Networked Knowledge:

Approaches to Analyzing Dynamic Networks of Knowledge in Wikis for Mass Collaboration

Dissertation

der Mathematisch-Naturwissenschaftlichen Fakultät

der Eberhard Karls Universität Tübingen

zur Erlangung des Grades eines

Doktors der Naturwissenschaften

(Dr. rer. nat.)

vorgelegt von

Diplom-Psychologe, Diplom-Kaufmann Iassen Halatchliyski

aus Sofia / Bulgarien

Tübingen

2014

Gedruckt mit Genehmigung der Mathematisch-Naturwissenschaftlichen Fakultät der Eberhard Karls Universität Tübingen.

Tag der mündlichen Qualifikation:	01.07.2015
Dekan:	Prof. Dr. Wolfgang Rosenstiel
1. Berichterstatterin:	Prof. Dr. Ulrike Cress
2. Berichterstatter:	Prof. Dr. Daniel D. Suthers

На моите родители — Саид и Христина Халачлийски

(To my parents)

Acknowledgements

It has been a privilege to conduct my research at the Knowledge Construction Lab of the Knowledge Media Research Center in Tübingen. With a proud eye at the result, I want to thank all the people from the institute who supported me during the years of toil in countless ways. To name just a few, Dr. Ulrike Cress has always been my sympathetic, patient and encouraging advisor. Manfred Knobloch kindly helped me with the data preparation.

The community of researchers in the learning sciences provided the grounds and inspiration for my current work. I thank them all for this and again list just a few of them, Dr. Jan van Aalst and Dr. Heinz Ulrich Hoppe, my outstanding collaborators.

This Ph.D. thesis was not just part of the lab work. It was an integral part of my life in the past years. Many people closely accompanied and fueled my striving. I am deeply grateful to my parents and my brother Deyan Halachliyski, and to my heart Pamela Gomez. I thank Boyko Amarov, Nikola Alexiev, Ulrike Hartmann, Tamar Fazii, Andree Strauß, Eva Schloter, Roman Klinger, Katharina Licht, Ognyan Petkov. The housing project Schellingstraße in Tübingen deserves particular mention.

I also thank the Wikipedia community for creating and maintaining this wonderful source of knowledge.

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

Background

Information age is a widely acknowledged description of our present time (Bell, 1973; Castells, 2011). It stems from the fast development and diffusion of digital technology in people's daily lives (cf. Negroponte, 1995). Knowledge has become the primary capital for full-value participation in the contemporary knowledge society (Bereiter, 2002; Stehr, 2001). Moreover, the speed of this development is accelerating, as new trends and innovations in computer technology appear on a daily basis.

The latest widespread adoption of social software marked a revolutionary turn. Web users used to be passive and independent consumers of information. While this practice still exists, almost each user of the Web 2.0 has now also become an active participant in the network, sharing own feelings, thoughts and knowledge (Jenkins, Clinton, Purushotma, Robinson, & Weigel, 2006). On the one side, these private contributions are intertwined in a dense web of interactions among the users. On the other side, openly accessible knowledge is now actively created by peer users (Kolbitsch & Maurer, 2006; Leuf, & Cunningham, 2001), who often do not have qualified expertise. These conditions closely correspond with social constructivist principles of learning (Palincsar, 1998) that are commonly accepted in educational science nowadays. Compared to schools, the informal context of social media is free from performance assessment and promotes the intrinsic motivation of the participants (Hung, Lim, Chen, & Koh, 2008). Full integration of formal and informal learning might be difficult, but the ways of learning in society have already experienced a considerable impact by social media (Barron, 2006; Dohn, 2009; Ravenscroft, 2009; Richardson, 2010).

Knowledge sharing could be hindered due to anonymity, lack of incentives or information exchange dilemmas (Cress, Barquero, Schwan, & Hesse, 2007). In spite of this, social media is obviously successful in attracting great numbers of people with different backgrounds and goals (Wasko & Faraj, 2005) to interact with each other about mutual interests. Although these interactions often seem to be transient and ad-hoc, they lead to the emergence of aggregate phenomena of self-organization that may span over prolonged periods of time such as mass collaboration (Cress, 2013; Tapscott & Williams, 2006), social movements (Gerbaudo, 2012), folksonomies (Mathes, 2004) and others. Networked knowledge, that is, interconnected information collectively created online, can be of almost scientific quality even under conditions of uncertain and inconsistent information (Giles, 2005; Oeberst,

Halatchliyski, Kimmerle, & Cress, 2014). The private contributions and interactions among the users are intertwined in a dense web of hyperlink references.

In this dissertation I address in depth a type of mass collaboration that directly relates to the development of networked knowledge. Under mass collaboration I mean the joint online activities of a multitude of different people who may or may not know each other personally. As there is no universal definition of *networked knowledge*, for the current purposes I will use the term to describe meaningfully and chronologically interrelated information that stems from different individuals. My work advances a new structural approach to networked knowledge emerging at a large scale within an online community as a complex system. Using a network analysis approach I identify structurally significant artifacts presenting ideas or topics and call them *pivotal knowledge*. My empirical studies relate the pivotal knowledge to the contributions of participants with specific roles in the community. I also use an established generative mechanism of network evolution to model the significance of pivotal knowledge structures for the dynamics of new developing knowledge. In an explorative account of knowledge dynamics, I further demonstrate a fine-grained approach to pivotal knowledge and pivotal contributions of authors by acknowledging the historical trajectory of development. From methodological point of view, this dissertation contributes to the emerging research field of learning analytics.

In the following sections of this chapter, I present the theoretical framework of my work and then give an outline of the empirical studies reported in the following chapters of this dissertation.

Theoretical foundation of networked knowledge

My research is identified with the interdisciplinary learning sciences and the recent social perspectives in cognitive psychology (cf. Smith & Semin, 2004; Thompson & Fine, 1999). In their explanation of the nature of knowledge and learning, these theories renounce the extreme mentalist focus on information processing *within* individuals. Intersubjectivity is highlighted instead as a phenomenon emerging in the interaction *between* individuals (Bonk & Cunningham, 1998; Suthers, 2006). Cognition is seen as situated in a sociocultural context (cf. Brown, Collins, & Duguid, 1989). Therefore, knowledge is not transferable like information but is actively constructed and shaped in its situated use together with other

people (Clancey, 2009; Vygotsky, 1978). Learning and knowledge are thus not regarded as private properties of individuals but as contextualized and continuous social interaction, a joint meaning-making discourse (Stahl, Koschmann, & Suthers, 2006).

Research in the field of computer-supported collaborative learning (CSCL) tackles the question how technology can enhance this process in which individual learning is coupled with collaborative knowledge building (cf. Scardamalia & Bereiter, 1994; Suthers, 2012). As cooperative learning in groups was shown to have beneficial outcomes (cf. Johnson & Johnson 1989; Slavin 1995), extensive research was directed at studying the boundary conditions and techniques for successful classroom collaboration (Cohen, 1994). The sequential collaborative interaction of small groups was examined in order to understand group cognition (Stahl, 2006).

The theory of situated learning (Lave & Wenger, 1991) draws attention to the long-term socialization of people participating in communities of practice (Wenger, 1998). Members gather experience by collaborating with other members and by using culturally established artifacts that capture the collective knowledge of a community. This view on learning is called a *participation metaphor* in contrast to the *acquisition metaphor* of learning of concepts and information (Sfard, 1998). Shared material and conceptual artifacts enable an iterative process of networked knowledge development over sustained periods of time. They mediate the interaction of participants, and their creation and transformation may also be the deliberate goal of collaboration. Thus, a third metaphor stresses the creation of knowledge (Paavola, Lipponen, & Hakkarainen, 2004) and incorporates artifacts with the dialogical understanding of collaboration between individuals into a trialogical interaction model (Paavola & Hakkarainen, 2005). The cultural-historical activity theory (Cole & Engeström, 1993) frames the broad context of cultural and historical development of artifacts and communities as an activity system. The cyclical process of change of the system is regarded as learning (Engeström, 2001). In adaptable learning systems, cognition can be seen as distributed among individuals as well as physical and symbolic artifacts (Hutchins, 1995). The actor-network theory (Latour, 1987) even ascribes equal importance to humans and non-human entities for the emergence of knowledge in a dynamic complex network.

Large-scale knowledge practices on the Internet have opened a whole new field of questions for research in CSCL (Stahl et al. 2006). Consequently, a systemic view has been adapted to the mass collaboration mediated by shared digital artifacts in Web 2.0. Online environments

such as wikis and folksonomies can be seen as social systems that are independent from the cognitive systems of their users (Kimmerle, Cress, & Held, 2010a). Both systems cross-fertilize each other in such a way that both the individual and the networked knowledge co-evolve (Cress & Kimmerle, 2008). An optimal condition for this process is a moderate level of incongruity between them. The knowledge building theory (Scardamalia & Bereiter, 1994, 2006) is another related approach, which is based on Popper's (1968) philosophical view on the gradual improvement of scientific knowledge. It illustrates how communities advance their collective knowledge by developing written conceptual artifacts (Bereiter, 2002) in a digital environment over a sustained period of time. All active participants take over collective responsibility (Scardamalia, 2002) for reaching deeper insight into the domain of interest of the community by sharing, discussing and build on each other's ideas.

Epiphenomena of mass collaboration around digital artifacts

The social web affords large-scale interaction dynamics among very heterogeneous masses of individuals. Direct interaction between all the participants is not feasible and is not a prerequisite for mass collaboration. Intersubjective understanding and coordinated activities are enabled through the use of shared digital workspaces. By creating artifacts in these workspaces, people externalize their heterogeneous knowledge and make it available to each other. Depending on the specific technological affordances for manipulation, the ideas expressed in artifacts can be revised, remixed, referred to and developed further in a collaborative process. Co-created artifacts can coordinate a long-term collaborative process (Paavola & Hakkarainen, 2009) with many different people who may anonymously work in parallel. This mechanism of mediated interaction is also referred to as *stigmergy*, where the artifacts created or modified by some individuals stimulate the subsequent activity of other individuals (Susi & Ziemke, 2001). It greatly amplifies the amount of interactions and leads to the emergence of epiphenomena.

Over time, decentralized virtual communities (Wellman & Gulia, 1999) are formed and set overarching goals and norms guiding the creative efforts. These shared social practices and rules are epiphenomena of the interaction among the participants (Engeström & Sannino, 2010). The created artifacts organized together as a digital knowledge base of interlinked contributions represent the networked knowledge of a community (cf. Bruckman, 2006). This is also an emergent (Theiner, Allen, & Goldstone, 2010) product of the wisdom of the crowds

(Surowiecki, 2005) or the collective intelligence (Levy, 1999) of the community as a whole. Although it develops on the basis of the activity of individuals, it is more than a collection of their individual ideas. Each single contribution needs to be adequately integrated into the existing networked knowledge. New knowledge for the community arises, as new concepts, connections and ideas are introduced to the knowledge base. It may not necessarily be scientifically new, just like the work of a knowledge-building school class (cf. Scardamalia & Bereiter, 1994, 2006). In a continuous development process of convergent and divergent contributions (Halatchliyski, Kimmerle, & Cress, 2011) over time, some ideas codified in the artifacts of a knowledge base may stand the test of time and become more prominent than others that fade away. Thus, mass collaboration goes along with development and improvement of ideas and artifacts according to goals and rules that emerge through self-organization in a community.

A complex system perspective on mass collaboration

Mass collaboration can be investigated considering – at a micro level – the mutual interactions of a large numbers of people and artifacts and – at a macro level – the self-organized, knowledge-building, online community. Given the difficulty to grasp the dynamic patterns of interplay of all relevant aspects of mass collaboration, a complex system perspective (Luhmann, 1984; Oeberst et al., 2014; von Foerster, 2003) provides a suitable framework for the research presented in this dissertation.

A complex system consists of locally interacting elements with heterogeneous characteristics and behavior. Over time, patterns of collective self-organization such as norms, division of roles, and other coordination mechanisms emerge on a large scale out of the local interaction of the elements and grow in sophistication (Kapur et al., 2007). Thus, from the contribution and discussion of ideas at the bottom level emerge structured knowledge and goal-oriented organization at the top level. These patterns then have a top-down impact on the local relations and interactions of participants and artifacts. A knowledge-related system autopoietically maintains a code of operation (Maturana & Varela, 1987) that consists of criteria for evaluating participative activities and for integrating or rejecting contributions. Thus, it directs the individual behavior and defines the acceptable knowledge. Existing knowledge controls the subsequent integration of new knowledge. Communities develop in

this way their own socially constructed and interpretative view on reality (cf. Berger & Luckmann, 1966; Kimmerle et al., 2013; von Glasersfeld, 1995).

Knowledge-related systems such as the scientific community demonstrate dialectics between structural patterns and dynamic processes (Lucio-Arias & Leydesdorff, 2009). *Static structures* arise from the tension between variation and selection of the elements such as scientists, publications and institutions. *Temporal dynamics* is created by forces of change and stabilization that operate over the course of history. Structure and dynamics can be identified at different levels of a system. In science, for example, researchers collaborate with each other, publish their work in a written form and build on each other's work by citing existing papers. Lucio-Arias and Leydesdorff (2009) also identified a second-order dynamics referring to scientific ideas that have a life on their own as part of a scientific discourse once they are published (cf. Bereiter, 2002; Popper, 1972). As scientists select their specific research questions, methods and the previous works to build on, global structural patterns of knowledge development emerge and stabilize over time. Thus, ideas may form a paradigm (Kuhn, 1962) that then again exerts top-down selection on the behavior of scientists. A paradigm represents a structure that is reified through the publication of consistent scientific work over time. Eventually, spontaneous breakthroughs, contradicting evidence and stabilization of alternative views may introduce a bottom-up change in the structure of science.

Network analysis approach to mass collaboration systems

The generative processes, conditions and patterns of development of networked knowledge and learning at the level of a community (Nonaka & Nishiguchi, 2000) can be appropriately investigated using a network approach. Mass collaboration thus implies the emergence of knowledge networks (Saviotti, 2009) in the context of online social networks (Lipponen 2002; Ryberg & Larsen 2008). A network is an abstract structure with certain patterns which consist of different sets of nodes such as individuals, artifacts and of their links. The concept has already been used to describe knowledge organization at different levels such as the semantic memory of individuals (e.g., Collins & Loftus, 1975), the interrelated ideas in a scientific community represented in papers citing each other (Garfield, 1972; Latour & Woolgar, 1979), or the Wikipedia knowledge base of interlinked artifacts (Voss, 2005). Networked knowledge essentially emerges from the specific semantic interconnections

between knowledge artifacts such as topical relations, problem-solution chains, discourses, etc. This structural approach (cf. Wellman, 1997) also allows dynamic analysis, as both the nodes and connections in a network are constantly changing.

A “new science of networks” (Barabási, 2002) unites research on networks from physical, biological, social and computer science offering a variety of tools and methods to measure, describe and visualize global network properties as well as relative positions of single nodes. Social network analysis (SNA; Wassermann & Faust, 1994) is increasingly adopted in CSCL research (e.g. Aviv, Erlich, Ravid, & Geva 2003; Cho, Stefanone, & Gay, 2002; de Laat, Lally, Lipponen, & Simons, 2007; Reffay, & Chanier, 2002) for analyzing log data on interactions among collaborating students. Bibliometric research (Glänzel, 2003) often applies network analysis techniques to networks of scientific papers that cite each other. Webometrics (Almind & Ingwersen, 1997; Björneborn & Ingwersen, 2004) adapts appropriate methods following a direct analogy between the analysis of scientific citations and of hyperlinks between webpages. Thus, network analysis methods can be used to meet the complexity introduced by the interaction of many network nodes in a knowledge creating system.

The network science has only lately started to expand the limited focus on measuring static structures in order to acknowledge the dynamics of complex networks. Temporal analyses are usually only descriptive and consider differences between network snapshots at particular moments in time (Mali, Kronegger, Doreian, & Ferligoj, 2012). During online mass collaboration, new networked knowledge is sequentially built upon the existing knowledge in an essentially temporal process. Aggregation across time based on coding and counting of events easily leads to a biased analysis of individual and community-level variables. Correspondingly, there is a strong need for temporal analysis methods in the learning sciences (Reimann, 2009; Mercer, 2008). Due to the analogy between scientific and online knowledge-building communities, established analytical approaches can be borrowed from bibliometrics and scientometrics. These research fields offer a variety of methods tailored for the quantitative analysis of knowledge artifacts, scientific work, and their authors. They can greatly enrich the newly emerging research in learning analytics (Siemens, 2012; Suthers & Verbert, 2013). One such method is the main path analysis (Hummon & Doreian, 1989) that examines temporally developing knowledge flows and uptakes (Suthers, 2006) in knowledge networks. It takes into account the structure of connections between artifacts together with the temporal order of development and has been applied to scientific citation networks and to

knowledge-building discourse in schools (Halatchliyski, Oeberst, Bientzle, Bokhorst, & van Aalst, 2012).

Examples of mass collaboration in Web 2.0

Among the Web 2.0 technologies, wikis are especially suitable for knowledge-building by enabling myriads of users to work in parallel, forming a community and co-creating a knowledge base of shared digital artifacts (Forte & Bruckman, 2006) as in the case of Wikipedia and Wikiversity, two mass collaboration projects of the Wikimedia Foundation. The mass collaboration process is open-ended, and the collective knowledge is constantly changing, as new articles are created and content is added or deleted. The participants also benefit in this process (Moskaliuk, Kimmerle, & Cress, 2009, 2012), so wikis can be used to support individual learning even in formal educational contexts (Konieczny, 2007). Open wikis like Wikipedia and Wikiversity are also suitable for research, as they provide the entire development history of the collective artifacts in which different opinions are integrated and conflicts are argued out. These wikis are tools for generating, connecting and revising networked knowledge rather than disseminating information (Purdy, 2009). Indeed, Wikipedia is not aimed at developing new knowledge, and the information added to it must not be novel according to its own "no original research" rule. Nevertheless, the externally sourced information is integrated in an original way (cf. Swarts, 2009) and presents a new product of emerging networked knowledge. Thus, Wikipedia's knowledge base is a novel product of the community and involves development processes that are typical of genuine knowledge-building communities (Cress & Kimmerle, 2008; Forte & Bruckman, 2006). Wikiversity is understood by its active members as an "open learning community" in which users can actively produce learning resources for a broad range of topics and thus learn while they participate.

Networked knowledge develops on many levels in wikis: article content is edited by adding, modifying or deleting parts of it and thus changing its textual structure; hyperlinks are extensively used to establish connections between articles; new articles are constantly created building up entire knowledge domains as well as connecting different domains. System rules and community practices are the backbone for such developments guiding individual activities and regulating the collaborative process (Niederer & van Dijck, 2010). They ensure the achievement of coherence and consensus from the diversity of views offered by the

participants. High-quality articles in Wikipedia (Wöhner & Peters, 2009) can thus be created by experienced participants in the community who lack a domain-specific expertise (Oeberst et al., 2014). Contributions that are not accordant with the rules are reverted and thus refused by the system. Vandalism in Wikipedia, for example, is fixed very fast (Viégas, Wattenberg, & Dave, 2004). These rules, their interpretation and application are subject to change over time through social negotiation too (Forte & Bruckman, 2008).

In sum, Wikipedia and Wikiversity are multifaceted wiki environments for mass collaboration around networked digital artifacts. They offer a unique field for studying the statics and dynamics of networks of emerging knowledge from the activity of contributors in a community that represents a complex system.

Overview of the dissertation

In the light of the foregoing, social media has a high practical relevance for the development of networked knowledge in contemporary society. Based on the theoretical grounding from the interdisciplinary learning sciences, the present dissertation will advance an approach for studying and understanding the principles that underlie the development of networked knowledge during online mass collaboration. I will measure networked knowledge by focusing on artifacts co-created in a community of learners. As they are both means and ends of collaboration (Dohn, 2009; Lipponen, Hakkarainen & Paavola, 2004), they are fundamental in the large-scale and long-term, stigmergic process. Networked knowledge is an epiphenomenon emerging in a complex system. Therefore, it can be appropriately studied by a network approach that acknowledges both its macro level structure and the micro level of single artifact relations and contributions by participants. Employing network analysis techniques, I will quantitatively model and evaluate real-life data from Wikipedia and Wikiversity in order to make statistical inferences. My studies will present test of hypotheses on causal relationships between static structures of pivotal knowledge, contribution activities of different groups of participants and dynamic processes of knowledge development over time. I will also include an explorative investigation of the multifaceted character of networked knowledge emerging through mass collaboration.

Chapter 2 discusses the relevance of large-scale mass collaboration for computer-supported collaborative learning (CSCL) research, adhering to a theoretical perspective that views

collective knowledge both as substance and as participatory activity. In an empirical study, using the German Wikipedia as a data source, I explore collective knowledge as manifested in the structure of artifacts created through the collaborative activity of authors with different levels of contribution experience. Wikipedia's interconnected articles are considered at the macro level as a network and analyzed using a network analysis approach. The focus of this investigation is the relation between the authors' experience and their contribution to two types of articles: *central pivotal articles* within the artifact network of a single knowledge domain and *boundary-crossing pivotal articles* within the artifact network of two adjacent knowledge domains. Both types of pivotal articles are identified by measuring the network position of artifacts based on network analysis indices of topological centrality. The results show that authors with specialized contribution experience in one domain predominantly contributed to central pivotal articles within that domain. Authors with generalized contribution experience in two domains predominantly contribute to boundary-crossing pivotal articles across the knowledge domains. Moreover, article experience (i.e., the number of articles in both domains an author had contributed to) is positively related to the contribution to both types of pivotal articles, regardless of whether an author had specialized or generalized domain experience. I discuss the implications of these findings for future studies in the field of CSCL.

In *Chapter 3* I followed a longitudinal network analysis approach to investigate the structural development of the knowledge base of Wikipedia and to explain the appearance of new knowledge. Building on the study in *Chapter 2*, the data consists of the articles and authors in same two adjacent knowledge domains. I analyze the development of networks of hyperlinked articles at seven snapshots from 2006 to 2012 with an interval of one year between them. Longitudinal data on the topological position of each article in the networks is used to model the appearance of new knowledge over time. Thus, the structural dimension of knowledge is related to its dynamics. Using multilevel modeling as well as eigenvector and betweenness measures, I explain the significance of pivotal articles that are central within one of the knowledge domains or boundary-crossing across both domains at a given point in time for the future development of new knowledge in the knowledge base.

Chapter 4 introduces the scientometric method of main path analysis and its explorative application in an exemplary study of the paths of knowledge development and the roles of contributors in Wikiversity. The study is a step forward in adopting and adapting network analysis techniques for analyzing collaboration processes in knowledge building

communities. Again, data from two scientific domains in an online learning community is used. By identifying a specific type of networks called *directed acyclic graphs*, the meaningfully interconnected knowledge artifacts are analyzed in consideration of their temporal sequence of development. Based on a fine-grained historical account of network dynamics, global coherence as well as pivotal moments of collaboration in the different domains are examined. A schema for the visualization of the results is introduced. The potential of the method is elaborated for the evaluation of the overall learning process in different domains as well as for the individual contributions of the participants. Different outstanding roles of contributors in Wikiversity are presented and discussed.

As a concluding part of this dissertation, *Chapter 5* summarizes the results of the three preceding empirical studies. It presents an integrative review of the theoretical and methodological strengths and limitations of the current research approach. Finally, implications for future research and practice are discussed.

It should be noted that the empirical *Chapters 2, 3* and *4* were written for independent publication in scientific journals. Beyond the explained methods and results, they also contain individual theoretical and discussion sections. As integrative parts of a dissertation project, some overlap in the presentation of the studies was unavoidable. The following list contains the publications in the order of the chapters in this dissertation.

Journal Articles and Submitted Manuscripts

- *Chapter 2* is based on: Halatchliyski, I., Moskaliuk, J., Kimmerle, J., & Cress, U. (2014). Explaining authors' contribution to pivotal artifacts during mass collaboration in the Wikipedia's knowledge base. *International Journal of Computer-Supported Collaborative Learning, 9*, 97-115.
- *Chapter 3* is based on: Halatchliyski, I., & Cress, U. (2014). How structure shapes dynamics: Knowledge development in Wikipedia - a network multilevel modeling approach. *PLoS ONE, 9*, e111958.
- *Chapter 4* is based on: Hatchliyski, I., Hecking, T., Göhnert, T., & Hoppe, H. U. (2014). Analyzing the main paths of knowledge evolution and contributor roles in an open learning community. *Journal of Learning Analytics, 1*, 72-93.

Chapter 2

Contribution to Pivotal Artifacts in Wikipedia

This chapter is based on:

Halatchliyski, I., Moskaliuk, J., Kimmerle, J., & Cress, U. (2014). Explaining authors' contribution to pivotal artifacts during mass collaboration in the Wikipedia's knowledge base. *International Journal of Computer-Supported Collaborative Learning, 9*, 97-115.

Introduction

Computers facilitate connectivity and coordination among large networks of people (Lipponen, 2002; Ryberg & Larsen, 2008) and enable them to form communities and build digital knowledge bases. Recently, Web 2.0 environments have greatly lowered the barriers to participative activities for all Internet users (Kapur et al., 2007). As a result, so-called *mass collaboration* has become a common phenomenon (Cress, 2013; Cress et al., 2013; Tapscott & Williams, 2006). With its specific affordances for knowledge-related activities (Lipponen, 2002; Pifarré & Kleine Staarman, 2011), mass collaboration presents a whole new field of study in computer-supported collaborative learning (CSCL; Scheuer, Loll, Pinkwart, & McLaren, 2010). Its essence resides not only in new technologies and enhanced connectivity but also in the fact that openly accessible knowledge is now increasingly shared by the masses of learners themselves. Large groups of participants interact from different places and different points in time via a shared virtual space, and their interaction revolves around the creation of shared artifacts. These artifacts often represent a digital knowledge base with a network structure. Direct social interaction for reaching common understanding is largely infeasible under these circumstances (Larusson & Alterman, 2009).

Mass collaboration bears three implications for CSCL research regarding collective knowledge bases: (1) The focus should incorporate the interplay between knowledge as substance (i.e., artifacts with meaningful content and interrelations) *and* as participatory activity (i.e., interactive contribution processes). (2) A knowledge base must be studied at the macro level, as it emerges in self-organized, long-term, interactive processes distributed across a large number of people. (3) The network perspective provides a multifaceted methodological approach to a knowledge base as a network of artifacts.

In the study reported here, I used data from the German version of the online encyclopedia Wikipedia, an outstanding example of artifact-based mass collaboration on the Web, to explore a collaboratively created knowledge base (for an extensive review of the large body of publications on the subject, see Okoli, Mehdi, Mesgari, Nielsen, & Lanamäki, 2012). It is a dynamic complex system of interconnected articles deliberately co-produced and modified by collaborative activities. With its large amount of data on the history of articles and authors' contributions, it offers a unique field for studying large-scale, open-ended collaborative processes. The contributions of two authors to the same article may take place years and hundreds of other authors' contributions apart. So—although authors often coordinate their

work over article talk pages, that is, discussion threads, and over numerous other channels (Pentzold & Seidenglanz, 2006)—a substantial part of the work is coordinated through the dynamically changing article itself. The written content mediates shared understanding on a specific topic, amalgamating views and styles of expression of a multitude of authors into a coherent exposition.

Although Wikipedia is not aimed at "inventing" new knowledge, or at providing a learning environment for the contributors, the processes that unfold in the online encyclopedia have been found to be essentially equivalent to scientific progress and knowledge-building discourse (Cress & Kimmerle, 2008; Forte & Bruckman, 2006; Kimmerle, Moskaliuk, Cress, & Thiel, 2011b; Swarts, 2009). The choice and argumentative composition of facts and citations from external sources produce an original knowledge artifact with every article. Obviously, Wikipedia is not just a trivial aggregation of external information, and its articles represent more than just links to the original sources. From the perspective of CSCL research, the complex knowledge-related collaborative activities on Wikipedia are interesting along with the developing knowledge base of mediating artifacts (Cress & Kimmerle, 2008; Halatchliyski et al., 2011), which is a novel product for the online community and the general public, irrespective of Wikipedia's "no original research" policy[1].

