V, E', \mathcal{V})$ with $E' = E \cup E''$, $E'' \subseteq \{\{v, w\} \,|\, \mathcal{V}(v) = \mathcal{V}(w)\}$ such that $\forall v \in V \forall i \in \{1, \ldots, k\} \setminus \mathcal{V}(v) \exists w : \mathcal{V}(w) = i \wedge \{v, w\} \in E$. A k-level graph M is said to be *stratified* into $G_1, \ldots, G_k$ where for $i \in \{1, \ldots, k\}$ $G_i = (V_i, E_i)$ is the ith component graph of M such that $V_i = \{v \in V \,|\, \mathcal{V}(v) = i\}$ and $E_i = \{e \in E'' \,|\, \mathcal{V}(\mathsf{in}(e)) = \mathcal{V}(\mathsf{out}(e)) = i\}$. Finally, a *directed multilevel graph* is a graph whose underlying graph is an undirected multilevel graph.

Remark. By analogy with the notion of a generalized tree one might think that multilevel graphs are hardly informative as there are many possibilities to build them out of a given graph. We reject this view for the very same reasons as we referred to when commenting on the definition of generalized trees (cf. Definition 3.1). The reason is that multilevel graphs are introduced as formal models of empirical systems which are structured far away from the principle of chance. Multilevel graphs are defined the way they are to grasp those restrictions which actually constrain the structure formation in real wikis.

Now, we instantiate the notion of a multilevel graph by means of wiki graphs with exactly three levels:

Definition 3.10. (Wiki Graph) A directed *wiki graph* is a directed 3-level graph $G_{\text{wiki}} = (V_{\text{w}}, E_{\text{w}}, \mathcal{V}_{\text{w}}, \mathcal{E}_{\text{w}}, H_{\text{w}})$ stratified into G_1, G_2, G_3 such that $H_{\text{w}} = (V, E, \hat{E}, \hat{A}, \{\mathcal{V}_i\}_{i \in I \cup \{0,1\}}, \{\mathcal{E}_j\}_{j \in J}, \omega)$ is a type hypergraph based on the vertex typing $\mathcal{V}_1 : V \to \mathrm{V}_1$ (according to Definition 3.4) and where

1. $G_1 = (V_1, E_1)$ is called *agent graph* whose vertices and arcs are called *agents* and *collaborations*, respectively. We set: $\forall v \in V_1 : \mathcal{V}_{\text{w}}(v) = \mathsf{agent}$.

2. $G_2 = (V_2, E_2)$ is called *document graph* whose vertices and arcs are called *documents* and *intertextual relations*, respectively. We set: $\forall v \in V_2 : \mathcal{V}_{\text{w}}(v) = \mathsf{document}$.

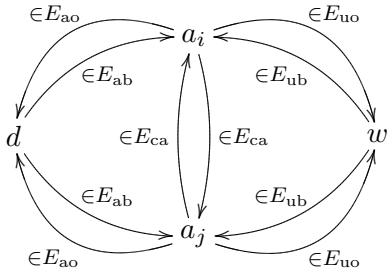
3. $G_3 = (V_3, E_3)$ is called *word graph* whose vertices and arcs are called *words* and *word associations*, respectively.[24] We set: $\forall v \in V_3 \colon \mathcal{V}_{\mathrm{w}}(v) = \mathsf{word}$.

We call G_{wiki} *(wiki) page-centered* if $V_2 = V$ and $E_2 = E$. Otherwise, if $V_2 = V'$ and $E_2 = A'$ for the directed graph $\mathsf{digraph}(H) = (V', A')$ induced by $H = (V, \hat{A})$ according to Definition 3.7, we call G_{wiki} *(wiki) document centered*. H_{w} is called *Support Document Network* (SDN) of G_{wiki} in order to stress its role as a mediator between the agent graph G_1 and the word graph G_3.[25] We utilize the following conventions for naming arcs $e \in E_{\mathrm{w}}$:

$$
\begin{array}{rrcll}
\textit{coauthor of:} & e \in E_{\mathrm{ca}} & = & E_{\mathrm{w}} \cap V_1 \times V_1 & \Rightarrow \quad \mathcal{E}_{\mathrm{w}}(e) = \mathsf{ca} \\
\textit{linked with:} & e \in E_{\mathrm{lw}} & = & E_{\mathrm{w}} \cap V_2 \times V_2 & \Rightarrow \quad \mathcal{E}_{\mathrm{w}}(e) = \mathsf{lw} \\
\textit{associated with:} & e \in E_{\mathrm{aw}} & = & E_{\mathrm{w}} \cap V_3 \times V_3 & \Rightarrow \quad \mathcal{E}_{\mathrm{w}}(e) = \mathsf{aw} \\
\textit{author of:} & e \in E_{\mathrm{ao}} & = & E_{\mathrm{w}} \cap V_1 \times V_2 & \Rightarrow \quad \mathcal{E}_{\mathrm{w}}(e) = \mathsf{ao} \\
\textit{authored by:} & e \in E_{\mathrm{ab}} & = & E_{\mathrm{w}} \cap V_2 \times V_1 & \Rightarrow \quad \mathcal{E}_{\mathrm{w}}(e) = \mathsf{ab} \\
\textit{user of:} & e \in E_{\mathrm{uo}} & = & E_{\mathrm{w}} \cap V_1 \times V_3 & \Rightarrow \quad \mathcal{E}_{\mathrm{w}}(e) = \mathsf{uo} \\
\textit{used by:} & e \in E_{\mathrm{ub}} & = & E_{\mathrm{w}} \cap V_3 \times V_1 & \Rightarrow \quad \mathcal{E}_{\mathrm{w}}(e) = \mathsf{ub} \\
\textit{cotext of:} & e \in E_{\mathrm{co}} & = & E_{\mathrm{w}} \cap V_2 \times V_3 & \Rightarrow \quad \mathcal{E}_{\mathrm{w}}(e) = \mathsf{co} \\
\textit{occurring in:} & e \in E_{\mathrm{oi}} & = & E_{\mathrm{w}} \cap V_3 \times V_2 & \Rightarrow \quad \mathcal{E}_{\mathrm{w}}(e) = \mathsf{oi}
\end{array}
$$

Based on these naming conventions the following commutativity constraints are demanded to be fulfilled by wiki graphs:

- *Agent-centered relations:* For all collaborating agents $a_i, a_j \in V_1$ there exists at least one document $d \in V_2$ and one word $w \in V_3$ such that the following diagram commutes:

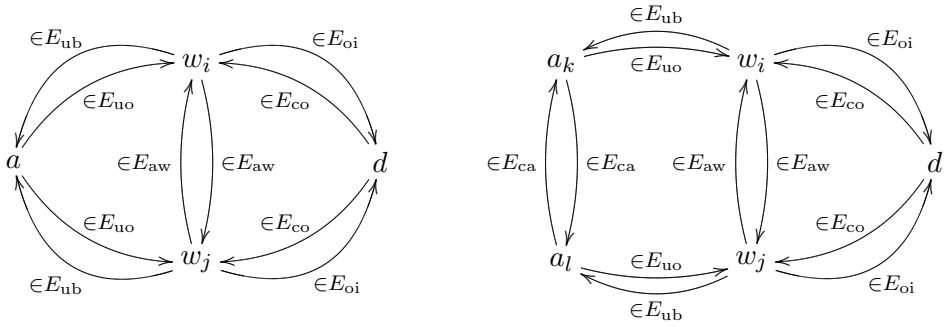


- *Word-centered relations:* For any pair of words $w_i, w_j \in V_3$ associated with each other (such that $\{\mathsf{in}(e), \mathsf{out}(e)\} = \{w_i, w_j\}$ for some association $e \in E_3$) there exists at least one agent $a \in V_1$ (or a pair of agents $a_k, a_l \in V_1$) and one document $d \in V_2$ such that the left (or the right) diagram

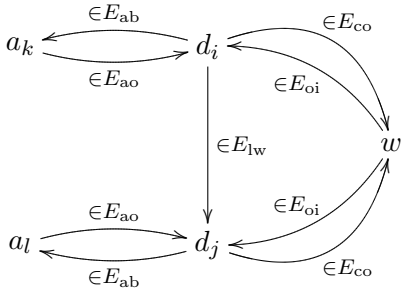
[24] A more general approach would be to speak of lexico-grammatical units instead of words.

[25] The mediator role of H_{w} is visualized in Figure 1 in Section 1.

(respectively) commutes:



- *Document-centered relations:* For all pairs of documents $d_i, d_j \in V_2$ linked with each other (such that $\mathsf{in}(e) = d_i$ and $\mathsf{out}(e) = d_j$ for some intertextual relation $e \in E_2$) there exist — possibly identical — agents $a_k, a_l \in V_1$ and at least one word $w \in V_3$ such that the following diagram commutes:



Finally, we call the underlying graph of G_{wiki} an *undirected wiki graph*.

Note that Definition 3.10 introduces wiki graphs as directed graphs. Their undirected counterpart is defined by abstracting from the orientation of agent, document and word relations. As our network-related classification experiments are mainly based on undirected graphs we apply this abstraction throughout the paper (making it explicit whenever we analyze directed instead of undirected networks).

The next and final step in defining wiki graphs is to introduce weighted arcs. This is not only motivated by the fact that associations of words and documents tend to be graded (ranging between maximal repugnance and affinity, respectively). Rather, recent developments in complex network theory extend the classical graph model of Watts and Strogatz (1998) and of Barabási and Albert (1999) in terms of weighted graphs (Serrano et al., 2006a). The following definition reflects this extension on the level of wiki graphs.

Definition 3.11. (Weighted Wiki Graph) A *weighted wiki graph* $G_{\mathrm{wiki}} = (V_{\mathrm{w}}, E_{\mathrm{w}}, \mathcal{V}_{\mathrm{w}}, \mathcal{E}_{\mathrm{w}}, H_{\mathrm{w}}, \omega_1, \omega_2, \omega_3)$ is a wiki graph with the support document network $H_{\mathrm{w}} = (V, E, \hat{E}, \hat{A}, \{\mathcal{V}_i\}_{i \in I \cup \{0,1\}}, \{\mathcal{E}_j\}_{j \in J}, \omega)$ and the weighting functions $\omega_i : E_i \to \mathbb{R}$ for each component graph $G_i = (V_i, E_i)$, $i \in \{1, 2, 3\}$, such that $\omega_2 = \omega$.

Example. *A wiki graph can be extracted from a wiki document network by extracting an agent, a document and a word graph as follows:*

- *An agent graph is extracted by inducing an agent vertex per user address (i.e. a user name or a TCP/IP address) where two agents are linked by multiple arcs in opposite direction if there is at least one page coauthored by them. Further, arcs between agents can be weighted by a function of the number of their coauthored pages. Obviously, there exist many other reference points of arc weighting. This may include, for example, a function of the proportion of co-authored texts in terms of authored words or sentences.*

- *A document graph is spanned, for example, by the network of wiki pages and their hyperlinks — in this case we consider a* page centered *wiki graph. Now, an arc between two pages can be weighted, for example, by a measure of the syntagmatic or paradigmatic similarity of their vocabularies (Mehler, 2002, 2005). An alternative is to start from the directed graph of wiki documents as explained in Section 3.2. In this case we would consider a* document centered *wiki graph. Note that both kinds of graphs can be extracted from the support document network of the input wiki graph. Note also that for any hyperlink l from a page A to a page B there exists a backlink so that we always get multiple arcs typed as* lw.

- *Finally, a word graph is extracted by applying some measure of lexical relatedness, similarity or association (Budanitsky and Hirst, 2006; Heyer et al., 2006; Ruge, 1995; Schütze, 1998) which operates on the corpus of textual nodes of the document graph. Thus, any two words are associated by multiple arcs if their association is confirmed by their usage in the document graph above a certain threshold of minimal affinity.*

Wiki graphs as introduced here hint at a fundamental difference of the present approach in comparison to previous studies on linguistic networks (cf. the survey in Mehler, 2008a): Not only that we start from a polypartite model. Rather, we also open the perspective on regularities of inter-level networking as mostly left out in other studies. The reason is that linguistic networking happens on ontologically different levels as well as between them. Thus, Definition 3.10 does not only provide a graph model for studying regularities of networking of either agents, documents or words (e,g, in terms of their small world status), but also of agents in relation to documents (authorship), agents in relation to lexical units (lexical usage regularities) or documents in relation to words (lexical cohesion). Linguistic networks are characterized by this multitude of networking and this makes them unique.

So far we have developed a formal framework of multilevel graphs. This is needed to clarify decisions obscured in previous approaches but to be made explicit when analyzing wikis. Formal models of the kind presented here do not only address the analysis of linguistic networks but also their synthesis as, e.g., in the area of computer-based simulations of language evolution (Mehler, 2008b). Wikis provide an excellent object for both of these research directions as they manifest all steps of the life cycle of social ontology formation and text production.

The next section introduces the web mining apparatus by which we automatically classify the component graphs of wiki networks.

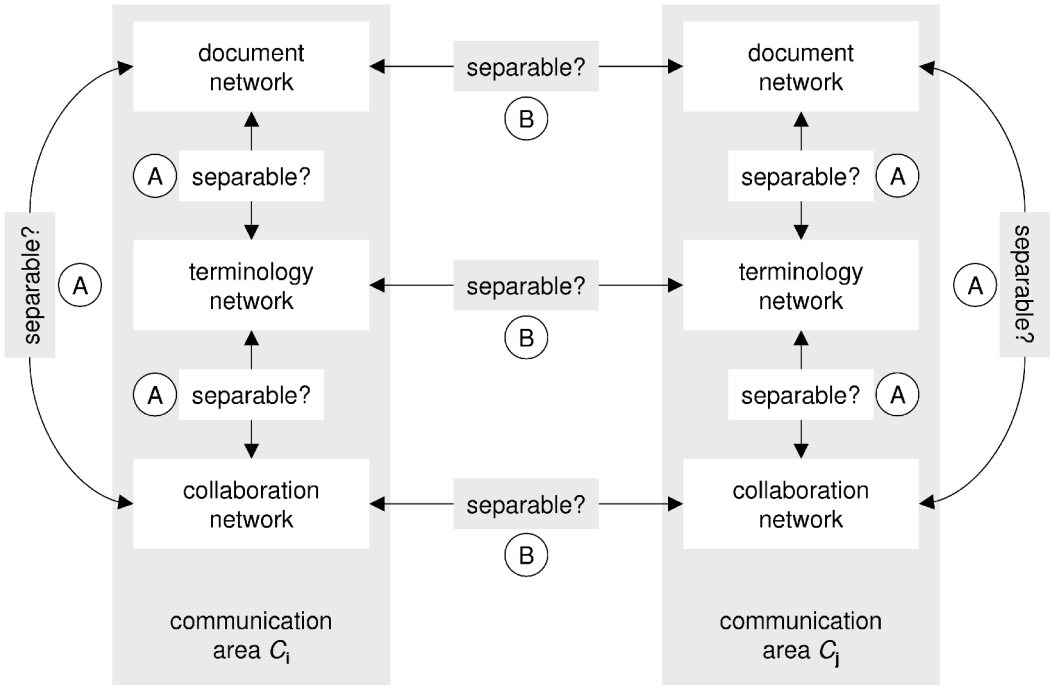


Figure 6: Classification scenarios of network analysis by example of wiki networks and their graph-theoretical representation.

4 Quantitative Network Analysis (QNA)

Based on the model of wiki graphs and its instantiation — by means of specifying the operative unit of the document graph and the wiki type system $\mathrm{T}_{\mathrm{wiki}}$ — we now present an algorithm of network classification based thereon. More specifically, we develop an algorithm for classifying the component graphs of wiki graphs. This is done regarding two different classification scenarios: On the one hand, we explore the separability of ontologically divergent networks — in our case these are agent, document and word graphs. On the other hand, we explore the separability of different communication areas by means of their wiki component graphs each of the same level, that is, either of the agent, document or word level. Thus, we examine two different separability hypotheses (cf. Figure 6):

- *Ontological separability:* Starting from a corpus of wiki networks of a certain communication area $\mathcal{C}$ and sampling their component graphs into a set X we ask for the separability of these ontologically divergent agent, document and word graphs (see Figure 6, Case A). That is, we aim to automatically classify the elements of X into the subset X_1 of agent, X_2 of document and X_3 of word graphs such that $X_1 \cap X_2 = X_1 \cap X_3 = X_2 \cap X_3 = \emptyset$ and $X = X_1 \cup X_2 \cup X_3$. As such perfect partitions are mostly out of reach we ask, more specifically, for the classification which best approximates this reference classification and use F-measure statistics to evaluate their difference (Hotho et al., 2005). Beyond the representation models of the input graphs (see below), the only information input to the classification algorithm is the number of classes.

- *Functional separability:* Secondly, we start from n communication areas $\mathcal{C}_1, \ldots, \mathcal{C}_n$ and sample either their agent, document or word graphs into the set X. In this case we ask for the separa-

bility of these ontologically homogeneous, though functionally heterogeneous graphs (cf. Figure 6, Case B). Once more we utilize the F-measure statistics in order to evaluate the classification and do this separately for the corpus of agent, document and word networks, respectively. As explained in Section 2 we consider wikis of leisure, knowledge, regional, scientific and technical communication (cf. Table 2).

- *Sensitivity analysis:* In each of these tasks in network classification we additionally ask for the best performing subset of quantitative features of the input graphs, that is, for those topological network characteristics which maximize the F-measure value of the corresponding classification experiment. In other words, we perform a feature-related sensitivity analysis. By means of this procedure we seek the most informative network characteristics which best separate either ontologically or functionally disparate networks. In this sense, we provide a means to measure the information value of topological network characteristics and exemplify this by wiki graphs. This procedure may also be used in other areas of complex network analysis where the information value of network characteristics is under consideration.

The notion of ontological separability directly relates to the network separability hypothesis of Section 1. Its counterpart in terms of functional separability extends this hypothesis to the area of functionally divergent networks of divergent areas of communication which, amongst others, differ in terms of the predominant purpose of communication (e.g. dissemination of information vs. self-portrayal). In order to build a representation model of wiki graphs whose instances are input to network classification along these two lines we utilize quantitative network characteristics. This approach follows Q*uantitative* S*tructure* A*nalysis* (QSA) which assumes that classifiers of textual units can be reliably learned by solely exploring quantitative features of their structure. Mehler et al. (2007a) show that this approach outperforms structure kernel-based approaches with a remarkable F-measure value. A central outcome of their study is that structure-oriented classifications are a real alternative to content-oriented approaches as, for example, the vector space approach (Salton, 1989). Pustylnikov (2007) as well as Pustylnikov and Mehler (2008) show that QSA yields reliable classifiers for a wide range of discourse types and, thus, proves operability in several linguistic areas. In this paper, we extend QSA by Q*uantitative* N*etwork* A*nalysis* (QNA). The connecting factor of QNA to QSA is the common emphasis on structure as a source of object identity. The point of departure is that in this paper large networks are considered instead of trees as done by QSA.

