Journal of Universal Computer Science, vol. 14, no. 4 (2008), 526-545
submitted: 3/10/07, accepted: 11/2/08, appeared: 28/2/08 © J.UCS

Analyzing Wiki-based Networks to Improve Knowledge Processes in Organizations

Claudia Müller
(University of Potsdam, Germany
cmueller@wi.uni-potsdam.de)

Benedikt Meuthrath
(University of Potsdam, Germany
bmeuthrath@wi.uni-potsdam.de)

Anne Baumgraß
(University of Potsdam, Germany
abaumgrass@wi.uni-potsdam.de)

Abstract: Increasingly wikis are used to support existing corporate knowledge exchange processes. They are an appropriate software solution to support knowledge processes. However, it is not yet proven whether wikis are an adequate knowledge management tool or not. This paper presents a new approach to analyze existing knowledge exchange processes in wikis based on network analysis. Because of their dynamic characteristics four perspectives on wiki networks are introduced to investigate the interrelationships between people, information, and events in a wiki information space. As an analysis method the Social Network Analysis (SNA) is applied to uncover existing structures and temporal changes. A scenario data set of an analysis conducted with a corporate wiki is presented. The outcomes of this analysis were utilized to improve the existing corporate knowledge processes.

Keywords: social software, wiki, knowledge work, network analysis, collaboration network
Categories: A.1, H.0, H.4.3, H.4.m, J.3, K.4.2, K.4.3, M.4

1 Introduction

Formerly centralized hierarchical organizations are increasingly transforming to decentralized and network organizational constructs, which promote more openness. In these systems which are subject to continuous change processes, people and available information are key resources [Barabási 03], [Tapscott and Caston 93]. Flat hierarchies and flexible structures, team orientation as well as a progressively growing dependency on knowledge are fundamental characteristics of the changing corporate landscape today [Cross and Parker 02]. These changes in organizational structures have also impacted corporate knowledge management. To ensure organizational learning, knowledge communities should be fostered [Romhardt 02]. These knowledge networks within an organization, which are not organizationally authorized, were initially analyzed in the context of Communities of Practices (CoP) [Wenger 98]. Information Technology (IT) is one important enabler for a successful knowledge community. In this context wikis are increasingly used for decentralized, collaborative knowledge management of corporate knowledge communities [Tapscott

and Williams 07], [Klobas 06]. The first wiki, implemented by Ward Cunningham was designed based on various design principles [Leuf and Cunningham 01]. These principles influence wiki software, the user behavior and the application of knowledge management (cp. Table 1). A wiki is "open", implying that each person (employee) can consume and edit every page. Information (externalized knowledge) in a wiki is mergeable or dividable depending on existing knowledge requirements. Also, information context is arbitrarily adaptable by linking specific articles. When users externalize their knowledge they can link it to existing or non-existing articles (displayed in another link color). In the latter case, a wiki enables users to formulate their existing knowledge needs. A wiki information space has initially no predefined structure. The structuring of information objects (articles) is specified by a collaborative self-organized process controlled by the users [Klobas 06]. All article versions are saved in a history. Usually the final product of knowledge creation is visible but not the actual process of knowledge creation. With the page history the process from the first draft to the final version is traceable. A critical issue of knowledge work is solved in this way.

Despite all known advantages it is still not proven whether wikis are an adequate knowledge management tool to foster knowledge communities in organizations.

Design principle	**Description**	**Impact on knowledge management**
Open	Each user can see and change all content.	Each employee is competent; knowledge is freely available and shareable.
Incremental	Content (article) is linked to article which does not exist yet.	Knowledge gaps are visible; efficient development of knowledge.
Organic	Development of structure and content is evolutionary.	Knowledge and its context is dynamic; developments depend on existing requirements.
Simple	Minor number of syntactical rules.	Low barriers of usages during knowledge documentation.
Universal	Creating, changing, and structuring of content follows the same principles.	No definition of knowledge management roles necessary.
Precise	Pages should be named clearly to avoid conflicting names.	The context of knowledge is considered.
Observable	The content development is retraceable by each user.	The origination and development of knowledge can be analyzed.
Convergent	Avoid duplications by linking existing content.	Redundant knowledge is mergeable.
Trust	Trust as central principle.	The success is dependent on a company's culture.

Table 1: Design principles of wikis and its impact on knowledge exchange processes [Cunningham 07], [Müller and Dibbern 06]

This paper presents a novel approach to analyze existing networks in wikis. The Social Network Analysis (SNA) is used to uncover existing structures and its temporal changes. An analysis of a corporate wiki shows first results of this approach. Two hypotheses constitute the base for the presented research in progress:

1. Social Network Analysis is an adequate method to analyze wiki information spaces
2. The collaboration network shows the nature of cooperation in a wiki and reveals special roles in the network.

Based on these hypotheses a scenario is presented, which introduces the theoretical concept in a corporate environment. This paper is organized as follows: the following section introduces SNA and possible application areas in knowledge management. Moreover, basic SNA metrics are explained based on a specific classification model. In the following paragraph four perspectives of wiki-specific networks and their characteristics are described. Finally, selected results of a conducted exploration of a collaboration network in a company are presented and results are discussed. The last section gives an insight into future tasks.

2 Applying social network analysis in knowledge management

Social network analysis (SNA) is an enabler to systematically describe and investigate network structures and their principles of order in organizations. Fundamentally, a social network is "a specific set of linkages among a defined set of persons, with the additional property that the characteristics of these linkages as a whole may be used to interpret the social behavior of these persons involved" [Mitchell 69]. The social environment is described by a pattern or regularity among the interacting persons. Social relationships of persons constitute the social capital of an organization. The social capital metaphor is that people who do better are usually better connected. SNA is increasingly applied for analyzing and reorganizing distributed communication and collaboration structures in knowledge management (e.g. [Cross and Parker 02]). Existing (cross-functional) relations are determined in their relation strength and intensity. Expertise within these knowledge networks can be located. The required data for network analysis can be gathered electronically or via interviews. A common approach utilizes user interaction with computer systems (e.g. email, blogs, and wikis) as applied in this contribution. To analyze these networks a transformation in a graph is necessary to enable the quantification and measurement of network properties [Wassermann and Faust 97, p. 93].