In order to tackle the large-scale dimensions, I employ the concept of a network and the approach of network analysis to the set of interconnected artifacts in two adjacent knowledge domains. My goal is to exemplify the application of network analysis to the structure of a knowledge base of an online community and relate it to the contribution activity of its authors. Focusing on an article's topological position in the artifact network, I differentiate between two types of pivotal articles, that is, articles that are important for the structure of the knowledge base. An article may be pivotal either in the sense of being *central* within a knowledge domain or in the sense of being *boundary-crossing* across two domains. I examine to what extent different types of editing experience within the knowledge base are important explanatory variables for the contribution to pivotal articles (see Halatchliyski, Moskaliuk, Kimmerle, & Cress, 2010; Sosa, 2011).

In the following I briefly recap theory trends in the field of CSCL, integrating views on both collective knowledge as substance and as participatory activity. Based on this theoretical foundation, I discuss the opportunities and challenges of studying collective knowledge in the

[1] http://en.wikipedia.org/wiki/Wikipedia:No_original_research

context of the recent phenomena of mass collaboration and knowledge base networks. I then introduce my research approach based on network analysis metrics, in order to deal with these new challenges in CSCL research. Subsequently, I provide findings from an empirical study on pivotal articles and their contributors, within the artifact network of two adjacent knowledge domains in Wikipedia. Finally, I discuss the implications of my findings for future CSCL research.

Perspectives on collective knowledge

Theories on collective processes of intersubjective meaning-making (Dillenbourg, Baker, Blaye, & O'Malley, 1996; Koschmann, 2002) have left behind individual cognition in order to focus on participation in community practices, negotiation of meanings, and building of shared understanding. Following the so-called *participation metaphor* (Sfard, 1998), learning and knowing are depicted as socially shared activities that cross the conceptual boundary from one to the other (see also Scardamalia & Bereiter, 1994). Knowing then consists of people's activities and practices that correspond with the specific physical and social context of a situation (Lave, 1988; Suchman, 1987). Accordingly, collaborative learning and knowing have been placed at the level of *group cognition* by Stahl (2006), emphasizing that they cannot be reduced to the level of cognitive representations and discussion statements of single individuals (see also Koschmann, 2002).

Stahl's (2006) model integrates these levels of individual learning and collective knowledge into an activity system consisting of artifacts, utterances, and interactions as focal points. The sequence of referencing and defining interactions of the individual participants in a particular context continually produces and modifies a network of shared interconnected meanings for the group. Meaning is grounded in the relative positions in this network of mutual references and is not statically attached to physical artifacts or even words. Nevertheless, the meaning-making process is supported by the use of artifacts and words, which have predefined meanings from past discourse activities and which may again become subject to recurrent negotiation by the group participants. Thus, collaboration involves participative interaction along with the creation and reuse of meaningful artifacts, which may often have a physical representation, or may be the focus of collaboration, as argued in the next section. Collective knowledge should then be defined not only as activity (i.e., knowing), but also as substance

(i.e., shared artifacts). Knowledge as substance generally manifests itself in the emergent pattern of shared interconnected meanings. Analogously, Wikipedia consists of the collaborative activities and practices of its authors *and* of the networked knowledge base, which can be thought of as snapshots of meaningful structure in constant development. The network structure of referenced artifacts can also be studied with attention to its dynamics.

Both participant interaction and the dynamics of developing artifact shape and content are complementary aspects of the meaning-making process (Hakkarainen, Ritella, & Seitamaa-Hakkarainen, 2011; Paavola & Hakkarainen, 2009). Two research endeavors explicitly acknowledge the relevance of collaborative creation, use, and transformation of artifacts as epistemic objects (see also Knorr-Cetina, 2001) in CSCL: The *metaphor of knowledge creation* (Paavola, Lipponen, & Hakkarainen, 2002, 2004) designates artifacts as the goal and the product of collaborative learning. The *co-evolution model of cognitive and social systems* (Cress & Kimmerle, 2008) shows how collective knowledge develops with the changing artifact content in the context of a wiki (see also Kimmerle, Moskaliuk, & Cress, 2011a; Moskaliuk et al., 2009, 2012). The development is presented as successive co-evolution cycles of internalization (i.e., individual learning) and externalization (i.e., creation of collective knowledge; Kimmerle et al., 2010a).

In the present chapter, my aim is to advance the perspective that—in contrast to the analysis of interaction sequences—artifacts and their meaningful interconnected structure offer a unique way of operationalizing knowledge-related processes in collectives. Maintaining the research focus at the intersubjective level, I extend the concept of collective knowledge to long-term processes and large-scale network structures.

Artifact-based mass collaboration

In line with the participation metaphor of situated learning and knowing, the predominant methodological approach in CSCL has been to study small groups of students in a neatly arranged situation: The students engage in synchronous discourse around a problem-solving task, and the sequence of their interactions represents a major research interest. Lipponen (2002), however, contested the popular definition of collaboration as “a coordinated, synchronous activity that is the result of a continued attempt to construct and maintain a shared conception of a problem” (Roschelle & Teasley, 1995, p. 70), because it puts narrow

constraints on the object of study. Suthers (2006) also stated that small groups do not deliver an exhaustive picture of collective knowledge processes. Jones, Dirckinck-Holmfeld, and Lindstrom (2006) argued for broadening the research focus on collaborative learning to include aspects of networked learning enabled by large-scale technological infrastructures on the Web. In fact, complex knowledge phenomena involve longer periods of time, larger and changing numbers of people, and fuzzy-structured settings (Kapur et al., 2007). In this spirit, any human achievement can be seen as a collaborative accomplishment—in terms of the metaphorical *dwarfs standing on the shoulders of giants*. Extending the view beyond problem-solving small groups enables a macro approach to the complexity of knowledge development across space, time, and collectives of people. This global level of human learning and knowledge creation has rarely been addressed by CSCL research (see Kafai & Peppler, 2011).

This large-scale perspective brings to the foreground the connecting role of artifacts in the collaborative process. Bearing in mind that most of the individuals among a vast number of participants cannot interact directly or do not even know each other, intersubjective understanding and coordinated activities are facilitated by artifacts. This is even more so when the individuals follow a common goal, as in the case of Wikipedia. Each individual must take account of the perspective of the others to contribute by building on the accomplishments of others. Collaborative artifacts represent crystallized knowledge that is preserved from past interactive situations, and that can be built on in the future, giving rise to phenomena like scientific understanding, social practices, and rules. This mechanism of indirect interaction is also referred to as *stigmergy*, where the artifacts created or modified by some individuals stimulate the subsequent activity of other individuals (Susi & Ziemke, 2001). Knowledge-related practices in Web 2.0 contexts fall under the participation metaphor of learning and additionally accentuate the creation of knowledge artifacts (Dohn, 2009). This view suggests integrating the two perspectives of artifacts as both means and ends of collaboration (see e.g., Kafai & Resnick, 2000) and also suggests studying the interplay between knowledge as substance and as participatory activity. In sum, artifact-based mass collaboration develops as a self-organized process around and through the created content, which reduces the need for direct coordinating interactions between the participants.

Networks of knowledge

The study of a mass collaboration knowledge base presupposes an approach that can encompass its macro structure and other large-scale and long-term characteristics. At the same time, it is desirable not to leave the level of artifacts, individuals, and small groups out of focus. According to the actor network theory (Latour, 2005), the analysis of social phenomena, of which mass collaboration is an example, should focus on the patterns of mutual influence in the network of actants (i.e., humans as well as artifacts endowed with equal agency). The fundamentals of such a multifaceted approach are provided by the *network* concept.

A network can be defined as a set of dynamically connected nodes that represent units of the same kind, such as persons or knowledge artifacts. The concept has already been used to describe knowledge organization at different levels. The semantic memory of individuals has classically been portrayed as a network of associated knowledge representations (e.g., Collins & Loftus, 1975). Stahl (2006) has advanced the idea of networks of references to explain how collective knowledge is created through group discourse activities. In the context of mass collaboration environments like Wikipedia, knowledge is organized in a network of interlinked artifacts (Voss, 2005).

Computer technology directly promotes the creation of networked knowledge in a number of ways. The Web itself represents a technological network that maintains hyperlinked information of various kinds. Due to the flexibility of hypertext the recipient can “jump” in multiple directions through the content and combine relevant aspects from different contexts, discerning new meanings (Moskaliuk & Kimmerle, 2009). The increased interactivity afforded by Web 2.0 applications also makes network structures and user-generated content prominent. Correspondingly, an increasing number of hyper-structured knowledge bases have emerged from the collaborative activity of a mass of individuals.

The network concept suitably highlights the emergent character of knowledge. According to the theory of conceptual integration and blending (Fauconnier & Turner, 2002), the creation of new meanings and knowledge can be thought of as recombination of different existing ideas. Knowledge essentially emerges from the specific way in which various meanings are connected, like nodes in a complex network that can build an infinite number of interconnection patterns. Although the network concept connotes a structural approach, it

does not imply a static view on knowledge. Networks are constantly changing as neither their nodes nor their links are enduring entities. Large-scale collective dynamics lead to the bottom-up development of patterns typical of complex systems (Kapur et al., 2007). These patterns then have a top-down impact on the local relations and interactions among individuals and knowledge artifacts.

Based on the network concept, network analysis (see Newman, 2010) offers methodological tools to begin dealing with the complex large-scale and long-term patterns in the knowledge base of a mass collaboration environment.

The network analysis approach

Network analysis is a multidisciplinary research approach for examining relational patterns among physical and digital, human and non-human entities. It includes a variety of methodological concepts and instruments to identify, describe, analyze, and visualize positions, relations, clusters of elements, and global network properties. The approach was greatly advanced by sociologists who studied networks of people under the term *social network analysis* (SNA; Wasserman & Faust, 1994), but their concepts and methods largely represent mathematical abstractions and are applicable to other kinds of networks. Some of the major applications are: detecting important actors, subgroups, and the actors bridging them; characterizing the position of different artifacts within a network; measuring information paths and flows.

SNA has become an increasingly common method in CSCL research (e.g., Aviv et al., 2003; Cho et al., 2002; Goggins, Valetto, Mascaro, & Blincoe, 2012; de Laat et al. 2007; Kimmerle et al., 2013; Palonen & Hakkarainen, 2000; Reffay & Chanier, 2002; Ryberg & Larsen, 2008). Analyses of online social networks are usually based on the logged collaborative interaction between learners that is mediated through a shared digital environment. For example, a network link between two people may mean that the one has read or responded to a contribution of the other, but more indirect relations like the co-presence in a discussion are also possible. Such analyses may yield information on the cohesiveness of learning groups and on the position of individual students relative to the others, at different points in time and overall.

As argued in the previous sections of this chapter, in addition to the knowledge-related activities of the participants, CSCL research should also incorporate the body of collaborative knowledge artifacts into the analysis. This is especially relevant for a mass collaboration environment, such as Wikipedia, that is directed at creating a knowledge base. The patterns in such a networked body of knowledge artifacts can be appropriately studied with network analysis methods in analogy to bibliometric citation analysis of scientific work (see Glänzel, 2003). Mass collaboration manifests itself in knowledge artifacts linked by hyperlinks, similar to scientific papers connected through citations. The emerging learning analytics discipline (Fournier, Kop, & Sitla, 2011; Siemens, 2012) might be a promising field for adapting borrowed bibliometric approaches to networked learning and mass collaboration (see, for example, Halatchliyski et al., 2010).

Only a few CSCL studies have analyzed networks of collaboratively created artifacts with content relations. Both Sha, van Aalst, and Teplovs (2010) and Oshima and Oshima (2007) applied automatic algorithms for the identification of semantic relations between the content of artifacts in order to define a so-called semantic network of contributions, to calculate general indices of the network and to cluster the topics of the contributions. Kimmerle, Moskaliuk, Harrer, and Cress (2010b) investigated the development of clusters in a network of Wikipedia articles related to the topic of schizophrenia over a period of five years. They found evidence for co-evolution of the artifact network and the contribution interest of authors over time. Halatchlyiski et al. (2010) examined an article network of two adjacent knowledge domains in Wikipedia and identified a group of experienced, boundary-spanning authors who influenced domain integration. The present study extends this approach by relating the concept of pivotal artifacts in a knowledge base to the activity characteristics of the contributing authors. Keegan, Gergle, and Contractor (2012) also used authors' editing experience to study the collaboration patterns on Wikipedia articles about breaking news.

In sum, the macro perspective on knowledge networks reveals a unique and largely unexplored field within CSCL research. Correspondingly, I argue that network analysis is an appropriate methodological approach when taking this perspective. In the following sections of this chapter, I present a study with Wikipedia data in which I employ two types of measures of topological centrality to identify pivotal articles in artifact networks: the one captures well-connected artifacts that have important positions within a single knowledge domain; the other accents boundary-crossing artifacts that have an interconnecting position

across two knowledge domains. Based on these indices I examine the relation between the authors' editing experience and their contribution to pivotal articles in the knowledge base.

Empirical study

Focusing on two adjacent knowledge domains in Wikipedia, the following study seeks to explain the contribution to pivotal articles in the artifact network of a knowledge base through the editing experience of its authors. Experience in this sense does not designate some scientific or professional expertise but simply the count of an author's content contributions to the investigated knowledge domains. *Pivotal* articles were those with a *central* network position within a single knowledge domain or those with a *boundary-crossing* network position across two knowledge domains. The study includes two levels of analysis: At the level of artifacts, I perform a network analysis on hyperlinked articles, which are categorized *a priori* in two adjacent knowledge domains. I test my hypotheses at the level of authors by relating their editing experience to their contribution to articles with pivotal network positions.

Level of artifacts: The body of knowledge artifacts in a mass collaboration environment may be divided into knowledge domains. The relevant artifacts in the current study were Wikipedia articles, and a *knowledge domain* was a set of articles that had been assigned to the same Wikipedia category, corresponding to a scientific discipline. Hence, my approach bears similarities to scientometric research on the scientific work in neighboring disciplines. The Wikipedia category system is a collaboratively created taxonomy with a nearly hierarchical structure of supra- and sub-categories. Any author can change what category is assigned to an article or a sub-category, and articles are often annotated with multiple categories (Kittur, Chi, & Suh, 2009). Accordingly, article categorization is an emergent characteristic of the mass collaboration environment and reflects the logic of the represented knowledge. It is independent of the article network structure and the authorship of articles. Based on the *a priori* Wikipedia categorization, I chose two knowledge domains for my study. I then distinguished *specialized articles*, which belonged to only one of the two knowledge domains, and *intersection articles*, categorized under both knowledge domains.

Exploring the network structure of a knowledge base at the macro level of knowledge domains, I focused on identifying articles with pivotal network positions. I distinguished between pivotal articles that are central within one knowledge domain and pivotal articles that

cross the boundary across two knowledge domains. In my reasoning, both types of articles may be important, supporting on the one hand the internal knowledge organization within a domain and, on the other hand, the interdisciplinary connections across domains.

Therefore, I defined two *separate networks* that corresponded to the hyperlinked specialized and intersection articles in each of the two domains. I also defined a *combined network*, including all the articles and their hyperlinks in both domains taken together. Network nodes represented articles, and network edges represented the hyperlinks not accounting for the direction (as I aimed at examining the relatedness of the articles and not browsing behavior).

Pivotal articles within a knowledge domain were operationalized by applying the eigenvector centrality index (Bonacich, 1972) to the articles in the separate networks. This measure characterizes the connectedness of an article relative to all the others in the network: Articles with more direct connections to other well-connected articles obtain higher values. These central articles contain knowledge that is highly significant for a domain. A similar measure is employed by the PageRank algorithm of the Google search engine for ranking the importance of web pages (Page, Brin, Motwani, & Winograd, 1998).

Pivotal articles that cross the boundary between two knowledge domains were operationalized by applying the betweenness centrality index (Freeman, 1979) to all the articles in the combined network[2]. This measure characterizes the bridging position of an article among the other articles in both domains: Articles that are repeatedly part of the shortest connection between pairs of other articles obtain higher values. These boundary-crossing articles link the two domains and enable knowledge transfer and integration across their boundaries.

Level of authors: This level was not a part of the network analysis, which only pertained to the articles and their hyperlinks. Based on the history of contributions to the articles included in the first level of the analysis, I determined the relevant authors and their experience. I used two aspects of experience—*article experience* (i.e., the count of individual articles in both domains an author had contributed to) and *domain experience*. Regarding domain experience, authors were classified into groups of *specialists* (i.e., authors who contributed to one of the domains but not to the other) and *generalists* (i.e., authors who contributed to both domains). As I investigated two domains, there were also two groups of specialists. Generalists were grouped into *intersection generalists* (i.e., authors who have contributed to at least one

[2] Both centrality indices were originally developed in SNA research and also used in various other networks (see Leydesdorff, 2007).

intersection article, which appeared in both domains) and *non-intersection generalists* (i.e., authors who have contributed to specialized articles in each of the domains but to none of the intersection articles). For purposes of illustration, Table 2.1 incorporates my definitions into an example. The rows represent articles which belong either to knowledge domain A, to knowledge domain B, or to both domains A and B (intersection articles). According to the definition of domain experience, author 1 is a specialist in A, author 2 is a specialist in B, author 3 is an intersection generalist, and author 4 is a non-intersection generalist. The last row in the table shows the article experience of each of the four authors as the count of articles an author has contributed to.

Table 2.1 Grouping of authors according to contribution activity and articles' categorization.

articles' a priori categorization	**author 1: specialist in A**	**author 2: specialist in B**	**author 3: intersection generalist**	**author 4: non-intersection generalist**
A	x			
A	x			x
A	x			
A & B			x	
A & B			x	
B		x		
B		x		x
article experience of an author	3	2	2	2

At the level of authors, I determined the relation between authors' experience and their contribution to pivotal articles by building on the measures of pivotal articles in the networks. I calculated author-level aggregate measures of the average centralities—once for eigenvector centrality and once for betweenness centrality—of the articles an author had contributed to. So, an author inherited the averaged centrality of the articles she or he had co-authored. I did this for the combined network as well as for each of the two separate networks independently. The *important* authors within a knowledge domain are those that have the highest aggregated eigenvector centrality, based on the articles they have contributed to. Correspondingly, the *boundary spanners* (Tushman & Scanlan, 1981) across two knowledge domains are those

authors that have the highest aggregated betweenness centrality based on the articles they have contributed to. They act as gatekeepers at the boundary between two knowledge domains, driving knowledge exchange and integration.

Hypotheses

The goal of the study was to simultaneously investigate the partial effect of authors' *article experience* and their *domain experience* as explanatory variables of their contribution to pivotal articles within and across knowledge domains. My hypotheses therefore concerned the author level of analysis.

While boundary-spanning contributors might not necessarily have a prominent role within the domains, by definition they should be experienced in both domains (Levina & Vaast, 2005). Consequently, I derive the following hypotheses:

Hypothesis 1a: Specialists contribute on average to more central, better-connected articles in each of the knowledge domains than generalists. Thus, specialists have a high aggregated *eigenvector* centrality derived from the *separate* domain networks compared with generalists.

Hypothesis 1b: Generalists act as boundary spanners and contribute on average to more boundary-crossing articles across both domains than specialists. So, generalists have a high aggregated *betweenness* centrality derived from the *combined* domain network compared with specialists.

Besides domain experience, I expect that authors' article experience (i.e., the count of articles in both domains an author has contributed to) is also a significant explanatory variable of the contribution to pivotal articles. According to the concept of legitimate peripheral participation (Lave & Wenger, 1991), experienced authors are expected to have a significant influence in a mass collaboration environment by contributing to pivotal articles within and across knowledge domains:

Hypothesis 2a: Authors' article experience is a significant predictor of the contribution to central, well-connected articles, so it is positively related to the aggregated *eigenvector* centrality of authors derived from the *separate* network of each of the knowledge domains.

Hypothesis 2b: Authors' article experience is a significant predictor of the contribution to boundary-crossing articles, so it is positively related to the aggregated *betweenness* centrality of authors derived from the *combined* network of both knowledge domains.

In order to estimate the partial effects of article and domain experience, Hypothesis 1a and Hypothesis 1b were simultaneously tested with one model for each of the two knowledge domains. Accordingly, Hypothesis 2a and Hypothesis 2b were simultaneously tested with one model for both domains taken together.

Data and method

I studied the contribution to central pivotal articles within and boundary-crossing pivotal articles across the two a priori delimited knowledge domains *psychology* and *education*, using the categorization system of Wikipedia. My data was sourced from an official dump file of the German Wikipedia (http://dumps.wikimedia.org/dewiki), containing a snapshot of its state as of January 16, 2012. I chose to study all articles categorized as topics of psychology (German: "Psychologie") or education ("Pädagogik"), as well as all their subcategories. The sample represented two knowledge domains with a similar number of articles and obvious content relations.

Level of artifacts: I considered three types of articles in the analysis: 5,085 specialized psychology articles, 4,696 specialized education articles, and 731 intersection articles (i.e., those categorized under both domains). Using eigenvector centrality I measured how well-connected and thus central an article was within each of the two separate domain networks (a total of 5,816 articles in the psychology network and 5,427 articles in the education network). The extent to which an article was boundary-crossing across both domains was measured with its betweenness centrality in the combined network (10,512 articles in total). The higher the eigenvector or betweenness centrality value of an article, the more pivotal was the position of the article within one of the domains or across the two domains. The network analysis measures were calculated with the *igraph* package for R (Csárdi & Nepusz, 2006).

Level of authors: I first excluded from the analysis contributions marked as minor, or made by anonymous authors or bots, deletions, reverts to a previous state of the articles, as well as contributions shorter than 150 characters. I also excluded the contributions of administrators and reviewers. Although they contribute a lot of content, their choice of articles and mode of contribution are different and depend on their Wikipedia control tasks. The remaining contributions were made by a total of 8,040 signed-in authors writing in one or both the domains. According to my taxonomy of author groups (see Table 2.1) I identified 3,980

psychology specialists, 2,762 education specialists, 1,002 intersection generalists, and 296 non-intersection generalists.

In the last stage of the analysis at the level of authors I aggregated article measures from the network analysis as an average over the articles an author had contributed to. This procedure resulted in two types of values: the eigenvector centrality of an author in a separate network, measuring how important the total contribution of an author within one domain was; and the betweenness centrality of an author in the combined network, measuring the extent to which an author contributed as a boundary spanner across domains. These aggregate measures enabled me to simultaneously investigate the partial significance of article and domain experience of an author as explanatory variables of his or her contribution to pivotal articles.

Results

Before I present the tests of the hypotheses (which concern the level of authors), I first provide the most relevant results from the analysis at the level of articles. Figure 2.1 depicts the combined network of articles in both knowledge domains. The grey dots represent education articles, the white ones psychology articles, and the black dots show intersection articles. The curved lines display the hyperlinks between the articles. The visualization was made with the Gephi platform (Bastian, Heymann, & Jacomy, 2009) using the OpenOrd layout algorithm that organizes the dots according to their interconnections. Thus, a number of dots that have direct connections to each other are represented as a cluster. Over ten repetitions of the algorithm the produced layouts were very similar.

It is interesting to note in the layout that both adjacent domains are clearly distinguishable as two separate parts in the combined network. The intersection articles are dispersed among both the education and the psychology parts of the network and do not form a homogenous network cluster. Some of the intersection articles have more connections to psychology articles and others are more tightly bound to education articles.

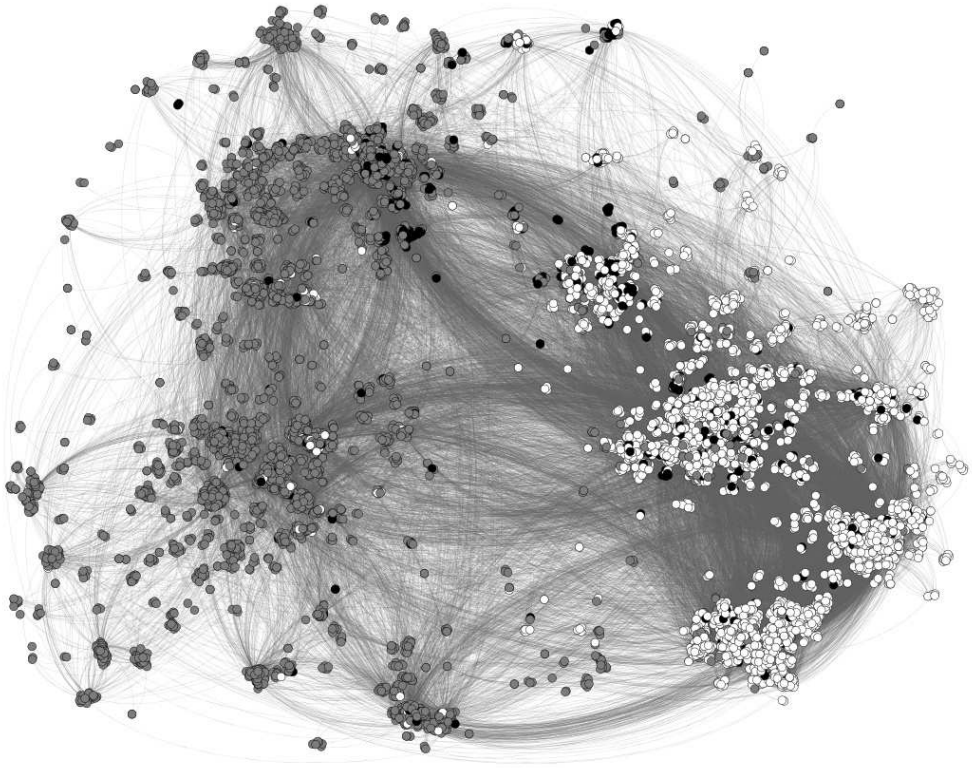

Figure 2.1 The combined network of Wikipedia articles in education and psychology.

I found moderate rank correlations between articles' eigenvector centrality in the education network and betweenness centrality in the combined network ($\tau = .53$, $p < .001$) as well as between eigenvector centrality in the psychology network and betweenness centrality in the combined network ($\tau = .43$, $p < .001$). In other words, boundary-crossing articles across the two domains are not necessarily central pivotal articles in either of the domains.

I corroborated this finding with independent-samples unequal-variances t-tests comparing the group of intersection articles with the specialized articles. Both betweenness and eigenvector centrality had distributions strongly skewed to the right, that is, only a few articles had high values, and the majority of them had very low values. I applied a logarithmic transformation to these variables in order to make them better fit the assumptions of the t-test. As expected from their definition, intersection articles were shown to be boundary-crossing articles in the combined network, with a significantly higher mean log betweenness centrality than that of specialized articles: $M_{int} = 7.01$, $SD = 3.36$ vs. $M_{spec} = 5.50$, $SD = 3.95$; $t(887.9) = 11.60$, $p < .001$. Thus, a specialized article was less likely to occupy a boundary-crossing position across the domains than an intersection article. In support of my reasoning on the moderate

correlations between eigenvector and betweenness centrality, intersection articles were shown to be less important in both separate networks, with a significantly lower mean log eigenvector centrality than that of specialized articles; in the education network: M_{int} = -4.95, SD =1.74 vs. M_{spec} = -4.64, SD = 1.44; $t(892.5) = -4.60$, $p < .001$; in the psychology network: M_{int} = -4.64, SD =1.44 vs. M_{spec} = -4.31, SD = 1.37; $t(928.6) = -5.72$, $p < .001$. Thus, a specialized article was more likely to occupy a central position in its domain than an intersection article.

I now turn to the analysis at the author level and the results of the main hypothesis tests. I excluded authors with article experience of less than 2, in order to enable fair comparisons between the groups of generalists and specialists. I did this because non-intersection generalists by definition have a minimal article experience of 2, as they have contributed to at least one education and one psychology article.[3] My sample was reduced to 1,663 authors (640 psychology specialists, 292 education specialists, 435 intersection generalists, and 296 non-intersection generalists). I used three ANCOVA models—two for the contribution to pivotal articles within each of the two domains and one for the contribution to pivotal articles across the domains. Both article experience and domain experience of an author were included in the models as predictors of the extent to which the author contributed to pivotal articles. Thus, their incremental predictive value could be simultaneously estimated. Again, I applied a logarithmic transformation to the continuous variables betweenness centrality, eigenvector centrality, and article experience, whose distributions were strongly skewed to the right, in order to make them better fit the preconditions of the ANCOVA. Article experience entered the models as a continuous predictor; domain experience was modeled as intercept dummy variables. The coefficients of these dummy variables directly indicated the differential effect of the generalist groups compared with a specific group of specialists for equal levels of article experience of specialists and generalists.