Generally speaking, instantiating QSA in terms of QNA demands to perform the following working steps (cf. Mehler et al. 2007a for a formalization of this procedure by example of tree-like structures):

1. *Segmentation:* The first step is to *segment* the constituents (e.g. nodes and links) of the target objects (i.e. networks). In the present paper this working step demands segmenting three sorts of component graphs of wiki networks (cf. Definition 3.10 and 3.11). This is done in order to gain units of measurement. Thus, vertices and edges are referred to as units of measurement which are typed according to the wiki ontology in Table 3, that is, according to the set $\mathrm{T}_{\mathrm{wiki}}$.

2. *Feature selection and validation:* Secondly, we select a set of quantitative features, validate them on the basis of the selected units of measurement and build distributions of these features

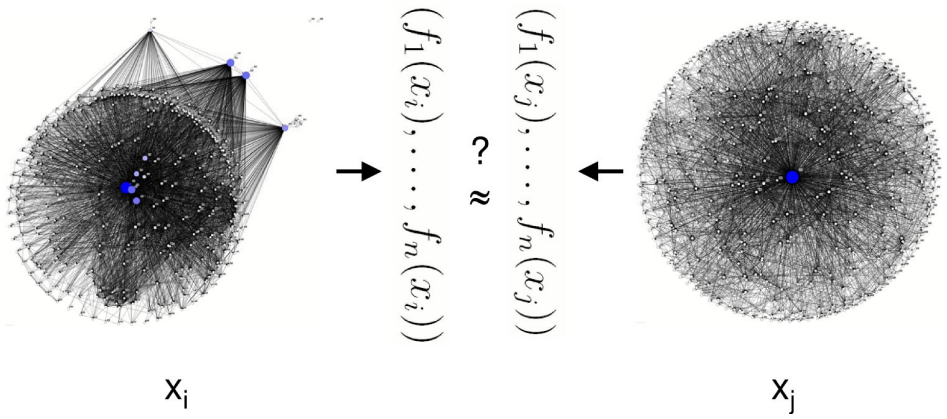


Figure 7: Mapping two input networks onto vectors of composite features as a prerequisite of validating their similarity.

by which the target objects are characterized. Think, for example, of betweenness centrality (Wasserman and Faust, 1999) as such a feature which for a given graph G is computed for each of its vertices so that G is mapped onto the distribution of the centrality values of its vertices. The quantitative features selected in this step are henceforth called *constituent features* as they are validated directly by the constituents of target objects.

3. *Feature aggregation & target object representation:* Thirdly, the distributions of constituent features are aggregated by some measures of location, dispersion, skewness or kurtosis. Additionally, they may be fitted by a distribution model. This is done in order to validate the model's parameters (as, e.g., the exponent of a power law) together with the corresponding coefficient of determination as additional network features. As a result, each target object is mapped for each of its distributions of constituent features onto one or more target object-related features henceforth called *composite features*. Next, each target object is represented by the vector of its composite feature as exemplified in Figure 7 by two target networks.

Finally, the feature vectors of target objects are input to supervised or unsupervised learning. Thus, QSA and QNA basically consist of selecting appropriate constituent features and aggregation functions which reliably separate the graph classes to be learnt. In this paper, we explore the following range of features:

- Firstly, we utilize features by which networks have been successfully described as small worlds (Newman, 2003; Pastor-Satorras and Vespignani, 2004; Watts, 2003).
- Secondly, we consider two composite features from earlier studies on structural properties of hypertexts (Botafogo et al., 1992; Dehmer, 2005a).
- Thirdly, we take a group of centrality features into account which have been explored in the field of *social* network analysis (Wasserman and Faust, 1999).

In this paper, we do not present yet another study of the SW-property of wiki networks. Rather, we seek quantitative features by which ontologically and functionally heterogeneous networks can be

reliably classified even if they are uniformly attributed as small worlds (cf. the network separability hypothesis of Section 1). The following enumeration describes the composite features used for this purpose. Thereby, $G = (V, E)$ denotes a component graph of a wiki graph. Each of the features being enumerated will be provided with an exemplary interpretation in terms of document networks which are spanned by intertextual relations. That is, intertextuality will be referred to as our *reference example* in order to illustrate the effect of the corresponding features.

Transitivity The *cluster* or *transitivity value* of a graph G is computed as the average cluster value of its vertices — $d_G(v_i)$ is the degree of vertex v_i (Watts and Strogatz, 1998):

$$C_1(G) = \frac{1}{n}\sum_{i=1}^{n} c_{ws}(v_i) = \frac{1}{n}\sum_{i=1}^{n} adj(v_i)/\binom{d_G(v_i)}{2} \in [0,1] \tag{1}$$

$adj(v_i)$ is the number of edges ending only at neighbors of v_i. We utilize c_{ws} as a *constituent* feature of vertices which is aggregated by C_1 as a *composite* feature of G. C_1 estimates the probability by which two vertices v, w are adjacent when linked with some $u \in V$. In the case of document networks, high cluster values indicate that documents which are adjacent to the same node are probably related on the same ground. In this case, intertextual relations diversify to a low degree in terms of the criteria by which they are spanned. If the criterion of intertextuality is, for example, *thematic relatedness*, then nodes linked from the same source tend to be themselves thematically related. In such a network intertextuality generates many triangles of interlinked documents. Note that this does not say anything about the existence or nonexistence of higher order document clusters. An alternative to C_1 is the cluster coefficient C_2 of Bollobás and Riordan (2003) which takes multiple edges into account. As we utilize C_1 and C_2 we get two composite features enumerated as F_1 and F_2 (see Table 8).

Path Distances The second feature being explored is the *average geodesic distance* of networks:

$$L(G) = \frac{1}{\binom{|V|}{2}} \sum_{\{v,w\}\in[V]^2} \delta(v,w) \tag{2}$$

where $[V]^2$ denotes the set of all subsets of 2 elements of V. Together with C_1, L defines the Watts & Strogatz model (WSM) of small worlds. As C_1, L is a composite feature using the mean as the operative aggregation function. In the case of thematically diversified document networks, small values of $L(G)$ indicate, for example, that the topic of documents changes rapidly when following a couple of intertextual relations. In this sense, L may help to distinguish document networks with rapidly diverging trails of intertextual relations — in the sense of the operative criterion of intertextuality — from networks with more convergent paths. Further, L may be explored as a starting point to estimate the probability by which recipients get "lost in the space of intertextuality" when following document relations. An alternative to L is the diameter $D(G)$ of G which is the length of the longest geodesic distance among its vertices. It uses the maximum as the operative aggregation function and will be considered as a composite feature of networks, too.

Assortativity Next, we consider *assortative mixing* which is computed by the following correlation coefficient (Newman and Park, 2003b):

$$r(G) = \frac{\frac{1}{m}\sum_i j_i k_i - \left[\frac{1}{m}\sum_i \frac{1}{2}(j_i + k_i)\right]^2}{\frac{1}{m}\sum_i \frac{1}{2}(j_i^2 + k_i^2) - \left[\frac{1}{m}\sum_i \frac{1}{2}(j_i + k_i)\right]^2} \in [-1, 1] \tag{3}$$

i denotes the edge ending at vertices v_j and v_k of degree $d(v_j) = j_i$ and $d(v_k) = k_i$, respectively, while $m = |E|$. $r(G)$ is a measure of the degree by which vertices with alike connectivity patterns tend to be linked (assortative) or not (disassortative mixing). Obviously, $r(G)$ is a composite feature which utilizes a correlation measure as the operative aggregation function while the degrees of interlinked vertices are explored as the constituent feature. At first glance, $r(G)$ promises a high separability with respect to ontologically divergent networks. The reason is that Newman and Park (2003b) show that social networks tend to assortative mixing while technical networks tend towards the opposite direction. In our reference example of intertextuality a high value of $r(G)$ indicates that highly connected entry points (with a somehow generic content) to the network tend to be linked with each other while segregated documents stay apart from them. This suggests a hierarchical topology where the more generic the content of two documents the larger the number of their intertextual relations, the more likely they are directly linked with each other, and vice versa, — somehow alike to the structure of an ontology.

Huge Connected Components The *fraction* $\mathsf{lcc}(G) = |W|/|V|$ of the number of vertices of the *Largest Connected Component* (LCC) of G in relation to its order is another standard characteristic of small worlds which are known for large values of lcc (Newman, 2003). In the case of our reference example such a large value indicates the existence of a large connected subspace of intertextually related documents with few thematically unrelated satellites. This also indicates a connected space of topics where disconnected themes induce only small document subspaces. Note that lcc does not say anything about the compactness of G as its LCC may either be given by a linear graph, by a completely connected graph or by something in-between.

Size Distribution Somehow related to $\mathsf{lcc}(G)$, but possibly more informative, is the exponent γ_S of the power law which best fits to the size distribution of the connected components of G. That is, we expect that the probability $P(n)$ of connected components of G of size n is distributed according to a power law

$$P(n) \sim n^{-\gamma} \tag{4}$$

We check this hypothesis in conjunction with the corresponding adjusted coefficient of determination in order to examine the degree of disconnectedness in input networks.

Local Connectivity A simple network feature is the *average degree* $\langle d_G \rangle = \frac{1}{|V|}\sum_{v_i \in V} d_G(v_i)$ of vertices in G — which is based on averaging as the aggregation function operating on the degree as a constituent feature of vertices. We alternatively compute this feature by the fraction $\epsilon(G) = |E|/|V|$. The idea to take $\epsilon(G)$ into account is to check the separability of simple features in competition to

complex ones. Obviously, the larger the number of edges in relation to a given number of vertices, the more compact the document space. It is worth to see wether $\epsilon(G)$ outperforms more complex features as, for example, the compactness measure (see below) in terms of network classification.

Connectivity Distribution Next we consider the network model of Barabási and Albert (1999), that is, we take the exponent $\gamma(G)$ of the power law fitted to the distribution of the degrees of the vertices of G into account. This is done by means of the model $P(k) \sim k^{-\gamma}$ where $P(k)$ is the probability of a vertex with degree k. Together with the corresponding adjusted coefficient of determination $\bar{R}^2_{\gamma(G)}$ this gets two more composite features of G based on the degree as a constituent feature of vertices aggregated by γ. In the case of our reference example, larger values of γ indicate — in conjunction with a value of $\bar{R}^2_{\gamma(G)} \gg 0.75$ — the existence of a Zipfian distribution (Tuldava, 1998) of intertextual relations, that is, a very small group of highly connected documents, a large set of rarely connected nodes and a rapid transition in-between these extremal cases as predicted by the downward slope of a power-law with a negative exponent. Note that γ alone does not tell much about small world graphs but characterizes the class of scale-free networks. The reason is that there are, for example, social networks which combine scale-freeness with community building with the effect that their information processing dramatically diverges from the one of purely scale-free random graphs (Volchenkov et al., 2002). The same may hold in the case of document networks so that γ is better studied in conjunction with other network features as done here.

Connectivity Correlation By the same procedure we consider the network model of Pastor-Satorras et al. (2001) (see also Serrano et al., 2006b). That is, we compute $\bar{k}_{nn}(k)$ as the average degree of the nearest neighbors of vertices with degree k. If $\bar{k}_{nn}(k)$ tends to grow with k the network exhibits assortative mixing so that edges tend to connect vertices of similar degrees. If in contrast to this $\bar{k}_{nn}(k)$ shrinks as k grows, the network is disassortative. In this case, edges tend to connect vertices of dissimilar degrees. Interestingly, Pastor-Satorras et al. (2001) observe many networks in which the distribution of $\bar{k}_{nn}(k)$ is successfully fitted by a power law with a negative exponent β such that we can write:

$$\bar{k}_{nn}(k) \sim k^{-\beta} \tag{5}$$

Obviously, networks of this sort show disassortative mixing to a very high degree. We will test this model by example of wiki graphs and therefore consider two additional composite features, that is, $\gamma_{\bar{k}_{nn}(k)}(G) = \beta$ and $\bar{R}^2_{\gamma_{\bar{k}_{nn}(k)}}(G)$.

Hierarchization Next, we take the model of hierarchical network structures of Ravasz and Barabási (2003) into account. It considers the average cluster value $C(k)$ of vertices of degree k. Ravasz and Barabási (2003) observe that neither in random graphs à la Erdős and Rényi (1959) nor in scale-free graphs à la Barabási and Albert (1999) $C(k)$ is distributed according to a power law. However, such a power law-like distribution

$$C(k) \sim k^{-\alpha} \tag{6}$$

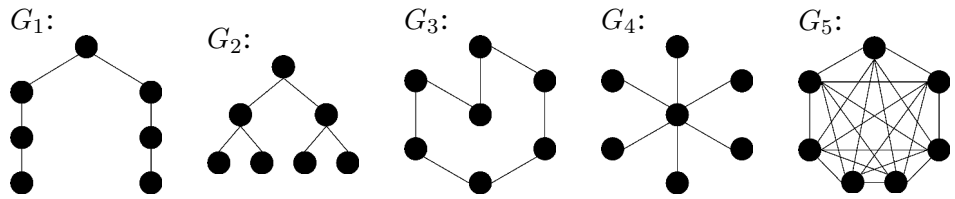


Figure 8: From left to right: a linear graph (G_1), tree-like graph (G_2), circle graph (G_3), star graph (G_4), and a completely connected graph (G_5) of the same order 7 (cf. Wasserman and Faust, 1999, 171). Table 7 summarizes some of their topological characteristics.

is observable in many real networks. Such networks do not only show scale-free degree distributions. Rather, their highly interlinked nodes tend to have smaller cluster values while rarely linked nodes tend to have higher ones. That is, the higher the degree of a vertex, the smaller its cluster value, the smaller the degree of group formation within the network. We test how far this model helps to classify ontologically and functionally different networks. Therefore, we compute $\gamma_{C(k)}(G)$ and $\bar{R}^2_{\gamma_{C(k)}}(G)$ as additional composite features.

Centrality Next, we consider *centrality measures* which evaluate the structural positions of the vertices of G. As before, these vertex-related features are aggregated in order to get graph-related ones. In the case of the betweenness centrality, we additionally consider the distribution of centrality values which — by analogy to the distributions discussed above — is aggregated by the exponent value of those power law which best fits to this distribution. More specifically, we consider the following centrality measures:

- *Degree centrality:* A simple graph-related centrality measure is the degree centrality of a graph (Feldman and Sanger, 2007):

$$\mathsf{DC}(G) = \frac{\sum_{v \in V} d_{\max}(V) - d(v)}{(|V|-1)(|V|-2)} \in [0,1] \tag{7}$$

  where $d_{\max}(V) = \max_{v \in V} d(v)$. $\mathsf{DC}(G)$ allows to characterize the topology of a graph in relation to the extremal cases of a circle graph $\mathsf{circ}(G)$ and a star graph $\mathsf{star}(G)$ of equal order (see Figure 8 and Table 7). The reason is that $\mathsf{DC}(\mathsf{circ}(G)) = 0$ while $\mathsf{DC}(\mathsf{star}(G)) = 1$. In order to relate DC to our reference example, we need to interpret circle and star graphs in terms of intertextuality: A star can be seen to indicate a hierarchical organization of a document network whose nodes are interrelated as rhematic units only by the topic (or theme) of their central thematic hub. In contrast to this a circle indicates thematically equally ordered document units without a thematic center manifested by a single text. Thus, the effective value of $\mathsf{DC}(G)$ says something about the status of G between these two extremal cases.

- The *(standardized) closeness centrality* (Wasserman and Faust, 1999) $\mathsf{CC}(v)$ of a vertex $v \in V$ is a function of the sum of its geodesic distances to the vertices of G:[26]

$$\mathsf{CC}(v) = \frac{|V|-1}{\sum_{w \in V} \mathsf{gd}(v,w)} \in [0,1] \tag{8}$$

[26] $\mathsf{CC}(v) \cdot (|V|-1)^{-1}$ is the classical closeness centrality. The factor $|V|-1$ is used to provide comparability among graphs of divergent order (cf. Wasserman and Faust, 1999).

$\mathsf{gd}(v, w)$ is the geodesic distance of v and w. A composite feature $\mathsf{CC}(G)$ is aggregated as the closeness centrality of G from this vertex-related constituent feature as follows:

$$\mathsf{CC}(G) = \begin{cases} \frac{\sum_{v \in V} \hat{\mathsf{CC}}_{\max}(V) - \hat{\mathsf{CC}}(v)}{|V|-1} & : \min_{v \in V} \mathsf{CC}(v) < 1 \\ 0 & : \min_{v \in V} \mathsf{CC}(v) = 1 \end{cases} \in [0, 1] \tag{9}$$

where $\hat{\mathsf{CC}}_{\max}(V) = \max_{v \in V} \hat{\mathsf{CC}}(v)$, $\hat{\mathsf{CC}}(v) = 1 - \frac{1-\mathsf{CC}(v)}{1-\min_{v \in V} \mathsf{CC}(v)}$ and $|V| > 1$. That is, we normalize the closeness centralities of single vertices in order to exploit the full interval $[0, 1]$ (where $\mathsf{CC}(G) = 1$ iff a single vertex is of maximal and all other of minimal centrality (as in the case of a star graph); on the other hand, $\mathsf{CC}(G) = 0$ iff the minimal centrality equals 1). In terms of our reference example, a document of maximal closeness centrality is the one with the minimal sum of geodesic distances to all other documents. This is a document which induces the least effort of thematic progression and change, respectively, when used as an entry point to the document network in order to reach any of the other documents. Note that we compute all centrality measures for the LCC of graphs only. As demonstrated in Table 7, CC spans an interval of centrality values which orders the graphs of Figure 8 the same way as done by the degree centrality DC. However, CC weights line graphs higher than DC, thereby taking indirectly linked nodes into account.