2.1 Basics of graph theory

Different disciplines and research areas have lead the network analysis to its current importance. An essential breakthrough for network analysis is based on the graph theory. This theory enables the formal description of a network. Mathematically nodes in a network are the vertices and relations are the edges in a graph. To analyze a network the respective network has to be transferred in a graph. A graph $G(N,L)$ consists of a finite number of vertices $N=\{n_1, n_2, ..., n_g\}$ and edges $L=\{l_1, l_2, ..., l_L\}$. In a graph there are g vertices and L edges [Wasserman and Faust 97]. Two vertices in a graph are adjacent if the edge $l_k=(n_i, n_j)$ is in the set of the edges L. A graph $G(N,L)$

can be presented as diagram. Its simplest form is a sociogram, the oldest and most known approach [Moreno 54]. Here nodes are persons and edges are relations between these persons. This visualization is applied to identify leading or isolated persons in groups, to find asymmetry or reciprocity in relationships and to describe indirect relations. The location of the points and the length of the lines depend on the used algorithm.

Edges in a graph can be undirected or directed. Whereas undirected graphs possess no direction, edges (arcs) of a directed graph cross in one direction only, e.g. hyperlinks in an HTML document. A directed graph or digraph consists of a set of vertices and arcs, whereas each arc is an ordered pair of distinct vertices $l_k=<n_i, n_j>$. The arc is directed from the sender n_i to the receiver n_j. In weighted graphs edges carry additional information like strength or intensity, e.g. frequency of interaction, rating of friendship. Here, a set of signs $S=\{s_1, s_2, ..., s_L\}$ attached to edges extend the graph description.

A graph might comprise of sub-graphs and/or components [Wasserman and Faust 97, p. 97]. A sub-graph can be defined using specific vertices or edges. Specific vertices or edges constitute a basis to construct a sub-graph, which is a part of the whole graph. Also, in a graph not all vertices are connected, and then these existing subsets are defined as components.

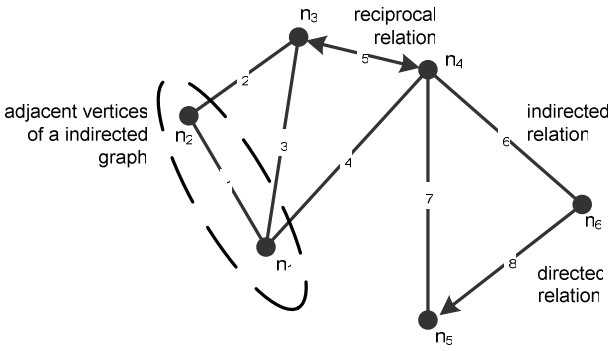

Figure 1: Basic characteristics of a graph [Müller 08]

A graph can also be represented as adjacency matrix. In an adjacency matrix there is a row and a column for each vertex. Entries of a matrix indicate if two vertices are adjacent.

2.2 Classifying network analysis metrics

A network is a fundamentally complex system. Based on changing external influences, these complex systems have a certain structural variability. This structural variability comes along with different phases of a network. Based on specific parameters transitions from one phase to another are possible based on the internal dynamic of the network. There are different structures in different phases. Each

phase is related to new macroscopic characteristics of a network. However, it is very important to differentiate between random fluctuations and phase transitions which strongly influence users' behavior. Therefore, an analysis of different orientations of networks is necessary. Four analysis methods are differentiated: attribute, position, structure, and dynamics analysis. Depending on the objective of analysis the specific method is applied. Figure 2 shows this classification.

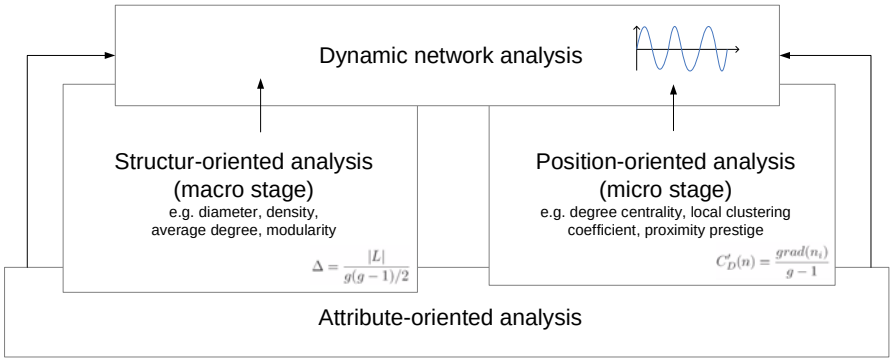

Figure 2: Scopes of network analysis with selected metrics [Müller 08]

A typical network analysis focuses on analyzing relational data [Scott 00, p. 2f.]. Nevertheless, a network offers in addition to relational data also attribute data. In an attribute analysis, the network data are combined and evaluated in relation to these conventional data types [Jansen 03, p. 51]. The attribute analysis as analysis of characteristics enables capturing the properties of vertices and edges and to relate them to the structure of the network. These data are investigated by applying the exploratory statistics [Tukey 77]. There is only minor information about the relations of these dates. One method is data mining where specific patterns are extracted from data [Fayyad et al. 96, p. 39]. A combination of relational and attribute data takes place in network data mining. Conventional explicit data types from data mining are augmented by relational data from network analysis to implicit relations between data sets [Galloway and Simoff 06]. Attribute analysis is especially important for wikis. Figure 3 shows the basic relation between data-base concepts of MediaWiki. Central are articles which are related by wiki-links. A category contains a number of articles and an article refers to a category. Also articles use images and templates and they are created by users. Changes in articles are recorded by revisions. These basic concepts are applied to define different metrics.

The wiki size is based on the number of pages, users and authors. The page count (PC) compromises all pages, that means discussion pages, personal pages, organizational pages, and content pages. This measure gives an impression of the wiki size in terms of article number. The article count (AC) measures only the content pages of a wiki and therefore only the pure content of the wiki (without files and organizational pages) are counted. Also the number of categories, images, and templates can be determined. Here, specific analyses are possible, e.g. how structured is the investigated wiki in terms of category usage. Another metric is the media count which only counts all files in a wiki information space, e.g. images, doc, pdf. This

measure is important to evaluate if the wiki is used as content management system, when the number of files is high compared to wiki pages. Each revision in the wiki information space increases the edit count (EC). But while each revision is counted, also very small ones are considered. Therefore minor changes should be added out. The edit count shows how lively a wiki is. The content size, which an author contributes to an article, is measured by the amount contribution (AmC). The AmC calculates the number of characters in a Byte (1 character = 2 Byte). Some authors can have a high edit count but a minor amount contribution. So, specific investigations of what work has been done in the wiki are possible. The utilization of articles can be identified based on the view count of an article (VC). The view count is a measure which sums the number of clicks on a wiki page. But the gathered value is only a benchmark, because not each page request is connected to an internalization of knowledge.