[3] Additional t-tests comparing the excluded intersection generalists and specialists with article experience of 1 corresponded with the results reported in the following.

The basic model for the three networks was:

$W_i = \alpha + \beta X_i + \gamma Y_i + \delta Z_i + \varepsilon_i$

where

W_i = predicted log betweenness or log eigenvector centrality of author i,

X_i = 1 if author i is an intersection generalist, 0 otherwise,

Y_i = 1 if author i is a non-intersection generalist, 0 otherwise,

Z_i = log article experience of author i,

ε_i = error term.

Hypothesis 1a assumed that specialists contribute on average to more central articles in each of the domains and thus have a higher aggregated *eigenvector* centrality derived from the *separate* domain networks compared with generalists. This assumption was partially supported for intersection generalists; in the education domain: $\beta = -0.15$, $t(1019) = -1.41$, $p = .159$; in the psychology domain: $\beta = -0.36$, $t(1367) = -5.04$, $p < .001$. It was fully supported for non-intersection generalists; in the education domain: $\gamma = -0.28$, $t(1019) = -2.45$, $p = .015$; in the psychology domain: $\gamma = -0.35$, $t(1367) = -4.35$, $p < .001$. Consequently, the overall effect of domain experience was marginally significant in the education domain ($F(2, 1019) = 2.99$, $p = .051$) and significant in the psychology domain ($F(2, 1367) = 16.27$, $p < .001$).

Hypothesis 1b assumed that generalists act as boundary spanners (i.e., contribute on average to more boundary-crossing articles) and thus have a high aggregated *betweenness* centrality derived from the *combined* domain network compared with education and psychology specialists taken together. This assumption was supported as well; for intersection generalists: $\beta = 0.54$, $t(1659) = 4.46$, $p < .001$; for non-intersection generalists: $\gamma = 0.31$, $t(1659) = 2.29$, $p = .022$; with a significant overall effect of domain experience: $F(2, 1659) = 10.45$, $p < .001$.

Hypothesis 2a assumed that article experience is a significant predictor of aggregated eigenvector centrality of the authors derived from the separate domain networks. This was supported for both knowledge domains; in the education domain: $\delta = 0.38$, $t(1019) = 5.19$, $p < .001$; in the psychology domain: $\delta = 0.30$, $t(1367) = 5.88$, $p < .001$.

Hypothesis 2b assumed that article experience is a significant predictor of aggregated betweenness centrality of the authors derived from the combined network of both knowledge domains taken together. This assumption was also supported ($\delta = 0.60$, $t(1659) = 6.75$, $p < .001$).

In sum, my hypotheses were largely confirmed except for a non-significant difference in the expected direction between education specialists and intersection generalists in the education domain (Hypothesis 1a). I found no significant interaction effects between article and domain experience, that is, there was no difference in the impact of article experience among the four groups of generalists and specialists. The reported results were confirmed by testing conservative ANCOVA models, using ranks (ordinal transformation) instead of the log transformed article experience, betweenness and eigenvector centrality.

Discussion

In the empirical study reported here my aim was to explain the authors' contribution to pivotal articles in the artifact network of two Wikipedia knowledge domains in relation to domain experience and article experience of the collaborating authors. Specialists (i.e., authors with contribution experience in only one of the domains) were expected to contribute on average to more central pivotal articles in each of the separate domains than generalists (i.e., authors with contribution experience in both domains). Generalists were expected to act as boundary spanners by contributing on average to more boundary-crossing pivotal articles across both domains than specialists. I further expected that article experience (i.e., the total number of articles an author has contributed to) was positively related both to the contribution to central articles within each of the two knowledge domains, and to the contribution to boundary-crossing articles across both knowledge domains.

The hypotheses of the study were supported by the empirical results. I found that both domain experience and article experience of an author are significantly related to the contribution to pivotal articles in the artifact network. Even the single non-significant result tended to be consistent with the hypothesis that education specialists would contribute to more central pivotal articles in the education domain than intersection generalists. Intersection generalists were defined as authors with at least one contribution to an intersection article. In this respect,

I found that intersection articles were boundary-crossing articles across domains and were responsible at least to some extent for the integration of knowledge across the domain boundaries. However, they were not so central and important within each of the two particular domains. Thus, the non-significant difference must be the consequence of other very central specialized articles in the education domain to which the intersection generalists had contributed. Even so, education specialists contributed on average to more central articles in the education domain than intersection generalists. Furthermore, as intersection articles turned out to be boundary-crossing articles, it is unsurprising that intersection generalists proved to be boundary spanners across the domains. However, non-intersection generalists also proved to be boundary spanners, confirming the significance of experience in both domains for the contribution to boundary-crossing articles.

Thus, my results suggest several principles of contribution to pivotal articles at domain level in a knowledge base: As I distinguished between pivotal articles that are central within a single knowledge domain and those that cross the boundaries between two domains, a difference between the authors who contributed to these two types of pivotal articles became evident. This division of roles in the mass collaboration process is related to the domain experience of the authors. Specialized experience in one domain goes together with contributions to central pivotal articles in that domain. Generalized experience in two domains goes together with contributions to pivotal articles that cross the boundaries between the domains. At the same time, the article experience of an author, regardless of the domain experience, is positively related to the contribution to both types of pivotal articles.

The reported results built on and enhanced my previous investigations (Halatchliyski et al., 2010) into knowledge construction in the context of a different pair of domains in Wikipedia. By differentiating two types of authors' experience I can now show that authors with experience in only one domain are not peripheral. These specialists play an important role in a mass collaboration environment, as their contribution is central within that knowledge domain. By isolating the relative significance of the explanatory variables domain experience and article experience, my understanding of the contribution to pivotal artifacts is now more differentiated. Generalists tend to contribute to boundary-crossing articles across domains but they are just as likely to contribute to very central articles within each of the domains, if they have a high article experience, that is, if these generalists contribute to a large number of articles. Accordingly, specialists tend to contribute to central articles within their domain but

they might also act as boundary spanners and contribute to boundary-crossing articles across domains, if they have a high article experience.

While drawn from the limiting perspective of two of the knowledge domains in Wikipedia, psychology and education, these results indicate a division of labor between generalists and specialists and a broad significance of the contribution experience of the collaborators. From a design point of view this speaks for the need of a general participation encouragement and empowerment of the long tail in networked environments. As a great number of the participants typically make few and isolated contributions, it is vital for the mass collaboration process to attract repeated contributions and commitment to pivotal artifacts. This can be facilitated at many levels of the design of an environment, from lowering the usability threshold of active participation to developing incentive systems to stimulate voluntary contributions.

Conclusion

This chapter conveys a two-fold contribution to CSCL research. It provides evidence for the significant relation between authors' experience and their contribution to pivotal artifacts at the level of knowledge domains in Wikipedia. It also provides an example of an integrative theoretical perspective within CSCL that views collective knowledge both as substance (i.e., collaborative artifacts) and as participatory activity (i.e., collaborative contributions). In accordance with this perspective, I took a multi-layered approach incorporating analysis at the level of artifacts and at the level of authors. My approach is appropriate for the self-organized, long-term and large-scale process of mass collaboration that produces a dynamic networked knowledge base of artifacts and their interconnections. Besides wikis, other multi-user virtual environments for learning, such as massive open online courses (MOOCs), or for gaming represent promising research contexts where my approach may be applicable. The condition is to identify a network of collaborative artifacts that is open to further interactive development by the participants. Such contexts may be different from formal education as learners self-regulate their motivation to participate and to achieve goals.

Considering that knowledge building in small-group settings also manifests itself in the creation of shared artifacts (Paavola & Hakkarainen, 2009; Scardamalia & Bereiter, 1994), it could be worthwhile to extend my approach to integrate the results of small-group and mass

collaboration research into a general theoretical framework. Surely, this would suggest combining the structural approach of network analysis, which is useful to discern abstract patterns, with content and interaction analysis techniques, which can supply richer interpretation of the observed patterns.

Another direction for future research would be to augment my approach with temporal aspects of knowledge development by analyzing an artifact network at different points in time. A dynamic network analysis has been shown to yield further insights into the essentially temporal collaborative process (see e.g., Chapter 4; Halatchliyski et al., 2012; Halatchliyski, Hecking, Göhnert, & Hoppe, 2013). Therefore, it would be interesting to examine the longitudinal aspects of the knowledge contained in pivotal articles in a knowledge base. As a structural backbone, such pivotal knowledge may be an important factor directing the development of new knowledge.

In line with the suggestions for further research and extension of the presented approach, I reassert my view that CSCL research should take a detailed account of the recent phenomenon of mass collaboration. The CSCL research community needs consider the increasing impact of mass collaboration on learning and knowledge creation. In my opinion, CSCL research would benefit from treating a collaborative artifact not only as a means of interaction support in small groups but also as a goal of the creation process within self-organized communities and networks. With the adoption of a network perspective, large-scale structures and long-term processes of knowledge development become accessible for investigation.

Interlude

The empirical study presented in this chapter is a first step to appreciate the potential of mass collaboration as a research area in computer-supported collaborative learning. Moreover, a framework for further analyses is now set. From a theoretical point of view, traditional CSCL approaches that focus on small-group interaction and participation can be seamlessly extended by regarding the substance of collaboratively created knowledge. Given the large-scale dimensions of online interaction, a suitable starting point to analyze knowledge as substance is not the detailed written content but its structural aspects. The availability of large and publicly accessible data sets – as in the case of Wikipedia – enables the use of powerful network analysis techniques that can capture macro-level structural patterns. The hyperlink structure of wikis suggests viewing their content as a network of interconnected articles. By measuring pivotal articles in the frame of two adjacent knowledge domains, the contributions of different Wikipedia authors can be evaluated. As the results of the cross-sectional analysis in this chapter reveal, contribution experience is a major characteristic of the authors who create pivotal articles within domains as well as across domains. The experience of the participants can be seen as an indicator of how well they have mastered the rules and goals of the community and found their place in the system. The pivotal articles in a knowledge base can be interpreted as a structural backbone of the emerging knowledge in a complex system. This speculation represents the object of investigation in the following Chapter 3. In the study presented there, I will longitudinally analyze the articles of the same domains as in the current cross-sectional analysis in order to draw more decisive conclusions on causality in the complex process of knowledge development in Wikipedia.

Chapter 3

Development of New Knowledge in Wikipedia

This chapter is based on:

Halatchliyski, I., & Cress, U. (2014). How structure shapes dynamics: Knowledge development in Wikipedia - a network multilevel modeling approach. *PLoS ONE, 9,* e111958.

Introduction

The social web affords natural interaction dynamics among a large number of participants. From the active participation of a multitude of users with different backgrounds and goals (Wasko & Faraj, 2005) emerge virtual communities (Wellman & Gulia, 1999) that define and follow their own overarching goals. The resulting process often takes the form of mass collaboration (Cress, 2013).

The interplay between the individual and the social in a self-organizing system of mass collaboration is based on the creation and use of shared digital artifacts that is enabled by Web 2.0 technologies (cf. Kolbitsch & Maurer, 2006; O'Reilly, 2005). Individuals externalize their knowledge into artifacts (Cress & Kimmerle, 2008), building a digital knowledge base with a network structure of interlinked contributions. This collective knowledge of a community (cf. Bruckman, 2006) is an emergent phenomenon (Theiner et al., 2010) of amalgamation of diverse contributions in a discourse-like process through referring, modifying and building on each other. Each new contribution needs to be adequately integrated into the existing structure. As new knowledge in the form of concepts, connections and facts is introduced to the knowledge base, the collective knowledge of the community develops in a continuous process.

The present chapter reports on a research endeavor to model and test the significance of a generative mechanism for the development of collective knowledge in Wikipedia. I relate the dynamics of knowledge to its structural dimension. I thus provide an example of a methodological approach to research questions concerning the structure and dynamics of knowledge in mass collaboration contexts.

Wikipedia is a prominent Web 2.0 community with pronounced knowledge-related activities. It follows a "no original research" rule, meaning that it accommodates only previously published facts. However, those facts stemming from external sources are then integrated in an original way (cf. Swarts, 2009). Thus, Wikipedia's knowledge base is a novel product of the community and undergoes development processes that are typical of genuine knowledge-building communities (Cress & Kimmerle, 2008; Forte & Bruckman, 2006). In Wikipedia, knowledge develops on many levels of the created artifact: Article content is edited by adding, modifying or deleting parts of it and thus changing its textual structure; hyperlinks are extensively used to establish connections between articles; new articles are constantly created

building up the knowledge domain as well as connecting different domains. System rules and community practices are the backbone for such developments and also experience changes over time themselves (Forte & Bruckman, 2008).

In the presented empirical study, I use a longitudinal network analysis approach to investigate the development of the knowledge base in Wikipedia by considering its structural properties. Focusing on two adjacent knowledge domains – psychology and education – and their intersection, I analyze the networks of knowledge consisting of interlinked articles and their development over 7 yearly snapshots from 2006 to 2012. Using multilevel modeling, I explain the significance of structurally pivotal articles (see Chapter 2) located within or across the domains at a given point in time for the future appearance of new knowledge in the knowledge base.

Measuring development in networks of knowledge

Online mass collaboration promises high potentials for development of one of the most important factors in society nowadays – knowledge. Extensive research has been done on the conditions for attracting and maintaining a critical mass of participants in virtual communities that are motivated to contribute actively (Iriberri & Leroy, 2009; Ling et al., 2005; Ridings & Gefen, 2004). Little is known about the complex patterns of self-organization when new knowledge is developed within a community. Indeed, the prediction of radical innovations as a research goal is impossible by definition (Lucio-Arias & Scharnhorst, 2012). Based on a historical account of the developed knowledge, however, it is possible to notice promising directions for further advancement. As I intend to show in the present chapter, it is also possible to model and measure relevant conditions and processes of knowledge development in a virtual community.

A suitable perspective on shared online knowledge is provided by the concept of a network. Big data sets of different collaborative networks such as interconnected scientific works, hyperlinked Wikipedia articles and many others are abundantly easily accessible on the Internet and have contributed to the rise of a “new science of networks” (Barabási, 2002). Webometric research, for example, adapts appropriate methods following a direct analogy between the analysis of scientific citations and of webpage hyperlinks as signs of knowledge relations or diffusion processes (Almind & Ingwersen, 1997). This analysis perspective

suggests that the meaning of a single scientific paper or webpage in such networks is structurally defined by the presence and absence of relations to other works and by its specific position in the network as a whole (Lucio-Arias & Leydesdorff, 2007). Therefore, well-connected and central works in a network tend to contain pivotal knowledge for the collaborative community. These network nodes are marked by high interest or quality and have an impact as hubs or brokers of knowledge (Park & Thelwall, 2003). Among the various measures of network centrality, eigenvector (Bonacich, 1972) is a popular indicator for global hubs and betweenness (Freeman, 1979) is a popular indicator for global brokers.

Network science has only lately started to expand its limited focus on static measures and structures in order to acknowledge the dynamics of complex networks. The simplest approach is a description of indicators changing over a specific time interval. Global and local network measures can be represented as a series of snapshots at different points in time. Temporal analyses of networks thus usually describe developments based on differences between snapshots (Mali et al., 2012). This has also been done for the articles and authors in Wikipedia (Buriol, Castillo, Donato, Leonardi, & Millozzi, 2006; Kittur, Chi, Pendleton, Suh, & Mytkowicz, 2007; Ortega, 2009). More complex approaches to network dynamics are necessary in order to explain the appearance of new knowledge based on change processes in existing knowledge or to explain the continuously changing position of existing ideas in a knowledge network in light of new emerging knowledge (Lucio-Arias & Leydesdorff, 2009). For the network of Wikipedia articles, such analysis could seek to establish a relation between the network position of existing interconnected articles, the change in their position over time, and the appearance of new knowledge in the form of new articles or new contributions to specific articles in the network.

The links of pivotal nodes in real-world networks are usually distributed according to a so-called power law, that is, there are very few hubs with a very high number of connections and a mass of network nodes with just a few connections. The more citations a paper has already received, the more new citations it is likely to receive. The “rich get richer” principle has been widely acknowledged in models of network growth as an explanation of such inequalities in the frequency distribution of pivotal, well-connected nodes in a network. For scientific networks, this principle was called “the Matthew effect” by Merton (1968) in reference to the Gospel of Matthew and also “cumulative advantage” by Price (1976) later on. Barabási and Albert (1999) finally coined the term “preferential attachment” and specified a network evolution model with a continuously rising number of nodes. According to this generative

mechanism, new nodes are linked with a higher probability to well-connected than to poorly connected nodes among the already existing ones.

In networks of Wikipedia articles, pivotal nodes have been regarded in the context of adjacent knowledge domains and related to contributions by experienced authors in the community (see Chapter 2). As the distribution of article hyperlinks follows a power law (Ortega, 2009), the probability that an article will receive new links is proportional to its degree, that is, the number of its existing connections with other articles. Correspondingly, the preferential attachment mechanism has been verified for Wikipedia and also for the Web as a network of websites (Capocci et al., 2006). Assuming that the elaborate system of rules in Wikipedia is strictly followed (cf. Oeberst et al., 2014), the hyperlink structure of Wikipedia articles, which is also regulated extensively[4], is a reliable representation of an extensive knowledge repository and reflects the internal structure of encyclopedic knowledge (cf. Gabrilovich & Markovitch, 2006). The preferential attachment rule could be extended to explain the process of knowledge development in Wikipedia. Thus, the appearance of new knowledge could be related to the existing structurally pivotal knowledge.

In the following, I present an empirical study with a longitudinal design that models the development of knowledge in the Wikipedia knowledge base. Employing a network analysis approach, I measure the topological position of articles within networks over a series of yearly snapshots in order to identify pivotal articles and their change over time. My goal is to test the significance of structurally pivotal articles for the subsequent appearance of new knowledge in future periods and thus to capture the interplay between structure and dynamics of knowledge.

Method

Data

I investigated the relevant factors for development of new knowledge in Wikipedia, focusing on the two related knowledge domains *psychology* and *education*. My data was sourced from

[4] http://en.wikipedia.org/wiki/Wikipedia:Manual_of_Style/Linking

an official dump file of the German Wikipedia[5], containing a snapshot of its state as of January 16, 2012.

All articles categorized as topics of psychology (German: "Psychologie") or education ("Pädagogik") as well as all their subcategories entered the study. The sample represented two knowledge domains with a similar number of articles and obvious content relations. Based on the content history of the past versions of these articles in the dump file, I took six successive snapshots of the two domains. Each of the snapshots referred to the same date, January 16, with an interval of one year between the snapshots. The first snapshot reflected the state of knowledge at the beginning of 2006, and the last (seventh) snapshot reflected the state of knowledge at the beginning of 2012.

Measures

I considered three types of articles in the analysis: specialized education articles, specialized psychology articles, and intersection articles (i.e., those categorized under both domains). Beginning by categorizing the final snapshot, which I regarded as the best developed, I traced back whether each categorized article existed in each preceding snapshot. The numbers of categorized articles over the years are shown in Table 3.1. For each article I recorded its year of creation and subtracted 2006 as a reference year from it (cf. Raudenbush & Chan, 1993). I took into account which articles were distinguished by the German Wikipedia community as featured articles for their exceptionally well-written content. In order to differentiate the controversiality of the article topics I used the algorithm developed by Yasseri and Kertész (2013). Explanatory variables that changed over the time span of the study were the total number of article edits at each snapshot year and the article age in years since creation. In order to make inferences about only the knowledge-related development of the articles, I first excluded from the analysis edits marked as minor, made by anonymous authors or bots, deletions, reversions to a previous article state of the article, as well as contributions shorter than 150 characters. I also excluded the contributions of administrators and reviewers. Although they contribute a large amount of content, their choice of articles and mode of contribution is driven by particular reasons reflecting their administrative responsibilities in

[5] http://dumps.wikimedia.org/dewiki

Wikipedia. Moreover, it has been shown (Kittur et al., 2007) that the percentage of contributions from such authors dramatically declined after 2004.

Table 3.1 Development of the number of categorized articles and of authors.

snapshot year	specialized psychology articles	specialized education articles	intersection articles	Σ articles	Σ authors
2006	2176	1357	325	3858	1776
2007	2911	1980	450	5341	3113
2008	3472	2556	526	6554	4265
2009	3908	3108	581	7597	5251
2010	4262	3595	626	8483	6104
2011	4660	4166	686	9512	7047
2012	5085	4696	731	10512	8002

I used network analysis in order to measure how pivotal each article was at a given point in time. Pivotal network position was regarded as an expression of an article's significance in the structural dimension of knowledge. For each of the seven snapshots of the knowledge domains, I extracted the current networks of knowledge at that time by parsing the relevant hyperlinks from the content of the last version of each article on each January 16. The networks of knowledge consisted of articles as nodes and of the hyperlinks between them transformed into undirected edges. Thus, the networks were aimed at representing the conceptual structure of knowledge in the interrelated articles and not the browsing and diffusion processes that only flow in the direction of the hyperlinks. For each snapshot, I constructed the two single domain networks, one for psychology and one for education, as well as the combined network of both domains. The extent to which an article was boundary-crossing across both domains was determined by measuring its betweenness (Freeman, 1979) in the combined network at each of the seven points in time. Analogously, in each of the two separate domain networks I used eigenvector centrality (Bonacich, 1972) to measure how well-connected and thus central each article was. The pivotal articles in each of the snapshots were those with higher values of either betweenness or eigenvector centrality. The network analysis measures were calculated with the igraph package for R (Csárdi & Nepusz, 2006).

New knowledge in Wikipedia appears in the form of either newly created articles or new edits that add information to existing articles. As I wanted to locate the development of new knowledge in the network of articles, I operationalized the concept of new knowledge for each article in each period in three ways: first, by counting its new neighboring articles that were created in that period, that is, directly hyperlinked articles with an age of less than one year; second, by calculating the change in the sum of edits to all the neighboring articles in that period; and third, by counting the number of new contributions an article received in that period.

Modeling approach

Our data consisted of article variables some of which were measured repeatedly and others that were time invariant characteristics. The longitudinal study design naturally lent itself to multilevel analysis, which is a state-of-the-art approach in educational research (Cress, 2008; Janssen, Erkens, Kirschner, & Kanselaar, 2011). The dependency of the repeated measures of the same articles was taken into account by differentiating two levels: the level of the measurement period and the level of the articles. Statistical calculations were executed with the *lme4* package for R (Bates, Maechler, & Bolker, 2013).

Hypotheses

The major goal of this longitudinal study was to explain the appearance of new knowledge in Wikipedia by focusing on the networks of hyperlinked articles within and across two domains. My hypotheses therefore address the variables betweenness and eigenvector centrality, which measure how pivotal the position of an article is in a network for a given period snapshot. Deriving from the preferential attachment rule (Barabási & Albert, 1999; Capocci et al., 2006) that an article receives new hyperlinks with a probability proportional to the number of its existing hyperlinks to other articles, I formulate the following hypotheses:

Hypothesis 1: Articles with a pivotal network position within or across knowledge domains become linked with a higher number of new neighboring articles during the subsequent period than do less pivotal articles.

Hypothesis 2: The neighbors of pivotal articles receive more new contributions than the neighbors of less pivotal articles during the subsequent period.

Hypothesis 3: Pivotal articles receive a higher number of contributions during the subsequent period than do less pivotal articles.

In order to more accurately evaluate the main effect in each of these hypotheses, it is simultaneously evaluated with the partial effects of a number of control variables: article type, year of creation, age, number of received contributions, number of neighbors, featured article distinction and article controversiality.

Each of the hypothesis tests were carried out using separate models for the psychology, the education and the combined networks.

Results

Descriptive trends

Before I present the statistical tests of the hypotheses, I first provide a descriptive overview of the development of the articles and authors in the domains between 2006 and 2012. This information outlines the state of Wikipedia before and during the longitudinal study interval and thus introduces to the context of my investigation. Table 3.1 shows a continuous increase in the number of articles and authors. A detailed investigation of the number of articles before the studied time interval revealed an increasing growth rate until the peak year 2005. The author growth rate in the studied domains rose until the peak year 2007, that is, for two years longer than that of the articles. In later periods, as depicted in Table 3.2, the number of both articles and authors had diminishing growth rates and reached a steady level of growth by 2008/2009.

Table 3.2 Yearly growth in the total number of articles and authors.

	2002/03	**2003/04**	**2004/05**	**2005/06**	**2006/07**	**2007/08**	**2008/09**	**2009/10**	**2010/11**	**2011/12**
articles	119	574	1658	1486	1483	1213	1043	886	1029	1000
authors	12	135	677	952	1337	1152	986	853	943	955

Considering the number of articles that received new contributions during a one-year period between two snapshots, Table 3.3 shows that it had been rising until 2007/2008 when it reached a stable level for specialized psychology articles and for intersection articles. The number of edited education articles per period rose throughout the studied interval, albeit slower since 2008.

Table 3.3 Development of the number of articles with new contributions per period.

	2006 - 2007	**2007 - 2008**	**2008 - 2009**	**2009 - 2010**	**2010 - 2011**	**2011 - 2012**
psychology	1285	1561	1541	1428	1451	1515
education	739	955	1026	1049	1071	1161
intersection	198	269	252	210	233	231
Σ	2222	2785	2819	2687	2755	2907

Regarding the network of articles, I observed a stable power law degree distribution in all the snapshots as displayed in Figure 3.1. Due to the growth in the network the degree distribution shifts upwards over time. The degree distributions of the psychology, education and intersection articles in the single domain networks show the same pattern.

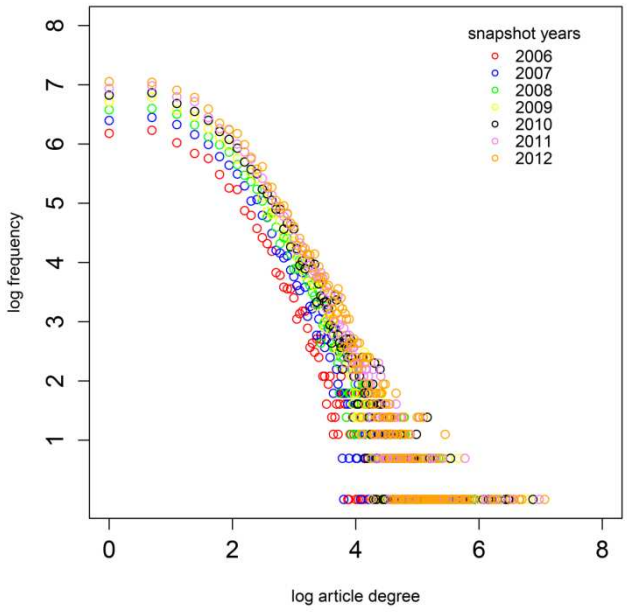

Figure 3.1 Degree distribution in the combined network of psychology and education articles in the seven snapshot years.

My data partly confirmed the results of Capocci et al. (2006) who demonstrated preferential attachment for the main English and Portuguese Wikipedia networks with decreasing linking probability for the articles with very high degree, that is, number of neighbors. As Figure 3.2 shows, the preferential attachment for the combined network of psychology and education articles becomes saturated for large values of the degree of an article. My data consists of discrete network snapshots in time, and I cannot observe well differences in the linking probability among articles with high degree that become linked to new articles in each period. In addition, Figure 3.2 reveals a decrease in the linking probability of low-degree articles and an increase in the linking probability of high-degree articles over the years.