- The *(standardized) betweenness centrality* $\mathsf{BC}(v)$ of a vertex $v \in V$ is a measure of the probability by which it is traversed by the shortest paths between two randomly chosen vertices of G (Wasserman and Faust, 1999; Brandes, 2001):

$$\mathsf{BC}(v) = \frac{2}{(|V|-1)(|V|-2)} \sum_{u \neq v \neq w \in V} \frac{\sigma_{uw}(v)}{\sigma_{uw}} \tag{10}$$

where $\sigma_{uw}(v)$ is the number of shortest paths between u, w which traverse v and σ_{uw} is the number of shortest paths between u, w. As before, we aggregate the betweenness centralities of single vertices to get a composite feature of graph centralization (cf. Feldman and Sanger, 2007):

$$\mathsf{BC}(G) = \frac{\sum_{v \in V} \mathsf{BC}_{\max(V)} - \mathsf{BC}(v)}{|V|-1} \in [0, 1] \tag{11}$$

where $\mathsf{BC}_{\max}(V) = \max_{v \in V} \mathsf{BC}(v)$. The higher the number of shortest paths between vertices of G which go through v, the higher the betweenness centrality of v. In terms of our reference example, this is a document which is traversed by many thematic progressions, that is, a document which is likely processed when readers follow the trails of intertextual relations between interlinked documents. The betweenness centrality $\mathsf{BC}(G)$ of a graph is maximal for a star graph where a single vertex has maximal, while all other vertices have minimal betweenness centrality. Consequently, BC rates tree graphs higher than line graphs, while completely connected graphs and circle graphs are of minimal betweenness centrality (cf. Table 7). In addition to computing $\mathsf{BC}(G)$ we also fit the distribution of betweenness centralities $\mathsf{BC}(v)$. The reason is that Barthélemy (2004) has shown that these values tend to be distributed according to a power

	G_i	$\mathsf{BC}(G_i)$	$\mathsf{CC}(G_i)$	$\mathsf{DC}(G_i)$	$\mathsf{GC}(G_i)$	$\mathsf{Cp}(G_i)$	$\mathsf{Cp}_{\mathsf{lcc}}(G_i)$	$\mathsf{Ch}(G_i)$
line graph:	G_1	.31	.18	.07	.15	.67	0	.29
tree graph:	G_2	.4	.27	.3	.3	.74	.23	.29
circle graph:	G_3	0	0	0	0	.8	.4	.33
star graph:	G_4	1	1	1	1	.86	.57	.29
cc graph:	G_5	0	0	0	0	1	1	1

Table 7: The betweenness (BC), closeness (CC), degree (DC) and graph centrality (GC) as well as the Botafogo compactness (Cp), the compactness of the LCC ($\mathsf{Cp}_{\mathsf{lcc}}$) and the cohesion (Ch) of the five graphs of Figure 8 (values are rounded to maximally two decimal places).

law if the connectivity pattern of the corresponding network is scale-free (Barabási and Albert, 1999). More specifically:

$$\mathsf{BC}_k \sim k^{\eta} \tag{12}$$

where BC_k is the average standardized betweenness centrality of vertices with degree k. Hence, there seems to be a relationship of the connectivity patterns of vertices and their centrality values. In order to check this hypothesis we extract the exponent $\gamma_{\mathsf{BC}}(G)$ of the power law fitted to the distribution of betweenness centralities in G together with the adjusted coefficient of determination $\bar{R}^2_{\gamma_{\mathsf{BC}}}(G)$ as two additional composite network features (see Table 8).

- Finally, the *graph centrality* (Hage and Harary, 1995) $\mathsf{GC}(v)$ of a vertex $v \in V$ is a function of the maximum of its geodesic distances to the vertices of G:

$$\mathsf{GC}(v) = \frac{1}{\max_{w \in V} \mathsf{gd}(v, w)} \tag{13}$$

By analogy to the closeness centrality, a composite feature $\mathsf{GC}(G)$ is aggregated as follows:

$$\mathsf{GC}(G) = \frac{\sum_{v \in V} \hat{\mathsf{GC}}_{\max}(V) - \hat{\mathsf{GC}}(v)}{|V| - 1} \in [0, 1] \tag{14}$$

where $\mathsf{GC}_{\max}(V) = \max_{v \in V} \hat{\mathsf{GC}}(v)$ and $\hat{\mathsf{GC}}(v) = 1 - \frac{1 - \mathsf{GC}(v)}{1 - \min_{v \in V} \mathsf{GC}(v)}$. The vertex of highest graph centrality (or lowest eccentricity) is the one with the shortest maximum geodesic distance among all pairs of vertices of G. In terms of our reference example, this is the document which induces the least effort of thematic progression and change when used as an entry point to the network to make accessible its thematically most remote document. As the other centrality features considered here, GC spans an interval between circle and completely connected graphs on the one hand and star graphs on the other hand. However, it tends to weight line and tree graphs by analogy to CC thereby departing from the opposite cases of BC and DC (cf. Table 7).

Compactness As a reference to classical hypertext theory we consider the compactness measure of Botafogo et al. (1992) which indicates cross-referencing in a network. As we classify undirected graphs we redefine this measure as follows (note that this definition bewares to compute the converted

distance matrix of a graph):

$$\mathsf{Cp}(G) = \frac{\left(\mathsf{Max}(G) - \left(\sum_{v \in V} \sum_{w \in V} \mathsf{gd}(v,w) + D_{\max}(G) \sum_{G' \in \mathsf{Com}(G)} |G'||V| - |G'|^2\right)\right)}{\mathsf{Max}(G) - \mathsf{Min}(G)} \in [0,1] \quad (15)$$

where $\mathsf{Com}(G)$ is the set of connected components of G, $D_{\max}(G)$ is the maximum value the diameter can assume in a graph of equal order as G, $\mathsf{Max}(G) = D_{\max}(G) \cdot (|V|^2 - |V|)$ and $\mathsf{Min}(G) = (|V|^2 - |V|)$. Further, $|G| = |V|$ is the order of $G = (V, E)$. Finally, $\mathsf{Com}(G)$ is the set of connected components of G. Note that we compute the compactness of G as a whole and not only of its LCC. The value of $\mathsf{Cp}(G)$ varies from 0 (which indicates a completely disconnected graph) to 1 (in the case of a completely connected graph). Other than the stratum measure of Botafogo et al. (1992) which explores the formation of hierarchical structures, Cp focuses on cross-referencing links which raise the degree of connectedness.[27] At the same time, Cp also evaluates disconnectedness. The reason is that the larger the number of connected components of G, the smaller the value of Cp. This associates Cp with the exponent γ_S of the size distribution of connected components as well as with the fraction lcc of the largest connected component so that we do not need to repeat this part of our reference example. However, other than this exponent, Cp is defined even in the case of graphs with a single connected component. Further, other than γ_S and lcc, Cp analyses the internal structure of these components. In order to stress this role of Cp we compute the compactness $\mathsf{Cp}_{\mathsf{lcc}}(G)$ of the LCC of G as an additional composite feature as follows:

$$\mathsf{Cp}_{\mathsf{lcc}}(G) = 1 - \frac{1 - \mathsf{Cp}(\mathsf{lcc}(G))}{1 - \mathsf{Cp}(\mathsf{lg}(\mathsf{lcc}(G)))} \in [0,1] \quad (16)$$

where $\mathsf{lg}(\mathsf{lcc}(G))$ is a line graph of equal order as $\mathsf{lcc}(G)$, $|\mathsf{lcc}(G)| = n$ and

$$\mathsf{Cp}(\mathsf{lg}(\mathsf{lcc}(G))) = \frac{\mathsf{Max}(\mathsf{lcc}(G)) - \sum_{k=1}^{n} \left(\sum_{i=1}^{k-1} i + \sum_{j=k}^{n-1} j - k + 1\right)}{\mathsf{Max}(\mathsf{lcc}(G)) - \mathsf{Min}(\mathsf{lcc}(G))} \quad (17)$$

In order to continue our reference example we can interpret $\mathsf{Cp}_{\mathsf{lcc}}(G)$ as follows: the shorter the geodesic distances of the vertices of the LCC of G, that is, the shorter the thematic trails between intertextually linked documents of this component, the higher the compactness of G. Note that Cp is not directly based on edges, but evaluates the geodesic distances of paths so that we expect a difference to the cohesion measure (see below) which disregards indirectly linked vertices. Note further that $\mathsf{Cp}_{\mathsf{lcc}}(G)$ does not say anything about the size of the LCC of G in relation to $|V|$. Thus, we additionally consider the distribution of the size of connected components as well as $\mathsf{Cp}(G)$ which takes all components of G into account. As can be seen from Table 7, $\mathsf{Cp}_{\mathsf{lcc}}(G)$ spans a continuum with the extremal cases of a line graph (of minimal LCC-related compactness) and a completely connected graph (of maximal LCC-related compactness) of equal order as G. Obviously, the compactness of a star graph is higher than that of a circle graph, which in turn is higher than that of a tree graph. Thus, we might say that $\mathsf{Cp}_{\mathsf{lcc}}$ behaves other than any of the centrality measures discussed above since it gives circle graphs a remarkably high value while completely connected graphs are, of course, of highest compactness, but lowest graph-related centrality. The order of the values of the graphs in Table 7 is not different between $\mathsf{Cp}_{\mathsf{lcc}}$ and Cp as long as we consider the LCC — their difference only occurs if we look at disconnected graphs.

[27] We do not compute St here as it has been defined for directed graphs.

Index	Feature	Short Description	Area
F_1	$C_1(G)$	the cluster coefficient of G	1
F_2	$C_2(G)$	the cluster coefficient of G	1
F_3	$L(G)$	the average geodesic distance of G	1
F_4	$D(G)$	the diameter of G	1
F_5	$r(G)$	the degree of assortative mixing of G	1
F_6	$\epsilon(G)$	the average degree of G	1
F_7	$\mathsf{lcc}(G)$	the fraction of the largest connected component of G	1
F_8	$\gamma(G)$	the γ of the power law of type $Ck^{-\gamma}$ which best fits to the degree distribution of G	1
F_9	$\bar{R}^2_{\gamma}(G)$	the corresponding adjusted coefficient of determination	1
F_{10}	$\gamma_S(G)$	the γ of $Cn^{-\gamma}$ which best fits to the size distribution of connected components of G	1
F_{11}	$\bar{R}^2_{\gamma_S}(G)$	the corresponding adjusted coefficient of determination	1
F_{12}	$\gamma_{\bar{k}_{nn}(k)}(G)$	the γ of the power law of type $Ck^{-\gamma}$ which best fits to the distribution of $\bar{k}_{nn}$ values of G	1
F_{13}	$\bar{R}^2_{\gamma_{\bar{k}_{nn}(k)}}(G)$	the corresponding adjusted coefficient of determination	1
F_{14}	$\gamma_{C(k)}(G)$	the γ of the power law of type $Ck^{-\gamma}$ which best fits to the distribution of $C(k)$ values of G	1
F_{15}	$\bar{R}^2_{\gamma_{C(k)}}(G)$	the corresponding adjusted coefficient of determination	1
F_{16}	$\mathsf{BC}(G)$	the betweenness centrality of G	2
F_{17}	$\gamma_{\mathsf{BC}}(G)$	the γ of the power law of type Ck^{γ} which best fits to the distribution of $\mathsf{BC}(k)$ values of G	1
F_{18}	$\bar{R}^2_{\gamma_{\mathsf{BC}}}(G)$	the corresponding adjusted coefficient of determination	1
F_{19}	$\mathsf{CC}(G)$	the closeness centrality of G	2
F_{20}	$\mathsf{DC}(G)$	the degree centrality of G	2
F_{21}	$\mathsf{GC}(G)$	the graph centrality of G	2
F_{22}	$\mathsf{Cp}(G)$	the compactness of G	3
F_{23}	$\mathsf{Cp}_{\mathsf{lcc}}(G)$	the compactness of the largest connected component of G	3
F_{24}	$\mathsf{Ch}(G)$	the cohesion of G	3

Table 8: The list of composite features considered in the present study. They fall into three groups, that is, features of *complex network theory* (1), *social network analysis* (2) or *hypertext structure analysis* (3) (as indicated in the last column).

Cohesion As an alternative to compactness we consider cohesion as a related, though less computationally complex measure. It has been used by Winne et al. (1994) for — as it is now called — web usage mining (cf. Dehmer, 2005a). Formally, the cohesion $\mathsf{Ch}(G)$ of a graph G is defined for an undirected graph G as follows:

$$\mathsf{Ch}(G) = \frac{\sum_{v \in V} d_G(v)}{|V|^2 - |V|} \in [0, 1] \tag{18}$$

Obviously, this measure does not explore paths, but only edges. In this sense we expect a difference of graph weighting as reflected by Table 7: circle graphs, for example, have a higher cohesion — but lower compactness — value than star graphs simply because of the additional link they contain. This measure allows us to end up our reference example with the hint that cohesive document networks are those with many directly linked vertices. As this is an unrealistic scenario we do not expect a high separability by this measure. However, our classification experiment might falsify this expectation.

Altogether, the measures discussed so far and the composite network features derived thereof (cf. Table 8) give a set

$$\mathbb{F} = \{F_1, \ldots, F_{24}\} \tag{19}$$

Algorithm 1 Selecting the subset of attributes in conjunction with the winner clustering procedure and its optimal parameter setting which maximize the F-measure value.

Require: a classification $\mathbb{L}$ of a set $\mathbb{G}$ of objects and two collections X, Y of representations of objects $G_i \in \mathbb{G}$ based on a set of features $\mathbb{F}$ (cf. Equation 20 and 21)

Ensure: the subset $\mathbb{F}^* \subseteq \mathbb{F}$ of attributes which maximize the F-measure value of a semi-supervised classification of $\mathbb{G}$ using $\mathbb{L}$ as the reference classification

```
1: for ∀𝔽′ ∈ 2^𝔽 do
2:     for X[𝔽′] ∧ Y[𝔽′] do
3:         for ∀CM_j ∈ {ClusteringMethod_m | m ∈ 𝕄} do
4:             for ∀S_k ∈ SetOfParameterSettings(CM_j) do
5:                 ComputeF-MeasureValue(Z[𝔽′], CM_j, S_k), Z ∈ {X, Y}
6:             end for
7:         end for
8:     end for
9: end for
```

of 24 composite features which fall into three different groups of features related to complex network theory, social network analysis and hypertext structure analysis (cf. Table 8). According to the generic procedure of QNA the next step is *object representation*. That is, if $\mathbb{G} = \{G_i \,|\, i \in I\}$ is our input corpus of component graphs to be processed each graph $G_i \in \mathbb{G}$ is mapped onto a separate feature vector $\mathbf{v}'_i = (F_1(G), \ldots, F_{24}(G))$ where $F_i \in \mathbb{F}$. This gives a collection

$$X = \{\mathbf{v}_i \,|\, G_i \in \mathbb{G}\} \tag{20}$$

of $|\mathbb{G}|$ object vectors as induced by the graphs $G_i \in \mathbb{G}$.

Starting from these object representations we look for the best performing classification of $\mathbb{G}$ with respect to a given reference classification $\mathbb{L}$. That is, we seek a classification method together with the subset $\mathbb{F}' \in 2^{\mathbb{F}}$ of composite features which results in a classification $\mathbb{D}$ of elements in $\mathbb{G}$ maximizing the F-measure value with respect to a known classification $\mathbb{L}$. This is done with the help of Algorithm 1 which includes the following selections:

1. *Feature subset selection* (Row 1): According to Algorithm 1, feature subset selection occurs by selecting elements out of $2^{\mathbb{F}}$. This is practically impossible if the number of composite features is large. In such cases, we alternatively perform an evolutionary algorithm in order to search the best performing subset of features.

2. *Input matrix selection* (Row 2): This criterion selects between the candidate object representation sets X (defined by Equation 20) and Y which is computed as follows: Let $\sigma : \mathbb{R}^n \to [0, 1]$ be a similarity measure (e.g. the cosine measure or the correlation coefficient), then we set

$$Y = \{\mathbf{y}'_i = (\sigma(\mathbf{x}_i, \mathbf{x}_1), \ldots, \sigma(\mathbf{x}_i, \mathbf{x}_{|X|}) \,|\, \mathbf{x_1}, \ldots, \mathbf{x_i}, \ldots \mathbf{x}_{|X|} \in X\} \tag{21}$$

   Note that Y is generated according to the two-level approach of similarity analysis of Rieger (2002). The aim of using Y as an alternative to X is to explore second order similarities as

input to classification algorithms. As matrices are input to classification, X and Y need to be projected on matrices first. This is done with the help of the feature set $\mathbb{F}^* = \{F_{l_1}, \ldots, F_{l_k}\} \subseteq \mathbb{F}$ where for F_{l_j}, $j \in \{1, \ldots, k\}$, l_j is the jth smallest index among $l_1, \ldots, l_k$. More specifically, $X[\mathbb{F}'] = (a_{ij})$ is a matrix based on the vector representations $\mathbf{x}_i \in X$ such that $a_{ij} = F_{l_j}(G_i)$. Further, $Y[\mathbb{F}'] = (b_{ij})$ where $b_{ij} = \sigma(\mathbf{x}_i, \mathbf{x}_{l_j})$.

3. *Cluster method & and parameter setting selection* (Row 3 and 4): In this step of Algorithm 1 we select among candidate cluster methods and their corresponding parameters. This includes, for example, to select a measure by which distances are computed among the objects to be classified. The details of this selection are dependent on the operative clustering method. In the present case, we utilize (single, complete, average and weighted) hierarchical agglomerative as well as partitive k-means clustering. As we seek a partitioning $\mathbb{D}$ which is most similar to $\mathbb{L}$ we need to further process the output of hierarchical agglomerations. This is done by means of a lower bound (Rieger, 1989) $\theta = \bar{\eta} + \frac{1}{2}\sigma$ of agglomeration where $\bar{\eta}$ is the mean and σ the standard deviation of the absolute value of the differences of the similarity levels of consecutive agglomeration steps. As a result, we get a threshold for selecting those agglomeration step whose similarity distance to the preceding step is greater than θ. We use the first step exceeding θ. This agglomeration step defines a unique partition of the corresponding dendrogram.

4. *F-Measure value computation* (Row 5): Finally, we validate the F-measure for the classification computed by the selected clustering method in relation to the golden standard $\mathbb{L}$.

Based on this algorithm we now apply QNA on our corpus of wiki networks in order to characterize these networks in quantitative terms (in Section 5) and to classify the networks based on their topological characteristics (in Section 6).