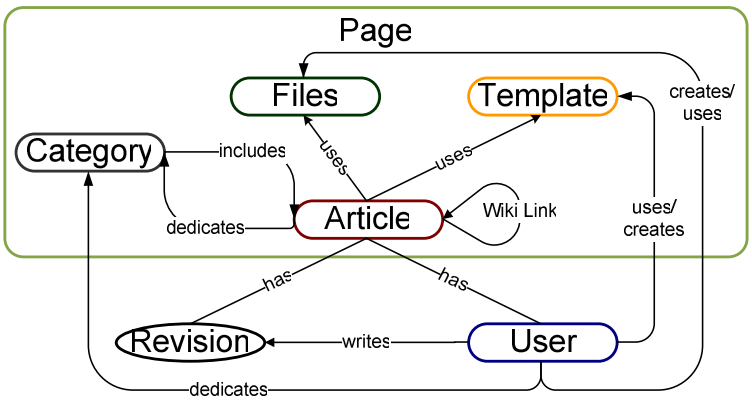

Figure 3: Concepts in MediaWiki and its relations

A position analysis investigates the micro stage of a network. The analysis aims for investigating single nodes of a network, their characteristics and their position in the network. One simple measure is degree of a node. The degree of a vertex $d(n_i)$ is the number of vertices that are incident with it and reflects the activity of a person in a social network [Wasserman and Faust 97, p. 100]. In the basic centrality concept three approaches can be differentiated: the degree centrality, the closeness centrality and the betweenness centrality. Here the first and the third concept are considered.

The activity of a person in a social network can be investigated with the metric degree centrality. This metric helps to identify the "most important" actor, meaning those that are extensively involved in relationships with other actors in the social network. A vertex is central in the sense of degree centrality, if a vertex has many relations to adjacent vertices. The degree centrality $C_D(n_i)$ is the relation of degree $grad(n_i)$ of a vertex to the number of vertices $|g|$ excluding the considered vertex in a graph:

$$C_D(n_i) = \frac{grad(n_i)}{g-1}$$

Whereas the betweenness centrality (C_B) based on the idea that one person is then important if this person lies often on a shortest path between two persons:

$$C_B(n_i) = \sum_{j<k} \frac{g_{jk}(n_i)}{g_{jk}}$$

Where g_{jk} is a shortest path between n_j and n_k and $g_{jk}(n_i)$ a shortest path that n_i lies on. In this case this specific person can "control" interactions between two nonadjacent persons.

With the structural analysis the macroscopic characteristics of a network can be studied. Measures to investigate the structure of the network are for instance density, average path length, and average degree. These metrics are used here to describe the network structure on a macroscopic level. Density of a graph G(g,L) is the ratio of relations L over number of maximal possible relations in a graph G [Wasserman and Faust 97, p. 101]. The number of possible relations is reached, when each vertices is connected to all other vertices. The number of relations is limited through the number of existing vertices and is maximal |g|/2 = g(g -1)/2. In this case the density of the graph is 1 and the graph is called complete. The density is calculated as follows:

$$\Delta = \frac{|L|}{g(g-1)/2}$$

The average path length of a graph is a measure for the average number of vertices which has to be passed in order to get from one vertex to another. The path is a sequence of vertices, while each vertex is connected to the previous and succeeded vertex. Short distances transmit information accurately and in a timely way. A long distance degree implies slower even distorted information transmission [Cross et al. 04].

The degree of a vertex is the number of edges that are connected to this vertex (adjacent nodes). This measure helps to evaluate the "attractiveness" of a vertex. The number of all neighbours of n is described as $g(n_i)$ whereas $grad(n_i)$ is the degree of n_i. A vertex is isolated, if $grad(n_i) = 0$. The average degree is calculated for all vertices in a graph:

$$d = \sum_{j=1}^{g} \frac{d(n_i)}{g}$$

Another metric is the cluster coefficient which frequently refers to transitivity [Newman 01, p. 408]. The clustering coefficient of a vertex i is the quotient from number of triangles $N_\Delta(i)$, which is connected to the vertex i and the number of triples $N_3(i)$, in which I is the centre:

$$C_i = \frac{N_\Delta(i)}{N_3(i)}$$

In Figure 4 the calculation of this metric is visualised based on an example. Here, the cluster coefficient of a vertex i has to be calculated (left side). At first the number of triangles is defined (middle). Then for each vertex which is connected to the vertex i, the number of triples is determined. On the right side shows this based on vertex k.

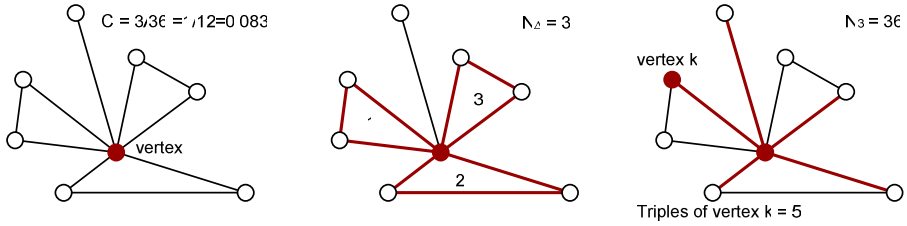

Figure 4: Example of calculating cluster coefficient for vertex i [Müller 08]

All measures of the introduced analysis can also be utilized to investigate dynamic characteristics of networks. The dynamic perspective of network analysis deals with the progression of state changes over time and enables the observation of events and its analysis. It is necessary to compare at least two different network states at different times to determine changes of any form in the network. Dynamic network analysis (DNA) enables analyzing network states, the vertices and their edges as well as changes in structure and configuration of the network [Carley 03]. The cumulative analysis enables the aggregation of a long time period of the network, including all changes and elements, which are or were part of the network during the period. The calculated measures and their changes are visualized in a diagram. Basically, the network dynamic and the network evolution can be differentiated [Stockman 97, p. 234]. The network dynamic is a general concept in which the change in time is described. In opposite, within network evolution adaptive processes are source of changes. These processes transform the network structure. There are two kind of analysis in DNA; cumulative analysis and sliding-window based analysis [Moody et al. 05]. Their difference lies in the kind of recording the network respective network parameters. The cumulative analysis enables the aggregation of a long time period of the network, including all changes and elements, which are or were part of the network during the period. The recorded strip can be interpreted as a complete image of the network. The image is growing over the period of time until it reaches the current state of the original network. In the sliding-window based approach a small period of time of the network or a sub-network is analyzed. The main objective of such investigation is to analyze just a subset of parameters or small sets of nodes, ties, etc., where the general time context can be disregarded [Moody et al. 05]. In our analysis, we apply the first approach to examine structural changes in wiki networks.