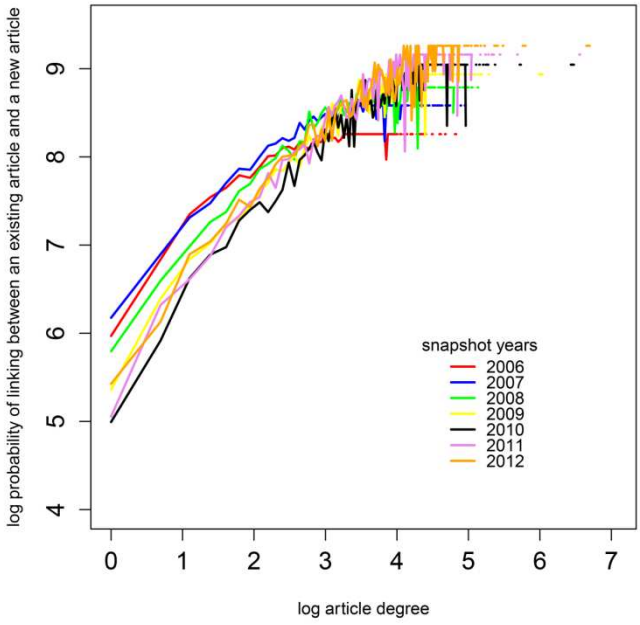

Figure 3.2 Preferential attachment in the combined network of psychology and education articles in the seven snapshot years.

The percentage distribution of links between the different types of articles in the network remained stable over the snapshots. The connections with articles outside the two domains of focus held the largest share in both domain networks: 91% in psychology and 84% in education. The connections between intersection articles and strictly psychology articles amounted to 8%, and between intersection articles and strictly education articles amounted to

15%. The connections between strictly psychology and education articles amounted to 1% in each of the domain networks.

In summary, the descriptive analysis revealed that the number of newly created articles during a one-year period was the first variable that stopped growing, with the growth rate diminishing around 2005. By 2009 the dynamics of articles and authors in the analyzed Wikipedia domains demonstrated stability. These results indicate that the studied interval was well-chosen and the hypotheses tests are not likely to be biased by exogenous disturbances of the dynamics of the mass collaboration system.

Hypotheses testing

My hypotheses concern the appearance of new knowledge in the article networks of two knowledge domains. According to Hypothesis 1, I first modeled the appearance of new knowledge as the number of newly created articles that become direct neighbors of an article in each of the three networks (i.e., psychology, education and combined) during each one-year period in the studied time interval.

The need for employing multilevel modeling was confirmed by the calculated design effects, which were all greater than 2 (cf. Peugh, 2010): 2.65 for psychology, 2.47 for education and 2.43 for the combined network. The dependent variable, the number of newly created articles as neighbors, is a count variable with a high percentage of zeros throughout the measurement instances: 69.1% in psychology, 74.0% in education and 72.0% in the combined network. Therefore I used logistic models that treat the number of new articles as a binary outcome variable, that is, they model the differences between cases with zero versus cases with a non-zero count of new articles as neighbors. The general model specification was:

$$K_{1ij} = \text{logistic}\ (\beta_0 + \beta_{0j} + \sum_{z=1}^{8} \beta_z X_{zij} + e_{ij})$$

where K_{1ij} denotes as 1 or 0 whether article *i* has received at least one newly created article as a neighbor between the snapshot periods *j-1* and *j*, β_0 is the global fixed intercept, β_{0j} is the random intercept for each of the seven snapshots, $\sum_{z=1}^{8} \beta_z X_{zij}$ is the linear combination of

the eight explanatory variables and their regression coefficients and e_{ij} is an error term. Table 3.4 shows the results of the regressions successively for the psychology, education and the combined network. The column *level of variable* indicates whether the variable is time invariant for each article or it is repeatedly measured in each period.

Table 3.4 Multilevel logistic models of newly created articles received as neighbors.

	Level of variable	**Estimate**	**Std. error**	**z value**	**Pr(>|z|)**
Combined					
(Intercept)		2.56	0.11	23.19	<2e-16***
creation year	article	-0.33	0.01	-25.27	<2e-16***
article age	period	-0.22	0.01	-21.57	<2e-16***
t-1 log betweenness	period	0.31	0.01	33.30	<2e-16***
t-1 log edit count	period	0.26	0.02	11.40	<2e-16***
education article	article	-0.17	0.04	-4.28	1.9e-05***
intersection article	article	0.19	0.07	2.87	0.0041**
featured article	article	0.04	0.19	0.20	0.8399
log controversiality	article	0.05	0.01	4.26	2.0e-05***
Psychology					
(Intercept)		1.55	0.10	15.33	<2e-16***
creation year	article	-0.47	0.02	-27.58	<2e-16***
article age	period	-0.33	0.01	-26.14	<2e-16***
t-1 log eigenvector	period	0.51	0.02	26.82	<2e-16***
t-1 log edit count	period	0.37	0.03	12.894	<2e-16***
intersection article	article	0.68	0.07	9.50	<2e-16***
featured article	article	-0.14	0.22	-0.61	0.5430
log controversiality	article	0.06	0.01	4.32	1.5e-05***
Education					
(Intercept)		-0.20	0.10	-1.89	0.0586.
creation year	article	-0.38	0.02	-19.97	<2e-16***
article age	period	-0.18	0.01	-12.47	<2e-16***
t-1 log eigenvector	period	0.27	0.02	15.69	<2e-16***
t-1 log edit count	period	0.47	0.03	13.80	<2e-16***
intersection article	article	0.70	0.08	9.04	<2e-16***
featured article	article	0.37	0.40	0.91	0.3605
log controversiality	article	0.08	0.02	4.15	3.3e-05***

The models for the three networks were congruent with each other. They all featured the same set of significant regressors. Regressors with a significant negative influence were article creation year and age, that is, the later the year of creation of an article in Wikipedia and the more years since its creation, the less likely it was that the article received any newly created articles as neighbors. In support of Hypothesis 1, both an article's previous period betweenness (t-1 log betweenness) in the combined network and an article's previous period eigenvector centrality (t-1 log eigenvector) in the psychology or education network were regressors with a significant positive influence, as was the number of contributions received up to the previous period in all three networks (t-1 log edit count). Intersection articles were significantly more likely than specialized psychology or education articles to receive newly created articles as neighbors. Featured articles were not significantly different from non-featured articles in their probability to receive newly created articles as neighbors. Article controversiality was a significant positive regressor.

To test Hypothesis 2, I next modeled the dynamics of new knowledge as the change in the total edit count of the neighboring articles of an article during one period. Again, multilevel modeling was necessary as the design effects were all greater than 2: 2.72 for psychology, 2.47 for education and 2.56 for the combined network. The distribution of the dependent variable permitted the use of linear multilevel models. The general model specification was:

$$K_{2ij} = \beta_0 + \beta_{0j} + \sum_{z=1}^{8} \beta_z X_{zij} + e_{ij}$$

where K_{2ij} is the change in the total edit count of the neighbors of article *i* between the snapshot periods *j-1* and *j*, β_0 is the global fixed intercept, β_{0j} is the random intercept for each of the seven snapshots, $\sum_{z=1}^{8} \beta_z X_{zij}$ is the linear combination of the eight explanatory variables and their regression coefficients and e_{ij} is an error term. The results are presented in Table 3.5 successively for the psychology, education and the combined network.

Table 3.5 Multilevel linear models of the change in the edit count of the neighboring articles.

	Level of variable	Estimate	Std. error	t value	Pr(>|z|)
Combined					
(Intercept)		131.95	4.09	32.29	<2e-16***
creation year	article	-7.73	0.48	-15.97	<2e-16***
article age	period	-8.13	0.35	-23.32	<2e-16***
Δ neighbors since t-1	period	20.42	0.11	180.30	<2e-16***
t-1 log betweenness	period	5.86	0.33	17.82	<2e-16***
t-1 log edit count	period	9.98	0.90	11.82	<2e-16***
education article	article	-37.04	1.67	-22.19	<2e-16***
intersection article	article	-9.28	2.91	-3.19	0.0014**
featured article	article	53.85	8.55	6.30	3.0e-10***
log controversiality	article	6.28	0.50	12.68	<2e-16***
Psychology					
(Intercept)		116.88	3.47	33.66	<2e-16***
creation year	article	-9.02	0.57	-15.79	<2e-16***
article age	period	-9.06	0.43	-21.28	<2e-16***
Δ neighbors since t-1	period	23.66	0.14	171.27	<2e-16***
t-1 log eigenvector	period	13.30	0.58	22.80	<2e-16***
t-1 log edit count	period	13.19	1.04	12.74	<2e-16***
intersection article	article	0.79	2.73	0.29	0.7732
featured article	article	58.11	9.01	6.45	1.1e-10***
log controversiality	article	6.61	0.55	11.92	<2e-16***
Education					
(Intercept)		53.14	2.43	21.90	<2e-16***
creation year	article	-6.38	0.43	-14.79	<2e-16***
article age	period	-5.85	0.33	-17.66	<2e-16***
Δ neighbors since t-1	period	14.87	0.13	112.62	<2e-16***
t-1 log eigenvector	period	4.52	0.36	12.70	<2e-16***
t-1 log edit count	period	10.54	0.83	12.63	<2e-16***
intersection article	article	36.77	2.09	17.56	<2e-16***
featured article	article	1.10	11.27	0.10	0.9225
log controversiality	article	4.04	0.56	7.23	4.9e-13***

Generally, the results correspond with those from the previous models of the number of new articles as neighbors. The models for the three networks again featured nearly the same set of significant regressors. Regressors with a significant negative influence were again article creation year and age. The change in the number of neighbors since the previous period functioned as a control variable and had a high positive t-value in explaining the variance of

the dependent variable, that is, the change in the total edit count of the neighbors. In support of Hypothesis 2, further regressors with a significant positive explanatory power were again the article's previous period betweenness in the combined network, the article's previous period eigenvector centrality in the psychology or education network, and the number of contributions received up to the previous period. Except in the psychology network, psychology articles had higher positive values of change in the edit count of the neighboring articles than intersection articles and intersection articles had higher positive values than education articles. Except in the education network, featured articles had significantly higher positive values of change in the edit count of the neighboring articles than non-featured articles. Article controversiality was again a significant positive regressor in these models.

Our third and last indicator of new knowledge was the number of new contributions an article receives during a one-year period (Hypothesis 3). Again, the calculated design effects required multilevel modeling: 2.39 for psychology, 2.16 for education and 2.27 for the combined network. In more than half of the data snapshots, the articles did not receive any new edits in the past year, so the dependent variable again contained an excess of zeros: 58.2% in psychology, 62.8% in education and 60.6% in the combined network. Using logistic models, I investigated the binary outcome, that is, the differences between occasions when articles received zero versus at least one new edit during a given period. The general model specification was:

$$K_{3ij} = \text{logistic}\,(\beta_0 + \beta_{0j} + \sum_{z=1}^{8} \beta_z X_{zij} + e_{ij})$$

where K_{3ij} denotes as 1 or 0 whether article *i* has received at least one new substantial contribution between the snapshot periods *j-1* and *j*, β_0 is the global fixed intercept, β_{0j} is the random intercept for each of the seven snapshots, $\sum_{z=1}^{8} \beta_z X_{zij}$ is the linear combination of the eight explanatory variables and their regression coefficients and e_{ij} is an error term. Table 3.6 shows the results of the regressions successively for the psychology, education and the combined network.

Table 3.6 Multilevel logistic models of an article receiving new edits.

	Level of variable	Estimate	Std. Error	z value	Pr(>|z|)
Combined					
(Intercept)		0.63	0.09	7.11	1.2e-12***
creation year	article	-0.39	0.01	-34.40	<2e-16***
article age	period	-0.31	0.01	-32.86	<2e-16***
t-1 log betweenness	period	0.09	0.01	12.10	<2e-16***
t-1 log edit count	period	0.65	0.02	33.02	<2e-16***
education article	article	0.01	0.03	0.20	0.8420
intersection article	article	0.01	0.06	0.34	0.7350
featured article	article	0.18	0.19	0.94	0.35
log controversiality	article	0.20	0.01	16.34	<2e-16***
Psychology					
(Intercept)		-0.02	0.07	-0.21	0.8303
creation year	article	-0.43	0.01	-32.38	<2e-16***
article age	period	-0.33	0.01	-30.33	<2e-16***
t-1 log eigenvector	period	0.05	0.01	4.43	9.6e-06***
t-1 log edit count	period	0.68	0.02	29.94	<2e-16***
intersection article	article	0.11	0.05	1.98	0.0475*
featured article	article	0.17	0.21	0.83	0.4093
log controversiality	article	0.19	0.01	14.08	<2e-16***
Education					
(Intercept)		-0.25	0.08	-3.31	0.0009***
creation year	article	-0.37	0.01	-26.28	<2e-16***
article age	period	-0.31	0.01	-24.51	<2e-16***
t-1 log eigenvector	period	0.03	0.01	2.92	0.0035**
t-1 log edit count	period	0.72	0.03	27.67	<2e-16***
intersection article	article	0.10	0.06	1.83	0.0678.
featured article	article	0.42	0.34	1.23	0.2160
log controversiality	article	0.22	0.02	11.34	<2e-16***

The results of this third perspective on new knowledge in the networks followed the pattern of the previous two. In all networks, article creation year and age were regressors with a significant negative influence on the dependent variable. Significant positive regressors were the previous period betweenness and eigenvector centrality in support of Hypothesis 3, as well as the number of received contributions up to the previous period. Article type was only marginally significant, with intersection articles being more likely than specialized psychology or education articles in both separate domain networks to receive new

contributions. Featured articles were not more likely to receive new contributions than non-featured articles. Finally, article controversiality had a significant positive explanatory power for my third indicator of new knowledge.

Discussion

The aim of this longitudinal empirical study was to explain the appearance of new knowledge in Wikipedia by taking the article network structure into account and by focusing on the pivotal articles in particular. Articles with a pivotal network position within or across two domains at a given point in time were expected to be a connecting factor for larger amounts of new emerging knowledge during the following year than less pivotal articles. The expectation was operationalized for three forms of new contributed knowledge in Wikipedia: the number of new articles as neighbors, the change in the total sum of edits of the neighbors and the number of new received contributions.

The hypotheses of the study were supported by the empirical results. The tests showed that pivotal articles, indicated by high betweenness in the combined network or by high eigenvector centrality in the separate domain networks, link to all three relevant forms of new emerging knowledge in Wikipedia at a significantly high rate. In spite of the differences in the network positions of these articles within and across the knowledge domains, both types of articles are pivotal for knowledge development.

According to my results shown in Figure 3.2, the probability of an article to receive newly created articles as neighbors increases with its degree. Thus, the results of testing Hypothesis 1 shown in Table 3.4 can also be interpreted as evidence of the relationship between the centrality indicators degree and betweenness, and between degree and eigenvector. Indeed, these are distinct indicators, all of which point out the relative significance of nodes in a network. A node with high degree also has a high probability to be part of the shortest connection between many of the other nodes. Thus, degree is related with betweenness, at least in non-fractal networks (Holme, Kim, Yoon & Han, 2002; Kitsak et al., 2007). Eigenvector is also related with degree as it extends the count of neighbors of a node by taking the degree of these neighbors into account (Hanneman & Riddle, 2005).

The explanatory power of pivotal articles for new emerging knowledge has been substantiated through the inclusion of a number of significant control variables in my models. These variables expand the preferential attachment mechanism (Barabási & Albert, 1999), which I have shown for the knowledge development in Wikipedia, by a number of other important effects. The model results showed that the later the starting year of an article and the older its age, the less likely it links to new emerging knowledge. Empirical evidence from scientometric studies also indicated the effect of aging of scientific work beyond the preferential attachment mechanism as the age of a scientific work is negatively related to its likelihood of receiving future citations, and thus to its impact on future knowledge development (Radicchi, Fortunato, & Vespignani, 2012). It seems that pivotal articles in Wikipedia were created in the early years of the studied domains. At that time, there was more intensive work on creating new articles and developing pivotal articles. The positive effect of the network position of pivotal articles on knowledge development is strong enough to supersede the negative effect of their higher age.

My results furthermore showed that articles with many contributions and intersection articles of two domains are largely more likely to link to new emerging knowledge. Independent of whether they are pivotal in the network, articles on topics that receive much attention have the potential to generate further knowledge development, at least for a short time. This supports the results of Wilkinson and Huberman (2007), who showed that popular and relevant article topics receive a high number of contributions and are likely to be of high quality. Featured article distinction can be regarded as an additional factor of knowledge development that only concerns the neighborhood of a featured article. Article controversiality was demonstrated to increase all three types of knowledge development.

It is important to note that my results about pivotal and new knowledge apply to specific knowledge domains and to a specific stage in the historical development of the German Wikipedia. Following my descriptive analysis of the studied time interval, the rates of growth in the number of articles, authors and contributions largely reached stable levels after a short interval of decrease. I would regard the preceding interval until 2006 as a different stage in the history of the German Wikipedia, as it grew exponentially. Suh, Convertino, Chi and Pirolli (2009) have observed the same for the English Wikipedia. As already mentioned, before the articles' growth rate began to diminish, it peaked in 2005 and thus 2 years earlier than the peak of the authors' growth rate. I see this as evidence that a new stage of the German Wikipedia's history, which was framed by the present study, was initiated by the stagnating

number of new articles. I call this stage of stable growth rates an equilibrium stage. Studying the development of scientific fields, Price (1963) regularly observed a similar saturation stage after knowledge had grown exponentially, and after opportunities for incremental developments had finished. As Wikipedia is an evolving complex system, it is unclear how stable its equilibrium might be and what internal or external processes might currently protect or endanger it.

In a study that first recognized this stagnation, Suh et al. (2009) noted three possible causes for it at the level of authors: conflicts between experienced and new authors; bureaucracy with rising coordination costs for the contributors; and deficient collaborative tools. For the level of articles, Suh et al. (2009) also conjectured that the number of available new encyclopedic topics that still had not been covered in Wikipedia might be declining. Halfaker, Geiger, Morgan and Riedl (2013) later doubted the relevance of this knowledge saturation hypothesis and pointed out that even if this were the case, there would still be a plenty of writing that could be done, as even some of the most important Wikipedia articles would suffer from bad quality. While my study did not directly test the hypothesis of whether worsened conditions of collaboration had slowed down the German Wikipedia's growth, my results indicate that this probably came as a later factor in a longer causal chain. Its origin seems to have been the reduced choice of new articles on accessible, well-known topics that could still be created.

The creation of new articles did not come to a halt but went back to a lower linear yearly rate. In the new equilibrium stage, articles in the studied domains of the German Wikipedia presumably required more specialized knowledge and greater cognitive efforts than in the earlier exponential growth stage. Table 3.3 also showed that the number of articles with new contributions per period continued growing without a decline and then became stable. A plausible reason for this is that some of the efforts for creating and expanding new articles were switched to other older articles.

The declining availability of popular topics that have not yet been written affected the numbers of new, inexperienced authors. As I have shown in my previous investigations (see Chapter 2; Halatchliyski et al., 2010), author's experience in contributing to different articles is needed in order to be able to contribute to pivotal articles that have reached advanced stages of development and make up the structure of a knowledge domain.

Conclusion

The interplay between structure and dynamics in knowledge-related networks has been pointed out as a promising area of research (Börner, Boyack, Milojević, & Morris, 2012). The present work applied powerful, longitudinal multilevel analysis and showed that structures that are pivotal within the static organization of knowledge are also pivotal for the dynamic development of new knowledge measured in three ways in Wikipedia. Thus, the results integrate with my previous investigations (see Chapter 2; Halatchliyski et al., 2010) of the contribution experience of authors that substantially promote the appearance of new knowledge by contributing to pivotal articles.

The presented results, however, also raise some critical thoughts about the mass collaboration system Wikipedia. Associated with the reduced availability of new topics, the online encyclopedia as a whole has reached a saturation stage after an initial exponential growth. Participation thresholds facing relatively inexperienced authors are continuously rising, and the work that remains can only be performed by a tiny percentage of authors who have acquired authority status (Shaw & Hill, 2014).

Although I cannot be sure about the transferability of the insights gained from Wikipedia, I found evidence that the structures and dynamics of knowledge development exhibit mechanisms similar to other knowledge-related realms like scientific work. This encourages me to look further into generally relevant conditions and processes and to embrace the challenge of the dynamics of knowledge, which is difficult to grasp (see e.g., Chapter 4; Halatchliyski et al., 2012). Series of ideas and actions are said to lead to the stabilization of historical trajectories and structural patterns over time (Lucio-Arias & Leydesdorff, 2009). I find it interesting and important to deepen our understanding of the dynamics of this major factor in society – knowledge.

Interlude

The previous Chapter 2 and the current Chapter 3 provided a cross-sectional and a longitudinal analysis of the articles and authors in two adjacent knowledge domains in Wikipedia. I presented a network analysis approach to the structural dimension of shared knowledge in a mass collaboration community. Using a multilevel statistical analysis of periodic snapshots of the studied networks, the current chapter showed that structural aspects of knowledge are related to causal mechanisms of its temporal dynamics. Pivotal articles identified by their topological position in the knowledge base give a static representation of the structural backbone of collective knowledge. At the same time, they represent an important factor of the long-term development of knowledge, as new knowledge that appears in future periods in the network predominantly links to the existing pivotal articles. This result complements the findings in Chapter 2 that pivotal articles are written by experienced Wikipedia contributors.

After examining these general mechanisms of online mass collaboration, for the practical purposes of educational science and learning analytics it is interesting to highlight the potentials of network analysis for a more immediate evaluation of the processes. Collaborative learning and knowledge building are essentially temporal processes of development. The following Chapter 4 will present the application of a scientometric network analysis method to the young online learning community Wikiversity. With the help of main path analysis, pivotal contributions will be identified not in the static structure of knowledge but directly in the dynamic trajectories of the evolution of knowledge.

Chapter 4

Main Paths of Knowledge Evolution in Wikiversity

This chapter is based on:

Hatchliyski, I., Hecking, T., Göhnert, T., & Hoppe, H. U. (2014). Analyzing the main paths of knowledge evolution and contributor roles in an open learning community. *Journal of Learning Analytics, 1*, 72-93.

Introduction

Nowadays, it is commonplace to perceive learning and knowledge building as closely related activities on the Web. Knowledge building is based on the creation of *epistemic artifacts* (Knorr-Cetina, 2001) that are shared in a community. Bereiter and Scardamalia (2003) point out that knowledge building is essential for learning but has a wider scope in that it is not necessarily limited to explicit learning scenarios. Scientific research is an example of a distributed knowledge-building activity that takes place in scientific communities and typically is not characterized as learning. According to Scardamalia and Bereiter (1994), the knowledge building pedagogy takes scientific research as a blueprint of the collaborative learning of students that needs to be facilitated. During a knowledge-building process, students discuss ideas and develop their shared knowledge in the manner of scientists. The philosophical foundation of this view dates back to Popper (1968), who explains the development of scientific knowledge as a constant process of emergence of new ideas and their gradual improvement or abandonment after discovering contradictory evidence. In fact, any learning community defines concepts and builds its knowledge base in a similar way (Stahl et al., 2006).

With the present chapter I offer an approach to analyzing learning processes organized in the form of online knowledge building. Online knowledge building is characterized by collaborative activities and the creation of shared artifacts within a community of learners. This form of collaborative learning is becoming increasingly popular on the Web and goes beyond formal educational contexts (see Chapter 2). As this is a relatively new phenomenon and it shifts the focus from the individual learner to the knowledge processes within a community, appropriate methodologies are expectedly complex and in a very early developmental stage.

Due to the relation between scientific production and learning in communities, I aim to show that both processes can be studied using the same analytical approaches. *Scientometrics* as a research field is particularly concerned with the quantitative measurement of scientific work, and so offers a variety of potentially fruitful approaches that are new to the area of learning analytics (Suthers & Verbert, 2013). Scientometric methods are tailored for the analysis of knowledge artifacts, most prominently publications, and their authors. One well-known method is the calculation of the h-index as a measure of scientific reputation (Hirsch, 2005). In the context of learning communities, however, individual excellence is not a primary

concern. Rather more interesting would be an approach to the long-term characteristics and the dynamics of interactive learning environments.

Hummon and Doreian (1989) have proposed a method to detect the main idea flows based on citation networks using a corpus of publications in DNA biology as an exemplar. My work reported in this chapter takes the *main path analysis* technique as a starting point in the analysis of a broad range of knowledge-building processes that take place in formal as well as informal collaborative settings. After an initial promising application of main path analysis to networks of knowledge artifacts created for educational purposes (Halatchliyski et al., 2012), I now want to elaborate on the adaptation and adequate formalization of the method. My guiding question in this endeavor is: What kind of insights can be gained from the main path analysis of knowledge creation in online learning communities? I will explore this question using data from Wikiversity[6] as an example. Wikiversity is understood by its active members as an "open learning community" in which users can actively produce learning resources for a broad range of topics in the form of web-based hypermedia. In my view, it represents a challenging and yet relevant field for exploring the potential of scientometric methodology to tackle the dynamics of computer-supported learning processes.

Background

Community learning

New knowledge in the world might be the accomplishment of an individual, but it is inconceivable without the body of previously existing knowledge that in turn has been established by many other individuals. Consequently, learning and development of new knowledge must be examined in the context of the community in which they take place.

Online communities like Wikiversity facilitate learning through the creation of a shared knowledge base in the form of digital artifacts such as texts, pictures or other multimedia. Users can passively learn by making use of the existing artifacts. Users can also actively learn by participating in the creation of new artifacts. The *knowledge building theory* suggests incorporating such activity in formal education (Scardamalia & Bereiter, 1994). Students are expected to benefit from self-motivated exploration of knowledge areas when they share and build on each other's findings in a collaborative online environment. During this long-term

[6] http://www.en.wikiversity.org

process, the shared community knowledge develops as ideas are constantly improved by the participants. Individual learning is an outcome of the knowledge development of the whole community.

The collaborative production of digital knowledge artifacts has become widespread since the emergence of Web 2.0. Widely and easily available tools such as wikis afford a long-term process of mass collaboration, as artifacts are built piece-by-piece and individual contributions have variable sizes. Moreover, a single contribution to an artifact can be revised or be built upon in order to produce newer versions. Every change to the shared artifacts of a wiki community can be logged as an individual contribution activity, but the ongoing development of the knowledge base is an emergent product of the community as a whole. Intersubjectivity and shared meaning-making are epiphenomena of the interaction among individuals in a community (Stahl et al., 2006). From the systemic view of the *co-evolution model* of individual learning and collaborative knowledge building (Cress & Kimmerle, 2008), a community and the participating individuals function as two different types of systems that co-evolve through mutual fertilization. Knowledge development is reflected in the changing shape and content of the artifacts.

Knowledge artifacts often hold connections among themselves that are marked by higher-level semantic structures like topical relations, problem-solution chains, discourses, etc. Regardless of whether these connections are deliberately made by the participants in a community or whether they are automatically produced by the online environment, hypermedia links bear meaning. This meaning is an integral part of the knowledge created by a community. It is also subject to change, as connections are added or deleted in parallel with the artifact development.

In sum, learning in a community represents a complex process that is dependent on the activities of many participants and supported by the use and development of artifacts as learning resources. The process evolves with the constant change of the shared knowledge base at the level of single resources or their interconnections.

Temporality of a learning process

The learning of an individual or of a whole community is a process that essentially develops over periods of time. New knowledge is built upon existing knowledge. A knowledge base develops gradually as its information content evolves. Single ideas become more concrete,

they can flow together or split into independent directions, marking a convergence or divergence in the development process (Halatchliyski et al., 2011). At a higher level of abstraction, the interconnections within the knowledge base also develop when new ideas are added to existing content, or when already existing connections are subsequently changed.

All these changes should be studied in order to understand the corresponding learning processes. Accordingly, the temporal dimension should be regarded as a main component of learning analytics. However, the modeling of the overall process of knowledge development is challenging, as the sequential relations between all the changes in the knowledge base need to be tracked. Any aggregation across time easily leads to a biased analysis of individual and community-level variables. A longitudinal study of different points in time is also an unsatisfactory option, as it misses out on the authorship of changes that have been made between the chosen time points. Especially difficult to grasp is the nonlinear flow of ideas that is characteristic to any learning process.

Previous work in the area of computer-supported learning has paid attention to the interactivity of collaborative processes and thereby implicitly to learning dynamics. Environment data logs have been used to describe and map interaction patterns. Their interpretation has often been supported by additional analysis of the content in the case of discussion board messages (see, for example, Hara, Bonk, & Angeli, 2000; Schrire, 2004). Suthers, Dwyer, Medina, and Vatrapu (2010) also presented a universal framework for describing interactivity in the form of uptakes between contributors independent of the environment that is used. Nevertheless, the field of learning analytics still needs a method to address the temporality of learning processes quantitatively. Aspects that need to be taken into account include: who influenced whom, which ideas were taken up in later stages and which were not, and how differently do the participants contribute to the overall learning process. The method should also be adaptable to the multiplicity of learning environments and communities that have emerged with Web 2.0.