5 Quantitative Features of Wiki Graphs — An Empirical Study

Communication Area	ID	L	URL	Download
▼ Leisure Communication (7)				
▷ Duckipedia Wiki	(1)	de	www.duckipedia.de	Jun 2007
▷ FFXI (Final Fantasy) Wiki	(2)	de	www.ffxi-wiki.de	Jan 2008
▷ FinalFantasy Wiki	(3)	de	wiki.finalfantasy.de	Jun 2007
▷ Geocaching Wiki	(4)	de	www.cachewiki.de	Jan 2008
▷ Grand Theft Auto Wiki	(5)	de	de.gta.wikia.com	Oct 2007
▷ The Elder Scrolls Wiki	(6)	de	www.wikiscrolls.de	Jan 2008
▷ Worldcup Wiki	(7)	de	de.worldcupwiki.org	Oct 2007
▼ Regional Communication (16)				
▷ Berlin Wiki	(8)	de	berlin.wikia.com	Sep 2007
▷ Bielefeld Wiki	(9)	de	www.bielepedia.de	Jan 2008
▷ Dortmund Wiki	(10)	de	www.wikido.de	Sep 2007
▷ Dresden Wiki	(11)	de	dresden.stadtwiki.de	Sep 2007
▷ Hamburg Wiki	(12)	de	www.hamburgwiki.de	Sep 2007
▷ Köln Wiki	(13)	de	www.colonipedia.de	Sep 2007
▷ München Wiki	(14)	de	www.muenchenwiki.de	Sep 2007
▷ Münster Wiki	(15)	de	wiki.muenster.org	Sep 2007
▷ Murgtal Wiki	(16)	de	www.weisenbach.de/wiki	Sep 2007
▷ Nordwest Wiki	(17)	de	nordwest-wiki.nwzonline.de	Sep 2007
▷ Passau Wiki	(18)	de	www.passau-wiki.de	Sep 2007
▷ Pforzheim Wiki	(19)	de	www.pfenz.de/wiki	Sep 2007
▷ Potsdam Wiki	(20)	de	www.potsdam-wiki.de	Sep 2007
▷ Rhein-Neckar Wiki	(21)	de	wiki.rhein-neckar.de	Sep 2007
▷ Wolfenbüttel Wiki	(22)	de	www.wf-guide.de	Sep 2007
▷ Wulfen Wiki	(23)	de	www.wulfen-wiki.de	Sep 2007
▼ Technical Communication (10)				
▷ Debian Wiki	(24)	mult	wiki.debian.org	Jan 2008
▷ Firefox Wiki	(25)	de	www.firefox-browser.de	Jan 2008
▷ Gentoo (Linux) Wiki	(26)	de	de.gentoo-wiki.com	Sep 2007
▷ Gimp Wiki	(27)	de	wiki.gimpforum.de	Jan 2008
▷ Linux Wiki	(28)	de	linuxwiki.de	Feb 2008
▷ Lin(ux) Wiki	(29)	de	de.linwiki.org	Feb 2008
▷ OpenOffice Wiki	(30)	de	www.ooowiki.de	Jan 2008
▷ Thunderbird Wiki	(31)	de	www.thunderbird-mail.de/wiki	Feb 2008
▷ Ubuntu-Forum Wiki	(32)	de	wiki.ubuntu-forum.de	Jan 2008
▷ wxWidgets Wiki	(33)	de	www.wxwidgets.org	Sep 2007
▼ Knowledge Communication (7)				
▷ Ameisen (Ant) Wiki	(34)	de	ameisenwiki.de	Jan 2008
▷ Glottopedia (Linguistics) Wiki	(35)	mult	www.glottopedia.org	Jan 2008
▷ Informationswissenschaft (Information Science) Wiki	(36)	de	server02.is.uni-sb.de/courses/wiki	Jan 2008
▷ Latein (Latin) Wiki	(37)	de/la	www.lateinwiki.org	Jan 2008
▷ Spieltheorie (Game Theory) Wiki	(38)	de	wikiludia.mathematik.uni-muenchen.de	Jan 2008
▷ Statistik (Statistics) Wiki	(39)	de	statwiki.wiwi.hu-berlin.de	Jan 2008
▷ Theater Wiki	(40)	de	wiki.odysseetheater.com	Jan 2008
▼ Scientific Communication (7)				
▷ ALA 2007 Annual Conference Wiki	(41)	en	wikis.ala.org/annual2007	Jan 2008
▷ Ontoworld.org (Semantic Web Community) Wiki	(42)	en	ontoworld.org/wiki	Sep 2007
▷ sem'base (Semantic Technologies Community) Wiki	(43)	de	www.sembase.at	Sep 2007
▷ WICSA (Software Architecture Community) Wiki	(44)	en	wwwp.dnsalias.org/wiki	Jan 2008
▷ Wikimania 2006	(45)	en	wikimania2006.wikimedia.org	Jan 2008
▷ Wikimania 2007	(46)	en	wikimania2007.wikimedia.org	Jan 2008
▷ Wikiversity	(47)	de	de.wikiversity.org	Jan 2008

Table 9: The corpus of wiki networks, their names, language codes, URLs and the corresponding dates of their download.

Leisure Communication

ID	$\|V\|$	$\|E\|$	C_1	C_2	L	D	r	ϵ	lcc	γ	$\bar{R}^2$	γ_S	$\bar{R}^2$	$\gamma_{\bar{k}_{nn}}$	$\bar{R}^2$	$\gamma_{C(k)}$	$\bar{R}^2$	BC	γ_{BC}	$\bar{R}^2$	CC	DC	GC	Cp	Cp_{lcc}	Ch	C_{re}	C_{ra}	L_{re}	L_{ra}
(1)	776	14654	.92	.23	2.19	4	-0.26	18.88	.99	-0.94	.95	-2	.93	-0.33	.72	-0.32	.54	.3	2.42	1	.46	.66	.23	.98	1	.02	.71	.02	20.55	2.27
(2)	43	275	.82	.37	1.94	3	-0.31	6.39	.91	-0.62	.81	-0.56	.73	-0.01	-0.1	-0.12	-0.01	.44	3.54	1	.56	.64	.23	.8	.92	.15	.61	.15	3.36	2.03
(3)	209	1577	.93	.17	2.06	4	-0.26	7.55	.98	-1.07	.85	-1.31	.77	-0.2	.22	-0.35	.41	.72	2.51	1	.66	.85	.22	.95	.98	.04	.63	.04	13.85	2.64
(4)	835	50567	.9	.58	2.05	5	-0.11	60.56	.99	-0.39	.71	-1.58	.86	-0.12	.41	-0.22	.48	.19	3.14	1	.44	.69	.08	.98	1	.07	.74	.07	6.89	1.64
(5)	526	5258	.94	.12	2.07	4	-0.24	10	.99	-0.69	.84	-1.32	.77	-0.28	.39	-0.35	.55	.71	2.33	1	.68	.9	.23	.98	.99	.02	.67	.02	26.31	2.72
(6)	371	5055	.87	.22	2.11	5	-0.25	13.62	.96	-0.9	.96	-1.46	.96	-0.18	.33	-0.38	.61	.49	2.55	1	.52	.76	.1	.93	.99	.04	.69	.04	13.61	2.27
(7)	4949	1180217	.97	.88	1.96	4	.08	238.48	1	-0.31	.77	-3.64	1	-0.15	.65	-0.08	.15	.38	3.23	1	.71	.92	.22	.99	1	.05	.75	.05	10.38	1.55

Regional Communication

ID	$\|V\|$	$\|E\|$	C_1	C_2	L	D	r	ϵ	lcc	γ	$\bar{R}^2$	γ_S	$\bar{R}^2$	$\gamma_{\bar{k}_{nn}}$	$\bar{R}^2$	$\gamma_{C(k)}$	$\bar{R}^2$	BC	γ_{BC}	$\bar{R}^2$	CC	DC	GC	Cp	Cp_{lcc}	Ch	C_{re}	C_{ra}	L_{re}	L_{ra}
(8)	276	3760	.86	.5	2.37	7	-0.15	13.62	.9	-0.55	.85	-2.55	1	-0.09	.25	-0.25	.41	.34	3.02	.99	.31	.4	.05	.8	.98	.05	.69	.05	10.13	2.15
(9)	82	466	.85	.24	2.01	3	-0.33	5.68	.93	-0.77	.88	0	0	-0.14	.07	-0.44	.24	.76	2.85	1	.7	.77	.2	.85	.96	.07	.59	.07	7.22	2.54
(10)	219	2493	.83	.33	2.15	5	-0.16	11.38	.95	-0.59	.8	-2.32	.96	-0.1	.11	-0.3	.48	.65	2.9	1	.54	.72	.05	.91	.98	.05	.68	.05	9.62	2.22
(11)	217	2075	.92	.3	2.16	4	-0.31	9.56	.97	-0.67	.87	-1.81	.9	-0.34	.61	-0.38	.5	.56	2.71	1	.55	.72	.24	.93	.98	.04	.66	.04	11.35	2.38
(12)	605	9033	.94	.28	2.17	4	-0.28	14.93	.99	-0.62	.89	0	0	-0.38	.74	-0.28	.41	.45	2.53	1	.52	.75	.23	.99	.99	.03	.7	.03	20.26	2.37
(13)	134	1032	.87	.38	2.38	4	-0.25	7.7	.9	-0.68	.87	-0.96	.96	-0.06	-0.01	-0.4	.4	.36	2.28	.98	.4	.44	.27	.81	.96	.06	.64	.06	8.7	2.4
(14)	2424	57862	.91	.17	2.26	5	-0.16	23.87	.98	-2.44	.89	-2.11	1	-0.27	.59	-0.3	.54	.32	2.26	1	.38	.6	.1	.96	1	.01	.72	.01	50.77	2.46
(15)	689	8297	.9	.32	2.61	6	-0.16	12.04	.91	-0.64	.87	-1.64	.99	-0.21	.55	-0.34	.55	.31	2.32	.99	.3	.36	.14	.83	.99	.02	.68	.02	28.61	2.63
(16)	82	832	.89	.4	1.96	3	-0.29	10.15	1	-0.81	.92	0	0	-0.27	.81	-0.34	.69	.59	3.06	1	.67	.85	.2	.99	.96	.12	.67	.12	4.04	1.9
(17)	819	29269	.92	.71	2.03	4	-0.05	35.74	.96	-0.44	.86	-2.46	.99	-0.32	.82	-0.23	.47	.74	2.88	1	.69	.84	.23	.92	1	.04	.73	.04	11.46	1.88
(18)	200	1922	.93	.17	1.99	3	-0.35	9.61	.99	-0.94	.83	-0.22	.79	-0.2	.27	-0.36	.5	.56	3.12	1	.73	.9	.18	.96	.99	.05	.66	.05	10.41	2.34
(19)	299	4275	.89	.22	2.11	4	-0.42	14.3	.98	-1.28	.86	0	0	-0.16	.29	-0.37	.6	.23	2.14	1	.45	.64	.22	.96	.99	.05	.69	.05	10.46	2.14
(20)	47	351	.83	.47	2.07	5	-0.26	7.47	.96	-0.62	.85	0	0	-0.17	.26	-0.35	.27	.4	2.98	.99	.43	.59	.05	.89	.93	.16	.63	.16	3.15	1.91
(21)	426	7282	.9	.22	2.13	4	-0.33	17.09	1	-0.76	.91	0	0	-0.28	.75	-0.32	.71	.23	2.5	1	.46	.68	.22	1	.99	.04	.7	.04	12.46	2.13
(22)	194	1636	.89	.17	2.09	4	-0.33	8.43	.97	-1.14	.97	0	0	-0.18	.24	-0.41	.54	.58	2.55	1	.61	.8	.23	.93	.98	.04	.65	.04	11.5	2.47
(23)	161	1645	.91	.28	2.02	3	-0.29	10.22	.97	-0.62	.82	-1.3	.77	-0.35	.58	-0.35	.5	.55	2.97	1	.66	.79	.18	.93	.98	.06	.67	.06	7.88	2.19

Technical Communication

ID	$\|V\|$	$\|E\|$	C_1	C_2	L	D	r	ϵ	lcc	γ	$\bar{R}^2$	γ_S	$\bar{R}^2$	$\gamma_{\bar{k}_{nn}}$	$\bar{R}^2$	$\gamma_{C(k)}$	$\bar{R}^2$	BC	γ_{BC}	$\bar{R}^2$	CC	DC	GC	Cp	Cp_{lcc}	Ch	C_{re}	C_{ra}	L_{re}	L_{ra}
(24)	398	13834	.9	.32	1.94	3	-0.34	34.76	.99	-0.81	.92	0	0	-0.16	.38	-0.28	.56	.16	2.87	1	.68	.85	.22	.98	.99	.09	.73	.09	5.72	1.69
(25)	2463	90877	.79	.3	2.56	6	-0.09	36.9	.95	-0.53	.81	-4.19	1	-0.08	.52	-0.26	.53	.13	2.29	1	.27	.32	.13	.9	1	.01	.73	.01	33.38	2.16
(26)	6877	255363	.94	.07	2	4	-0.13	37.13	1	-1.16	1	0	0	-0.57	.99	-0.34	.7	.54	3.43	1	.75	.99	.22	1	1	0	.73	0	92.6	2.44
(27)	947	12997	.92	.25	2.43	5	-0.2	13.72	.96	-1.29	.85	-2.05	1	-0.33	.78	-0.24	.38	.29	1.64	.99	.35	.48	.11	.93	1	.01	.69	.01	34.5	2.62
(28)	2893	109687	.95	.35	2.16	5	-0.08	37.92	.99	-0.51	.77	-1.42	.98	-0.3	.65	-0.19	.4	.64	2.43	1	.57	.86	.11	.98	1	.01	.73	.01	38.15	2.19
(29)	41	577	.8	.69	1.68	3	-0.14	14.07	.93	-0.44	.73	0	0	-0.01	.59	-0.21	.48	.19	6.4	.8	.48	.44	.07	.84	.94	.35	.69	.34	1.46	1.4
(30)	346	4236	.94	.26	2.09	4	-0.25	12.24	.99	-0.61	.85	0	0	-0.48	.47	-0.29	.5	.63	2.37	1	.61	.84	.24	.98	.99	.04	.68	.04	14.13	2.33
(31)	19	127	.89	.62	1.74	2	-0.33	6.68	1	-0.86	.93	0	0	-0.17	.5	-0.31	.57	.52	4.85	1	.75	.76	1	.96	.87	.37	.62	.35	1.42	1.55
(32)	71	899	.86	.47	1.87	3	-0.34	12.66	.99	-0.52	.8	0	0	-0.09	.07	-0.3	.41	.35	3.67	1	.68	.76	.12	.96	.96	.18	.69	.18	2.8	1.68
(33)	620	9384	.91	.25	2.1	5	-0.16	15.14	.97	-0.59	.81	-1.89	.99	-0.25	.43	-0.3	.53	.61	2.13	1	.57	.79	.04	.94	1	.02	.7	.02	20.48	2.37

Knowledge Communication

ID	$\|V\|$	$\|E\|$	C_1	C_2	L	D	r	ϵ	lcc	γ	$\bar{R}^2$	γ_S	$\bar{R}^2$	$\gamma_{\bar{k}_{nn}}$	$\bar{R}^2$	$\gamma_{C(k)}$	$\bar{R}^2$	BC	γ_{BC}	$\bar{R}^2$	CC	DC	GC	Cp	Cp_{lcc}	Ch	C_{re}	C_{ra}	L_{re}	L_{ra}
(34)	757	25559	.94	.63	2.04	4	-0.15	33.76	.98	-0.45	.87	-1.67	.98	-0.4	.54	-0.17	.33	.45	2.74	1	.63	.83	.23	.95	1	.05	.73	.05	11.21	1.88
(35)	60	412	.92	.46	2.05	3	-0.34	6.87	.98	-0.7	.9	0	0	-0.05	-0.03	-0.08	-0.01	.67	4.39	1	.59	.74	.17	.95	.94	.12	.62	.11	4.37	2.12
(36)	111	901	.88	.24	1.97	3	-0.4	8.12	.95	-1.36	.87	0	0	-0.16	.18	-0.32	.33	.44	3.1	1	.71	.76	.06	.9	.97	.07	.64	.07	6.84	2.25
(37)	492	4114	.77	.37	2.76	6	-0.09	8.36	.81	-0.7	.9	-2.58	1	-0.1	.6	-0.35	.49	.51	2.01	1	.34	.29	.13	.66	.99	.02	.65	.02	29.42	2.92
(38)	420	16256	.88	.64	1.99	4	-0.03	38.71	.94	-0.42	.73	-0.38	.79	-0.07	.12	-0.21	.43	.43	3.48	1	.65	.74	.23	.88	.99	.09	.73	.09	5.43	1.65
(39)	20	224	.86	.77	1.56	3	-0.13	11.2	1	-0.53	.65	0	0	.15	.29	-0.16	.44	.13	8.9	.72	.43	.46	.12	.97	.91	.59	.68	.56	.89	1.24
(40)	60	430	.69	.87	1.74	2	.22	7.17	.58	-0.6	.94	-2.64	.98	.17	0	.03	-0.19	.75	0	0	.79	.25	1	.34	.93	.12	.63	.12	4.19	2.08

Scientific Communication

ID	$\|V\|$	$\|E\|$	C_1	C_2	L	D	r	ϵ	lcc	γ	$\bar{R}^2$	γ_S	$\bar{R}^2$	$\gamma_{\bar{k}_{nn}}$	$\bar{R}^2$	$\gamma_{C(k)}$	$\bar{R}^2$	BC	γ_{BC}	$\bar{R}^2$	CC	DC	GC	Cp	Cp_{lcc}	Ch	C_{re}	C_{ra}	L_{re}	L_{ra}
(41)	422	16154	.92	.76	2.24	5	.05	38.28	.98	-0.45	.67	-0.32	.87	.08	.37	-0.2	.44	.4	2.91	.99	.42	.6	.09	.96	.99	.09	.73	.09	5.51	1.66
(42)	1449	17319	.81	.19	2.53	5	-0.21	11.95	.82	-0.68	.9	-2.45	1	-0.35	.68	-0.4	.61	.28	1.91	1	.33	.32	.11	.68	1	.01	.68	.01	60.62	2.93
(43)	102	752	.85	.37	2.22	4	-0.22	7.37	.83	-0.7	.85	-0.99	.95	-0.18	.16	-0.39	.42	.43	2.4	.98	.42	.42	.05	.69	.96	.07	.63	.07	6.92	2.32
(44)	134	2186	.84	.52	2.13	4	-0.1	16.31	.97	-0.5	.76	-0.81	.61	-0.02	.26	-0.29	.45	.28	3.81	.99	.42	.53	.23	.93	.97	.12	.7	.12	4.11	1.75
(45)	762	45228	.88	.61	2.16	5	-0.11	59.35	.96	-0.4	.76	-3.12	1	-0.01	.11	-0.2	.46	.14	2.86	.98	.45	.61	.06	.93	1	.08	.74	.08	6.42	1.62
(46)	1075	114277	.93	.85	2	5	.08	106.3	.98	-0.39	.82	-3.39	.99	.03	.02	-0.11	.23	.23	4.82	1	.5	.7	.04	.96	1	.1	.74	.1	5.06	1.5
(47)	886	26914	.91	.39	2.17	4	-0.19	30.38	.99	-0.5	.79	0	0	-0.23	.74	-0.28	.55	.2	2.34	1	.35	.56	.27	.98	1	.03	.72	.03	14.58	1.99

Table 10: Composite features of 47 agent networks of five different communication areas.