2.3 Specifying wiki networks

A wiki with all related activities forms a specific information space in a company. This information space consists of different networks. Based on a classification made by Carley [Carley 04], wiki-specific networks can be arranged in four categories (Figure 5): social perspective (who knows who), knowledge perspective (who knows what), information perspective (what refers to what), and temporal perspective (what was done before). Relationships in one network usually imply relationships in another [Carley 04]. Therefore, to understand all knowledge processes in a wiki an analysis of these different network types is necessary.

The first group contains collaboration networks and discussion networks. Both types are social networks (cp. Figure 5). A social network consists of persons as nodes and their relationships as edges. A *collaboration network* is used to investigate the nature and extend of collaboration between persons, also known as co-authorship network. In wikis it is a network of authors of wiki articles and enables the analysis of information exchange between communities. Collaboration is based on mutually referred asynchronous generated contributions in a knowledge process. In a collaboration network vertices are authors, whereas the edges are constructed on common edits of an article. Whenever two or multiple authors worked on the same article in a wiki, they are connected to each other. The central assumption is that changes of different authors on an article imply collaboration. A strong tie exists, if two authors have contributed comparatively more then other authors (in terms of content) or they work often collaboratively on one or more articles (in terms of frequency). By analyzing collaboration networks the time component has a high impact because cooperation is possible over a long period. Consequently, the effect of aging on network structures has to be considered [Zhu et al. 03]. With the second network in this perspective - the *discussion network* - a topic specific communication process on a wiki talk page can be studied. Vertices are authors of specific postings and edges are related postings. The hierarchical order of these postings is defined by the time of edits. An activity level of discussions in a wiki information space is measurable. The last network in this perspective is the *message exchange network*. Each registered user can exchange messages. This exchange is carried out by an edit on a personal talk page (like an entry in a guest book). Users get informed about new messages on their personal talk pages for instance via email. Nodes are users and relations exist based on exchanged messages in this network. This network supports analyzing and evaluating the amount and characteristics of communication processes in wikis.

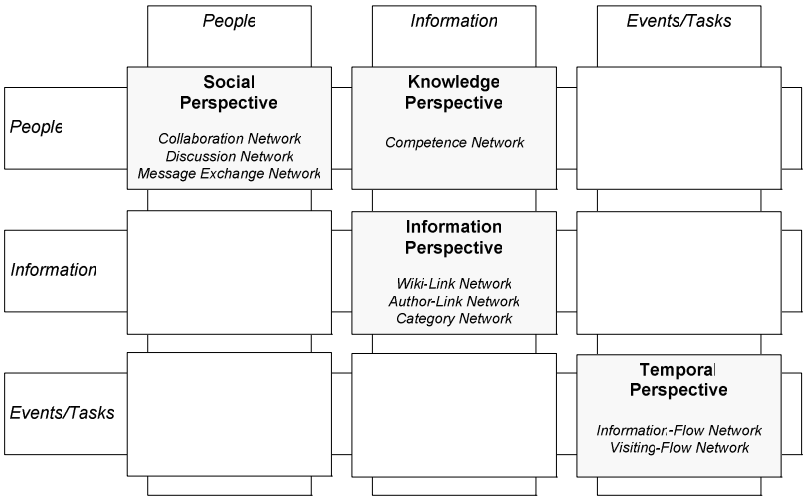

Figure 5: Specification of wiki networks depending on different perspectives [Müller and Meuthrath 07]

The second perspective comprises of knowledge networks (cp. Figure 5). Wiki specific realizations are *competence networks*. This type of network shows existing thematic competences of authors. It is a two-mode network which means there are two different types of nodes – authors and articles. The directed edge between two vertices (article, author) represents the editing activity of a certain author for a given article. An investigation of a competence network gives an impression which author works on which article very often and/or has contributed a lot to one article. This is an index of existing knowledge in a specific topic area. A main assumption in this network is that if authors edit articles they externalize their knowledge as well. Minor changes, e.g. typographical errors, are left out of these considerations. An interpretation of competence networks is more complicated than in the networks already introduced due to their character.

The information perspective defines information networks (cp. Figure 5). In the wiki context there are wiki-linked networks, author-linked networks, and category networks. A *wiki-linked network* consists of articles as vertices and wiki-links as edges. It is a directed network since wiki-links are not reciprocal. The temporal analysis of wiki-linked networks enables the investigation of structural developments of themes in wikis. The evolutionary process of emerging and "dying" topics is visible in this way. A structural analysis based on degree centrality shows highly connected areas in this network. Identifying topic clusters is another structural approach. Existing theme affiliations can be detected by text analysis. The second network in this perspective is the *author-linked network*. It focuses on articles with a high modification rate. The vertices are articles and edges represent changes on both articles by a certain author. The weight of such a relation increases with the number of authors which work on both articles. Opposite to the wiki-linked network, these network centers emerge based on thematic relations of authors and not of articles. Furthermore, this network describes the "attractiveness" of a single article in the information space. One disadvantage of this network is, that authors, who are active with one article only are neglected. Network usefulness decreases, if the number of such authors is high in the investigated information space. The third network is a simple hierarchical network – the *category network*. In wiki information spaces content is often structured in categories. Besides a hierarchical structure of categorization there are also cross-links between the various category entries. One single category consists of thematic similar topics. The objective of this network is to describe these categories and their utilization (the more articles a category has the higher is the diameter of nodes). The category network comprises category pages as vertices and main-category/sub-category relationship as directed edges. Based on a temporal analysis the development of articles in specific categories can be studied.

The fourth perspective comprises information-flow and visiting-flow networks. Especially in specific topic areas, an analysis of the temporal development of articles and their interdependencies is interesting. The *information-flow network* sequences wiki articles. The nodes are articles and the relations are defined based on the article history and/or the day of creation. Therefore, these relations show the temporal ordering of article creation or change. The measuring of an information flow from one article to the other is not trivial and time is not a sufficient parameter. Rather the content of articles has to be considered. This means a relation should be defined by transferred text from one article to the other. In the wiki context it may be interesting

to gather information about how the wiki is used not only by its contributors but also by its readers. The *visiting-flow network* sequences wiki pages by its temporal order by means of the *referrer field* [Fielding et al. 99, p. 140] that can be found in the web server log. A referrer is the page a reader visits directly before the actual page. This information is sent by most browsers and is logged by web and proxy servers in default configuration. Assuming that there are two starting points (external and no referrer) the referrer information is sufficient to generate a directed network with wiki pages as nodes and their referrer relationship as edges, weighted by its frequency. To clean up results editing and resource access (style sheets, java scripts etc.) should be filtered.