Different forms of sequential analysis of learner actions have also been developed in order to detect and understand the best practices of orchestration of tools and content in the learning process (Cakir, Xhafa, Zhou, & Stahl, 2005; Jeong, 2003; Perera, Kay, Koprinska, Yacef, & Zaïane, 2009). Frequently occurring sequences of actions or events reveal connections between the learning history as captured in log files and learning performance. Such analysis should help warn learners against inefficient strategies and also better adapt the environment and the learning materials to their needs (Zaïane & Luo, 2001). Although it certainly accounts

for the temporal dimension and thereby gives deeper insight into the learning process, sequential analysis as a data-mining technique relies on the *a priori* definition of activity and event categories. The necessary coding scheme always represents a potential weak point in the analysis as it predetermines the level of abstraction and the scope of possible patterns that can be found. The method also lacks the possibility of utilizing information on the relations between specific participants or artifacts. The latter lend themselves to analysis with a network perspective.

Social network analysis (SNA) has been used in various areas, including computer-supported collaborative learning (Aviv et al., 2003; de Laat et al., 2007; Harrer, Malzahn, Zeini & Hoppe, 2007; Reffay & Chanier, 2002). The basic approach relies on representing communication events as links between the actors in the network. Applied to networks of knowledge artifacts on the Web, SNA can be an efficient approach to knowledge and its collaborative development by analyzing the meaningful structure of connections between knowledge artifacts (see Chapter 2). The resulting network structure will very much depend on the time span during which these events are collected (Zeini, Göhnert, & Hoppe, 2012). However, the target representation no longer represents temporal characteristics. For this reason, SNA has been criticized for eliminating time. Although advances are being made to analyze the development of networks, these rarely address true network dynamics. Process temporality represents a major dimension of online learning and should not be ignored in an analysis. In this chapter I present a network analysis technique that can explicitly address learning dynamics in the context of an open learning community.

Analytical approaches to knowledge development

Actor-artifact networks

The knowledge-building process develops around the creation of knowledge artifacts. A specific version of a so-called two-mode-network can be built on the basis of the relation between the actor (or author) and the artifact (or product). In the SNA methodology (Wasserman & Faust, 1994), such two-mode networks are also called affiliation networks. In the pure form, these networks are assumed to be bi-partite, that is, only links alternating between actor-artifact (“created/modified”) and artifact-actor (“created/modified-by”) would be allowed. Using simple matrix operations, such bi-partite two-mode networks can be

"folded" into homogeneous (one-mode) networks. Here, for example, two actors would be associated if they have acted upon the same artifact (Suthers & Rosen, 2011). The relation between the actors was mediated by the artifact. A typical example of such a transformation is offered by co-publication networks based on co-authorship. Similarly, one can derive relationships between artifacts by considering agents (engaged in the creation of two different artifacts) as mediators.

The "pure" view of actor-artifact relations as bi-partite networks has a clear mathematical-operational structure. However, there are good reasons to extend this approach: Both actors and artifacts may be interrelated in other ways than by this type of cross-wise mediation. For instance, social relations between actors may operate independently of the artifact mediation. Semantic relations may be salient between knowledge artifacts, as in the "semantic web". Mika (2007) was one of the first to elaborate on methods and potential gains of blending social and semantic network structures. Other approaches allocate actors and artifacts on different layers of a multi-layer model with homogeneous relation within each layer and heterogeneous relations in between (Reinhardt, Moi, & Varlemann, 2009; Suthers & Rosen, 2011). Such multi-relational representations may appear superior in expressiveness; however, operations in such structures are more difficult to define.

As with any other network representation, actor-artifact networks also fail to capture the notion of time explicitly. However, time may be implicitly modeled in the network relations. In the scientometric field, this is the case for citation networks: If publication X cites publication Y, I can safely assume that Y is older than X. The ensuing network structure does not contain cycles (excluding specific rare cases of cross-citation). The main path analysis method builds on such acyclic citation networks and can also be adapted to the dynamics of networks of knowledge artifacts built in the process of online collaborative learning.

Main path analysis

The main path analysis (Hummon & Doreian, 1989) is a network analysis technique for the scientometric study of scientific citations over a period of time. Its major application is the identification of key publications in the development of a scientific field. While many scientometric methods, such as the analyses of co-citation and co-authorship networks, stress the semantic structure of scientific work, main path analysis additionally considers the temporal structure of development. Temporality is accounted for through the very definition

of a *directed acyclic graph* (DAG) where nodes are single publications and directed edges represent citations between publications. The direction of an edge corresponds to the flow of knowledge from the cited and older publication to the citing and newer publication. Therefore, these links incorporate both the dimension of content relations and the temporal order of the contributions.

A DAG always contains at least one node with no ingoing edges (i.e., a source) and at least one node with no outgoing edges (i.e., a sink). In the citation network of scientific publications within one field, often one important publication is chosen as a starting point for the development of the field. This publication represents the first source. Later on, other sources that may not have cited previous publications in the field can become prominent and highly cited. Sink nodes, in contrast, represent either unimportant or very new publications that have not been cited at the time of analysis.

The main path can be described as the most used path among all possible paths of successive edges from the source nodes to the sink nodes in a citation network. This most used path can be found by a two-step procedure: first, the traversal counts for each edge are calculated as the number of different paths between each source and sink nodes that go through this edge and, second, an algorithm is used to identify the main path based on the edge traversal counts.

This chapter employs the search path count (SPC) algorithm (Batagelj, 2003), which introduces one fictitious source node and one fictitious sink node and links these to each of the actual source and sink nodes, respectively. In the example in Figure 4.1 the fictitious source and sink nodes are 1 and 10. Their only purpose is to simplify the original procedure (Hummon & Doreian, 1989) of weight calculation for the edges connecting the real nodes. Starting at the fictitious source node (1), the main path is identified by successively following the edge with the maximal weight to the next node until the fictitious sink node (10) is reached. At node 7 in Figure 4.1, there are two possible alternatives to reach the next node, because both outgoing edges have the same traversal weight. In this case the main path branches.

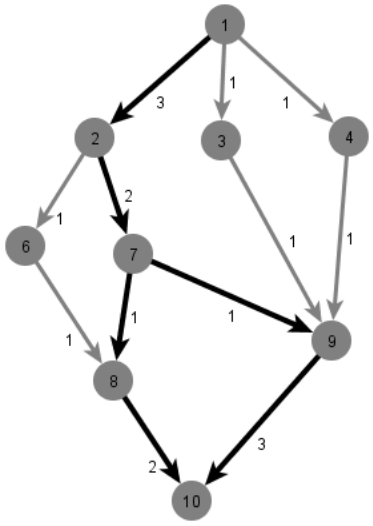

Figure 4.1 Example of a main path calculation.

The SPC algorithm might present too strict an approach to the idea of main path, depending on the nature of the graph. For the case when the analysis requires a broader view on the main contributions in a field, Liu and Lu (2012) suggested lowering the search constraint by defining a threshold. In each step one chooses not only the edges with the maximum weight but also edges with weight above a certain percentage of the maximum weight. In the present study, I applied a slightly modified procedure to identify the *multiple* main paths (Liu & Lu, 2012): After calculating the traversal weight of each node, I considered all the nodes with a weight above a certain threshold as part of the multiple main paths. This strategy facilitates the identification of multiple main paths of important but thematically disparate contributions that may not necessarily build one connected component.

Methods related to the main path analysis represent a structural approach that is appropriate for addressing the dynamics of online community learning. Depending on the nature of hyperlinks, a DAG may trace the flow of influences between ideas or the change in meanings that accompanies knowledge development. The technique allows identifying the most influential contributions and their authors in the course of the construction of a community knowledge base over time. It also facilitates the characterization of the overall discourse trajectory in collaborative learning (Halatchliyski et al., 2012).

Empirical study

The context of the Wikiversity data

Wikiversity is an online learning environment operating on a wiki technology since 2006. Like its larger and older sister projects Wikipedia and Wikibooks, Wikiversity is offered in many languages and directed at any Web user. It is not designed as the online version of an academic organization providing courses or exam certificates. It is rather an experimental open space for collaborative learning to be used by any groups of participants according to their learning goals. A major feature is the openness of the created artifacts and of the community practices to accept constructive suggestions and participation by any interested user. Thus, Wikiversity follows a learner-centered approach (Bonk & Cunningham, 1998).

As a constantly developing so-called *open learning community*, Wikiversity accumulates a rather diverse body of many types of learning resources that are loosely structured in scientific topics from accounting to zoology. The pages categorized under any one Wikiversity category are often set up by different users and may serve different purposes. There are separate articles but also pages connected as bigger projects or organized as courses. Nevertheless, there are often hyperlink interconnections between the different pages and contributors often join multiple projects, sometimes years after their initial start. Because of the openness, there is a great variety of participation modes within and between the different topic categories.

The development of participation is an essential part of the learning process for users. In fact, users who become more involved with the community extend their participation to many unrelated scientific topics. Even when experienced users stay within the borders of one scientific category, their contributions increasingly follow the dynamics of the shared online environment and go beyond the starting individual goals. Such possible starting goals might be, for example, the arrangement of materials for a clearly delineated course as a teacher or the participation in such a course as a student, often in connection with offline lectures in parallel. Similar scenarios of online learning and teaching in Wikiversity do occur but are not representative of the idea that the community envisions, because this form of participation is not particularly collaborative. In the long run, the learning of individuals should become interconnected, producing an interwoven socio-epistemic fabric of a community that is constantly open to new constructive contributions.

Because of the non-homogeneous learning practices and artifacts, the Wikiversity data represents a real challenge for a learning analytics specialist. In the following, I present my approach for discerning major patterns of learning activities and profiles of contributing participants.

Extraction and preparation of wiki data

As already mentioned, the main path analysis was originally developed as a method to investigate the main discourse structure of scientific fields, using networks of publications linked by citations. However, the analysis method is not restricted to this field of application. The first author and colleagues have already demonstrated how it can be applied in the educational context of computer-supported classroom discussions (Halatchliyski et al., 2012). Moreover, it can be applied to any kind of directed acyclic graph (DAG). In this chapter I show how to employ the main path analysis approach to examine the development of interconnected learning resources related to a knowledge domain in the context of a wiki environment.

All analyses presented in this chapter are based on data from an official dump file[7] of the English Wikiversity from February 20, 2012. I did not use the complete wiki data but employed the concept of MediaWiki[8] categories in order to identify the body of artifacts related to a specific knowledge domain. Each wiki page can be categorized under one or more headings. The categories are themselves structured into subcategories. The actual data gathering process usually starts with extracting the complete subcategory structure by following the hierarchy starting at a given top-level category. In a second step all pages that are organized into at least one of the categories found in this structure are identified. It is not mandatory that each wiki page be categorized, but approximately 70 percent of all articles in the English Wikiversity belong to at least one category. Thus, I assume that my procedure yields a representative selection of the major learning resources in a knowledge domain. The chance of considering pages that are unrelated to a domain, which can happen when complete subcategory structures are extracted, also needs to be eliminated. One example is the category “electrical engineering” which contains “Wikiversity” as a subcategory with its large number

[7] http://dumps.wikimedia.org/enwikiversity
[8] http://www.mediawiki.org

of administrative pages that are factually unrelated to electrical engineering. Therefore, a list of subcategories for exclusion from the extraction process needs to be predefined.

As a next step, a directed acyclic graph is constructed, describing the complete flow of knowledge within a single domain in a wiki. Networks of hypermedia resources in a wiki are analogous to networks of publications that are interconnected by citations. Wiki pages can be regarded as publications that are connected by hyperlinks instead of citations. Both citations and hyperlinks indicate a flow of knowledge with a direction from a source (i.e., a cited paper or a hyperlinked page) to a target (i.e., a citing paper or a hyperlinking page).

The temporal stability of publications is crucial for the generation of a DAG from citation networks. Moreover, only works that have already been published can be cited. In contrast to scientific publications and citations featured in their content, which are published once and then remain static from that point on, wiki pages evolve over time under the collaborative efforts of community members. Furthermore, it is quite natural that one wiki page is hyperlinked to a second page and, at the same time, the second page links back to the first one, thus introducing a cycle. In order to overcome these problems I used the Wikiversity revision logs and the page versions after each revision contained in the dump.

Regarding stability over time, revisions of a wiki page behave like classical publications. They are created (published) at a certain point in time and do not change later on. A change to a wiki page will result in a new revision and thus a modified content of that page but not in a modification of the former revision. This approach suggests using page revisions instead of wiki pages as nodes in a DAG extracted from wiki data. I distinguish between two types of directed edges in such graphs: update edges and hyperlink edges.

Update edges can be introduced between any two directly subsequent revision nodes that belong to the same page. Update edges are directed from the older revision to the newer, updated revision and, thus, represent knowledge flow over the course of the collaborative process on a single wiki page.

Hyperlink edges can be traced between two revision nodes that belong to different pages with a hyperlink pointing from one to the other. A wiki hyperlink almost exclusively points to a page and not to a specific revision and it can be interpreted as an inversely directed knowledge flow, so in the proposed DAG hyperlink edges go in a direction *opposite* to the direction of the hyperlinks in the wiki. A knowledge flow between two wiki pages is elicited at the moment of the hyperlink creation between them. Thus, a hyperlink edge in the DAG

starts at the latest revision of a hyperlinked page relative to the creation time of the relevant hyperlink and points to the first revision of the target page containing that hyperlink.

The described construction procedure results in a two-relational DAG that features update edges between revisions of a single page on the one hand and hyperlink edges between revisions of two related pages on the other hand. The procedure also guarantees that all update and all hyperlink edges are directed from a preceding revision to a succeeding revision in time. An example for such a DAG can be seen in Figure 4.2. In order to visualize the main paths of idea flows in a wiki I use the visual metaphor of a “swim lane” diagram introduced in Figure 4.2. The page titles are shown in the left part of the diagram. All revisions of one page are represented as nodes connected by update edges and ordered in a horizontal line. The update edges of different pages are drawn parallel to one another, forming horizontal “swim lanes”. Hyperlink edges between different pages are depicted as diagonal lines crossing the swim lanes. All edges point from left to right depicting the knowledge flow over time. Time is represented on the horizontal axis along the swim lanes. For any pair of nodes that belong to the same or to different pages, the node closer to the left represents the earlier of the two revisions. Node size reflects the traversal weight of a revision as calculated by the main path analysis. The more important a revision is within the paths of ideas, the larger the node is that represents it.

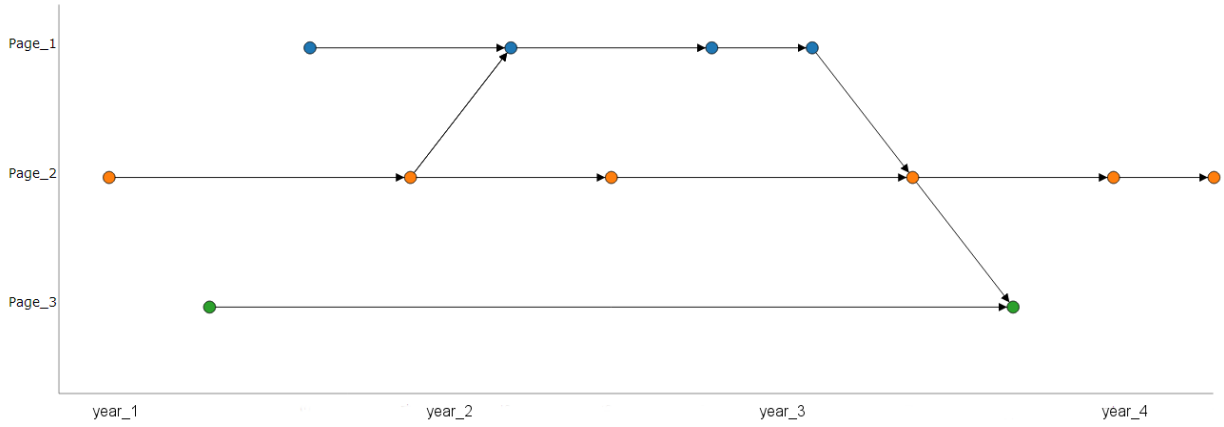

Figure 4.2 Swim lane diagram of a sample DAG of three articles with update and hyperlink edges.

Results of the main path analyses

Using the described method to build a DAG from wiki data, I analyzed the main paths in the two scientific domains *biology* and *electrical engineering* in Wikiversity. Both chosen categories represent well-developed domains in Wikiversity and serve as example datasets of different scales to illustrate my analysis method. Table 4.1 first gives a basic description of the two domains based on the revision logs in the dump.

Table 4.1 Descriptive characteristics of the studied domains.

domain	pages in total	pages on multiple main paths (90%)	pages on main path	edits in total	edits on multiple main paths (90%)	edits on main path	authors in total	authors on multiple main paths (90%)	authors on main path
Biology	1268	58	8	9404	949	111	925	118	6
El. Engin.	398	34	6	4672	442	130	687	103	42

The three data blocks in Table 4.1 contain the number of pages, edits and authors in the chosen Wikiversity categories. Each block shows the total count of each variable, as well as their distribution on the main path according to the employed SPC method and on the multiple main paths with 90 percent threshold (i.e., containing all nodes with a traversal weight above the 90th percentile).

Although the biology domain is much larger than electrical engineering in terms of page count, the latter domain is marked by a proportionally higher number of edits and authors. A clearly higher percentage of the pages in biology seem to be peripheral to the development of this domain. A similar number of authors in biology have produced roughly double the number of edits and pages on the multiple main paths in electrical engineering. This comparison reveals a higher average productivity of the authors on the multiple main paths in the biology domain. From the reverse point of view, this means that the multiple main paths in the biology domain were developed less collaboratively than those in the electrical engineering domain. Lastly, the main path in both domains is of similar length of edits and pages, but in electrical engineering, it is created by proportionally many more authors. Next, I present in detail the main path and the multiple main paths in both domains.

Main paths in the biology domain

The result of the main path analysis with the SPC method is depicted in Figure 4.3 as a swim lane diagram.

The main path consists of pages from an online course on the applications of evolutionary principles that was held in 2009. The articles are well orchestrated, indicating a course syllabus of topics that build on one another. With only six contributors in total (see Table 4.1) and only two of them contributing more than two changes to the pages, the course represents a top-down approach to the design of instructional materials for a relatively passive group of learners. The revision logs reveal that the course materials did not initiate further development of the topic, as only three edits have been made since the second half of 2009, namely to the article on applications in physics (see Figure 4.3).

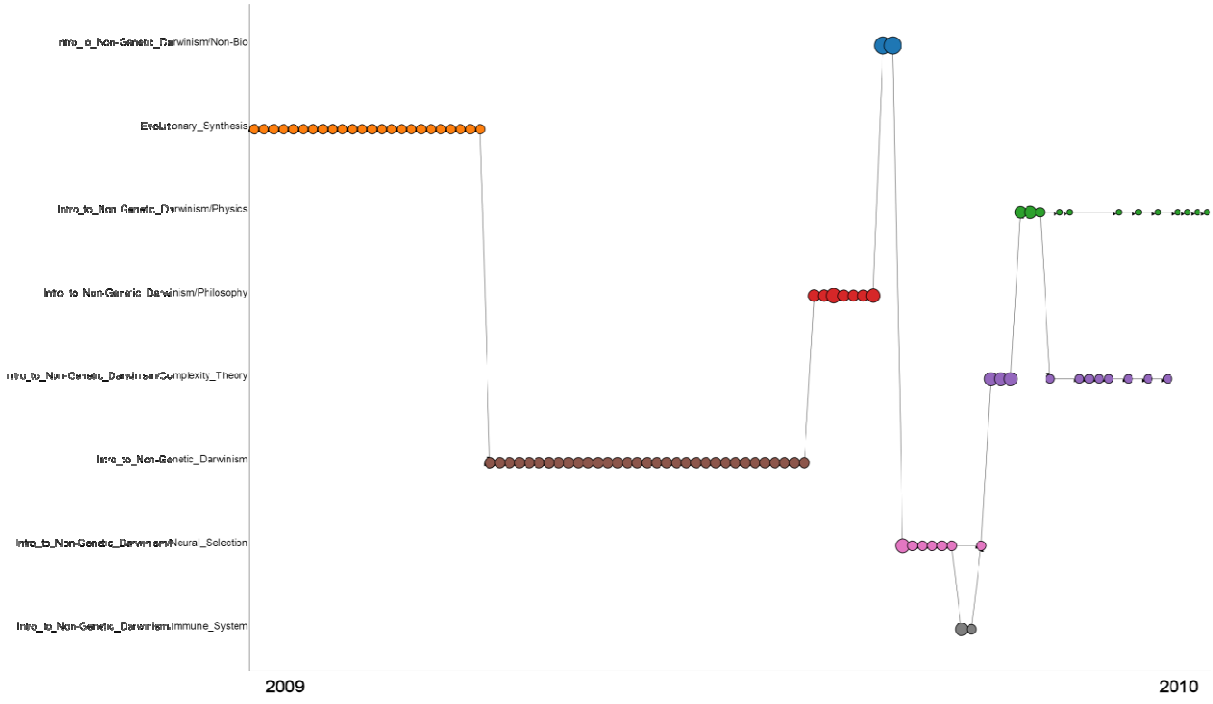

Figure 4.3 Simple main path in the biology domain.

Figure 4.4 Multiple main paths in the biology domain.

In order to broaden the range of important topics in the further analysis of the biology domain, I identified the multiple main paths as explained in previous sections of this chapter. Figure 4.4 shows the resulting swim lane diagram with additional branches of nodes and edges. Only ten percent (90th percentile threshold) of the article revisions with the highest traversal weight appear as part of the multiple main paths. Among them are all revisions presented as the main path in Figure 4.4.

Besides the discussed main path of the online course on evolutionary principles, several other topics appear as new separate branches: a cluster on sustainability and renewable energy from 2007 and 2008; two pages from a course on complex systems from 2011; an article about gynecological interviews gradually developed from 2007 to 2011; a small cluster on UFO research from 2006 and 2007; a larger and long-spanning cluster containing well-developed learning project pages about vitalism and consciousness, RNA interference, stem cells, life origins, human genetics, dominant group and the connected basic biological concepts. Both branches containing the topics of vitalism and human genetics were first developed independently and later on flowed into the larger cluster. The main trajectory of that cluster starts with the topics RNA interference and cell improvement and ends with the topic dominant group.

The overall picture of the learning process in this domain suggests a heterogeneous evolution of ideas organized into separate topics. This conforms to the picture of groups of learners that followed different clearly defined interests in biology with little inter-group collaboration, except for the larger cluster of projects building on basic shared learning resources such as the general article on biology. The biology domain seems representative for the diverse and partly disconnected culture of online learning in the whole Wikiversity community.

Main paths in the electrical engineering domain

Figure 4.5 shows the swim lane diagram of the SPC main path in the domain.

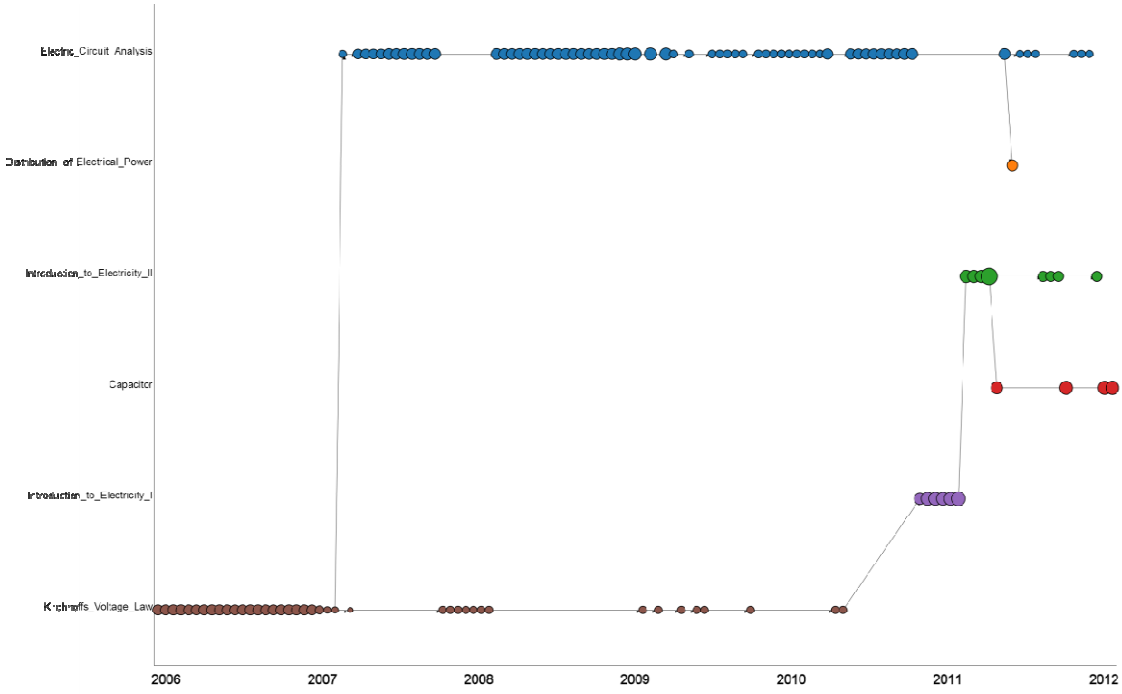

Figure 4.5 Simple main path in the electrical engineering domain

As with the main path in the biology domain, the core of the main path is the main page of an online course on electric circuits. In contrast to the course on evolutionary principles, this electric engineering course has been developed over a longer period from 2007 to 2009 and thus goes beyond the format of a course in the formal educational sense. The main path also contains an older resource from 2006 about voltage law that was later included in the course

syllabus, as well as newer introductory resources on electricity from 2010 and 2011 that also referred to voltage law.

The interconnected and well-maintained articles indicate the core and narrowly interrelated topics in the domain. The creation of these core materials is an example of a truly collaborative learning process with many participating contributors (42 authors as shown in Table 4.1) over longer periods of time. The produced materials are structured as courses in order to facilitate any passive user encountering the topic, but the interesting learning process of the community of contributors is manifested in the collaborative creation of the study material itself.

As in the biology domain, I took a detailed look into the broader range of important topics in electrical engineering by analyzing the multiple main paths traced by ten percent of the article revisions with the highest traversal weight (90th percentile threshold). Figure 4.6 shows the resulting swim lane diagram that contains several new branches and additional nodes besides all the edits on the main path from Figure 4.5.

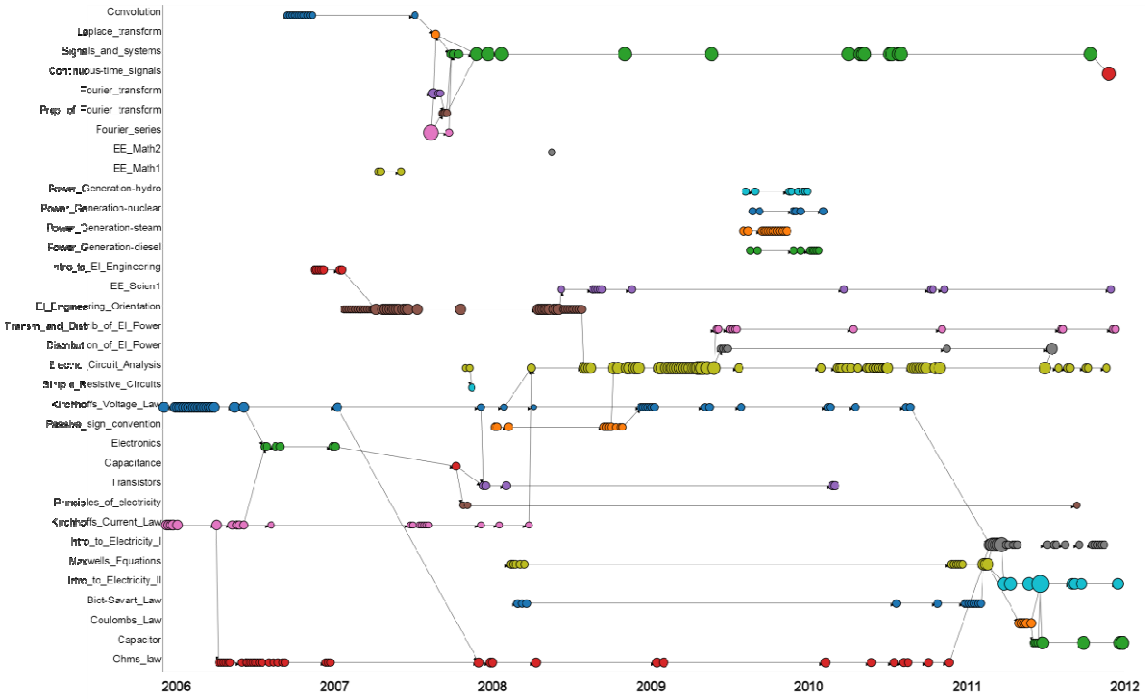

Figure 4.6 Multiple main paths in the electrical engineering domain.