Leisure Communication

ID	$\|V\|$	$\|E\|$	C_1	C_2	L	D	r	ϵ	lcc	γ	$\bar{R}^2$	γ_S	$\bar{R}^2$	$\gamma_{\bar{k}_{nn}}$	$\bar{R}^2$	$\gamma_{C(k)}$	$\bar{R}^2$	BC	γ_{BC}	$\bar{R}^2$	CC	DC	GC	Cp	$\mathrm{Cp}_{\mathrm{lcc}}$	Ch	C_{re}	C_{ra}	L_{re}	L_{ra}
(1)	2121	58894	.53	.03	1.99	2	-0.22	27.77	1	-1.58	.97	0	0	-0.84	.99	-0.75	.95	.27	2.67	1	.99	.99	1	1	1	.01	.72	.01	38.19	2.3
(2)	4066	166394	.87	.02	1.99	2	-0.54	40.92	1	-2	1	0	0	-0.96	1	-0.69	.97	.07	6.09	1	1	.99	1	1	1	.01	.73	.01	49.68	2.24
(3)	772	28324	.91	.08	1.95	2	-0.68	36.69	1	-3.3	.99	0	0	-0.94	1	-0.57	.95	.06	8.3	1	.98	.95	.98	1	1	.05	.73	.05	10.52	1.85
(4)	324	2762	.8	.06	1.97	2	-0.35	8.52	1	-1.48	.97	0	0	-0.88	1	-0.55	.92	.47	2.32	1	.99	.98	1	1	.99	.03	.65	.03	19	2.7
(5)	4944	85488	.6	.01	2.02	4	-0.15	17.29	1	-1.77	.98	0	0	-0.9	.82	-0.7	.72	.47	2.21	1	.74	.98	.22	1	1	0	.7	0	142.96	2.98
(6)	892	14854	.51	.1	1.98	2	-0.11	16.65	1	-0.96	.98	0	0	-0.8	.99	-0.5	.69	.85	2.48	1	.99	.98	1	1	1	.02	.7	.02	26.78	2.41
(7)	3870	209466	.62	.05	1.99	2	-0.2	54.13	1	-0.89	.94	0	0	-0.83	1	-0.59	.97	.12	2.23	1	.99	.99	1	1	1	.01	.74	.01	35.75	2.07

Regional Communication

ID	$\|V\|$	$\|E\|$	C_1	C_2	L	D	r	ϵ	lcc	γ	$\bar{R}^2$	γ_S	$\bar{R}^2$	$\gamma_{\bar{k}_{nn}}$	$\bar{R}^2$	$\gamma_{C(k)}$	$\bar{R}^2$	BC	γ_{BC}	$\bar{R}^2$	CC	DC	GC	Cp	$\mathrm{Cp}_{\mathrm{lcc}}$	Ch	C_{re}	C_{ra}	L_{re}	L_{ra}
(8)	492	5132	.71	.05	1.98	2	-0.3	10.43	1	-1.75	.98	0	0	-0.86	1	-0.64	.94	.46	2.43	1	.99	.98	1	1	.99	.02	.67	.02	23.58	2.64
(9)	3347	38338	.67	0	2	2	-0.45	11.45	1	-3.65	.98	0	0	-1.01	.99	-0.83	.98	.37	3.13	1	1	1	1	1	1	0	.68	0	146.1	3.33
(10)	369	2346	.66	.04	1.98	2	-0.29	6.36	1	-2.52	.95	0	0	-0.81	1	-0.31	.3	.86	2.79	1	.99	.99	1	1	.99	.02	.61	.02	29.02	3.2
(11)	2200	27260	.41	.02	1.99	2	-0.1	12.39	1	-1.38	.99	0	0	-0.91	1	-0.75	.98	.95	2.32	1	1	1	1	1	1	.01	.68	.01	88.78	3.06
(12)	10561	177836	.7	.02	2	2	-0.16	16.84	1	-1.13	.99	0	0	-0.93	1	-0.27	.23	.48	2.26	1	1	1	1	1	1	0	.7	0	313.59	3.28
(13)	516	4824	.68	.04	1.98	2	-0.29	9.35	1	-2.28	.94	0	0	-0.92	1	-0.69	.94	.48	2.23	1	.99	.99	1	1	.99	.02	.66	.02	27.6	2.79
(14)	5240	74626	.69	.01	2	2	-0.29	14.24	1	-2.07	.99	0	0	-0.97	1	-0.87	.94	.32	2.18	1	1	1	1	1	1	0	.69	0	183.97	3.22
(15)	1772	30474	.88	.03	1.99	2	-0.49	17.2	1	-1.98	.99	0	0	-0.97	1	-0.57	.86	.19	2.3	1	1	.99	1	1	1	.01	.7	.01	51.52	2.63
(16)	1554	20802	.66	.03	1.99	2	-0.25	13.39	1	-1.61	.98	0	0	-0.88	1	-0.62	.93	.45	2.43	1	1	.99	1	1	1	.01	.69	.01	58.04	2.83
(17)	538	14622	.85	.09	1.95	2	-0.54	27.18	1	-2	.99	0	0	-0.85	1	-0.55	.94	.1	2.27	1	.98	.95	.99	1	1	.05	.72	.05	9.9	1.9
(18)	3531	81556	.76	.03	1.99	2	-0.24	23.1	1	-1.37	1	0	0	-0.91	1	-0.5	.8	.24	2.2	1	1	.99	1	1	1	.01	.72	.01	76.44	2.6
(19)	3069	64706	.63	.05	1.99	2	-0.14	21.08	1	-1.07	.99	0	0	-0.85	1	-0.38	.64	.45	2.38	1	1	.99	1	1	1	.01	.71	.01	72.78	2.63
(20)	653	8168	.83	.04	1.98	2	-0.48	12.51	1	-2.45	.98	0	0	-0.95	1	-0.75	.96	.24	4.34	1	.99	.98	1	1	1	.02	.69	.02	26.1	2.57
(21)	2591	77404	.79	.03	1.99	2	-0.35	29.87	1	-1.77	.99	0	0	-0.9	1	-0.43	.79	.14	2.2	1	1	.99	1	1	1	.01	.72	.01	43.37	2.31
(22)	920	9902	.64	.03	1.99	2	-0.25	10.76	1	-2.47	.94	0	0	-0.93	1	-0.74	.95	.48	2.22	1	1	.99	1	1	1	.01	.67	.01	42.74	2.87
(23)	241	1704	.81	.05	1.97	2	-0.43	7.07	1	-1.27	.94	0	0	-0.89	1	-0.72	.95	.47	2.32	1	.99	.98	1	1	.99	.03	.63	.03	17.04	2.8

Technical Communication

ID	$\|V\|$	$\|E\|$	C_1	C_2	L	D	r	ϵ	lcc	γ	$\bar{R}^2$	γ_S	$\bar{R}^2$	$\gamma_{\bar{k}_{nn}}$	$\bar{R}^2$	$\gamma_{C(k)}$	$\bar{R}^2$	BC	γ_{BC}	$\bar{R}^2$	CC	DC	GC	Cp	$\mathrm{Cp}_{\mathrm{lcc}}$	Ch	C_{re}	C_{ra}	L_{re}	L_{ra}
(24)	416	9254	.56	.15	1.95	2	-0.2	22.25	1	-0.65	.82	0	0	-0.59	.98	-0.56	.92	.35	2.5	1	.97	.95	1	1	.99	.05	.71	.05	9.35	1.94
(25)	5175	49072	.74	0	2	3	-0.26	9.48	1	-1.77	.99	0	0	-0.94	.99	-0.81	.56	.58	3.65	1	.75	1	.24	1	1	0	.66	0	272.87	3.8
(26)	3169	48812	.82	.01	2	3	-0.44	15.4	1	-1.95	.99	0	0	-0.8	1	-0.52	.89	.24	3.03	1	.99	1	.01	1	1	0	.7	0	102.87	2.95
(27)	1175	15058	.58	.12	1.99	2	-0.14	12.81	1	-0.66	.89	0	0	-0.83	.96	-0.13	.05	.95	3.09	1	1	.99	1	1	1	.01	.69	.01	45.84	2.77
(28)	696	17952	.94	.06	1.96	2	-0.74	25.79	1	-3.52	.99	0	0	-0.88	1	-0.62	.95	.09	2.3	1	.98	.97	.99	1	1	.04	.72	.04	13.49	2.01
(29)	98	554	.72	.12	1.94	2	-0.38	5.65	1	-0.82	.87	0	0	-0.71	.99	-0.25	.17	.81	2.94	1	.98	.96	1	.99	.97	.06	.59	.06	8.67	2.65
(30)	639	11542	.59	.09	1.97	3	-0.2	18.06	1	-1.73	.81	0	0	-0.74	.99	-0.54	.93	.42	2.43	1	.75	.97	.23	1	1	.03	.71	.03	17.69	2.23
(31)	218	3572	.89	.12	1.92	2	-0.61	16.39	1	-2.15	.99	0	0	-0.84	1	-0.51	.91	.15	2.66	1	.96	.93	.98	1	.99	.08	.7	.08	6.65	1.93
(32)	313	3684	.65	.09	1.96	2	-0.29	11.77	1	-1.23	.98	0	0	-0.72	.99	-0.62	.94	.43	2.56	1	.98	.97	1	1	.99	.04	.68	.04	13.3	2.33
(33)	545	9460	.97	.05	1.97	2	-0.84	17.36	1	-6.13	.99	0	0	-0.97	1	-0.57	.96	.12	7.63	1	.99	.97	.99	1	1	.03	.7	.03	15.7	2.21

Knowledge Communication

ID	$\|V\|$	$\|E\|$	C_1	C_2	L	D	r	ϵ	lcc	γ	$\bar{R}^2$	γ_S	$\bar{R}^2$	$\gamma_{\bar{k}_{nn}}$	$\bar{R}^2$	$\gamma_{C(k)}$	$\bar{R}^2$	BC	γ_{BC}	$\bar{R}^2$	CC	DC	GC	Cp	$\mathrm{Cp}_{\mathrm{lcc}}$	Ch	C_{re}	C_{ra}	L_{re}	L_{ra}
(34)	636	5748	.81	.05	1.99	2	-0.33	9.04	1	-1.31	.98	0	0	-0.89	1	-0.72	.91	.48	2.32	1	.99	.99	1	1	1	.01	.66	.01	35.19	2.93
(35)	1634	16584	.59	.02	1.99	2	-0.26	10.15	1	-2.14	.97	0	0	-0.76	.99	-0.51	.66	.74	2.97	1	1	1	1	1	1	.01	.67	.01	80.5	3.19
(36)	536	3740	.65	.04	1.99	2	-0.19	6.98	1	-1.81	.92	0	0	-0.9	1	-0.32	.34	.95	2.36	1	1	.99	1	1	.99	.01	.62	.01	38.41	3.23
(37)	2771	37772	.85	.02	2	2	-0.34	13.63	1	-1.49	.99	0	0	-0.77	.99	-0.38	.67	.4	2.52	1	1	1	1	1	1	0	.69	0	101.64	3.03
(38)	338	2778	.64	.09	1.98	2	-0.2	8.22	1	-0.73	.87	0	0	-0.81	.99	-0.54	.42	.9	2.52	1	.99	.98	1	1	.99	.02	.65	.02	20.56	2.76
(39)	583	5282	.86	.03	1.98	2	-0.49	9.06	1	-2.04	.97	0	0	-0.79	.98	-0.54	.9	.34	2.94	1	.99	.99	1	1	1	.02	.66	.02	32.17	2.89
(40)	784	9672	.66	.04	1.98	2	-0.22	12.34	1	-2.23	.92	0	0	-0.89	1	-0.68	.94	.47	2.29	1	.99	.99	1	1	1	.02	.68	.02	31.77	2.65

Scientific Communication

ID	$\|V\|$	$\|E\|$	C_1	C_2	L	D	r	ϵ	lcc	γ	$\bar{R}^2$	γ_S	$\bar{R}^2$	$\gamma_{\bar{k}_{nn}}$	$\bar{R}^2$	$\gamma_{C(k)}$	$\bar{R}^2$	BC	γ_{BC}	$\bar{R}^2$	CC	DC	GC	Cp	$\mathrm{Cp}_{\mathrm{lcc}}$	Ch	C_{re}	C_{ra}	L_{re}	L_{ra}
(41)	108	314	.43	.03	1.97	2	-0.54	2.91	1	-1.37	.97	0	0	-0.93	1	-0.18	-0.03	.98	2.46	1	1	.99	1	.99	.97	.03	.36	-1	18.57	4.39
(42)	4705	82504	.77	.04	2.01	4	-0.25	17.53	1	-1.09	.97	0	0	-0.93	.6	-0.14	.08	.32	2.23	1	.74	.99	.22	1	1	0	.7	0	134.16	2.95
(43)	358	8740	.8	.14	1.93	2	-0.46	24.41	1	-1.55	.99	0	0	-0.79	.99	-0.52	.86	.12	2.48	1	.97	.94	.98	1	.99	.07	.72	.07	7.33	1.84
(44)	127	668	.97	.05	1.96	2	-0.76	5.26	1	-2	.92	0	0	-0.81	1	-0.51	.81	.45	2.45	1	.99	.97	.99	.99	.98	.04	.57	.04	12.07	2.92
(45)	559	16250	.96	.08	1.95	2	-0.81	29.07	1	-4.88	.93	0	0	-0.92	1	-0.25	.92	.07	13.1	1	.98	.95	.98	1	1	.05	.72	.05	9.61	1.88
(46)	966	25698	.92	.05	1.97	2	-0.62	26.6	1	-2.68	1	0	0	-0.94	1	-0.67	.94	.09	6.68	1	.99	.97	.99	1	1	.03	.72	.03	18.16	2.1
(47)	1159	7428	.72	.02	1.99	2	-0.21	6.41	1	-1.49	.97	0	0	-0.94	1	-0.54	.58	.98	2.43	1	1	1	1	1	1	.01	.61	-1	90.42	3.8

Table 11: Composite features of 47 document graphs of five different communication areas.

Regional Communication

ID	$\|V\|$	$\|E\|$	C_1	C_2	L	D	r	ϵ	lcc	γ	$\bar{R}^2$	γ_S	$\bar{R}^2$	$\gamma_{\bar{k}_{nn}}$	$\bar{R}^2$	$\gamma_{C(k)}$	$\bar{R}^2$	BC	γ_{BC}	$\bar{R}^2$	CC	DC	GC	Cp	$\mathrm{Cp}_{\mathrm{lcc}}$	Ch	C_{re}	C_{ra}	L_{re}	L_{ra}
(8)	1241	305212	.6	.58	1.88	4	.13	245.94	.99	-0.3	.61	0	0	.17	.54	-0	.13	.02	1.61	.84	.32	.46	.02	.97	1	.2	.75	.2	2.52	1.29
(9)	1295	123556	.58	.79	2.28	5	.38	95.41	.9	-0.35	.72	0	0	.35	.63	.12	.13	.06	2.71	.87	.25	.27	.1	.81	1	.07	.74	.07	6.79	1.57
(10)	323	12484	.46	.48	2.1	4	.21	38.65	.99	-0.4	.66	0	0	.2	.54	.06	.26	.02	1.41	.84	.2	.27	.02	.98	.99	.12	.73	.12	4.18	1.58
(11)	2816	533502	.36	.53	2.05	4	.49	189.45	1	-0.38	.64	0	0	.4	.69	.34	.35	0	.8	.72	.13	.21	.01	.99	1	.07	.75	.07	7.43	1.51
(12)	4299	2215594	.6	.35	1.88	2	-0.21	515.37	1	-0.67	.85	0	0	-0.21	.93	-0.3	.83	.02	2.92	1	.94	.88	0	1	1	.12	.75	.12	4.17	1.34
(13)	636	47212	.55	.59	2.24	5	.1	74.23	.99	-0.37	.76	0	0	.16	.44	.06	.05	.05	1.2	.78	.27	.4	.09	.98	.99	.12	.74	.12	4.28	1.5
(14)	4669	3302716	.55	.46	1.85	4	-0.06	707.37	1	-0.34	.63	0	0	.03	.07	-0.11	.31	.01	2.94	.99	.43	.64	.21	1	1	.15	.75	.15	3.3	1.29
(15)	2323	1391294	.65	.52	1.74	3	-0.12	598.92	1	-0.55	.67	0	0	-0.11	.78	-0.27	.68	.01	3.09	1	.82	.73	0	1	1	.26	.75	.26	1.94	1.21
(16)	1623	259178	.43	.43	2.01	4	.27	159.69	.99	-0.35	.61	0	0	.25	.77	-0.01	.28	.01	1.09	.85	.18	.29	.03	.98	1	.1	.75	.1	5.08	1.46
(17)	2914	2938568	.74	.6	1.65	3	-0.15	1008.43	1	-0.52	.62	0	0	-0.14	.67	-0.27	.66	0	3.52	.99	.78	.65	0	1	1	.35	.75	.35	1.45	1.15
(18)	3061	756838	.39	.34	1.97	5	.16	247.25	1	-0.34	.51	0	0	.15	.81	-0.17	.57	.01	1.96	.98	.09	.19	.08	1	1	.08	.75	.08	6.19	1.46
(19)	2492	570084	.38	.36	1.94	4	.06	228.77	1	-0.33	.51	0	0	.06	.54	-0.1	.25	.01	2.52	.98	.12	.23	0	1	1	.09	.75	.09	5.45	1.44
(20)	1224	156396	.3	.29	1.92	3	.1	127.78	1	-0.33	.46	0	0	.1	.84	-0.08	.7	.01	1.87	.99	.14	.26	.24	1	1	.1	.74	.1	4.79	1.47
(21)	2921	1672682	.63	.44	1.8	2	-0.2	572.64	1	-0.58	.76	0	0	-0.2	.95	-0.27	.76	.01	2.93	1	.89	.8	1	1	1	.2	.75	.2	2.55	1.26
(22)	1127	144658	.41	.4	1.95	4	.19	128.36	1	-0.41	.63	0	0	.2	.84	-0.03	.25	.01	1.45	.94	.22	.37	.22	1	1	.11	.74	.11	4.39	1.45
(23)	1042	233940	.65	.6	1.84	4	-0.06	224.51	1	-0.31	.63	0	0	.05	.1	-0.04	.02	.02	2.97	.88	.38	.55	.23	.99	1	.22	.75	.22	2.32	1.28