3 SONIVIS:Tool

Besides the specification of a common analysis model, a tool - SONIVIS:Tool - is developed (for more information see www.sonivis.org). SONIVIS:Tool is a Java-based, open-source software, which based on eclipse rich client platform (RCP). Eclipse was proprietarily developed by IBM and published under an open source license in 2001. RCP enables the creation of generic applications with a native graphical user interface. It offers a simple dynamic component based model to develop plug-ins, which contains the functionality of the application. Initially, the Information Visualization Cyber Infrastructure (IVC) framework was used to integrate the SONIVIS:Tool functionality on top of. This framework was adapted to fit the needs and by now there is only little left of it. In the future there will be a completely new architecture and IVC will not be used any more. Network visualization is implemented using the prefuse visualization toolkit.

There are predefined graphical analyses which offer a quick overview on actual wiki states or developments, e.g. author activity levels, edit growth, collaboration index. Additionally, various network analysis metrics are provided on a microscopic and macroscopic level. Metric calculation is done by GNU R, an open source software environment for statistical computing and graphics.

SONIVIS:Tool essentially follows the model-view-controller (MVC) pattern, that means data, functionality and user interface are separated. The architecture (cp. Figure 6) provides an eclipse extension point for connector plug-ins to implement other data sources (like other wikis or blogs etc.) and extension points for other analysis. Figure 6 uses Fundamental Modeling Concepts (FMC) to model the architecture. The FMC is a modeling technique to support the communication about information processing systems. It is not tied to a programming paradigm and uses a simple notation which can be used easily for ad-hoc creation of models. FMC defines three fundamental aspects of information processing systems [Knöpfel et al. 05]:

1. *Compositional Structure* is mapped to a block diagram and used in Figure 6 to describe the structure of the system in terms of active and passive components. Squares are used to model agents that are active and process information, while rounded figures are used to model storages and small circular figures to model communication channels, which are both passive components to keep or transport information. This structure describes which agent can access what data and communicates with other agents via channels or shared storages.

2. *Behaviour* is modelled by petri nets and describes the behaviour of agents that operate on data, and their reaction to requests which they receive via channels.
3. *Data / value structure* is mapped to Entity / Relationship diagrams to describe the structure and the relationships between data in storages.

The compositional structure of SONIVIS:Tool is modelled in Figure 6 to describe the fundamental architecture. The "Connector Layer" contains a plug-in that implements the connector extension point, network loaders, which implement the network definitions in chapter 2.3 in a domain specific way, and calculators, which are restricted to domain specific data. Manager components in the "Application Layer" handle things like a connection to GNU R, connectors and calculators, which are present to runtime, or events. The "Presentation Layer" consists of five complementary views: One each for visualizing the network, network metrics, node metrics, some statistic plots and a best of list of nodes, where the user can choose a metric to show the top ten results.

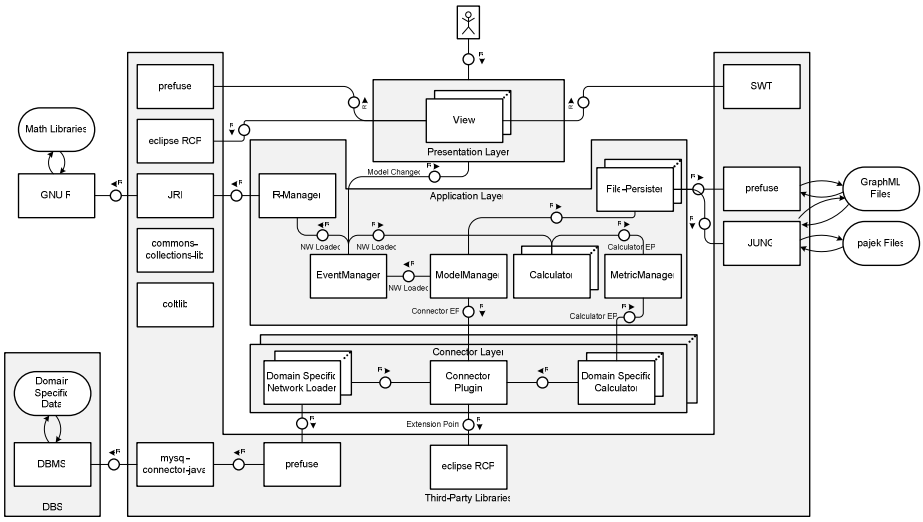

Figure 6: SONIVIS:Tool architecture in FMC

After loading the selected databases automatically basic wiki measurements are conducted and visualized. These metrics are categorized in the following areas: wiki averages, wiki activity, wiki amount and wiki ratio. In addition, first statistical measurements (e.g. ratio user/author) and temporal measurements (e.g. author growth, number of articles, and rate of change) are visualized graphically. These measurements offer users a fast overview about the existing developments in the investigated wiki.

During visualization of wiki networks amongst others the visualized period can be defined. When the network data are non-anonymous, node-names are displayed as the node IDs. Furthermore the weight of the relations between the nodes is shown depending on the selected network. Parallel to a visualization of the network specific

metrics on the macro level are measured and represented. The same applies to node metrics, which describe the micro-level. Here the user has the option to visualize the metrics in diagrams as well. A best-of view calculates a ranking between all network nodes based on specific metrics.