A new cluster of pages from 2006 and 2007 appearing now on the main paths covers the topic of signals and systems. The remaining separate pages of the main paths relate to mathematical tests and to a course on electrical power generation from 2008, all written by the same single author, as indicated by the revision logs.

The core cluster of the discussed main path now consists of many new articles covering basic electrical laws. On the main paths also appear pages from other topics structured as courses: on orientation to the domain and on transmission and distribution of electrical power. The important position in the DAG of the electric circuits topic in between the early orientation to electrical engineering and the later introduction to electricity explains why it is part of the simple main path in Figure 4.5. Although the enlarged core cluster consists of different courses and groups of topics, I found strong cross-participation of contributors across the pages in the cluster as I consulted the revision logs. In addition to the pages being thematically close, the cross-collaboration of authors presents an additional reason for the emergence of this large connected cluster.

Overall, this study showed that electrical engineering was a more compact and coherent domain than biology in the Wikiversity community. Many contributors collaborated over longer periods of time and a large number of pages, creating highly interrelated learning resources. Thus, materials organized as online courses were authored by a large number of people and serve general interests instead of that of a limited number of students for a limited period of time. The electrical engineering domain is an example of a self-organized learning community with enough time to build collaborative structures of practices and artifacts. Evaluated by main path analysis, the development resulted in more tightly interwoven topics than in the biology domain. Overall, the method revealed one large cluster of articles in both domains, as well as a few smaller ones, representing the core knowledge in those domains. This method allows for a subsequent analysis of the development of the topics over time and of the distribution of participation of their authors.

Author profiles and roles

After the overview of the main paths in the two domains I turn to the analysis of the authors contributing to pages off as well as on the main paths. Here, I used the main path analysis results in combination with the revision logs in the dump. As already mentioned, Wikiversity is an open virtual space and so there is no standard guideline on how authors should interact

and use the environment. However, my data revealed differences in the contribution activity profiles of authors that can be interpreted in terms of a division of roles in the process of collaborative learning in a Wikiversity knowledge domain. I started by calculating for each author the number of edits and different edited pages and focused on the profiles of prominent authors who stood out among the large group of low contributors. Forty-six percent of the authors in the biology domain and 51% of the authors in the electrical engineering domain had minimal participation, just making a single edit without hyperlinks in the DAG. Respectively, 30% and 27% of the authors in the two domains who had at least one contribution on the multiple main paths did not make any other contribution. This highly skewed distribution of participation in online environments is a well-documented fact (Rafaeli & Ariel, 2008). More specifically, I see that the authors that have a contribution on the main paths are generally less likely to make only a single contribution. According to this evidence, main path contributions can be interpreted to indicate high involvement in the community.

According to my interpretations of the profiles of active authors, I identified several categories of contributors: first, the role of *specialists*, who made many edits to only one or a few pages; second, the role of *maintainers* with a relatively high number of edited pages and a relatively low number of edits; third, the role of *leaders* with an outstanding number of edits and edited pages. As I show in the following, the interpretation of these roles was only accurate after taking the results of the main path analysis into account.

The investigated articles, and thus the contributions to them, are not of equal importance to the collaborative learning process of the community. Many articles are short stubs not interlinked with any other articles within the corresponding category. Such isolated and largely unimportant articles are not part of the main paths in a domain. Therefore, the results of the main path analyses in both domains of the study can enhance the analysis of the author roles by qualifying the number of contributions that lie on the main paths. As mentioned above, the SPC method of identifying a single main path leads to a strong focus on a small number of revisions and articles on a narrow topic. Hence, in this chapter the author profiles are related to the extracted multiple main paths described in the previous sections of this chapter. Using the main path analysis in this way, a more adequate view on activity and division of roles of authors is achieved.

Author roles in biology

The three analyzed author roles in the biology domain are presented in the rows of Table 4.2 through the contribution profiles of distinctive sample authors. Each role is subdivided into type A and type B according to whether any of the contributions of an author are part of the main paths. The author activity in total and on the main paths is grouped in blocks containing the number of edits, edited pages and edits with hyperlinks. As explained in previous sections of this chapter, hyperlinks represent knowledge flows between pages. Thus, the edits introducing a hyperlink and the edits referred through a hyperlink by another edit are important and should be regarded separately.

Table 4.2 Sample authors with a distinct role in the biology domain.

author profile	author ID	edits in total	edits on multiple main paths	pages in total	pages on multiple main paths	hyperlinked/ / hyperlinking edits	edits with links on multiple main paths
specialist A	278565	468	0	1	0	0 / 0	0 / 0
specialist B	348476	10	10	1	1	0 / 0	0 / 0
maintainer A	9357	35	0	31	0	0 / 0	0 / 0
maintainer B	21778	43	9	41	8	0 / 1	0 / 0
leader A	263421	1966	0	729	0	0 / 0	0 / 0
leader B	20	552	154	112	20	31 / 35	25 / 20

The first rows, the specialist A with ID 278565 has the third highest number of edits in the domain, but these edits were all made to the same single page, moreover, none of them is part of the multiple main paths. This example shows that output quantity — the number of contributions does not necessarily correspond to output quality — the importance for the evolution of discourse in a Wikiversity knowledge domain. The example of author 348476 adds to this finding. With ten edits in the domain in total, this is the most prolific author among the type B specialists — authors who are specialized in one single page and have at least one edit on the main paths. The low rate of activity of such specialists with important contributions would normally suggest that they should be regarded as low contributors. In the next rows, the type A and B maintainers 9357 and 21778 similarly show a low to middle rate of contribution. Maintainers mostly make small formal changes that are unrelated to the

content of the edited Wikiversity pages. They correct spelling mistakes, organize the categorization and sometimes also set hyperlinks, as does author 21778. Such authors typically contribute to very different domains and topics at the same time. Most of their contributions that appear on the main paths can be regarded as coincidental as they fall within a chain of important updates of the page content made by other authors. Table 4.2 further shows that the most prolific contributor and a type A leader in the biology domain, author 263421, didn't make a single important contribution on the main paths. A closer look into the data revealed that this author used Wikiversity to build a database on specific genes. This voluminous project was not much related to the other core topics in biology. Type B leaders such as author 20, whose edits sometimes appear on the main paths, seem to play the most important role in the domain. Besides having the highest number of contributions on the main paths, this author also has the highest number of edits with hyperlinks. Further analyses of the data showed that authors with edits on the main paths tend to have more contributions and especially more interlinked edits than authors without edits on the main paths. Indeed, by the design of the method itself, hyperlinked and hyperlinking edits are more likely to occur on the main paths.

Author roles in electrical engineering

Table 4.3 presents the analysis of author roles in the electrical engineering domain following the structure of Table 4.2.

Table 4.3 Sample authors with a distinct role in the electrical engineering domain.

author profile	author ID	edits in total	edits on multiple main paths	pages in total	pages on multiple main paths	hyperlinked/ / hyperlinking edits	edits with links on multiple main paths
specialist A	858	44	0	1	0	0 / 0	0 / 0
specialist B	292570	6	6	1	1	0 / 0	0 / 0
maintainer A	3705	19	0	17	0	0 / 0	0 / 0
maintainer B	8437	34	8	27	4	0 / 0	0 / 0
leader A	32	245	0	75	0	1 / 0	0 / 0
leader B	19038	867	114	133	14	20 / 35	8 / 8

As argued above, the two domains are marked by a number of differences. Nevertheless, the studied author roles are identifiable in the same way in both domains, so the inferences about the authors in biology made in the previous subsection also apply for the authors in electrical engineering. The only difference worth mentioning is that author 19038, a type B leader in Table 4.3, has the highest number of contributions among all authors in the domain and at the same time has contributed the highest number of edits on the main paths. This case still corresponds to the conclusion that important authors are distinguished not just by a high number of edits but also by significant contributions appearing on the main paths.

Technological implementation

The analysis processes described in this chapter have been integrated into the network analytics workbench of my coauthors (Göhnert, Harrer, Hecking, & Hoppe, 2013). A form of this workbench was used in the recent EU project "SISOB"[9], which had the goal of measuring the influence of science on society based on the analysis of (social) networks of researchers and created artifacts. One area of research in this project was *knowledge sharing*. Thus the analysis techniques based on main path analysis presented in this chapter were also of essential value in the project context.

I conceive workbenches as a general type of software environment designed to serve active and skilled users, without assuming the users to be computer experts. I have decided to develop a network analytics workbench as a web-based environment for several reasons, such as ease of deployment, access and update, and independence of the local computing facilities and devices. An important part of my experience with network analysis and network analysis tools is the need to combine several tools even for a single analysis process. The use of several tools sometimes also results in the need for conversion between the different data formats used by these tools. Therefore one important goal behind the development of the network analytics workbench is the integration of multiple tools and conversion mechanisms into one interface.

[9] http://sisob.lcc.uma.es

The workbench provides readily available processing chains for known use cases and furthermore allows for setting up new ones. The user interface (UI) is built upon a pipes-and-filters metaphor for processing chains in order to reduce the complexity of the underlying system for users who are not computer experts. An example of the UI that has been created using the WireIt[10] JavaScript library can be seen in Figure 4.7. In using the pipes-and-filters metaphor and being web-based, the workbench is similar to mashup projects like YAHOO pipes[11].

In contrast to these projects, the actual processing of data in the workbench is not part of the user interface code itself but is done by a multi-agent system controlled by the workbench. The multi-agent system approach allows for combining several mostly independent tools into one workflow. These tools can be either pre-existing or newly developed. Examples of existing tools that have been successfully integrated into the workbench are the network text analysis tool AutoMap (Diesner & Carley, 2005), the network analysis tool Pajek (Batagelj & Mrvar, 1998) and a wrapper for the R language[12]. Examples for newly developed components are a MediaWiki extraction component based on the mechanism presented in this chapter and a main path analysis filter also used for the analyses presented in this chapter. The communication between the web-based user interface and the agents is based on the SQLSpaces (Weinbrenner, Giemza, & Hoppe, 2007), an implementation of the tuple space architecture (Gelernter, 1985). From the user interface a description of the constructed workflow is posted into the SQLSpaces server, which contains a message for each agent (filter) type that is part of the workflow. These messages contain information about the input data and the parameter configuration of that filter.

Figure 4.7 shows one of the workflows used for the analyses described in this chapter. The first filter is used to provide input for the following filters. In this case the filter connects to a MediaWiki database with Wikiversity data and creates a DAG for a given category from it. The extraction process follows the approach outlined in the previous sections of this chapter. The filter accepts two parameters. The name of the category for which the DAG should be extracted is a mandatory parameter. The second parameter accepts a list of categories to be excluded from the search and is optional. The next filter in the workflow presented here just duplicates all input into two parallel outputs. Thus, it allows for performing different analyses on the same possibly preprocessed input data in one workflow. In this example the two

[10] http://neyric.github.com/wireit/docs
[11] http://pipes.yahoo.com/pipes
[12] http://www.r-project.org

outputs are used to perform main path analysis and analysis of author profiles in the same category of a wiki, as presented in the result sections of this chapter. On the left side, the Main Path Analysis filter allows for selecting a weighting scheme to be used in the main path analysis and for defining a threshold for the multiple main path analysis. The results of this filter are then visualized using the swim lane metaphor also used throughout this chapter. The other branch of the duplicator leads into the Main Path Role Assignment filter, which generates Tables 4.2 and 4.3 used for the author profile analysis. These tables are then fed into the Result Downloader, which allows for downloading these results onto the local machine for further usage.

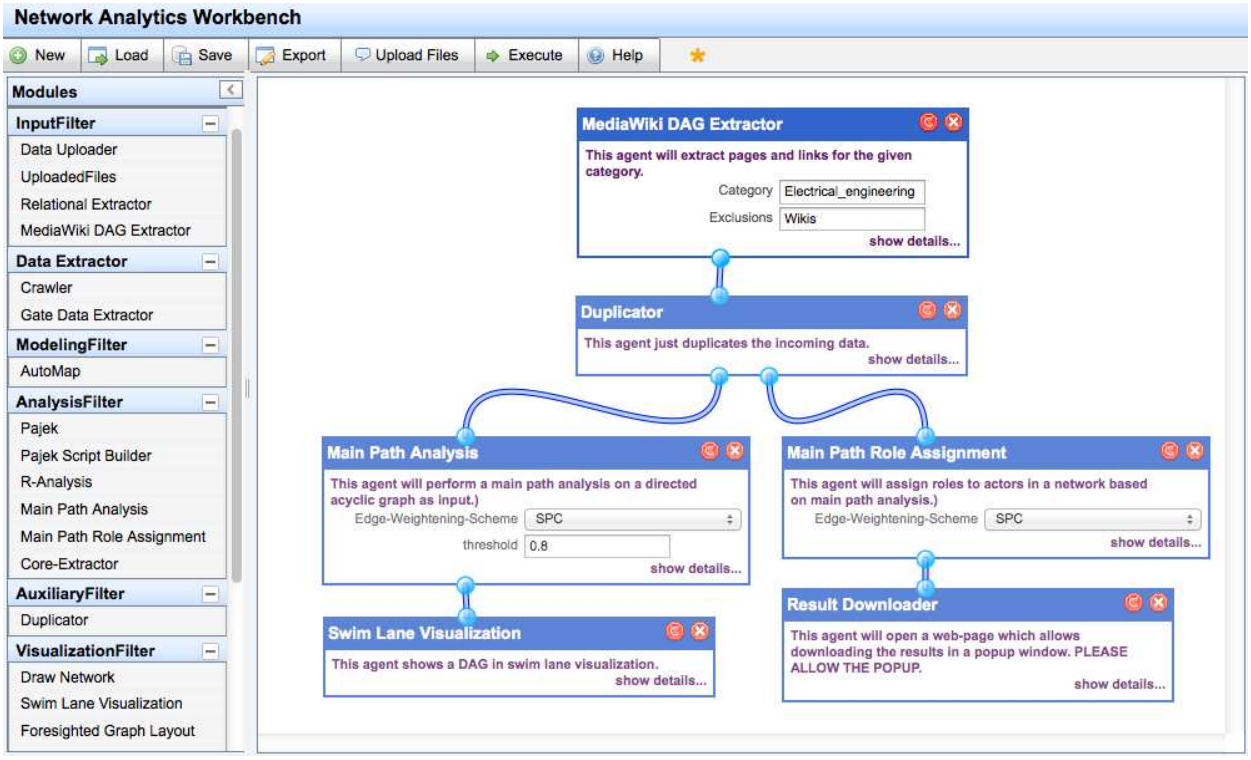

Figure 4.7 Screenshot of the network analytics workbench.

Conclusion

With the help of the main path analysis I detected the core topics in the two Wikiversity domains of biology and electrical engineering. While biology had much broader scope, the collaboration of the authors was weaker. The resulting main paths had a similar size and structure to the main paths in electrical engineering, which was a small coherent domain with

a relatively large group of authors and a higher necessity for collaboration. Thus, the small ratio of main path versus other articles in the biology domain compared to the electrical engineering domain could be explained through differences in the level of collaboration among the authors revealed by the revision logs.

The exemplary results of the presented empirical study may be useful for the Wikiversity community as a whole. As it seems, some scientific domains like biology might benefit from strengthening of collaboration. Additional analyses may be helpful to choose appropriate directions for development, but my results point to the need for better coordination of the disparate topics in this domain. The main path analysis can also orient participants by showing them the importance of the topic they are working on. It can also reveal important reference points to other core topics in the field. A beginning contributor can be aided by a presentation of the main paths with the decision to add to an existing strand of knowledge development or to start a new peripheral one. An advanced participant in the community may benefit from the analysis as a historical reconstruction of the shared knowledge-building process, in order to compare his or her own visions and goals with the actual knowledge development of the community and to discover topical gaps necessitating further efforts. With some additional work to adapt and standardize the analysis and the necessary interventions relative to the specific goals within an educational context, the main path analysis can be used to support and even take the load off a teacher or coordinator of knowledge building.

Our approach presented in this chapter is the first application of scientometric methodology for analyzing the flow of ideas in the context of an open learning wiki environment. Using the examples of the biology and electrical engineering domains in Wikiversity, I showed how main path analysis can be employed to analyze the collaborative creation of various knowledge artifacts and the learning processes of the online community. My methods have been embedded into a web-based analytics workbench that supports the definition and re-use of analysis modules in a user-friendly visual environment.

The chapter presented a procedure for creating directed acyclic graphs from wiki data and for illustrating the obtained main idea flows in swim lane diagrams. The employed visualization technique allows for a unified view of knowledge flows in a network of artifacts with multiple relationships. The main path analysis results were helpful in understanding the differences in the collaborative structure of two scientific domains in Wikiversity. The results further facilitated the characterization of different roles that authors have in the community. I found that the total rate of contribution was not a sufficient criterion for identifying the most

important authors in a domain. But, as the role of maintainers demonstrates, some contributions on the main paths may also not testify to the importance of an author. Instead, the total number of contributions should be evaluated in combination with the number of contributions that appear on the main paths.

For my future work, I plan to elaborate on the characterization of contributions and contributors with respect to the main paths of development in other educational knowledge-building scenarios. It appears promising to provide moderators, teachers, tutors, or the productive teams themselves with results of such analyses, in order to support reflective practices (Schön, 1983). This will raise further challenges regarding visualization and cognitive ergonomics.

Interlude

The study presented in this chapter showed how to adapt the scientometric method of main path analysis to the evaluative investigation of online mass collaboration in the knowledge-building community Wikiversity. Beyond the cross-sectional and longitudinal analyses of static article networks presented in the previous Chapter 2 and 3, the current approach takes the temporal sequence of each contribution to the articles and their meaningful interconnections into account. The evaluation focuses not on the collaborative artifacts as whole units but on each single contribution to them. As pivotal contributions are identified those that lie on the main paths of the evolving knowledge in a specific domain. These can be the core topics and ideas or other important moments of collaboration for the studied time interval.

The employed method bears high potential for a real-time evaluation of collaborative processes that can be used for supportive interventions by moderators or teachers, or for self-regulative purposes by the community as a whole or by single contributors. The examples in this chapter demonstrate only some of the possible aspects that can be explored such as topical coherence of the contributions, structure and intensity of collaboration, topical gaps that present contribution opportunities, important roles of contributors. This is definitely a fertile field for future research in learning analytics.

The empirical part of this dissertation concludes with the dynamic network perspective on mass collaboration presented in the current chapter. The three studies demonstrated different network analysis approaches to general mechanisms and practical questions of the structure and dynamics of collective knowledge in online communities. In Chapter 5, a general discussion of the theoretical, methodological and empirical contributions of the current dissertation will be provided.

Chapter 5

General Discussion

The aim of this dissertation was to advance an approach for studying and understanding the principles that underlie knowledge development during mass collaboration. I used different network analysis techniques to model the interplay between structure and dynamics of collective knowledge while taking the contribution activity of participants in the online communities Wikipedia and Wikiversity. The resulting quantitative models allowed hypothesis-based statistical tests of the relation between pivotal artifacts and contributions of experienced authors. Longitudinal analysis enabled causal interpretation of the impact of pivotal contributions on the subsequent development of knowledge. Finally, a network analysis of the main paths of knowledge development was shown to provide fine-grained and immediate evaluation of the pivotal contributions from a temporal perspective on the collaborative process.

This chapter highlights the main aspects that emerge from the synopsis of the entire dissertation. I will first summarize the main empirical and theoretical aspects of the studies in the order of their presentation in the preceding chapters. Then, the strengths and limitations of my approach will be discussed. I will finally derive the major implications for future research and practice and will provide an encompassing conclusion.

Summary of the main findings

The knowledge base in Wikipedia consists of networks of hyperlinked articles categorized in different knowledge domains. Chapter 2 showed how these networks can be analyzed as static structures in order to identify articles with outstanding topological position. Such articles were called pivotal articles. For one thing, these were the central articles within a specific knowledge domain. For another, pivotal articles were also the boundary-crossing articles across two knowledge domains. By thus modeling the structural representation of knowledge in a mass collaboration environment, I incorporated a second level of analysis considering the authors contributing this knowledge. In this way, my integrative theoretical perspective on collective knowledge both as substance (i.e., collaborative artifacts) and as participatory activity (i.e., collaborative contributions) was employed in the empirical study. The most remarkable result was the significant relationship between authors' experience and their contribution to pivotal articles. Authors mainly gained experience in the community by contributing to different articles. There was also evidence of a division of labor, as authors with experience in only one of the studied domains predominantly contributed to central

articles within this domain, and authors with experience in both domains predominantly contributed to boundary-crossing articles across the domains.

Mass collaboration *communities such as Wikipedia build knowledge bases with a complex structure. The pivotal elements in this structure heavily depend on the contributions of experienced members of the community. Designing sophisticated mechanisms to stimulate repeated contributions to different artifacts is of vital importance for a sustained mass collaboration.*

Building on the structural approach to the knowledge base in Wikpedia employed in the previous chapter, the study presented in Chapter 3 investigated a generative mechanism of knowledge dynamics over a period of six years. I used established network analysis metrics to identify the pivotal articles in each periodic snapshot of the studied networks. With the help of powerful, longitudinal, multilevel models I was able to prove that pivotal articles are significantly more likely than other articles to link to the new knowledge that appeared in subsequent periods in a network. New knowledge in Wikipedia was measured as the number of new articles as neighbors, the change in the total sum of edits of the neighbors and the number of new received contributions. Thus, articles that are pivotal within the static organization of knowledge are also pivotal for its dynamic development. I further showed that the German Wikipedia has entered a saturation stage of non-exponential growth.

Embracing the *challenge of understanding the dynamics of collective knowledge in mass collaboration communities, a structural analysis can provide valuable insights. Knowledge structure and dynamics are in a constant interaction with each other.*

The original application of the scientometric method main path analysis to the knowledge base in Wikiversity portrayed in Chapter 4 further ways of grasping the temporal dimension of mass collaboration. Considering the complexity of interactions between many participants and artifacts, knowledge processes essentially develop over longer periods of time and go along with a continuous change of the shared knowledge base. The temporal sequence and the relations between these changes can be analyzed avoiding any biasing aggregation in order to identify pivotal contributions on the main paths of the evolving collective knowledge. With this network analysis technique, I focused on pivotal artifacts in the dynamic sense of building on many preceding contributions and influencing many subsequent contributions. The results allowed also structural comparisons of the studied domains of activity regarding topical coherence and intensity of collaboration. By taking the authors of the pivotal

contributions into account, the main path analysis further facilitated the characterization of different roles in the community. As in the study in Chapter 2, the network analysis results were combined with further data on the participants' activity in order to enhance the interpretation.

Main path analysis, as a network analysis technique that focuses on the dynamics of collective knowledge, provides valuable immediate evaluation of mass collaboration in learning communities. Its results can be an orientation for inexperienced as well as for advanced contributors and facilitators of the process. Thus, I recommend it as a method with good potentials for the emerging field of learning analytics.

Taken together, the current dissertation presents related methodological approaches to the interplay of structure and dynamics of collective knowledge emerging in mass collaboration contexts. Using data on complete knowledge domains in Wikipedia and Wikiversity, my work provides quantitative models of the complex and mutually determining influence of knowledge structures and of contribution activity of participants on the process of knowledge development. In the following sections of this chapter, the strengths and limitations of my research will be critically discussed.

Strengths and limitations

The most distinguishing feature of the presented empirical work in this dissertation is unquestionably the innovative methodological approach. It consists of powerful and state-of-the-art techniques for the analysis of big data, which has currently become easily accessible on the Internet. Although the social network analysis has been used to analyze relationships between learners for quite some time (e.g. Aviv et al., 2003; Cho et al., 2002; de Laat et al., 2007; Reffay & Chanier, 2002), my approach is more precise in defining networks of knowledge artifacts with only very clear type of links between each other such as the hyperlinks. Thus, I borrowed some of the innovative scientometric methods developed for studying scientific work and applied them to mass collaboration artifacts in combination with data on contribution activity of community participants. The three studies in the previous chapters all illustrate a different analysis design: cross-sectional in Chapter 2, longitudinal in Chapter 3 and continuous time in Chapter 4. This variety speaks of the potentials for

employing network analysis techniques beyond the widespread way of descriptive statistics and visualizations.

Considering the contributions of this work to theory, the interdisciplinary grounding of the research is a clear strength. Although it was not a specific goal of the dissertation to develop existing theory further, the boundary spanning character of the analyzed questions allowed me to contribute relevant connections between theoretical perspectives. CSCL research could, for example, benefit from extending its traditional view on knowledge as a participatory activity by adding a scientometric perspective on knowledge as substance, that is, as created and shared interconnected artifacts. In this way, the mass collaboration phenomenon becomes accessible for research promising valuable insights and extensions for the theories of learning and knowledge-building. My conception of collective knowledge and community processes can be seen as a fruitful, albeit distant contribution to the cognitive psychological perspectives on individual learning and communication in groups (cf. Clark & Brennan, 1991). In order to integrate the different levels and units of analysis, my work employs a complex systems perspective and regards some of the studied phenomena as emergent. It is based on other systemic perspectives (cf. Cress & Kimmerle, 2008) in the learning sciences but also extends them with structural and dynamic aspects.

The three empirical studies are concerned with the mass collaboration phenomenon and investigate in detail knowledge-related questions. My results thus contribute to the understanding of the intangible but presently very relevant concept of knowledge. Collective knowledge demonstrates properties of static substance and can be approached by analyzing shared digital artifacts. However, it is not a collection of pieces of information but emerges at a collective level from the individual contributions. Knowledge has en essentially dynamic nature and can be fully appreciated only by taking the dimension of time into account. Regarding the individual contributors to the knowledge in a community, the results of my work unanimously indicate that the work experience in the community is a highly significant factor and maybe more important than individual knowledge expertise (cf. Oeberst et al., 2014).

Research based on real-life data typically cannot analyze complex phenomena in all their facets. Potentially important factors and relations are left out of the research focus in order to render the investigation manageable. Several distinct aspects have not been considered by the empirical work in this dissertation.

First, mass collaboration interactions in communities certainly have different modalities. Wikipedia and Wikiversity for example afford discussions on the articles (cf. Niederer & van Dijck, 2010), user profile pages (Schwämmlein & Wodzicki, 2012) and personal communication between users next to the main collaboration on the articles. As I was interested in the collective knowledge and its development, I focused on the contribution of content to articles and did not consider the interpersonal communication between the authors. Arguably, even off-topic relations can be informative for understanding the processes within a community and the interrelated development of knowledge. This is however difficult to study in a mass collaboration context, as only part of the communication is logged within the technological environment. Wikipedians, for example, organize personal meetings and use other technology such as IRC to communicate.

Second, my research focus on the created articles does not take their content into account except for identifying the knowledge domains they belong to. My approach was to analyze their structural properties, which is innovative for CSCL research. Most studies of collaborative knowledge building have followed a qualitative approach to the discourse content (Chi, 1997; Gunawardena, Lowe, & Anderson, 1997; Henri, 1992; Sacks, 1992; Stahl, 2006). Coding and interpreting thousands of articles and millions of changes in a mass collaboration context is infeasible, but a combination of content and structural approaches would be clearly desirable to develop for future research.