Technical Communication

ID	$\|V\|$	$\|E\|$	C_1	C_2	L	D	r	ϵ	lcc	γ	$\bar{R}^2$	γ_S	$\bar{R}^2$	$\gamma_{\bar{k}_{nn}}$	$\bar{R}^2$	$\gamma_{C(k)}$	$\bar{R}^2$	BC	γ_{BC}	$\bar{R}^2$	CC	DC	GC	Cp	$\mathrm{Cp}_{\mathrm{lcc}}$	Ch	C_{re}	C_{ra}	L_{re}	L_{ra}
(24)	2052	1768740	.79	.7	1.58	2	-0.18	861.96	1	-0.67	.72	0	0	-0.09	.55	-0.16	.49	0	4.59	.98	.71	.58	1	1	1	.42	.75	.42	1.19	1.13
(25)	4280	11777860	.81	.77	1.36	2	-0.17	2751.84	1	-0.77	.59	0	0	-0.13	.75	-0.21	.68	0	0	0	.5	.36	1	1	1	.64	.75	.64	.78	1.06
(26)	2689	4797732	.8	.77	1.34	2	-0.13	1784.21	1	-0.68	.51	0	0	-0.1	.8	-0.2	.74	0	0	0	.49	.34	1	1	1	.66	.75	.66	.75	1.05
(27)	1474	1793254	.89	.89	1.17	2	-0.09	1216.59	1	-0.53	.38	0	0	-0.03	.14	-0.05	.29	0	1.54	.6	.28	.17	1	1	1	.83	-1	.82	.61	1.03
(28)	3687	5916120	.73	.63	1.57	2	-0.19	1604.59	1	-0.99	.74	0	0	-0.15	.84	-0.26	.79	0	0	0	.7	.56	0	1	1	.44	.75	.44	1.15	1.11
(29)	376	85028	.79	.76	1.4	2	-0.06	226.14	1	-0.77	.65	0	0	-0.03	.1	-0.15	.5	0	4.81	.93	.57	.4	1	1	1	.6	.75	.6	.83	1.09
(30)	2951	3201526	.71	.6	1.63	2	-0.18	1084.9	1	-0.96	.85	0	0	-0.14	.66	-0.25	.58	0	3.26	.99	.78	.63	1	1	1	.37	.75	.37	1.36	1.14
(31)	855	180636	.58	.5	1.76	3	-0.09	211.27	1	-0.44	.63	0	0	-0.05	.62	-0.14	.4	.03	2.93	.99	.65	.71	.18	1	1	.25	.75	.25	2.02	1.26
(32)	1106	447228	.59	.56	1.63	2	-0.02	404.36	1	-0.94	.77	0	0	0	.41	-0.14	.48	0	2.89	.96	.64	.56	0	1	1	.37	.75	.37	1.37	1.17
(33)	2179	1408824	.55	.52	1.7	3	-0.04	646.55	1	-0.63	.66	0	0	.03	.32	-0.08	.33	0	2.46	.99	.7	.66	.05	1	1	.3	.75	.3	1.68	1.19

Scientific Communication

ID	$\|V\|$	$\|E\|$	C_1	C_2	L	D	r	ϵ	lcc	γ	$\bar{R}^2$	γ_S	$\bar{R}^2$	$\gamma_{\bar{k}_{nn}}$	$\bar{R}^2$	$\gamma_{C(k)}$	$\bar{R}^2$	BC	γ_{BC}	$\bar{R}^2$	CC	DC	GC	Cp	$\mathrm{Cp}_{\mathrm{lcc}}$	Ch	C_{re}	C_{ra}	L_{re}	L_{ra}
(41)	1202	411940	.64	.54	1.72	3	-0.17	342.71	1	-0.33	.56	0	0	-0.09	.49	-0.17	.55	.01	3.38	.99	.68	.67	.08	1	1	.28	.75	.28	1.75	1.22
(42)	6076	8984148	.68	.51	1.76	3	-0.2	1478.63	1	-0.35	.59	0	0	-0.15	.64	-0.17	.64	0	0	0	.83	.75	.01	1	1	.24	.75	.24	2.06	1.19
(43)	746	101998	.5	.46	1.89	5	.08	136.73	1	-0.29	.47	0	0	.06	.67	-0.19	.69	.03	2.27	.96	.18	.34	.09	1	1	.18	.74	.18	2.73	1.35
(44)	1147	1032100	.86	.85	1.22	2	-0.1	899.83	1	-0.65	.44	0	0	-0.07	.56	-0.16	.55	0	1.68	.71	.34	.22	1	1	1	.79	-1	.79	.64	1.04
(45)	2235	960150	.66	.64	1.95	6	-0.08	429.6	.99	-0.3	.64	0	0	.08	.18	-0	-0	.02	5.03	.9	.35	.53	.1	.98	1	.19	.75	.19	2.6	1.27
(46)	2966	2379012	.61	.66	1.74	4	.29	802.09	1	-0.3	.56	0	0	.28	.74	.07	.07	.01	3.49	.72	.4	.52	.22	1	1	.27	.75	.27	1.85	1.2
(47)	2759	807114	.5	.51	1.98	4	.06	292.54	1	-0.33	.61	0	0	.14	.47	.02	0	.01	3.19	.92	.24	.4	.01	1	1	.11	.75	.11	4.72	1.39

Table 12: Composite features of 33 word graphs of five different communication areas.

In order to give a quantitative account of our wiki corpus we now perform a network analysis of this data. There are two reasons behind this procedure:

- Firstly, we want to shed light on the topological characteristics of these networks. This is done by focusing on the features F_{1-24} of Table 8. The aim is to show that — according to the *Network Correlation Hypothesis* (NCH) of Section 1 — ontologically different networks correlate with respect to their small world-property — at least on the agent and document level.
- Secondly, we want to show that agent and document networks differ. That is, despite of being alike in terms of their small world-status, these networks diverge topologically. This argument is in the line of the *Network Separability Hypothesis* (NSH) of Section 1. It is studied in detail in Section 6.

Section 5.1 starts with describing the network corpus itself as well as its text technological representation and processing. Next, Section 5.2 discusses some of the networks' topological characteristics.

5.1 The Wiki Corpus

Table 9 lists the 47 wiki networks of the 5 communication areas analyzed in this paper. Their numerical features are reported in the Tables 10–12. Note that each wiki network induces a separate agent, document and word network so that we actually face 141 different component graphs according to Definition 3.10. The distribution of network instances per communication area is obviously skewed — it ranges from the 16 special wikis of regional communication to the 7 special wikis of scientific communication. Likewise, the sizes of the agent and document graphs vary tremendously in all of the 5 wiki subcorpora of the 5 communication areas.

Some remarks on wiki graph extraction: According to Definition 3.10 we refer to page-centered wiki graphs and set $\mathrm{T}_{\text{wiki}} = \{\texttt{Article}, \texttt{Category}, \texttt{Portal}\}$. Thus, we extract graphs induced by content-related pages only. On the one hand, this is justified by the fact that special wikis use quite different namespaces so that reducing T_{wiki} to this set retains their comparability. On the other hand, Mehler (2006) shows that the small world-property can already be checked by this subset of wiki pages so that it is not necessary to additionally consider collaboration management or network management types.

Collaborations and intertextual relations are extracted as described by the example of Definition 3.10. That is, two agents are seen to be linked if there is at least one page co-authored by them while hyperlinks are explored to induce intertextual relations. Note that we mainly analyze undirected component graphs — in the case of word associations and collaborations this is evident. In the case of document graphs we interpret the backlink of a link as a multiple edge so that we can abstract from the orientation of these graphs. Further, we do not generate weighted wiki graphs so that all links are seen to be of equal weight.[28]

Word graphs are computed according to the association measure of Mehler (2008b). That is, for the vertex set of a word graph V_3 and the sequence $S_n = (x_1, \ldots, x_n)$ of n texts, the association of two

[28] This is a starting point for extending the present approach in future work.

lexical items $v_i, v_j \in V_3$ is computed as

$$\alpha(v_i, v_j, S_n) = \sum_{k=1}^{n} \left(f_{ik} \frac{k}{F_{ik}} \right) \left(f_{jk} \frac{k}{F_{jk}} \right) = \sum_{k=1}^{n} k^2 \frac{f_{ik} f_{jk}}{F_{ik} F_{jk}} \tag{22}$$

where F_{ik} is the number of texts in S_k in which v_i occurs and f_{ik} is the frequency of v_i in x_k. This association measure computes weighted links between words. As said before, we abstract from this weighting. However, we apply a threshold in order to exclude low-weighted associations. This is done by deleting all edges whose weight $\alpha(v_i, v_j, S_n)$ is smaller than $\mu - \frac{1}{8}\sigma$.[29] Note that we compute word associations only for wikis of regional, technical and scientific communication, thereby testing the separability of all, agent, document and word networks.

A remark on the text-technology used here: we utilize the HyGraphDB (Gleim et al., 2007) for downloading and managing wikis. It is used to process the serializations of wiki graphs by means of the *Graph eXchange Language* (GXL) (Holt et al., 2006). The GXL maps graph structures of a wide range of complexity. A central benefit of the HyGraphDB is the expressiveness of this underlying data model which allows to capture web and text documents as well as multimodal documents.[30] Using the HyGraphDB, the wiki networks were downloaded over a longer period of time which ranges from June 2007 (in the case of the Duckipedia) to February 2008 (in the case of the Wikiversity) (cf. Table 9).

Looking back on the size of the wiki networks in the Tables 10–12 we see that the number of their vertices (i.e. agents or documents, respectively) and edges (collaborations or intertextual relations) hardly serve as reliable network characteristics. Rather, we need to consider the complete set of attributes of Table 8. This is done in the following section.

5.2 Topological Network Characteristics

The Tables 10–12 show the values of the composite features (cf. Table 8) of the agent, document and word networks of five communication areas (cf. Table 2). The results coincide as they predominantly characterize the networks as small worlds — in terms of the model of Watts and Strogatz (1998) and of Barabási and Albert (1999). This is seen as follows:

- Watts and Strogatz (1998) present four standard values for judging the small world status of a graph G: the cluster values C_{re}, C_{ra} and the average geodesic distances L_{re} and L_{ra} of a corresponding regular and random graph of equal order.[31] These standard values are computed by the rules summarized in Table 13. Watts and Strogatz predict that clustering and distances of vertices in small worlds behave as follows: $C_{re}(G) \sim C_1(G) \gg C_{ra}(G)$ and

[29] This value is empirically established and needs further justification. At the present stage of word graph extraction it is justified as a bound below which the word graph gets to dense and above which it is thinned out to much in order to exhibit the small world-property. Note also that we use a frequency threshold in order to filter out hapax legomena and other low frequency words.

[30] The HyGraphDB (Gleim et al., 2007) has been developed as part of the X1 project of the Sonderforschungsbereich 673 *Alignment in Communication* and of the A4 project of the Research Group 437 *Text Technological Information Modeling*. Cf. www.sfb673.org/x1 for more information on the HyGraphDB.

[31] The *order* $|G|$ of a graph G is the number of its vertices.

Graph	Clustering	Estimation	Distance	Estimation
regular	high	$k < \frac{2}{3}n \Rightarrow C_{re} = \frac{3(k-2)}{4(k-1)}$	long	$L_{re} = \frac{n}{2k} \gg 1$
SW	high	see Equation 1	short	see Equation 2
random	low	$C_{ra} = k/n$	short	$L_{ra} \sim \frac{\ln(n)}{\ln(k)}$

Table 13: Cluster values and average distances in small worlds compared to regular graphs and random graphs. Estimators are given for corresponding random and regular graphs subject to the condition that $n \gg k \gg \ln n \gg 1$ (Watts and Strogatz, 1998) where $n = |V|$ and $k = d(G)$. $k \gg \log n$ ensures that the corresponding random graph is connected (Baldi et al., 2003).

$L_{re}(G) \gg L(G) \sim L_{ra}(G)$. Comparing these expected values with those observed in our network corpus we find that the latter two conditions are predominantly met.[32] The only exception is given by lexical networks which — possibly due to their highly parameterized computation (remember the thresholds in use) — do not allow to apply these conditions. This hints at a classificatory power of cluster values at least what regards ontologically divergent networks.

- Scale-freeness is also confirmed. Barabási and Albert (1999) show that networks as the WWW tend to be scale-free in the sense that the degrees of their vertices are distributed according to a power law (cf. Section 4). In such networks a very small group of vertices is highly and a large group rarely connected with a Zipfian transition between these extremal cases. For agent networks this means, e.g., that a small group of persons is responsible for the majority of collaborations. For document networks this means that a small group of documents serves for the largest part of cohesion. These are documents which are most frequently passed when users follow paths in hypertext. Both interpretations are confirmed by our analysis of social and textual networks (see Column γ in the Tables 10 and 11): we find that the degrees of vertices of agent and document networks are — with a small number of exceptions — distributed such that power laws can be successfully fitted (see the corresponding values of the adjusted coefficient of determination $\bar{R}^2$ in Tables 10–12 which are mostly larger than 0.8). This means that if we randomly select a vertex from such graphs, it is very likely linked with only one or two other vertices. In other words: collaborations as well as intertextual relations are rather rare events.[33] However, by the Tables 10–12 we also see that the connectivity distribution is less skewed in agent networks than in document networks: not only are the γ values smaller, but also the values of $\bar{R}^2$ indicate poorer fits. Finally, the word networks in Table 12 do not show Zipfian degree distributions — in contrast to what is reported, for example, by Ferrer i Cancho and Solé (2001). This is certainly due to the different network induction methods used in both studies.

A note on power law fitting: we perform fitting according to Newman's (2005) method. Thus, we do not cut input distributions — *neither on the left, nor on the right margin*. Note that assuming different values $k_{\min}$ and $k_{\max}$ as the degrees *from which* and *up to which* power laws are fitted might account

[32]Note that a value of -1 of $C_{re}(G), C_{ra}(G), L_{re}(G)$ or $L_{ra}(G)$ indicates that the condition for its estimation as reported in Table 13 does not hold.

[33]As we will see this does *not* mean that the graphs are marginally compact.

for differences in power law statistics reported on the same networks. Further, we do not perform grouping which is often used to overcome the problem of 'fraying tails' in empirical distributions. Rather, we always explore the whole data set and fit to the complementary cumulative distribution derived from the ranked degree data as described by Newman (2005). Further, we fit to get that value of γ which maximizes the adjusted coefficient of determination $\bar{R}^2$ for the input distribution.

Now, let us look beyond classical characteristics of small worlds. This relates, for example, to Newman's coefficient r of assortative mixing. According to the Tables 10–12 we see that (irrespective of minor differences) agent, document and word networks tend to disassortative mixing. This observation contrasts the prediction made by Newman and Park (2003a) who claim assortative mixing for social networks. This is only observable for a subset of lexical networks of the area of regional communication which indicates a classificatory power of this feature. However, we do not observe this pattern on the level of collaboration networks as induced here.

Can we distinguish wiki networks further on by some other quantitative characteristics? This holds, for example, for power laws fitted to the empirical distributions of $\bar{k}_{nn}(k)$ and $C(k)$. We observe that while document networks predominantly tend to disassortativity, the underlying agent networks behave less obvious. In their cases fitting power laws to $\bar{k}_{nn}(k)$ is mostly unreliable (in the sense of small values of the adjusted coefficient of determination where $\bar{R}^2_{\gamma_{\bar{k}_{nn}(k)}}(G) \ll 1$).[34] In these cases we can hardly assume a Zipfian distribution of mixing patterns as a function of the degree k — neither in the sense of assortative nor of disassortative mixing. Thus, agent networks are (other than document networks) all but not definite in terms of their disassortativity.

This finding is also confirmed by agent networks when regarding clustering as a function of vertex degree k. That is, collaboration networks are not hierarchical in the sense of Ravasz and Barabási (2003). However, document networks tend to be this (even though with some exceptions): here we observe cases (as, e.g., the city wiki of Bielefeld with the network ID 9) which have a hierarchical or, more concretely, modular topology. As $\bar{R}^2_{\gamma_{C(k)}}(G) \gg .7$ we can reliably assume for such networks that instances of the very small set of highly interlinked documents tend to be linked with low-degree nodes. Such networks are organized around a tiny set of highly interlinked nodes, a large periphery of hardly interlinked documents and a sparsely linked middle field in-between. In summary we see that at least $\gamma_{\bar{k}_{nn}(k)}, \gamma_{C(k)}$ and the corresponding coefficients of determination open the door to classifying networks which are homogeneously attributed as small worlds. This is systematically examined in Section 6.