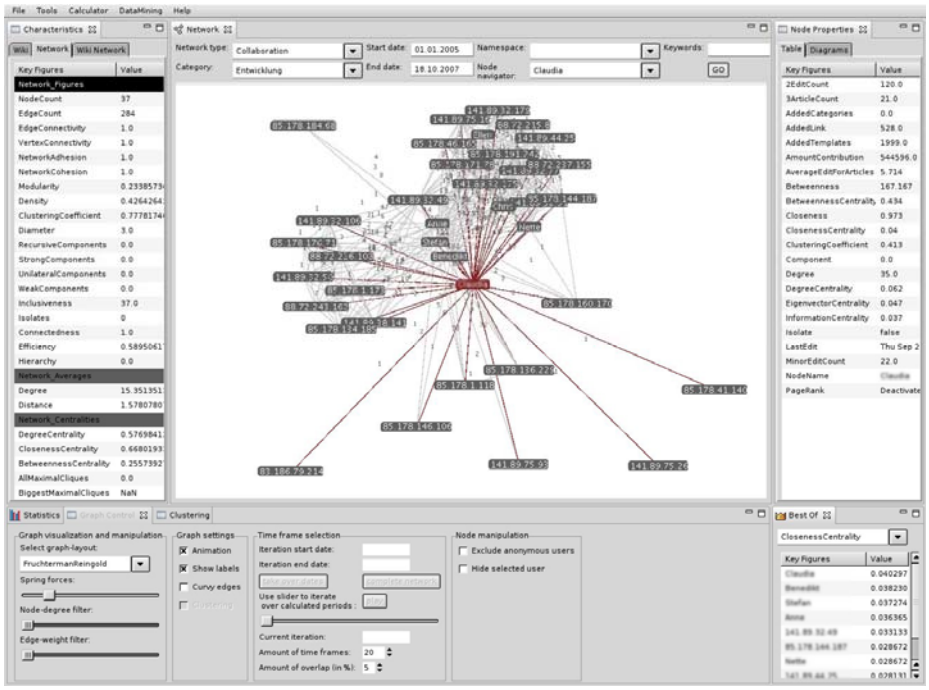

Figure 7: User Interface SONIVIS:Tool

Additional functions like the calculator and the clustering are offered. The application of the calculator enables pre-calculations of the wiki-database. All calculated metrics and networks are automatically transferred in a XML-format. At a later date these calculation can be loaded and manipulated in SONIVIS:Tool. The clustering extension facilitates the application of e.g. the k-means algorithm. The network can be partitioned in single cluster, visualized by convex covers, which for instance presents a participation level.

4 Analyzing and evaluating wiki networks to improve knowledge work

The predefined wiki networks are applied to enhance existing network analysis approaches. Using these networks provides a comprehensive insight into existing knowledge processes in a wiki in terms of people interaction, topic emergence, communication activities and their dynamic changes. Here, the collaboration network is visualized and analyzed to emphasize this new approach. The data set used is

based on a wiki that was launched in the summer of 2005. It serves in a software company as a knowledge transfer platform between 12 business units and one central unit with about 1,000 employees. The wiki is a replacement for the former intranet solution where only a dozen employees were able to change content. The objectives of implementing a wiki as intranet solution were

- Designing a knowledge exchange platform
- Exchanging best practices between departments
- Documenting operational technical information
- Reducing volatile knowledge
- Developing a "condensate" of knowledge

In July 2007 the corporate wiki had 353 registered users and 261 authors, which created about 10,000 pages with more than 35,000 revisions (cp. Table 2).

Metric	**31.12.05**	**23.07.06**	**30.12.06**	**24.07.07**
User Count (UC)	49	142	252	355
Author Count (AuC)	49	92	176	263
Page Count (PC)	919	2,418	5,904	10,072
Article Count (AC)	407	937	1,419	1,954
Media Count (MC)	261	913	2,892	5,398
Article Edit Count (EC)	2,363	5,630	11,004	18,744
Article View Count (VC)	133,483	198,072	260,373	284,240
Amount Contribution (AmC)	~5 MB	~17 MB	~37 MB	~82 MB

Table 2: Growth of wiki information space

According to Figure 3 there is a difference between Page Count and Article Count. This wiki is partly used like a file system – about half of the pages are files such as pictures, documents etc (cp. Table 2 PC and MC). There is also a page for every file in the wiki (cp. Figure 3) and this might be collaborative created, but that can only be tracked, if every version of the file is uploaded to the wiki. Article Edit Count and Amount Contribution measures the extent of wiki content production whereas the View Count measures the consumption of information in a wiki. In a knowledge management manner it is satisfying to see that articles were not only edited but also read (cp. Table 2 EC & AmC vs. VC).

Figure 8 yields a first insight into the development of the collaboration network. At the beginning of the investigated timeframe (07/05) there is one area (sub-network) where specific vertices are closer than the others. There are also nodes with no connection to the main network. This picture changes over the investigated period (07/06). The network is growing. More and more authors join the network and existing collaborations get stronger in certain areas. In the third snapshot (12/06) there is one specific sub-network showing strong relations, but this group has grown compared to the year before. There are still a number of lightly connected nodes in the periphery of the collaboration network. These authors do not collaborate much, maybe because they work only on specific articles in the wiki. The last visualization (07/07) shows three subgroups. A conducted analysis reveals that these subgroups have been formed around specific projects. Therefore, it will be interesting to

examine how the wiki structure is changing because of external organizational changes.

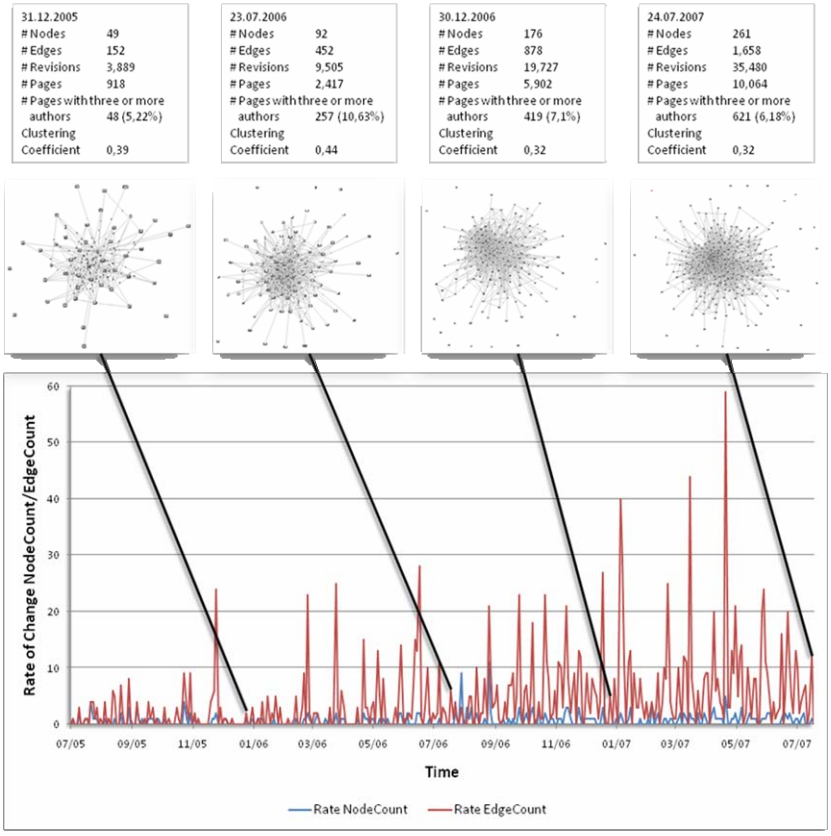

Figure 8: Temporal development of a collaboration network

A pure visualization of a network is not sufficient for understanding the whole evolutionary process. Hence, specific metrics should be used to analyze the network structure on a macroscopic and microscopic level. Figure 9 shows on the left the temporal development of the network's density and on the right side the temporal development of the average path length. Both measures allow an evaluation of the network on a macroscopic level. The density declines continuously during the considered period. However, Figure 8 shows that there are specific zones in the network with a density of almost 1 (so called cliques). The average path length increases during the evaluated period. Although the density is extremely small (0.05) in 07/07 the average number of nodes between any two nodes in the network (average path length) is not that high (2.26). There must be *hubs* (also called *short cuts*) in the network, which connect far-off parts of the network or, in small networks, nearly each node. This phenomenon is often noticed in social networks and called *small world*. Consequently, the spread of information in this network needs nearly the same time in

the whole analysis period although the network is not as dense as in the beginning. A correlation analysis was constitutively conducted based on these results to reveal existing relations between the collected measures.