Third, regarding the contribution measures I used, there is another limitation of my approach. I considered only net additions of content larger that one average sentence in the studies. Small changes as well as deletions of article content were left out, although they might also bear some insight into the development of collective knowledge. The reasons for my decision were that small article changes tackle appearance but not meaning, and that deletions are often motivated by a destructive vandalism. Thus, I sacrificed a broad description of the various modes of participation in the communities for a focus on the large-scale development of collective knowledge.

It should be finally noted that the presented results in this dissertation might have a limited validity in other contexts. I considered only two knowledge domains in Wikipedia and in Wikiversity. Both communities are large enough that there could be different subcultures with own collaborative practices. Also the development of different knowledge domains might vary. The analyzed data cover specific time intervals from the community history. Online

communities as complex systems often have different stages of development in which different mechanisms of knowledge development may be relevant. As both studied communities are based on a wiki technology, there might be some difficulties to transfer the presented research approach to another kind of technologically supported communities.

As knowledge development through mass collaboration is a novel phenomenon, the limitations of the methods used in this dissertation do not impair the relevance of its approach and findings for future research in the learning sciences and for the development of practical applications.

Implications for future research and practice

Due to its innovative focus and approach, the present dissertation opens up new horizons for investigating knowledge development during mass collaboration in future. As this is an emerging field for CSCL research there is not much that has been done yet in this direction. Past CSCL research can certainly be built on in order to identify relevant insights about collaboration in small groups and to test them in mass collaboration contexts. The process of creation and change of shared artifacts seem to represent a relevant focus (Paavola & Hakkarainen, 2009). The significance of small groups as optimal collaborative units might also be traceable in larger communities and networks. Moreover, mass collaboration might yield more beneficial outcomes, not only for the collective knowledge of a community, but also for the learning individual. The most important research goal would be to understand how to optimally use computer technology in order to support individual learning in social contexts as well as collective knowledge development at the community level (cf. Lipponen et al., 2004). In view of the large amounts of big data available for research, the fields of learning analytics and educational data mining has been formed in the last years (Siemens & Baker, 2012; Suthers & Verbert, 2013). They have a pronounced quantitative focus, and the approach presented in this dissertation can be seen as one of the starting points in the area.

Between the macro-level network perspective of knowledge-building communities and the micro level analysis approaches to small-group discourse, there is a broad range of interactions that require innovative analysis approaches. For example, the feedback loop between the contribution of an individual, the subsequent dynamics of the collective knowledge and the repeated contribution of the same individual may reveal unexplored

factors and mechanisms that can be later used for practical support of mass collaboration. It is a challenging and open question how to connect global structural and temporal metrics with the specific individual decisions how and where to contribute and with the micro patterns of conflict and coordination between participants. Further connections with psychological research on motivation and social interaction might be fruitful in this respect. Generally, it would be interesting to understand how the contributions of an individual participant develop over shorter time intervals. The main path analysis presented in Chapter 4 offers a possible start that can be built on. The possibility of stimulating newcomers to gather a diverse contribution experience, which has been shown in Chapters 2 as beneficial for a community, deserves further research (cf. Kraut, Burke, Riedl, & Resnick, 2012).

Besides the interaction between micro and macro levels, the stages of development of a complex system of mass collaboration also deserve systematic attention. The self-organizing processes of formation of rules and practices in a community and their interplay present a further interesting macro detail. In sum, there is a great need for a systematic evaluation of the different possible aspects of collective knowledge as it emerges in present day online contexts. Approaches for grasping structural (i.e., network analysis), interactional (i.e., sequential discourse analysis) and content (i.e., computational linguistics, see, for example, Rosé et al., 2008; Teplovs & Fujita, 2013) dimensions at different levels of analysis should be brought to work in combination (cf. Halatchliyski et al., 2010).

In the learning sciences, as a field that is heavily determined by technological development, research and practice go close together. Therefore, there are several practical implications that can be derived from the present work. Major goals in designing mass collaboration environments might be the attraction of a high number of active participants and the production of highly valuable outcomes. In the present work, I have shown how network analysis can be applied to identify pivotal artifacts in the structure of a collective knowledge base. This information may be used to provide recommendations to experienced as well as inexperienced participants in a community as where they can most suitably contribute. With results from a main path analysis, there are additional opportunities for immediate and differentiated orientation of potential contributors. Besides the difficulty to attract community newcomers (cf. Kraut et al., 2012), my results show that low experienced contributors typically need time and efforts before they understand and adopt the practices of a community. The integration of suitable analysis results into modern awareness tools (cf. Dehler, Bodemer, Buder & Hesse, 2011) could provide invaluable support for the mass

collaboration process rendering higher productivity and more valuable outcomes. Following the idea of formative assessment (Chan & van Aalst, 2004), the individual contributor could also benefit from the improved learning experience.

The integration of formal and informal learning contexts has unofficially started. Even though it may seem improbable or far in the future, the schools of the present-day might completely lose their significance as an institution. On this way to a united knowledge building society (Scardamalia, 2002), the main question may not be how to orchestrate classroom learning with the use of modern tools, but instead how to support the participation in the global knowledge-related mass collaboration.

Conclusion

The contributions of the present dissertation are manifold: Different methodological approaches were developed for the analysis of structural patterns and development processes of collective knowledge in mass collaboration contexts. Large real-life data sets from the online communities Wikipedia and Wikiversity were evaluated using network analysis techniques. The obtained results revealed interaction mechanisms between static structures of knowledge, the dynamics of its further development and the contribution activity of individual participants. The contribution experience in a community has been worked out as an important factor with implications for practical design of mass collaboration environments. The work provides a starting point in the quantitative research field of learning analytics. With its theoretical view on knowledge as substance and as participatory activity based on a complex systems perspective, the work also contributes to the theoretical development of the learning sciences and CSCL in particular.

References

Almind, T. C., & Ingwersen, P. (1997). Informetric analyses on the world wide web: methodological approaches to 'webometrics'. *Journal of Documentation, 53*(4), 404-426.

Aviv, R., Erlich, Z., Ravid, G., & Geva, A. (2003). Network analysis of knowledge construction in asynchronous learning networks. *Journal for Asynchronous Learning Networks, 7,* 1-23.

Barabási, A. L. (2002). Linked: The new science of networks. Perseus, Cambridge, MA.

Barabási, A. L., & Albert, R. (1999). Emergence of scaling in random networks. *Science*, *286*(5439), 509-512.

Barron, B. (2006). Interest and self-sustained learning as catalysts of development: A learning ecology perspective. *Human Development, 49*(4), 193-224.

Bastian M., Heymann S., & Jacomy M. (2009). Gephi: An open source software for exploring and manipulating networks. *Proceedings of the third international conference on weblogs and social media* (pp. 361-362). Menlo Park, CA: AAAI Press.

Batagelj, V. (2003). Efficient algorithms for citation network analysis. arXiv: Computer Science. [http://arxiv.org/abs/cs/0309023]

Batagelj, V., & Mrvar, A. (1998). Pajek: A program for large network analysis. *Connections, 21,* 47-58.

Bates, D., Maechler, M., & Bolker, B. (2013). *lme4: Linear mixed-effects models using S4 classes*. R package version 0.999999-2. [http://cran.r-project.org/src/contrib/Archive/lme4]

Bell, D. (1973). *The coming of post-industrial society: A venture in social forecasting*. New York: Harper Colophon Books.

Bereiter, C. (2002). *Education and mind in the knowledge age.* Hillsdale, NJ: Erlbaum.

Bereiter, C., & Scardamalia, M. (2003). Learning to work creatively with knowledge. In E. De Corte, L. Verschaffel, N. Entwistle, & J. van Merriënboer (Eds.), *Powerful learning environments: Unravelling basic components and dimensions* (pp. 73-78). Oxford: Elsevier Science.

Berger B. L., & Luckmann T. (1966). *The Social Construction of Reality: A Treatise in the Sociology of Knowledge.* New York: Anchor.

Björneborn, L., & Ingwersen, P. (2004). Toward a basic framework for webometrics. *Journal of the American Society for Information Science and Technology, 55*(14), 1216-1227.

Bonacich, P. (1972). Factoring and weighting approaches to status scores and clique identification. *Journal of Mathematical Sociology, 2,* 113-120.

Bonk, C., & Cunningham, D. (1998). Searching for learner-centred, constructivist, and sociocultural components of collaborative educational learning tools. In C. J. Bonk, & K. S. King (Eds.), *Electronic collaborators: Learner-centred technologies for literacy, apprenticeship, and discourse* (pp. 25-50). Mahwah, NJ: Erlbaum.

Börner, K., Boyack, K. W., Milojević, S., & Morris, S. (2012). An introduction to modeling science: Basic model types, key definitions, and a general framework for the comparison of process models. In A.Scharnhorst, K. Borner, & P. van den Besselaar (Eds.), *Models of Science Dynamics. Understanding Complex Systems* (pp. 3-22). Heidelberg: Springer.

Brown, M. S., Collins, A., & Duguid, P. (1989). Situated cognition and the culture of learning. *Educational Researcher, 18,* 32-42.

Bruckman, A. (2006). Learning in online communities. In R. K. Sawyer (Ed.), The Cambridge handbook of the learning sciences (pp. 461–472). New York: Cambridge University Press.

Buriol, L., Castillo, C., Donato, D., Leonardi, S., & Millozzi, S. (2006). Temporal evolution of the wikigraph. In *Proceedings of the 2006 IEEE/WIC/ACM International Conference on Web Intelligence* (pp. 45-51).

Cakir, M., Xhafa, F., Zhou, N., Stahl, G. (2005). Thread-based analysis of patterns of collaborative interaction in chat. In *International conference on AI in Education, Amsterdam, The Netherlands*, (pp. 120-127).

Capocci, A., Servedio, V. D., Colaiori, F., Buriol, L. S., Donato, D., Leonardi, S., & Caldarelli, G. (2006). Preferential attachment in the growth of social networks: The internet encyclopedia Wikipedia. *Physical Review E, 74,* 036116: 1-6.

Castells, M. (2011). *The information age: Economy, society, and culture: Vol. 1. The rise of the network society*. (2nd ed.). Oxford: Wiley Blackwell.

Chan, C. K. K. & van Aalst, J. (2004). Learning, assessment and collaboration in computer-supported environments. In J. W. Strijbos, P. A. Kirchner, & R. L. Martens (Eds.), *What we know about CSCL and implementing it in higher education* (pp. 87–112). Dordrecht, The Netherlands: Kluwer Academic Publishers.

Chi, M. T. H. (1997). Quantifying qualitative alalysis of verbal data: A practical guide. *The Journal of the Learning Sciences, 6,* 271-315.

Cho, H., Stefanone, M., & Gay, G., (2002). Social network analysis of information sharing networks in a CSCL community. In G. Stahl (Ed.), *Computer support for collaborative learning: Foundations for a CSCL community. Proceedings of the computer-supported collaborative learning conference* (pp. 43-50). Mahway, NJ: Lawrence Erlbaum Associates.

Clancey, W. J. (2009). Scientific Antecedents of Situated Cognition. In P. Robbins, & M. Aydede (Eds.), *The Cambridge handbook of situated cognition* (pp. 11–34). New York, NY: Cambridge University Press.

Clark, H. H., & Brennan, S. E. (1991). Grounding in communication. In L. B. Resnick, J. Levine, & S. D. Teasley (Eds.), *Perspectives on socially shared cognition* (pp. 127–149). Washington, DC: American Psychological Association.

Cohen, E. G. (1994). Restructuring the classroom: Conditions for productive small groups. *Review of Educational Research, 64,* 1-35.

Cole, M. & Engestrom, Y. (1993). A cultural-historical approach to distributed cognition. In G. Salomon (Ed.), *Distributed cognitions: Psychological and educational considerations*. New York: Cambridge University Press.

Collins, A. M., & Loftus, E. F. (1975). A spreading-activation theory of semantic processing. *Psychological Review, 82*, 407-428.

Cress, U. (2008). The need for considering multi-level analysis in CSCL research. An appeal for the use of more advanced statistical methods. *International Journal of Computer-Supported Collaborative Learning, 3*, 69-84.

Cress, U. (2013). Mass collaboration and learning. In R. Luckin, P. Goodyear, B. Grabowski, S. Puntambekar, J. Underwood, & N. Winters (Eds.), *Handbook on Design in Educational Technology* (pp. 416-425). London: Taylor and Francis.

Cress, U., Barquero, B., Schwan, S., & Hesse, F. W. (2007). Improving quality and quantity of contributions: Two models for promoting knowledge exchange with shared databases. *Computers & Education, 49*, 423-440.

Cress, U., Barron, B., Halatchliyski, I., Oeberst, A., Forte, A., Resnick, M., & Collins, A. (2013). Mass collaboration - an emerging field for CSCL research. In N. Rummel, M. Kapur, N. Nathan, & S. Puntambekar (Eds.), *To see the world and a grain of sand: Learning across levels of space, time and scale: CSCL 2013 Proceedings* (Vol. I, pp. 557-563). Madison, USA: International Society of the Learning Sciences.

Cress, U., & Kimmerle, J. (2008). A systemic and cognitive view on collaborative knowledge building with wikis. *International Journal of Computer-Supported Collaborative Learning, 3*, 105-122.

Csárdi, C., & Nepusz, T. (2006). The igraph software package for complex network research. *InterJournal Complex Systems, 1695*.

de Laat, M., Lally, V., Lipponen, L., & Simons, R.-J. (2007). Investigating patterns of interaction in networked learning and computer-supported collaborative learning: A role for social network analysis. *International Journal of Computer-Supported Collaborative Learning, 2*, 87-103.

Dehler, J., Bodemer, D., Buder, J., & Hesse, F. W. (2011). Guiding knowledge communication in CSCL via group knowledge awareness. *Computers in Human Behavior, 27*(3), 1068-1078.

Diesner, J., & Carley, K. (2005). Revealing social structure from texts: Meta-matrix text analysis as a novel method for network text analysis. In V. K. Naraynan, & D. J. Armstrong (Eds.), *Causal Mapping for Information Systems and Technology Research: Approaches, Advances, and Illustrations* (pp. 81-108). Harrisburg, PA: Idea Group Publishing.

Dillenbourg, P., Baker, M., Blaye, A., & O'Malley, C. (1996). The evolution of research on collaborative learning. In H. Spada, & P. Reiman (Eds.), *Learning in humans and machine: Towards an interdisciplinary learning science* (pp. 189-211). Oxford: Elsevier.

Dohn, N. (2009). Web 2.0: Inherent tensions and evident challenges for education. *International Journal of Computer-Supported Collaborative Learning, 4*, 343-363.

Engeström, Y. (2001). Expansive learning at work: Toward an activity theoretical reconceptualization. *Journal of education and work, 14*, 133-156.

Engeström, Y., & Sannino, A. (2010). Studies of expansive learning: Foundations, findings and future challenges. *Educational Research Review, 5*(1), 1-24.

Fauconnier, G., & Turner, M. (2002). *The way we think: Conceptual blending and the mind's hidden complexities.* New York: Basic Books.

Forte, A., & Bruckman, A. (2006). From Wikipedia to the classroom: Exploring online publication and learning. In S. A. Barab, K. E. Hay, & D. T. Hickey (Eds.), *Proceedings of the 7th international conference of the learning sciences* (pp. 182-188). Mahwah, NJ: Erlbaum.

Forte, A., & Bruckman, A. (2008). Scaling consensus: Increasing decentralization in Wikipedia governance. In *Proceedings of the 41st Annual Hawaii International Conference on System Sciences* (pp. 157-157), IEEE.

Fournier, H., Kop, R., & Sitla, H. (2011). The value of learning analytics to networked learning on a personal learning environment. In P. Long, G. Siemens, G. Conole, & D. Gasevic (Eds.), *Proceedings of the 1st international conference on learning analytics and knowledge, LAK'11* (pp. 104-109). New York: ACM Press.

Freeman, L. C. (1979). Centrality in social networks: Conceptual clarification. *Social Networks, 1*, 215-239.

Gabrilovich, E., & Markovitch, S. (2006). Overcoming the brittleness bottleneck using Wikipedia: Enhancing text categorization with encyclopedic knowledge. In *AAAI* (Vol. 6, pp. 1301-1306).

Garfield, E. (1972). Citation Analysis as a Tool in Journal Evaluation. *Science*, *178*(4060), 471-479.

Gelernter, D. (1985). Generative communication in Linda. *ACM Transactions on Programming Languages and Systems (TOPLAS), 7*(1), 80-112.

Gerbaudo, P. (2012). *Tweets and the streets: social media and contemporary activism*. Pluto Press.

Giles, J. (2005). Internet encyclopaedias go head to head. *Nature*, *438*, 900-901.

Glänzel, W. (2003). Bibliometrics as a research field. A course on the theory an application of bilbiometric indicators. Budapest: Magyar Tudományos Akadémia Könyvtára. [http://nsdl.niscair.res.in/jspui/bitstream/123456789/968/1/Bib_Module_KUL.pdf]

Goggins, S., Valetto, G., Mascaro, C., & Blincoe, K. (2012). Creating a model of the dynamics of socio-technical groups. *User Modeling and User-Adapted Interaction, 22*, 1-35.

Göhnert, T., Harrer, A., Hecking, T., & Hoppe, H. U. (2013). A workbench to construct and re-use network analysis workflows: Concept, implementation, and example case. In *Proceedings of the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, ASONAM 2013, Niagara Falls, Canada, 25-28 August*. ACM, New York, NY, USA, 1464-1466.

Gunawardena, C. N., Lowe, C. A., & Anderson, T. (1997). Analysis of a global online debate and the development of an interaction analysis model for examining social construction of knowledge in computer conferencing. *Journal of Educational Computing Research, 17*, 397-431.

Hakkarainen, K., Ritella, G., & Seitamaa-Hakkarainen, P. (2011). Epistemic mediation, chronotope, and expansive knowledge practices. In H. Spada, G. Stahl, N. Miyake, & N. Law (Eds.), *Connecting computer-supported collaborative learning to policy and practice: CSCL2011 conference proceedings* (Vol. II, pp. 948-949). Hong Kong: International Society of the Learning Sciences.

Halatchliyski, I., & Cress, U. (2014). How structure shapes dynamics: Knowledge development in Wikipedia - a network multilevel modeling approach. *PLoS ONE, 9*, e111958.

Halatchliyski, I., Hecking, T., Göhnert, T., & Hoppe, H. U. (2013). Analyzing the flow of ideas and profiles of contributors in an open learning community. In D. Suthers, K. Verbert, E. Duval, & X. Ochoa (Eds.), *Proceedings of the 3rd international conference on learning analytics and knowledge, LAK'13* (pp. 66-74). New York: ACM Press.

Halatchliyski, I., Hecking, T., Göhnert, T., & Hoppe, H. U. (2014). Analyzing the main paths of knowledge evolution and contributor roles in an open learning community. *Journal of Learning Analytics, 1*, 72-93.

Halatchliyski, I., Kimmerle, J., & Cress, U. (2011). Divergent and convergent knowledge processes on Wikipedia. In H. Spada, G. Stahl, N. Miyake, & N. Law (Eds.), *Connecting computer-supported collaborative learning to policy and practice: CSCL2011 conference proceedings* (Vol. II, pp. 566-570). Hong Kong: International Society of the Learning Sciences.

Halatchliyski, I., Moskaliuk, J., Kimmerle, J., & Cress, U. (2010). Who integrates the networks of knowledge in Wikipedia? *Proceedings of the 6th international symposium on wikis and open collaboration* (Article No. 1). New York: ACM Press.

Halatchliyski, I., Moskaliuk, J., Kimmerle, J., & Cress, U. (2014). Explaining authors' contribution to pivotal artifacts during mass collaboration in the Wikipedia's knowledge base. *International Journal of Computer-Supported Collaborative Learning, 9*, 97-115.

Halatchliyski, I., Oeberst, A., Bientzle, M., Bokhorst, F., & van Aalst, J. (2012). Unraveling idea development in discourse trajectories. In J. van Aalst, K. Thompson, M. J. Jacobson, & P. Reimann (Eds.), *The future of learning: Proceedings of the 10th international conference of the learning sciences* (Vol. II, pp. 162-166). Sydney, NSW, Australia: International Society of the Learning Sciences.

Halfaker, A., Geiger, R. S., Morgan, J. T., & Riedl, J. (2013). The rise and decline of an open collaboration system: How Wikipedia's reaction to popularity is causing its decline. *American Behavioral Scientist, 57*, 664–688.

Hanneman, R. A., Riddle. M. (2005). *Introduction to social network methods*. Riverside, CA: University of California, Riverside. [http://faculty.ucr.edu/~hanneman]

Hara, N., Bonk, C.J., & Angeli, C. (2000). Content analysis of online discussion in an applied educational psychology course. *Instructional Science, 28,* 115-152.

Harrer, A., Malzahn N., Zeini S., & Hoppe H. U. (2007). Combining Social Network Analysis with Semantic Relations to Support the Evolution of a Scientific Community. In *Mice, Minds, and Society - The Computer Supported Collaborative Learning Conference Proceedings, CSCL 2007, New Brunswick, USA, 16-21 July, 2007* (pp. 267-276). Mahwah, NJ: Lawrence Erlbaum Associates.

Henri, F. (1992). Computer Conferencing and Content Analysis. In A. Kaye (Ed.), *Collaborative Learning through Computer Conferencing: The Najaden Papers* (pp. 117-136). Berlin: Springer.

Hirsch, J. E. (2005). An index to quantify an individual's scientific research output. *Proceedings of the National Academy of Sciences of the United States of America, 102,* 16569–16572.

Holme, P., Kim, B. J., Yoon, C. N., Han, S. K. (2002). Attack vulnerability of complex networks. *Physical Review E, 65,* 056109.

Hummon, N. P., & Doreian, P. (1989). Connectivity in a Citation Network: The Development of DNA Theory. *Social Networks, 11,* 39-63.

Hung, D., Lim, K. Y., Chen, D. T. V., & Koh, T. S. (2008). Leveraging online communities in fostering adaptive schools. *International Journal of Computer-Supported Collaborative Learning, 3,* 373-386.

Hutchins, E. (1995). *Cognition in the wild.* Cambridge, MA: MIT Press.

Iriberri, A., & Leroy, G. (2009). A life-cycle perspective on online community success. *ACM Computing Surveys (CSUR), 41*(2), 11:1–11:29.

Janssen, J.J.H.M., Erkens, G., Kirschner, P.A., & Kanselaar, G. (2011). Multilevel analysis in CSCL research. In S. Puntambekar, G. Erkens, & C. E. Hmelo-Silver (Eds.), *Analyzing interactions in CSCL: Methods, approaches and issues* (pp. 187-205). New York: Springer.

Jenkins, H., Clinton K., Purushotma, R., Robinson, A.J., & Weigel, M. (2006). *Confronting the challenges of participatory culture: Media education for the 21st century*. Chicago, IL: The MacArthur Foundation.

Jeong, A. C. (2003). The sequential analysis of group interaction and critical thinking in online. *The American Journal of Distance Education, 17,* 25-43.

Johnson, D. W., & Johnson, R. T. (1989). *Cooperation and competition: Theory and research*. Edina, MN: Interaction Book Company.

Jones, C., Dirckinck-Holmfeld, L., & Lindstrom, B. (2006). A relational, indirect, meso-level approach to CSCL design in the next decade. *International Journal of Computer-Supported Collaborative Learning, 1,* 35–56.

Kafai, Y. B. &, Peppler, K. A. (2011). Beyond small groups: New opportunities for research in computer-supported collective learning. In H. Spada, G. Stahl, N. Miyake, & N. Law (Eds.),

Connecting computer-supported collaborative learning to policy and practice: CSCL2011 conference proceedings (Vol. I, pp. 17-24). Hong Kong: International Society of the Learning Sciences.

Kafai, Y., & Resnick, M. (2000). *Constructionism in practice: Designing, thinking, and learning in a digital world*. Mahwah, NJ: Erlbaum.

Kapur, M., Hung, D., Jacobson, M., Voiklis, J., Kinzer, C., & Chen, D.-T. (2007). Emergence of learning in computer-supported, large-scale collective dynamics: A research agenda. In C. A. Clark, G. Erkens, & S. Puntambekar (Eds.), *Proceedings of the international conference of computer-supported collaborative learning* (pp. 323-332). Mahwah, NJ: Erlbaum.

Keegan, B., Gergle, D., & Contractor, N. (2012). Do editors or articles drive collaboration? Multilevel statistical network analysis of Wikipedia coauthorship. *Proceedings of the 2012 ACM conference on computer-supported cooperative work* (pp. 427-436). New York: ACM Press.

Kimmerle, J., Cress, U., & Held, C. (2010a). The interplay between individual and collective knowledge: Technologies for organisational learning and knowledge building. *Knowledge Management Research and Practice, 8,* 33-44.

Kimmerle, J., Moskaliuk, J., & Cress, U. (2011a). Using wikis for learning and knowledge building: Results of an experimental study. *Educational Technology & Society, 14*, 138-148.

Kimmerle, J., Moskaliuk, J., Cress, U., & Thiel, A. (2011b). A systems theoretical approach to online knowledge building. *AI & Society: Journal of Knowledge, Culture and Communication, 26*, 49-60.

Kimmerle, J., Moskaliuk, J., Harrer, A., & Cress, U. (2010b). Visualizing co-evolution of individual and collective knowledge. *Information, Communication and Society, 13,* 1099-1121.

Kimmerle, J., Thiel, A., Gerbing, K.-K., Bientzle, M., Halatchliyski, I., & Cress, U. (2013). Knowledge construction in an outsider community: Extending the communities of practice concept. *Computers in Human Behavior, 29*, 1078-1090.

Kitsak, M., Havlin, S., Paul, G., Riccaboni, M., Pammolli, F., & Stanley, H. E. (2007) Betweenness centrality of fractal and nonfractal scale-free model networks and tests on real networks. *Physical Review E, 75,* 056115.

Kittur, A., Chi, E. H., Pendleton, B. A., Suh, B., & Mytkowicz, T. (2007). Power of the Few vs. Wisdom of the Crowd: Wikipedia and the Rise of the Bourgeoisie. In *alt.CHI 2007*, San Jose, CA.

Kittur, A., Chi, E. H., & Suh, B. (2009). What's in Wikipedia? Mapping topics and conflict using socially annotated category structure. In *Proceedings of the 27th international conference on human factors in computing systems* (p. 1509). New York: ACM Press.

Knorr-Cetina, K. (2001). Objectual practice. In T. R. Schatzki, K. Knorr-Cetina, & E. von Savigny (Eds.), *The practice turn in contemporary theory* (pp. 175-188). London/New York: Routledge.

Kolbitsch, J., & Maurer, H. (2006). The transformation of the Web: How emerging communities shape the information we consume. *Journal of Universal Computer Science, 12*, 187-213.

Konieczny, P. (2007). Wikis and Wikipedia as a teaching tool. *International Journal of Instructional Technology and Distance Learning, 4*(1), 15-34.

Koschmann, T. (2002). Dewey's contribution to the foundations of CSCL research. In G. Stahl (Ed.), *Computer support for collaborative learning: Foundations for a CSCL community. Proceedings of the computer-supported collaborative learning conference* (pp. 17-22). Mahway, NJ: Lawrence Erlbaum Associates.

Kraut, R.E., Burke, M., Riedl, J., & Resnick, P. (2012). The challenges of dealing with newcomers. In R. E. Kraut, & P. Resnick (Eds.), *Evidence-based social design: Mining the social sciences to build online communities.* (pp. 179-230) Cambridge, MA: MIT Press.

Kuhn, T. S. (1962). *The Structure of Scientific Revolutions*. Chicago: University of Chicago Press.

Larusson, J., & Alterman, R. (2009). Wikis to support the "collaborative" part of collaborative learning. *International Journal of Computer-Supported Collaborative Learning, 4*, 371-402.

Latour, B. (1987). *Science in action: How to Follow Scientists and Engineers through Society.* Cambridge, MA: Harvard University Press.

Latour, B. (2005). *Reassembling the social: An introduction to actor-network-theory*. New York: Oxford University Press.

Latour, B., & Woolgar, S. (1979). *Laboratory life: The social construction of scientific facts.* Princeton University Press.