Next, we have a look at compactness as a measure of classical hypertext theory. Obviously, agent, text and word networks tend to be extremely compact indicating short geodesic distances between any randomly chosen pair of their nodes (cf. the Column Cp). This observation is confirmed by the remarkably small average geodesic distances in these networks. We also observe a slight difference between social and semiotic networks: lexical and textual networks tend to be less compact than the former. This is the place to prevent a misunderstanding indicated by the Tables 10–12: because of space limits all their values have been rounded to at most two decimal places so that a compactness value of 1 does not necessarily indicate a completely connected graph. On the contrary, the compactness values of all graphs evaluated in this study are smaller than 1. That is we observe an effect of

[34] (Unrealistically) negative values of the adjusted coefficient of determination indicate completely unsuccessful fittings — they are produced by the software which has been used to make the fittings, that is, MATLAB.

rounding which directly affects the compactness measure as its values are mostly near to 1. However, this does not affect network classification which operates on a higher precision. Anyhow, there is not very much variance in the compactness values reported here so that we do not expect a high separability by using these values. The picture gets a little bit more sophisticated when looking for the compactness $\mathsf{Cp}_{\mathsf{lcc}}(G)$ of the largest connected component of a network G. As this measure uses another standardization than Cp the values of both are not directly comparable. Once more we observe very high values of $\mathsf{Cp}_{\mathsf{lcc}}$ so that this measure is — by analogy to Cp — hardly informative. However, there is the tendency that semiotic networks have a higher LCC-compactness than social networks.[35]

So what can we say about network characteristics which stem from social network theory? More specifically: *What can we say about the centrality status of the networks under consideration?* Because of their spectrum the centrality values of agent networks reported in Table 10 hardly demonstrate a tendency. However, we see that the centrality values of word networks are mostly in-between the centrality values of the corresponding agent and document networks. Apparently, centrality measures are important what regards the separation of ontologically divergent networks. Whether this observation is confirmed by our network classification experiment will be shown in Section 6. Note that the values of the centrality measures indicate a similarity between document networks and star graphs as exemplified by the wiki of the Firefox browser in Figure 2. This seems to be a stronger tendency of document graphs while agent networks tend to have smaller centrality values.

As this paper is *NOT* about a descriptive analysis of network characteristics we stop our short investigation of the Tables 10–12 (leaving out many interesting observations and distinctions to be made by single topological network characteristics) and pass over to the topic of network classification.[36] In summary, the Tables 10–12 show that wiki networks tend to have a small world-like topology — irrespective of the communication area considered. There are two implications related to this observation:

- Firstly, the *Network Correlation Hypothesis* (NCH) of Section 1 is supported in the sense that the small world-property of document and agent networks coincides. However, we do not get reliable evidence for the absence of this property on the document level subject to its absence on the agent level or vice versa. Thus, the NCH is only partly confirmed.

- Further, the fact that we predominantly observe the small world-property makes it hardly informative with respect to the divergence of networks of different provenance. We neither observe such a divergence in the case of ontologically different (agent or document) networks nor in the case of functionally heterogenous networks (of different communication areas) by checking small word-models.[37] Thus, it seems that these kinds of analyses do not relate to the network separability hypothesis of Section 1.

[35] A correlation analysis of the values of these and related measures (e.g. L) would be appropriate but is out of reach in the present paper because of space limits.

[36] A profound analysis of the variance and other dispersion parameters of the values reported in the Tables 10–12 is clearly indicated but out of reach in the present paper — it will be part of future work. The present paper is limited to a summary examination of the topological features of Table 8 in the framework of network classification.

[37] As they are bound to too many parameters we disregard word networks at this time.

Notwithstanding the latter observation, the topological features of Table 8 have a potential for network classification. This is checked in the next section which revises the questionable information value of complex network analysis.

6 Unsupervised Learning of Network Structure

In this section we classify tripartite wiki graphs in the framework of QNA. Our starting point is the network separability hypothesis of Section 1 and its concretization according to Section 4. This relates to the ontological (Section 6.1) and functional separability (Section 6.2) of the constituent graphs of wiki networks.

6.1 Classification Scenario I: Ontological Separability

With respect to *ontological* separability the NSH says that networks of divergent social-semiotic provenance are distinguishable according to their laws of networking. In this sense we should be able to classify them even if we completely abstract from what their nodes and links denote, that is, we completely disregard their content but only rely on their structure. This scenario corresponds to classification scenario A of Figure 6. We test this variant of the NSH with respect to the word, document and agent networks of the area of regional, technical and scientific communication. We also test it in the area of knowledge and leisure communication by focusing on document and agent networks. The results of these five different classification experiments are reported in Table 14. We see that if we take the complete set of features $\mathbb{F}$ (cf. Equation 19) into account classification is always perfect (as indicated by an F-value of 1) except in the area of technical communication. In Scenario A, D and E we also get the result that the quality of classification is not influenced by varying the operative clustering method (amongst others, ward, weighted and median linkage give all perfect results). Thus we retain the hypothesis that the ontogenetically divergent areas of lexical, textual and social networking are perfectly separable by focusing on their topological features. In other words: though being homogeneously attributed as small worlds, networks of lexical units have different topologies than those being composed of textual units. Further, these two kinds of semiotic units follow different laws of networking than agents seen as social units.

However, the picture gets more sophisticated if we distinguish between first and second order feature space representations, that is, between projecting feature vectors from the set X and Y (cf. Equations 20 and 21), respectively. On the one hand, we see that when projecting second order similarity values as mapped by vectors $\mathbf{y} \in Y$ the classification downgrades. That is, if we represent each network by its topological similarity to all other networks of the same classification experiment (as done when computing the object representations $\mathbf{y} \in Y$), classification gets worse. In these cases, the separability of the networks is — in contrast to our expectation — reduced. In all classification scenarios reported in Table 14 the goodness of classification (as measured by F-values) suffers from working on Y-based instead of X-based network representations. Thus, we can disregard this second order representation model and focus instead on first order object representations built directly from the topological features summarized in Table 8.

So what do we get if we look for optimal subsets of topological features of $\mathbb{F}$ *(see Equation 19) when performing the classification experiments A–E?* As goodness of classification is already equal to its maximum in Experiments A, C, D and E computing this optimal subset only hints at some potential to minimize the effort of network representation. Evidently, this holds for the Scenarios A, C, D and E: As reported in Table 15 there always exists such a subset of optimal features which outputs an F-value at least as high as the one produced by using the complete feature set $\mathbb{F}$. Moreover, in Scenario B in which $\mathbb{F}$ fails to generate a perfect classification we additionally see that this classification is produced by the subset $\mathbb{F}_{\mathsf{opt}} = \{F_1, F_3, F_6, F_7, F_8, F_{10}, F_{12}, F_{13}, F_{14}, F_{16}, F_{18}, F_{21}, F_{24}\} \subset \mathbb{F}$. That is, $|\mathbb{F}_{\mathsf{opt}}| = 13$ features suffice to produce a much better classification than the one induced by the whole feature set. This classificatory potential of a feature subset selection is also evident when looking for second order network representations: Table 14 reports for all Scenarios A–E an increase in goodness of classification if an evolutionary search for an optimal feature subset is included. Thus, we can infer that the topological characteristics enumerated in Table 8 vary with respect to their classificatory power. There are topological features (e.g., in Scenario B this relates to the diameter D, the coefficient of assortativity r, the degree centrality DC and the compactness Cp, respectively) which prevent a perfect classification when being included into network representation. These are features which make appear networks alike even if they are ontologically divergent. As reported in Table 15 on Scenario B this relates to all features which do not belong to the corresponding optimal subset of features.

By Table 15 we also get the information that Watts & Strogatz cluster value C_1 together with the fraction of the largest connected component lcc, the adjusted coefficient of determination $\bar{R}^2_{\gamma_S}$ of the power law fitted to the size distribution of connected components, the exponent $\gamma_{\bar{k}_{nn}(k)}$ of the power law fitted to the distribution of $\bar{k}_{nn}(k)$ values, the betweenness centrality BC and the compactness of the largest connected components $\mathsf{Cp}_{\mathsf{lcc}}$ are most frequently separable within the 10 classification scenarios of Table 14. Further, we get the information that the average geodesic distance is more informative about class membership of ontologically divergent networks than the diameter of a graph, while assortative mixing, graph and degree centrality values as well as the compactness of a graph are less informative topological features. According to Table 15 the least informative feature is the exponent of the power law fitted to the distribution of the betweenness centrality values of a graph — this result may hint at the universality of this feature as studied by Barthélemy (2004). Because of space limits in the present paper, a systematic sensitivity analysis is out of reach and will be part of future work. However, the next section gives some hints whether our findings also hold in the area of classifying functionally divergent networks.

A last note on the choice of the cluster method: Although we have chosen among hierarchical (agglomerative with subsequent partitioning) and partitive k-means clustering, we observe that hierarchical methods are mostly more successful. Moreover, Ward's method of linkage dominates the range of successful agglomerative clusterings (thereby outperforming the methods of single, complete, average, centroid, weighted and median linkage). Starting from this observation we constrain the evolutionary search for optimal feature subsets to methods of hierarchical clustering (leaving out the k-means method).

6.2 Classification Scenario II: Functional Separability

In this section we look at the NSH from the point of view of functionally heterogeneous, but ontologically homogeneous networks. More specifically, we perform two classification experiments: Firstly, we pool all agent networks of the five communication areas considered here and compute the best partition thereof based on unsupervised learning. Secondly, we do the same by example of the corresponding document networks. By this procedure we aim to estimate the degree by which it is possible to predict the function of a wiki (i.e. its membership to a certain communication area) by solely examining its structure.

According to Table 16 we see that this functional variant of the NSH performs much worse than its ontological counterpart examined in the preceding section. On the level of agent networks as wells of document networks the corresponding baselines are outperformed. However, in the case of agent networks we only get an F-measure value of $.49842$ while in the case of the corresponding document networks the F-measure value is raised to $.60386$ (once more, operating on Y-based object representations (cf. Equation 21) does not better the classification).

Nevertheless, the results are good enough in the sense that they do not falsify the hypothesis about the functional separability of wiki networks, even if they are far away from the optimal results of ontology-related network classifications reported in Section 6.1. This finding is confirmed to a much higher degree by document networks than by agent networks. *In other words:* the component graphs of wiki networks are more distinct in ontological terms than in functional terms whereby document networks can be functionally classified more reliable than agent networks. The results of Table 16 do not falsify these two hypotheses. Note that z-score transformations of the input variables do not change this picture. Further, we do not get a better F-measure value when additionally subdividing the corpus of wikis of scientific communication into the one of conferences and of community building. Other than expected there is nothing peculiar about conference wikis in comparison to other wikis of scientific communication which address long-term community building — at least our wiki corpus does not support this distinction.

As the best classifications of the networks under consideration are not produced by the feature set $\mathbb{F}$ as a whole, but by some of its subsets it is worth to have a closer look at them (cf. Table 17). When comparing Table 17 with Table 15 we see that features which support functional separability differ somehow from those which support ontological separability. In Table 17 the exponent γ of the degree distribution is most successful in conjunction with the average geodesic distance, the diameter, the average degree and some other network features. These results are based on four experiments only so that we shall not overestimate them. However, we get a first hint that functional network classifications rely on different topological features than ontological ones. Further experiments have to check whether this finding is falsified or not.

In summary, our results on functional network classifications are indefinite enough so that future experiments have to decide between two contrary cases: On the one hand, there may exist other topological features which support functional separability of networks of different communication areas to a much higher degree than observed here. On the other hand, this search might fail so that we either have to question our classification of communication areas or to accept that they are not distinguished by laws of social-semiotic networking. It is worth to study these questions further on as

they break fresh ground in a rapidly growing research area.

7 Conclusion

In this paper we have presented a formal framework for representing, analyzing and classifying linguistic networks on the level of agent, document and lexico-grammatical units. We have exemplified this model by wiki networks in the growing area of web-based collaborative writing. Our findings suggest a correlation of small world-topologies at least on the level of social and textual networking. These findings also indicate the distinguishability of ontologically and functionally divergent networks — the former to a much higher degree than the latter. There is a wide range of theoretical and practical impacts of this research. First of all, our approach opens the door to structure-oriented machine learning in the area of large complex networks. Thus it might be used in automatically exploring or maintaining web-based resources (e.g. in semisupervised scenarios of collaborative writing). Further, building social-semantic services in digital libraries (beyond existing search technologies) which obey the laws of linguistic networking can be seen as a further application scenario. Last but not least, this paper also addresses research on language evolution and language change. The reason is that we provide empirical data on the co-evolvement of social-semiotic networks which can be used for evaluating simulation models in this area (Mehler, 2008b). Generally speaking, because of the ever-growing areas of web-based communication we are in need of basic research in this field. The present paper contributes to this research on the level of linguistic conceptualizations, formal modeling and empirical analysis.

Acknowledgement

Financial support of the German Research Foundation (DFG) through the SFB 673 *Alignment in Communication* (`www.sfb673.org`) — Project A3 *Dialogue Games and Group Dynamics* — and the Research Group 437 *Text Technological Information Modeling* — Project A4 *Induction of Document Grammars for the Representation of logical Hypertext Document Structures* — at Bielefeld University is gratefully acknowledged.

References

Amaral, L. A. N., Scala, A., Barthélémy, M., and Stanley, H. E. (2000). Classes of small-world networks. *Proceedings of the National Academy of Sciences*, 97(21):11149–11152.

Baldi, P., Frasconi, P., and Smyth, P. (2003). *Modeling the Internet and the Web*. Wiley, Chichester.

Barabási, A.-L. and Albert, R. (1999). Emergence of scaling in random networks. *Science*, 286:509–512.

Barrat, A., Barthélemy, M., Pastor-Satorras, R., and Vespignani, A. (2004). The architecture of complex weighted networks. *Proc. Nat. Acad. Sci. USA*, 101(11):3747–3752.

Barthélemy, M. (2004). Betweenness centrality in large complex networks. *European Physical Journal B*, 38:163–168.

Berge, C. (1989). *Hypergraphs: Combinatorics of Finite Sets*. North Holland, Amsterdam.

Biber, D. (1995). *Dimensions of Register Variation: A Cross-Linguistic Comparison*. Cambridge University Press, Cambridge.

Björneborn, L. (2004). *Small-World Link Structures across an Academic Web Space: A Library and Information Science Approach*. PhD thesis, Royal School of Library and Information Science, Department of Information Studies, Denmark.

Blanchard, P. and Krüger, T. (2004). The cameo principle and the origin of scale free graphs in social networks. *Journal of statistical physics*, 114(5-6):399–416.

Bollobás, B. and Riordan, O. M. (2003). Mathematical results on scale-free random graphs. In Bornholdt, S. and Schuster, H. G., editors, *Handbook of Graphs and Networks. From the Genome to the Internet*, pages 1–34. Wiley-VCH, Weinheim.

Botafogo, R. A., Rivlin, E., and Shneiderman, B. (1992). Structural analysis of hypertexts: Identifying hierarchies and useful metrics. *ACM Transactions on Information Systems*, 10(2):142–180.

Brandes, U. (2001). A faster algorithm for betweenness centrality. *Journal of Mathematical Sociology*, 25(2):163–177.

Brinker, K. (1992). *Linguistische Textanalyse. Eine Einführung in Grundbegriffe und Methoden*. Erich Schmidt, Berlin.

Budanitsky, A. and Hirst, G. (2006). Evaluating WordNet-based measures of lexical semantic relatedness. *Computational Linguistics*, 32(1):13–47.

Capocci, A., Servedio, V. D. P., Colaiori, F., Buriol, L. S., Donato, D., Leonardi, S., and Caldarelli, G. (2006). Preferential attachment in the growth of social networks: the case of wikipedia. *Phys. Rev. E*, 74:036116.

Chakrabarti, S. (2002). *Mining the Web: Discovering Knowledge from Hypertext Data*. Morgan Kaufmann, San Francisco. Das Buch behandelt das Teilgebiet bzw. das Anwendungsgebiet des Web Mining.

Dehmer, M. (2005a). Data Mining-Konzepte und graphentheoretische Methoden zur Analyse hypertextueller Daten. *LDV-Forum*, 20(1):113–141.

Dehmer, M. (2005b). *Strukturelle Analyse Web-basierter Dokumente*. Multimedia und Telekooperation. DUV, Berlin.

Dehmer, M., Mehler, A., and Emmert-Streib, F. (2007). Graph-theoretical characterizations of generalized trees. In *Proceedings of the 2007 International Conference on Machine Learning: Models, Technologies & Applications (MLMTA'07), June 25-28, 2007, Las Vegas*.

Diestel, R. (2005). *Graph Theory*. Springer, Heidelberg.

Emmert-Streib, F., Dehmer, M., and Kilian, J. (2005). Classification of large graphs by a local tree decomposition. In Arabnia, H. R. and Scime, A., editors, *Proceedings of DMIN'05, International Conference on Data Mining, Las Vegas, Juni 20-23*, pages 200–207.

Erdős, P. and Rényi, A. (1959). On random graphs. *Publicationes Mathematicae*, 6:290–297.

Feldman, R. and Sanger, J., editors (2007). *The Text Mining Handbook*. Cambridge University Press, Cambridge.

Ferrer i Cancho, R. and Solé, R. V. (2001). The small-world of human language. *Proceedings of the Royal Society of London. Series B, Biological Sciences*, 268(1482):2261–2265.

Fraas, C. (1997). Bedeutungskonstitution im Diskurs – Intertextualität über variierende Wiederaufnahme diskursiv zentraler Konzepte. Eine exemplarische Analyse. In und Ulla Fix, J. K., editor, *Textbeziehungen: linguistische und literaturwissenschaftliche Beiträge zur Intertextualität*, pages 219–234. Stauffenburg, Tübingen.

Gallo, G., Longo, G., Pallottino, S., and Nguyen, S. (1993). Directed hypergraphs and applications. *Discrete Appl. Math.*, 42(2-3):177–201.

Gleim, R., Mehler, A., and Eikmeyer, H.-J. (2007). Representing and maintaining large corpora. In *Proceedings of the Corpus Linguistics 2007 Conference, Birmingham (UK)*.

Hage, P. and Harary, F. (1995). Eccentricity and centrality in networks. *Social Networks*, 17:57–63.

Halliday, M. A. K. (1978). *Language as Social Semiotic: the Social Interpretation of Language and Meaning*. Edward Arnold Publishers, London.

Hammwöhner, R. (1997). *Offene Hypertextsysteme: das Konstanzer Hypertextsystem (KHS) im wissenschaftlichen und technischen Kontext*. Universitätsverlag, Konstanz.

Heyer, G., Quasthoff, U., and Wittig, T. (2006). *Text Mining: Wissensrohstoff Text*. W3L, Herdecke.

Hjelmslev, L. (1969). *Prolegomena to a Theory of Language*. University of Wisconsin Press, Madison.

Holt, R. C., Schürr, A., Elliott Sim, S., and Winter, A. (2006). GXL: A graph-based standard exchange format for reengineering. *Science of Computer Programming*, 60(2):149–170.

Hotho, A., Nürnberger, A., and Paaß, G. (2005). A Brief Survey of Text Mining. *LDV-Forum*, 20(1):19–62.