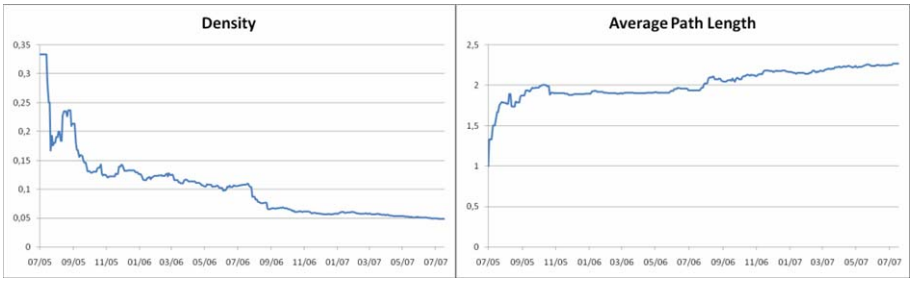

Figure 9: Density and Average Path Length

A positive correlation between article count and average degree (0.85) and a negative correlation between density and average path length (-0.92) was determined. The first result is attributed to a minor collaboration in the whole network but a high collaboration in a subgroup. This is in accordance with the results of visualization (cp. Figure 8). A negative correlation implies that the average distance increased due to emerging subgroups. Simultaneously, the density decreases but not at the same rate.

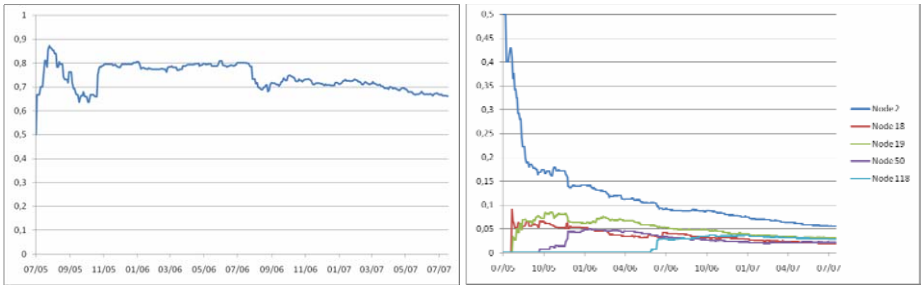

Figure 10: Temporal change of degree centrality on a macro- and microscopic level

This result and the network's degree centrality (cp. Figure 10 left side), which measures a high heterogeneity of degree centrality (0.66), is also an indication for existing hubs in the network. In a collaboration network hubs are persons with a high level of activity because they are connected to a lot of persons through different articles. In early stages of a wiki so called Wiki-Champions have a critical importance for the development of a wiki information space. They are recognized as early adopters who understand how to use a wiki very well, encourage others, and get people involved by informally training them and being available for ongoing support. They serve as a model in using wikis (http://www.wikipatterns.com). In this case one extremely active person can be identified in the beginning. As time moves on, this person looses its special position gradually. But there is no reason for concern because the activity level in the wiki is increasing overall. Also at the same time

other persons gain importance in the network (cp. Figure 10). More and more people join the wiki-Champion and engage more people in the wiki.

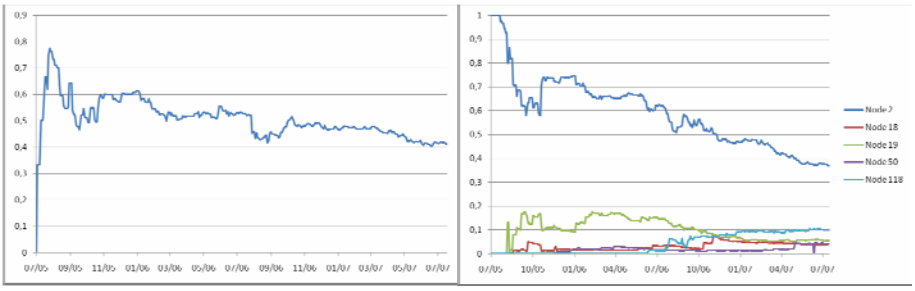

Figure 11: Temporal change of betweenness centrality on a macro- and microscopic level

Whereas degree centrality reveals persons with a high level of collaboration, betweenness centrality is a measure for the impact a person may have on the flow of information in a collaboration network [Newman 01a]. A person with high betweenness centrality is influential and a removal results in the largest increase of average path length and therefore in less efficiency in terms of information spreading. The analysis on a macroscopic level reveals a heterogeneous betweenness centrality (cp. Figure 11 left side). On a microscopic level there is one node with by far the largest measure (cp. Figure 11 "Node 2", the top curve). This person is highly collaborative and on the same time influential in terms of information spreading in the wiki. On the one hand this can be positively interpreted, because this person is able to spread important information in a systematic way. On the other hand this might be a danger for the development of the network, because this person can control the information that passes him. Additionally, if he drops out the structure of collaboration, this wiki may change dramatically.

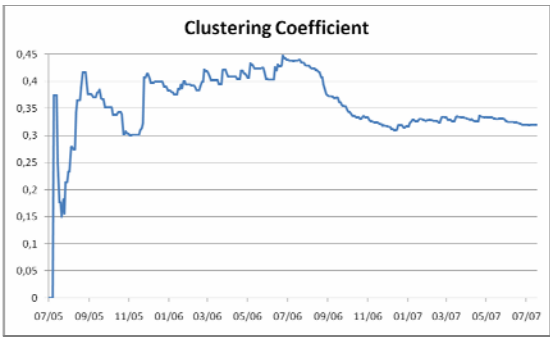

Figure 12: Temporal change of clustering coefficient on a macroscopic level

Network analysis of collaboration is not only able to reveal information about persons and their position in a network but also to point out the collaboration's

structure. The *clustering coefficient* for example is a measure, which describes the probability of transitive triangles. In graph theory clustering coefficient for networks with quite a few nodes tends to be zero. But social networks possess a clustering coefficient greater than zero, because there is a finite (and probably quite large) probability that two people will be acquainted if they have another acquaintance in common [Newman 01]. In the analyzed wiki this can be proofed (cp. Figure 12). In terms of collaboration the clustering coefficient may be a positive sign, because it is always good, if more than two people work on an article. But this is not the only interpretation of this measure – it could be that the three authors of a transitive triangle worked pair wise on one article and this tends to result in a high clustering coefficient. And compared with the number of articles that were edited by three or more authors (cp. Figure 8) this measure is not the only factor for high clustering coefficient, even though there is a correlation between those measures.