Lave, J. (1988). *Cognition in practice: Mind, mathematics and culture in everyday life (learning in doing).* Cambridge: Cambridge University Press.

Lave, J., & Wenger, E. (1991). *Situated learning: Legitimate peripheral participation*. Cambridge: Cambridge University Press.

Leuf, B., & Cunningham, W. (2001). *The wiki way. Quick collaboration on the web*. Boston: Addison-Wesley.

Levina, N., & Vaast, E. (2005). The emergence of boundary spanning competence in practice: Implications for implementation and use of information systems. *MIS Quarter*ly, *29*, 335-363.

Levy, P. *(*1999*). Collective Intelligence: Mankind's Emerging World in Cyberspace.* Cambridge, MA: Perseus Books.

Leydesdorff, L. (2007). "Betweenness centrality" as an indicator of the "interdisciplinarity" of scientific journals. *Journal of the American Society for Information Science and Technology, 58*, 1303-1309.

Ling, K., Beenen, G., Ludford, P., Wang, X., Chang, K., Li, X., Cosley, D., Frankowski, D., Terveen, L., Rashid, A. M., Resnick, P. & Kraut, R. (2005). Using social psychology to motivate contributions to online communities. *Journal of Computer-Mediated Communication, 10*(4).

Lipponen, L. 2002: Exploring foundations for computer-supported collaborative learning. In G. Stahl (Ed.), *Computer support for collaborative learning: Foundations for a CSCL community. Proceedings of the computer-supported collaborative learning conference* (pp. 72-81). Mahway, NJ: Lawrence Erlbaum Associates.

Lipponen, L., Hakkarainen, K. & Paavola, S. (2004). Practices and orientations of computer supported collaborative learning. In J. Strijbos, P. Kirschner & R. Martens (Eds.), *What we know about CSCL, and implementing it in higher education* (pp. 31-50). Boston, MA: Kluwer.

Liu, J. S., & Lu, L. Y. Y. (2012). An integrated approach for main path analysis: Development of the Hirsch index as an example. *Journal of the American Society for Information Science and Technology*, *63*, 528-542.

Lucio-Arias, D., & Leydesdorff, L. (2007). Knowledge emergence in scientific communication: from “fullerenes” to “nanotubes”. Scientometrics, 70(3), 603-632.

Lucio-Arias, D., & Leydesdorff, L. (2009). The dynamics of exchanges and references among scientific texts, and the *autopoiesis* of discursive knowledge. *Journal of Informetrics, 3*(3), 261-271.

Lucio-Arias, D., & Scharnhorst, A. (2012). Mathematical approaches to modeling science from an algorithmic-historiography perspective. In A. Scharnhorst, K. Borner & P. van den Besselaar (Eds.), *Models of science dynamics, Understanding complex systems.*

Luhmann N. (1984). *Soziale Systeme: Grundriß einer allgemeinen Theorie*. Frankfurt am Main: Suhrkamp

Mali, F., Kronegger, L., Doreian, P. & Ferligoj, A. (2012). Dynamic scientific co-authorship networks. In A.Scharnhorst, K. Borner & P. van den Besselaar (Eds.), *Models of science dynamics, Understanding complex systems* (pp. 195-232). Heidelberg: Springer.

Mathes, A. (2004). Folksonomies-cooperative classification and communication through shared metadata. *Computer Mediated Communication, 47*, 1-13.

Maturana, H. R, & Varela, F. J. (1987). The tree of knowledge: The biological roots of human understanding.Boston, MA: New Science Library, Shambhala Publications.

Mercer, N. (2008). The seeds of time: Why classroom dialogue needs a temporal analysis. *The Journal of the Learning Sciences*, *17*, 33-59.

Merton, R. K. (1968). The Matthew effect in science. *Science*, *159*(3810), 56-63.

Mika, P. (2007). *Social networks and the semantic Web*. New York: Springer.

Moskaliuk, J., & Kimmerle, J. (2009). Using wikis for organizational learning: Functional and psycho-social principles. *Development and Learning in Organizations, 23*, 21-24.

Moskaliuk, J., Kimmerle, J., & Cress, U. (2009). Wiki-supported learning and knowledge building: Effects of incongruity between knowledge and information. *Journal of Computer Assisted Learning, 25*, 549-561.

Moskaliuk, J., Kimmerle, J., & Cress, U. (2012). Collaborative knowledge building with wikis: The impact of redundancy and polarity. *Computers & Education, 58,* 1049-1057.

Negroponte, N. (1995). *Being digital*. London: Vintage Books.

Newman, M. (2010). *Networks: An introduction*. Oxford: Oxford University Press.

Niederer, S., & Van Dijck, J. (2010). Wisdom of the crowd or technicity of content? Wikipedia as a sociotechnical system. *New Media & Society, 12*(8), 1368-1387.

Nonaka, I., & Nishiguchi, T. (2000). *Knowledge emergence: Social, technical, and evolutionary dimensions of knowledge creation*. Oxford University Press.

O'Reilly, T. (2005). *What is Web 2.0? Design patterns and business models for the next generation of software*. [http://oreilly.com/web2/archive/what-is-web-20.html]

Oeberst, A, Halatchliyski, I., Kimmerle, J., & Cress, U. (2014). Knowledge construction in Wikipedia: A systemic-constructivist analysis. *The Journal of the Learning Sciences, 23*, 149-176.

Okoli, C., Mehdi, M., Mesgari, M., Nielsen, F.Å., & Lanamäki, A. (2012). The people's encyclopedia under the gaze of the sages: A systematic review of scholarly research on Wikipedia. Working paper. [http://ssrn.com/paper=2021326]

Ortega, F. (2009). *Wikipedia: A Quantitative Analysis*. Ph.D. dissertation, Universidad Rey Juan Carlos, Madrid. [http://www.researchgate.net/publication/200773248_Wikipedia_A_quantitative_analysis]

Oshima, J., Oshima, R., & Knowledge Forum ® Japan Research Group (2007). Complex network theory approach to the assessment on collective knowledge advancement through scientific discourse in CSCL. In C. A. Clark, G. Erkens, & S. Puntambekar (Eds.), *Proceedings of the international conference of computer-supported collaborative learning* (pp. 563-565). Mahwah, NJ: Erlbaum.

Paavola, S., & Hakkarainen, K. (2005). The knowledge creation metaphor–An emergent epistemological approach to learning. *Science & Education, 14*(6), 535-557.

Paavola, S., & Hakkarainen, K. (2009). From meaning making to joint construction of knowledge practices and artefacts – A trialogical approach to CSCL. In C. O'Malley, D. Suthers, P. Reimann, & A. Dimitracopoulou (Eds.), *Computer supported collaborative learning practices: CSCL2009 conference proceedings* (pp. 83-92). Rhodes, Greece: International Society of the Learning Sciences.

Paavola, S., Lipponen, L., & Hakkarainen, K. (2002). Epistemological foundations for CSCL: A comparison of three models of innovative knowledge communities. In G. Stahl (Ed.), *Computer support for collaborative learning: Foundations for a CSCL community. Proceedings of the computer-supported collaborative learning conference* (pp. 24-32). Mahway, NJ: Lawrence Erlbaum Associates.

Paavola, S., Lipponen, L., & Hakkarainen, K. (2004). Models of innovative knowledge communities and three metaphors of learning. *Review of Educational Research, 74*, 557-576.

Page, L., Brin, S., Motwani, R., & Winograd, T. (1998). The PageRank citation ranking: Bringing order to the eb. Technical report. Stanford University. [http://ilpubs.stanford.edu:8090/422/1/1999-66.pdf]

Palincsar, A. (1998). Social constructivist perspectives on teaching and learning. *Annual review of psychology*, *49*, 345-375.

Palonen, T., & Hakkarainen, K. (2000). Patterns of interaction in computer-supported learning: A social network analysis. In B. Fishman, & S. O'Conner-Divelbiss (Eds.), *Proceedings of the fourth international conference on the learning sciences.* Mahwah, NJ: Lawrence Erlbaum.

Panciera, K., Halfaker, A., & Terveen, L. (2009). Wikipedians are born, not made: a study of power editors on Wikipedia. In *Proceedings of the ACM 2009 international conference on supporting group work (GROUP '09).* (pp. 51-60). New York: ACM Press.

Park, H. W., & Thelwall, M. (2003). Hyperlink Analyses of the World Wide Web: A Review. *Journal of Computer-Mediated Communication, 8*(4).

Pastor-Satorras R., Vespignani A. (2004). *Evolution and structure of the Internet: A statistical physics approach*. Cambridge University Press, Cambridge.

Pentzold, C., & Seidenglanz, S. (2006). Foucault@Wiki: First steps towards a conceptual framework for the analysis of wiki discourses. In *Proceedings of the 2006 international symposium on Wikis* (pp. 59–68). New York: ACM Press.

Perera, D., Kay, J., Koprinska, I., Yacef, K., & Zaïane, O. R. (2009). Clustering and sequential pattern mining of online collaborative learning data. *IEEE Transactions on Knowledge and Data Engineering, 21*, 759-772.

Peugh, J. L. (2010). A practical guide to multilevel modeling. *Journal of School Psychology, 48*, 85–112.

Pifarré, M., & Kleine Staarman, J. (2011). Wiki-supported collaborative learning in primary education: How a dialogic space is created for thinking together. *International Journal of Computer-Supported Collaborative Learning*, *6*, 187-205.

Popper, K. R. (1968). Epistemology without a knowing subject. *Studies in Logic and the Foundations of Mathematics*, *52*, 333-373.

Popper, K. R. (1972). *Objective knowledge: An evolutionary approach*. Oxford: Oxford University Press.

Price, D. J. D. S. (1963). *Little science, big science*. New York: Columbia University Press.

Price, D. J. D. S. (1976). A general theory of bibliometric and other cumulative advantage processes. *Journal of the American Society for Information Science*, *27*(5), 292-306.

Purdy, J. P. (2009). When the tenets of composition go public: A study of writing in Wikipedia. *College Composition & Communication*, *61*, 383.

Radicchi, F., Fortunato, S., & Vespignani, A. (2012). Citation networks. In A.Scharnhorst, K. Borner, & P. van den Besselaar (Eds.), *Models of science dynamics, Understanding complex systems* (pp. 233-257). Heidelberg: Springer.

Rafaeli, S., & Ariel, Y. (2008). Online motivational factors: Incentives for participation and contribution in Wikipedia. In A. Barak (Ed.), *Psychological aspects of cyberspace: Theory, research, applications* (pp. 243–267). Cambridge, UK: Cambridge University Press.

Raudenbush, S. W., & Chan, W. S. (1993). Application of a hierarchical linear model to the study of adolescent deviance in an overlapping cohort design. *Journal of Consulting and Clinical Psychology*, *61*(6), 941-951.

Ravenscroft, A. (2009). Social software, Web 2.0 and learning: status and implications of an evolving paradigm. *Journal of Computer Assisted Learning, 25*(1), 1-5.

Reffay, C., & Chanier, T. (2002) Social network analysis used for modeling collaboration in distance learning groups, In S. A. Cerri, G. Gouardères, & F. Paraguaçu (Eds.), *ITS 2002 Lecture Notes in Computer Science, 2363* (pp. 31-40). Berlin, Heidelberg: Springer.

Reimann, P. (2009). Time is precious: Variable-and event-centred approaches to process analysis in CSCL research. *International Journal of Computer-Supported Collaborative Learning*, *4*, 239-257.

Reinhardt, W., Moi, M., & Varlemann, T. (2009). Artefact-Actor-Networks as tie between social networks and artefact networks. In *Proceedings of the 5th International Conference on Collaborative Computing: Networking, Applications and Worksharing, CollaborateCom 2009, Washington DC, USA, 11-14 November*. IEEE Computer Society.

Richardson, W. (2010). *Blogs, wikis, podcasts, and other powerful web tools for classrooms* (3rd ed.). Thousand Oaks, CA: Corwin.

Ridings, C. M., & Gefen, D. (2004). Virtual community attraction: Why people hang out online. *Journal of Computer-Mediated Communication*, *10*(1).

Roschelle, J., & Teasley S. D. (1995). The construction of shared knowledge in collaborative problem solving. In C. E. O'Malley (Ed.), *Computer-supported collaborative learning.* (pp. 69-197). Berlin: Springer.

Rosé, C., Wang, Y.-C., Cui, Y., Arguello, J., Stegmann, K., Weinberger, A., Fischer, F. (2008). Analyzing collaborative learning processes automatically: Exploiting the advances of computational linguistics in CSCL. *International Journal of Computer-Supported Collaborative Learning*, *3*(3), 237-271.

Ryberg, T., & Larsen, M. C. (2008). Networked identities: Understanding relationships between strong and weak ties in networked environments. *Journal of Computer-Assisted Learning, 24,* 103-115.

Sacks, H. (1992). *Lectures on conversation*. Oxford, UK: Blackwell.

Saviotti, P. (2009). Knowledge networks: Structure and dynamics. In: A. Pyka, A. Scharnhorst, *Innovation networks: Developing an integrated approach* (pp. 19-41). Heidelberg: Springer Verlag.

Scardamalia, M. (2002). Collective cognitive responsibility for the advancement of knowledge. In B. Smith (Ed.), *Liberal education in a knowledge society* (pp. 67–98). Chicago: Open Court.

Scardamalia, M., & Bereiter, C. (1994). Computer support for knowledge-building communities. *Journal of the Learning Sciences, 3*, 265-283.

Scardamalia, M., & Bereiter, C. (2006). Knowledge building: Theory, pedagogy and technology. In K. Sawyer (Ed.), *Cambridge handbook of the learning sciences* (pp. 97–118). Cambridge, UK: Cambridge University Press.

Scheuer, O., Loll, F., Pinkwart, N., & McLaren, B. (2010). Computer-supported argumentation: A review of the state of the art. *International Journal of Computer-Supported Collaborative Learning, 5*, 43-102.

Schön, D. (1983). *The Reflective Practitioner, How Professionals Think In Action.* London: Temple Smith.

Schrire, S. (2004). Interaction and cognition in asynchronous computer conferencing. *Instructional Science, 32*, 475-502.

Schwämmlein, E., & Wodzicki, K. (2012). What to Tell About Me? Self-Presentation in Online Communities. *Journal of Computer-Mediated Communication, 17*(4), 387-407.

Sfard, A. (1998). On two metaphors for learning and the dangers of choosing just one. *Educational Researcher, 27,* 4-13.

Sha, L., van Aalst, J., & Teplovs, C. (2010). A visualization of group cognition: Semantic network analysis of a CSCL community. In K. Gomez, L. Lyons, & J. Radinsky (Eds.), *Learning in the disciplines: Proceedings of the 9th international conference of the learning sciences* (Volume 1, pp. 929-936). International Society of the Learning Sciences: Chicago, IL.

Shaw, A., & Hill, B M. (2014). Laboratories of oligarchy? How the iron law extends to peer production. *Journal of Communication, 64*, 215–238.

Siemens, G. (2012). Learning analytics: Envisioning a research discipline and a domain of practice. In S. Buckingham Shum, D. Gasevic, & R. Ferguson (Eds.), *Proceedings of the second international conference on learning analytics and knowledge, LAK'12* (pp. 4-8). New York: ACM Press.

Siemens, G., & Baker, R. S. (2012). Learning analytics and educational data mining: Towards communication and collaboration. In S. Buckingham Shum, D. Gasevic, & R. Ferguson (Eds.), *Proceedings of the second international conference on learning analytics and knowledge, LAK'12* (pp. 252-254). New York: ACM Press.

Slavin, R. E. (1995). *Cooperative learning: Theory, research, and practice* (2nd ed.). Boston: Allyn and Bacon.

Smith, E. R., & Semin, G. R. (2004). Socially situated cognition: Cognition in its social context. *Advances in experimental social psychology, 36*, 53-117.

Sosa, M. E. (2011). Where do creative interactions come from? The role of tie content and social networks. *Organization Science*, *22*, 1-21.

Stahl, G. (2006). *Group cognition: Computer support for building collaborative knowledge.* Cambridge, MA: MIT Press.

Stahl, G., Koschmann, T., & Suthers, D. (2006). Computer-supported collaborative learning: An historical perspective. In R. K. Sawyer (Ed.), *Cambridge handbook of the learning sciences*, Cambridge, UK: Cambridge University Press.

Stehr, N. (2001). *The fragility of modern societies. Knowledge and risk in the information age.* Thousand Oaks, CA: Sage.

Stvilia, B., Twidale, M. B., Smith, L. C., & Gasser, L. (2008). Information quality work organization in Wikipedia. *Journal of the American society for information science and technology, 59*(6), 983-1001.

Suchman, L. (1987). *Plans and situated actions: The problem of human-machine communication.* Cambridge, UK: Cambridge University Press.

Suh, B., Convertino, G., Chi, E. H., & Pirolli, P. (2009). The singularity is not near: slowing growth of Wikipedia. In *Proceedings of the 5th International Symposium on Wikis and Open Collaboration* (Nr. 8). ACM: New York, NY, USA.

Surowiecki, J. (2005). The Wisdom of Crowds. New York: Anchor Books.

Susi, T., & Ziemke, T. (2001). Social cognition, artifacts, and stigmergy: A comparative analysis of theoretical frameworks for the understanding of artifact-mediated collaborative activity. *Cognitive Systems Research*, *2*, 273-290.

Suthers, D. (2006). Technology affordances for intersubjective meaning-making: A research agenda for CSCL. *International Journal of Computers Supported Collaborative Learning*, 1, 315-337.

Suthers, D. D. (2012). Computer-Supported Collaborative Learning. In Norbert M. Seel (Ed.), Encyclopedia of the Sciences of Learning. New York: Springer.

Suthers, D. D., Dwyer, N., Medina, R., & Vatrapu, R. (2010). A framework for conceptualizing, representing, and analyzing distributed interaction. *International Journal of Computer-Supported Collaborative Learning*, *5*, 5-42.

Suthers, D., & Rosen, D. (2011). A unified framework for multi-level analysis of distributed learning. In P. Long, G. Siemens, G. Conole, & D. Gasevic (Eds.), *Proceedings of the 1st international conference on learning analytics and knowledge, LAK'11* (pp. 64-74). New York: ACM Press.

Suthers, D., & Verbert, K. (2013). Learning analytics as a "middle space". In D. Suthers, K. Verbert, E. Duval, & X. Ochoa (Eds.), *Proceedings of the 3rd international conference on learning analytics and knowledge, LAK'13* (pp. 1-4). New York: ACM Press.

Swarts, J., (2009). The collaborative construction of "fact" on Wikipedia. *Proceedings of the 27th ACM international conference on design of communication* (pp. 281-288). New York: ACM Press.

Tapscott, D., & Williams, A. D. (2006). *Wikinomics: How mass collaboration changes everything.* New York: Portfolio.

Teplovs, C., & Fujita, N. (2013). Socio-dynamic latent semantic learner models. In D. D. Suthers, K. Lund, C. P. Rosé, C. Teplovs, & N. Law (Eds.), *Productive multivocality in the analysis of group interactions* (pp. 383-396). New York, NY: Springer.

Theiner, G., Allen C., & Goldstone, R. L. (2010). Recognizing Group Cognition. Cognitive Systems Research, 11, 378-395.

Thompson, L., & Fine, G. A. (1999). Socially shared cognition, affect, and behavior: A review and integration. *Personality and Social Psychology Review*, *3*, 278-302.

Tushman, M. L., & Scanlan, T. J. (1981). Boundary spanning individuals: Their role in information transfer and their antecedents. *The Academy of Management Journal*, *24*, 289-305.

Viégas, F. B., Wattenberg, M., & Dave, K. (2004). Studying cooperation and conflict between authors with history flow visualizations. In *Proceedings of the SIGCHI conference on Human factors in computing systems* (pp. 575-582). New York: ACM Press.

von Foerster, H. (2003). *Understanding understanding: essays on cybernetics and cognition*. New York: Springer.

von Glasersfeld E. (1995). *Radical constructivism: a way of knowing and learning*. Bristol, PA: Falmer Press, Taylor & Francis.

Voss, J. (2005). Measuring Wikipedia. In P. Ingwersen, & B. Larsen (Eds.), *Proceedings of 10th international conference of the international society for scientometrics and informetrics* (pp. 221-231). Stockholm: Karolinska University Press.

Vygotsky, L.S. (1930/1978). *Mind in society: The development of higher psychological processes.* Cambridge, MA: Harvard University Press.

Wasko, M. M., & Faraj, S. (2005).Why should I share? Examining knowledge contribution in networks of practice. *MIS Quarterly*, *29*, 35-58.

Wassermann, S., & Faust, K. (1994). *Social network analysis: Methods and application.* Cambridge, UK: Cambridge University Press.

Weinbrenner, S., Giemza, A., & Hoppe H. U. (2007). Engineering heterogeneous distributed learning environments using tuple spaces as an architectural platform. In *Proceedings of the 7th IEEE International Conference on Advanced Learning Technologies, ICALT 2007, Niigata, Japan, 18-20 July* (pp. 434-436). IEEE Computer Society.

Wellman, B. (1997). Structural analysis: From method and metaphor to theory and substance. *Contemporary Studies in Sociology*, *15*, 19-61.

Wellman, B., & Gulia, M. (1999). Net Surfers Don't Ride Alone: Virtual Community as Community In B. Wellman (Ed.), *Networks in the global village* (pp. 331-367). Boulder, CO: Westview Press.

Wenger, E. (1998). *Communities of practice: Learning, meaning, and identity*. Cambridge university press.

Wilkinson, D. M., & Huberman, B. A. (2007). Cooperation and quality in wikipedia. In *Proceedings of the 2007 International Symposium on Wikis* (pp. 157-164). ACM Press.

Wöhner, T., & Peters, R. (2009). Assessing the quality of Wikipedia articles with lifecycle based metrics. In *Proceedings of the 5th International Symposium on Wikis and Open Collaboration*. New York: ACM Press.

Yasseri, T., Kertész, J. (2013). Value production in a collaborative environment. *Journal of Statistical Physics, 151,* 414-439.

Zaïane, O., Luo, J. (2001). Web usage mining for a better web-based learning environment. In *Proceedings of Conference on Advanced Technology for Education* (pp. 60–64), Banff, Alberta, Canada.

Zeini, S., Göhnert, T., & Hoppe, H.U. (2012). The impact of measurement time on subgroup detection in online communities. In *Proceedings of the IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, ASONAM 2012, Istanbul, Turkey, 26-29 August*. IEEE Computer Society.

Summary

Contemporary Web 2.0 technologies facilitate the establishment of large online communities of mass collaboration. In shared workspace environments, millions of people interact without knowing each other. The outcome is openly accessible and constantly developing collective knowledge in the form of a more or less organized knowledge base of digital artifacts. With this dissertation I advance a differentiated approach for studying and understanding the principles that underlie knowledge development under these conditions.

The work builds on a theoretical consideration of collaborative learning and knowledge building stemming from the interdisciplinary learning sciences and research on computer-supported collaborative learning (CSCL) in particular. Knowledge is understood as substance with static structure that changes over longer periods of time through the activity of community participants in analogy to the progress of scientific ideas in different domains. A complex systems perspective is used to explain knowledge as an emergent phenomenon that amounts to more than the additive collection of individual contributions. This macro level of processes and structures in a community determines to a large extent how new contributions are made and thus how knowledge develops.

Based on these conceptualizations, the present dissertation empirically examines large real-life data sets from the online communities Wikipedia and Wikiversity. Knowledge is captured as a network of interconnected articles in different knowledge domains. The topological position of the articles in the networks is evaluated through established network analysis metrics in order to indentify pivotal articles that form the static structural backbone of the collective knowledge. A cross-sectional analysis demonstrates that pivotal articles tend to be written by authors with extensive contribution experience in the community. In a longitudinal study, a mechanism of knowledge development is evidenced according to which pivotal articles attract new knowledge that appears in the network in subsequent periods. Thus, structure and dynamics of collective knowledge are mutually determining. A continuous time approach to studying their interplay is presented using the scientometric method of main path analysis. It consists in evaluating how pivotal the position of each contribution to the knowledge base is relatively to the historical trajectory of knowledge development. This method allows a more immediate analysis of the collaborative process and connects the micro level of individual contributions with the macro level of collective knowledge development.

In sum, this dissertation provides a straightforward contribution to the analysis, understanding

and facilitation of informal contexts of knowledge production, which become increasingly important also for formal learning policy and practice. My work further makes the relevant novel phenomenon of online mass collaboration accessible for theoretical and empirical consideration in CSCL research and also contributes a valuable methodological approach for the new research filed of learning analytics.

Deutsche Zusammenfassung

Moderne Web 2.0-Technologien ermöglichen das Entstehen von großen Online-Communities der Massenkollaboration. In der virtuellen Umgebung mit gemeinsamen Arbeitsbereichen interagieren Millionen von Menschen, ohne sich gegenseitig zu kennen. Das Ergebnis ist offen zugängliches und sich ständig entwickelndes kollektives Wissen in der Form einer mehr oder weniger organisierten Wissensbasis aus digitalen Artefakten. Mit dieser Dissertation lege ich einen differenzierten Ansatz für die Untersuchung und das Verständnis der Prinzipien vor, die dem Wissensfortschritt unter diesen Bedingungen zugrunde liegen.

Die Arbeit baut auf einer theoretischen Betrachtung der kollaborativen Prozesse des Lernens und Wissensproduktion auf, die von den interdisziplinären Learning Sciences und insbesondere von der Forschung im Bereich des computerunterstützen kollaborativen Lernens (CSCL) stammt. Wissen wird als Substanz mit statischer Struktur verstanden, die sich über längere Zeiträume durch die Aktivität von Community-Teilnehmern ändert, in Analogie zum Fortschritt der Ideen in verschiedenen wissenschaftlichen Gebieten. Die Perspektive komplexer Systeme wird eingenommen, um den emergenten Charakter des Wissens zu erklären, der über die Ansammlung einzelner Beiträge hinausgeht. Diese Makroebene der Prozesse und Strukturen in einer Gemeinschaft bestimmt zu einem großen Teil, wie neue Beiträge vorgenommen werden und somit wie sich das Wissen entwickelt.

Basierend auf diesen Konzeptualisierungen untersucht die vorliegende Dissertation empirisch große reale Datensätze aus den Online-Communities Wikipedia und Wikiversity. Wissen wird als ein Netzwerk von miteinander verbundenen Artikeln aus verschiedenen Wissensbereichen erfasst. Die topologische Position der Artikel in den Netzen wird durch etablierte Netzwerkanalyse-Metriken bewertet, um die grundlegenden Artikel zu ermitteln, die das statische strukturelle Rückgrat des kollektiven Wissens bilden. Eine Querschnittsanalyse zeigt, dass die grundlegenden Artikel eher von Autoren mit umfangreicher Beitragserfahrung in der Gemeinschaft geschrieben werden. In einer Längsschnittstudie wird ein Mechanismus des Wissensfortschritts belegt, nach dem die grundlegenden Artikel das neue Wissen anlocken, das sich in den Folgeperioden im Netzwerk manifestiert. Dementsprechend bedingen sich Struktur und Dynamik von kollektivem Wissen gegenseitig. Um ihr Zusammenspiel zu untersuchen wird ein Ansatz vorgestellt, der die kontinuierliche Zeit mit der szientometrische Methode der Main Path Anylsis (Hauptpfad-Analyse) berücksichtigt. Es wird ausgewertet, wie grundlegend die Position eines jeden Beitrags zur Wissensbasis ist, in

Abhängigkeit von der historischen Zeitschiene des Wissensfortschritts. Diese Methode ermöglicht eine unmittelbare Analyse des kollaborativen Prozesses und verbindet die Mikroebene der einzelnen Beiträge mit der Makroebene der kollektiven Wissensproduktion.

Zusammenfassend bietet diese Dissertation einen eindeutigen Beitrag zur Analyse, Verständnis und Förderung von informellen Kontexten der Wissensproduktion, die auch zunehmend für die Politik und Praxis von formellem Lernen an Bedeutung gewinnen. Darüber hinaus macht meine Arbeit das relevante und neuartige Phänomen der Online-Massenkollaboration zugänglich für die theoretische und empirische Forschung in der CSCL und steuert gleichzeitig wertvolle methodische Ansätze für das neue Forschungsfeld der Learning Analytics.