Jurafsky, D. and Martin, J. H. (2000). *Speech and Language Processing: an Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. Prentice Hall, Upper Saddle River.

Kuhlen, R. (2004). Kollaboratives Schreiben. In Bieber, C. and Leggewie, C., editors, *Interaktivität – ein transdisziplinärer Schlüsselbegriff*, pages 216–239. Campus-Verlag, Frankfurt.

Landow, G. P. (1987). Relationally encoded links and the rhetoric of hypertext. In *Proceedings of the First ACM Conference on Hypertext*, pages 331–343. ACM.

Manning, C. D. and Schütze, H. (1999). *Foundations of Statistical Natural Language Processing*. MIT Press, Cambridge, Massachusetts.

Martin, J. R. (1992). *English Text. System and Structure*. John Benjamins, Philadelphia.

Mehler, A. (2002). Hierarchical orderings of textual units. In *Proceedings of the 19th International Conference on Computational Linguistics (COLING '02), August 24 – September 1, 2002, Taipei, Taiwan*, pages 646–652, San Francisco. Morgan Kaufmann.

Mehler, A. (2005). Lexical chaining as a source of text chaining. In Patrick, J. and Matthiessen, C., editors, *Proceedings of the 1st Computational Systemic Functional Grammar Conference, University of Sydney, Australia*, pages 12–21.

Mehler, A. (2006). Text linkage in the wiki medium – a comparative study. In Karlgren, J., editor, *Proceedings of the EACL Workshop on New Text – Wikis and blogs and other dynamic text sources, April 3-7, 2006, Trento, Italy*, pages 1–8.

Mehler, A. (2007a). Evolving lexical networks. A simulation model of terminological alignment. In Benz, A., Ebert, C., and van Rooij, R., editors, *Proceedings of the Workshop on Language, Games, and Evolution at the 9th European Summer School in Logic, Language and Information (ESSLLI 2007), Trinity College, Dublin, 6-17 August*, pages 57–67.

Mehler, A. (2007b). Structure formation in the web. A graph-theoretical model of hypertext types. In Witt, A. and Metzing, D., editors, *Linguistic Modeling of Information and Markup Languages. Contributions to Language Technology*, Text, Speech and Language Technology. Springer, Dordrecht.

Mehler, A. (2008a). Large text networks as an object of corpus linguistic studies. In Lüdeling, A. and Kytö, M., editors, *Corpus Linguistics. An International Handbook of the Science of Language and Society*, pages 328–382. De Gruyter, Berlin/New York.

Mehler, A. (2008b). On the impact of community structure on self-organizing lexical networks. In Smith, A. D. M., Smith, K., and Ferrer i Cancho, R., editors, *Proceedings of the 7th Evolution of Language Conference (Evolang 2008), March 11-15, 2008, Barcelona*, pages 227–234. World Scientific.

Mehler, A., Geibel, P., and Pustylnikov, O. (2007a). Structural classifiers of text types: Towards a novel model of text representation. *LDV Forum – Zeitschrift für Computerlinguistik und Sprachtechnologie*, 22(2):51–66.

Mehler, A., Waltinger, U., and Wegner, A. (2007b). A formal text representation model based on lexical chaining. In *Proceedings of the KI 2007 Workshop on Learning from Non-Vectorial Data (LNVD 2007) September 10, Osnabrück*, pages 17–26, Osnabrück. Universität Osnabrück.

Melnikov, O., Tyshkevich, R., Yemelichev, V., and Sarvanov, V. (1994). *Lectures on Graph Theory*. BI-Wissenschaft, Mannheim/Leipzig.

Miller, G. A. and Charles, W. G. (1991). Contextual correlates of semantic similarity. *Language and Cognitive Processes*, 6(1):1–28.

Newman, M. E. J. (2003). The structure and function of complex networks. *SIAM Review*, 45:167–256.

Newman, M. E. J. (2005). Power laws, Pareto distributions and Zipf's law. *Contemporary Physics*, 46:323–351.

Newman, M. E. J. and Park, J. (2003a). The origin of degree correlations in the internet and other networks. *Phys. Rev. E*, 68:026121.

Newman, M. E. J. and Park, J. (2003b). Why social networks are different from other types of networks. *Physical Review E*, 68:036122.

Pastor-Satorras, R., Vázquez, A., and Vesipignani, A. (2001). Dynamical and correlation properties of the internet. *Physical Review Letters*, 87(25):268701.

Pastor-Satorras, R. and Vespignani, A. (2004). *Evolution and Structure of the Internet*. Cambridge University Press, Cambridge.

Power, R., Scott, D., and Bouayad-Agha, N. (2003). Document structure. *Computational Linguistics*, 29(2):211–260.

Pustylnikov, O. (2007). Guessing Text Type by Structure. Proceedings of the ESSLLI Student Session 2007.

Pustylnikov, O. and Mehler, A. (2007). Structural differentiae of text types. A quantitative model. In *Proceedings of the 31st Annual Conference of the German Classification Society on Data Analysis, Machine Learning, and Applications (GfKl)*.

Pustylnikov, O. and Mehler, A. (2008). Towards a uniform representation of treebanks: Providing interoperability for dependency tree data. In *Proceedings of First International Conference on Global Interoperability for Language Resources (ICGL 2008), Hong Kong SAR, January 9-11*.

Ravasz, E. and Barabási, A.-L. (2003). Hierarchical organization in complex networks. *Physical Review E*, 67:026112.

Rieger, B. B. (1989). *Unscharfe Semantik: Die empirische Analyse, quantitative Beschreibung, formale Repräsentation und prozedurale Modellierung vager Wortbedeutungen in Texten*. Peter Lang, Frankfurt a. M.

Rieger, B. B. (2002). Semiotic cognitive information processing: Learning to understand discourse. A systemic model of meaning constitution. In Kühn, R., Menzel, R., Menzel, W., Ratsch, U., Richter, M. M., and Stamatescu, I. O., editors, *Perspectives on Adaptivity and Learning*, pages 347–403. Springer, Berlin.

Ruge, G. (1995). *Wortbedeutung und Termassoziation. Methoden zur automatischen semantischen Klassifikation*. Olms, Hildesheim.

Salton, G. (1989). *Automatic Text Processing: The Transformation, Analysis, and Retrieval of Information by Computer*. Addison Wesley, Reading, Massachusetts.

Schütze, H. (1998). Automatic word sense discrimination. *Computational Linguistics*, 24(1):97–123.

Serrano, M. Á., Boguñá, M., and Pastor-Satorras, R. (2006a). Correlations in weighted networks. *Physical Review*, 74(055101(R)):1–4.

Serrano, M. Á., Boguñá, M., Pastor-Satorras, R., and Vespignani, A. (2006b). Correlations in complex networks. In Caldarelli, G. and Vespignani, A., editors, *Structure and Dynamics of Complex Networks, From Information Technology to Finance and Natural Science*, chapter 1. World Scientific.

Steels, L. (2002). Grounding symbols through evolutionary language games. In Cangelosi, A. and Parisi, D., editors, *Simulating the Evolution of Language*, chapter 10, pages 211–226. Springer, London.

Steels, L. (2006). Collaborative tagging as distributed cognition. *Pragmatics & Cognition*, 14(2):287–292.

Steyvers, M. and Tenenbaum, J. (2005). The large-scale structure of semantic networks: Statistical analyses and a model of semantic growth. *Cognitive Science*, 29(1):41–78.

Storrer, A. (2002). Coherence in text and hypertext. *Document Design*, 3(2):156–168.

Thibault, P. J. (1997). Intertextuality. In Asher, R. E., editor, *The encyclopedia of language and linguistics*, pages 1751–1754. Pergamon Press, Oxford.

Tuldava, J. (1995). *Methods in quantitative linguistics*. Wissenschaftlicher Verlag, Trier.

Tuldava, J. (1998). *Probleme und Methoden der quantitativ-systemischen Lexikologie*. Wissenschaftlicher Verlag, Trier.

Volchenkov, D., Volchenkova, L., and Blanchard, P. (2002). Epidemic spreading in a variety of scale free networks. *Physical Review E*, 66(4):046137.

Wasserman, S. and Faust, K. (1999). *Social Network Analysis. Methods and Applications*. Cambridge University Press, Cambridge.

Watts, D. J. (2003). *Six Degrees. The Science of a Connected Age*. W. W. Norton & Company, New York/London.

Watts, D. J. and Strogatz, S. H. (1998). Collective dynamics of 'small-world' networks. *Nature*, 393:440–442.

Winne, P., Gupta, L., and Nesbit, J. (1994). Exploring individual differences in studying strategies using graph theoretic statistics. *Alberta Journal of Educational Research*, 40:177–193.

Zlatic, V., Bozicevic, M., Stefancic, H., and Domazet, M. (2006). Wikipedias: Collaborative web-based encyclopedias as complex networks. *Physical Review E*, 74:016115.

A	F-Value	#C	Clustering	Method
$\mathsf{X}[\mathbb{F}_{\mathsf{opt}}]$	1.0	3	hierarchical	median
$\mathsf{X}[\mathbb{F}]_1$	1.0	3	hierarchical	median
$\mathsf{X}[\mathbb{F}]_2$	1.0	3	hierarchical	ward
$\mathsf{X}[\mathbb{F}]_3$	1.0	3	hierarchical	weighted
$\mathsf{Y}[\mathbb{F}_{\mathsf{opt}}]$	1.0	3	hierarchical	ward
$\mathsf{Y}[\mathbb{F}]_1$	.79167	3	k-means	cityblock
$\mathsf{Y}[\mathbb{F}]_2$	.78591	3	hierarchical	ward
$\mathsf{Y}[\mathbb{F}]_3$	.78591	3	hierarchical	centroid
Baseline	.4359	3	random	—

B	F-Value	#C	Clustering	Method
$\mathsf{X}[\mathbb{F}_{\mathsf{opt}}]$	1.0	3	hierarchical	ward
$\mathsf{X}[\mathbb{F}]_1$	.82222	3	hierarchical	weighted
$\mathsf{X}[\mathbb{F}]_2$	.78022	3	hierarchical	ward
$\mathsf{X}[\mathbb{F}]_3$	.77128	3	hierarchical	complete
$\mathsf{Y}[\mathbb{F}_{\mathsf{opt}}]$	.96658	3	hierarchical	complete
$\mathsf{Y}[\mathbb{F}]_1$	.7619	3	hierarchical	centroid
$\mathsf{Y}[\mathbb{F}]_2$	.7619	3	hierarchical	single
$\mathsf{Y}[\mathbb{F}]_3$	.73333	3	hierarchical	ward
Baseline	.46127	3	random	—

C	F-Value	#C	Clustering	Method
$\mathsf{X}[\mathbb{F}_{\mathsf{opt}}]$	1.0	3	hierarchical	ward
$\mathsf{X}[\mathbb{F}]_1$	1.0	3	hierarchical	ward
$\mathsf{X}[\mathbb{F}]_2$	.75439	3	hierarchical	weighted
$\mathsf{X}[\mathbb{F}]_3$	.75439	3	hierarchical	average
$\mathsf{Y}[\mathbb{F}_{\mathsf{opt}}]$	1.0	3	hierarchical	complete
$\mathsf{Y}[\mathbb{F}]_1$	.65741	3	hierarchical	complete
$\mathsf{Y}[\mathbb{F}]_2$	.60819	3	hierarchical	ward
$\mathsf{Y}[\mathbb{F}]_3$	.60784	3	k-means	cityblock
Baseline	.49433	3	random	—

D	F-Value	#C	Clustering	Method
$\mathsf{X}[\mathbb{F}_{\mathsf{opt}}]$	1.0	2	hierarchical	ward
$\mathsf{X}[\mathbb{F}]_1$	1.0	2	hierarchical	ward
$\mathsf{X}[\mathbb{F}]_2$	1.0	2	hierarchical	complete
$\mathsf{X}[\mathbb{F}]_3$	1.0	2	hierarchical	average
$\mathsf{Y}[\mathbb{F}_{\mathsf{opt}}]$	1.0	2	hierarchical	single
$\mathsf{Y}[\mathbb{F}]_1$	.7754	2	k-means	correlation
$\mathsf{Y}[\mathbb{F}]_2$	.7754	2	k-means	cosine
$\mathsf{Y}[\mathbb{F}]_3$	.7754	2	k-means	cityblock
Baseline	.60583	2	random	—

E	F-Value	#C	Clustering	Method
$\mathsf{X}[\mathbb{F}_{\mathsf{opt}}]$	1.0	2	hierarchical	ward
$\mathsf{X}[\mathbb{F}]_1$	1.0	2	hierarchical	ward
$\mathsf{X}[\mathbb{F}]_2$	1.0	2	hierarchical	weighted
$\mathsf{X}[\mathbb{F}]_3$	1.0	2	hierarchical	complete
$\mathsf{Y}[\mathbb{F}_{\mathsf{opt}}]$	1.0	2	hierarchical	ward
$\mathsf{Y}[\mathbb{F}]_1$	.71429	2	hierarchical	ward
$\mathsf{Y}[\mathbb{F}]_2$	.70833	2	k-means	correlation
$\mathsf{Y}[\mathbb{F}]_3$	.70833	2	k-means	cosine
Baseline	.61571	2	random	—

Table 14: Testing ontological network separability: F-measure values for separating agent, document and word networks in Scenarios A (*regional communication*), B (*technical communication*) and C (*scientific communication*) as well as separating agent and document networks in Scenarios D (*knowledge communication*), and E (*leisure communication*). According to Algorithm 1, F-values are computed for direct (X) and indirect (Y) object representations. For $Z \in \{X, Y\}$, $Z[\mathbb{F}_{\mathsf{opt}}]$ is the selection of the subset $\mathbb{F}_{\mathsf{opt}} \in \mathbb{F}$ (cf. Equation 19) of features which optimize the F-value according to Algorithm 1. Further, $Z[\mathbb{F}]_i$ is the F-measure value of rank i (in descending order) of classifying the corresponding corpus of networks based on the complete feature set $\mathbb{F}$. The column entitled *#C* denotes the number of classes which is input to the classification algorithm. The column entitled *Clustering* denotes the winner method together with the operative similarity measure denoted by the column *Method* (cf. Algorithm 1). Finally, the corresponding baseline scenario — based on random clustering averaged over 1000 runs — is shown in the last row of each table.

Setting	C_1	C_2	L	D	r	ϵ	lcc	γ	$\bar{R}^2$	γ_S	$\bar{R}^2$	$\gamma_{\bar{k}_{nn}}$	$\bar{R}^2$	$\gamma_{C(k)}$	$\bar{R}^2$	BC	γ_{BC}	$\bar{R}^2$	CC	DC	GC	Cp	Cp_{lcc}	Ch
A :: X	1	2					7		9		11	12			15			18		20			23	
A :: Y	1	2	3						9				13		15	16			19			22	23	24
B :: X	1		3			6	7	8		10		12	13	14		16		18			21			24
B :: Y		2	3	4			7			10	11		13			16			19	20			23	24
C :: X	1			4	5	6	7		9		11	12											23	24
C :: Y	1		3		5						11	12				16			19			22	23	
D :: X	1	2					7		9		11	12			15			18		20			23	
D :: Y	1		3	4			7			10	11			14		16	17		19			22	23	24
E :: X	1	2				6	7	8		10	11	12	13	14	15	16	17	18	19		21			
E :: Y		2	3		5			8	9	10	11	12	13		15	16			19				23	
$\sum$	8	6	6	3	3	3	7	3	5	5	8	7	5	3	5	7	2	4	6	3	2	3	8	5

Table 15: Optimal subsets of topological features in structure-based classifications of networks regarding their membership to ontological classes.

F	F-Value	#C	Clustering	Method
$\text{X}[\mathbb{F}_{\text{opt}}]$	.49842	5	hierarchical	ward
$\text{X}[\mathbb{F}]_1$	.44839	5	hierarchical	weighted
$\text{X}[\mathbb{F}]_2$	.42897	5	hierarchical	ward
$\text{X}[\mathbb{F}]_3$	.42177	5	hierarchical	median
$\text{Y}[\mathbb{F}_{\text{opt}}]$	.49015	5	hierarchical	complete
$\text{Y}[\mathbb{F}]_1$	.44662	5	hierarchical	complete
$\text{Y}[\mathbb{F}]_2$	.43308	5	hierarchical	centroid
$\text{Y}[\mathbb{F}]_3$	.43308	5	hierarchical	average
Baseline	.3697	5	random	—

G	F-Value	#C	Clustering	Method
$\text{X}[\mathbb{F}_{\text{opt}}]$	.60386	5	hierarchical	ward
$\text{X}[\mathbb{F}]_1$	.48085	5	k-means	cityblock
$\text{X}[\mathbb{F}]_2$	.46011	5	k-means	sqEuclidean
$\text{X}[\mathbb{F}]_3$	.45796	5	hierarchical	complete
$\text{Y}[\mathbb{F}_{\text{opt}}]$	.53441	5	hierarchical	weighted
$\text{Y}[\mathbb{F}]_1$	.45008	5	hierarchical	weighted
$\text{Y}[\mathbb{F}]_2$	.41761	5	hierarchical	ward
$\text{Y}[\mathbb{F}]_3$	.41528	5	hierarchical	centroid
Baseline	.3697	5	random	—

Table 16: Testing functional network separability: Scenario F denotes an experiment with *agent* networks pooled from five different communication areas (cf. Table 2), while Scenario G denotes an experiment in which the corresponding *document* networks were pooled from the same areas. Columns have the same meaning as in Table 14.

Setting	C_1	C_2	L	D	r	ϵ	lcc	γ	$\bar{R}^2$	γ_S	$\bar{R}^2$	$\gamma_{\bar{k}_{nn}}$	$\bar{R}^2$	$\gamma_{C(k)}$	$\bar{R}^2$	BC	γ_{BC}	$\bar{R}^2$	CC	DC	GC	Cp	Cp_{lcc}	Ch
F :: X			3	4		6		8	9			12		14		16							23	
F :: Y	1		3	4	5	6		8	9	10		12	13	14	15	16	17	18	19	20		22	23	24
G :: X		2	3	4		6		8				12		14			17	18				22		
G :: Y							7	8	9				13				17	18	19	20	21		23	
$\sum$	1	1	3	3	1	3	1	4	3	1	0	3	2	3	1	2	3	3	2	2	1	2	3	1

Table 17: Optimal subsets of topological features in structure-based classifications of networks regarding their membership to functional classes.