Based on these selected results the SNA enables the analysis of existing collaboration in a wiki and reveals specific wiki roles such as wiki Champions and gives an insight in trends that may be problematic.

5 Outlook

This paper describes methods on to how to analyze knowledge activities in terms of collaboration in a wiki. The network analysis can be used to enable a measurement of an existing network. Due to the complexity of such systems four groups of networks are defined to analyze existing interdependencies. This work was done by a research group - the SONIVIS:Team (http://www.sonivis.de). In the future, all defined networks will be implemented and enriched with further analysis like text mining or clustering to enhance analysis work and to improve our understanding of social processes in a wiki information space. The SONIVIS:Tool will be based on a new architecture to reach scalability and a simple extension mechanism. The team is currently working on a reorganization of the presentation layer in order to simplify the handling and visualization of the analysis data.

References

[Barabási 03] Barabási, A.-L.: “Linked”; Plume Printing, New York (2003).

[Carley 03] Carley, K.M.: “Dynamic Network Analysis”; In: Brelger, R., Carley, K. M. and Pattison, P.: “Dynamic Social Network Modeling and Analysis: Workshop Summary and Papers”, National Academy Press, Washington, D.C., (2003), p. 133-145.

[Carley 04] Carley, K. M.: “Dynamic Network Analysis”; In: Brelger, R., Carley, K. M., Pattison, P. (eds.): “Dynamic Social Network Modeling and Analysis: Workshop Summary and Papers”, National Academy Press, (2004), 133-145.

[Cross et al. 02] Cross, R., Parker, A., Borgatti, S. P.: “A bird’s-eye view: Using social network analysis to improve knowledge creation and sharing”;
URL: http://www-1.ibm.com/services/us/imc/pdf/g510-1669-00-a-birds-eye-view-using-social-network-analysis.pdf (2002).

[Cross and Parker 04] Cross, R. L., Parker, A.: "The Hidden Power of Social Networks: Understanding how work really gets done in organizations"; Harvard Business School Press, Cambridge (2004).

[Cunningham 07] Cunningham, W.: "Wiki Design Principles";
URL: http://c2.com/cgi/wiki?WikiDesignPrinciples (2007).

[Fayyad et al. 96] Fayyad, U., Piatetsky-Shapiro, G., Smyth, P.: "From Data Mining to Knowledge Discovery in Databases"; American Association for Artificial Intelligence Magazine (1996), p. 37-54.

[Fielding et al. 99] Fielding, R. et al: "RFC 2616. Hypertext Transfer Protocol -- HTTP/1.1". URL: http://tools.ietf.org/html/rfc2616 (1999).

[Galloway and Simoff 06] Galloway, J., Simoff, S. J.: "Network data mining: methods and techniques for discovering deep linkage between attributes"; In: Proceedings of APCCM (2006), p. 21-32.

[Jansen 03] Jansen, D.: "Einführung in die Netzwerkanalyse" (in German); 2nd Edition, Leske + Budrich, Opladen (2003).

[Klobas 06] Klobas, J.: "Wikis: Tools for Information Work and Collaboration"; Chandos Publishing, Oxford (2006).

[Knöpfel et al. 05] Knöpfel, A., Gröne, B., Tabeling, P.: "Fundamental Modeling Concepts. Effective Communication of IT Systems"; John Wiley & Sons Ltd., Chichester (2005).

[Leuf and Cunningham 01] Leuf, B., Cunningham, W.: "The Wiki Way - Quick Collaboration on the Web"; Addison-Wesley, New York (2001).

[Mitchell 69] Mitchell, C.: "Social Networks in urban situations: Analyses of personal relationships in Central African towns"; University Press, Manchester (1969), p. 2.

[Moody et al. 05] Moody, J., McFarland, D., Bender-deMoll, S.: "Dynamic Network Visualization"; In: American Journal of Sociology, Vol. 110, 4, (2005), p. 1206-1241.

[Moreno 54] Moreno, J. L.: "Die Grundlagen der Soziometrie" (in German); Westdeutscher Verlag, Köln, Opladen (1954).

[Müller and Dibbern 06] Müller, C., Dibbern, P.: "Selbstorganisiertes Wissensmanagement in Unternehmen auf Basis der Wiki-Technologie – ein Anwendungsfall" (in German); HMD – Praxis der Wirtschaft, 252, (2006), p. 45-54.

[Müller and Meuthrath 07] Müller, C., Meuthrath, B.: "Analyzing Wiki-based networks to improve knowledge processes in organizations"; In: Proceedings of I-Know'07, (2007), p. 103-110.

[Müller 08] Müller, C.: Graphentheoretische Analyse der Evolution von Wiki-basierten Netzwerken für selbstorganisiertes Wissensmanagement. Phd-Thesis. University of Potsdam, Potsdam (2008).

[Newman 01] Newman, M. E. J.: "Scientific collaboration networks. I. Network construction and fundamental results"; Phys. Rev. E, American Physical Society, 64:016131 (2001).

[Newman 01a] Newman, M. E. J.: "Scientific collaboration networks. II. Shortest paths, weighted networks, and centrality" Phys. Rev. E, American Physical Society, 64:016132 (2001).

[Romhardt 02] Romhardt, K.: „Wissensgemeinschaften. Orte des lebendigen Wissensmanagements" (in German); Versus Verlag, Zürich (2002).

[Scott 00] Scott, J. “Social Network Analysis”; 2nd Edition. SAGE Publications, London (2000).

[Stockman 97] Stockman, F. N., Doreian, P.: “Evolution of Social Networks: Processes and Principles”; In: Evolution of Social Networks, Gordon & Breach, Amsterdam (1997), p. 233-250.

[Tapscott and Williams 07] Tapscott, D., Williams, A. D.: “Wikinomics“; Penguin Books Ltd., London (2007).

[Tapscott and Caston 93] Tapscott, D., Caston, A.: “Paradigm Shift: The new promise of information technology“; McGraw-Hill Companies, New York (1993).

[Tukey 77] Tukey, J. W.: “Exploratory Data Analysis”; Addison-Wesley, Massachusetts (1977).

[Wasserman and Faust 97] Wasserman, S., Faust, K.: “Social network analysis: methods and applications“; Cambridge University Press, Cambridge (1997).

[Wenger 98] Wenger, E.: “Communities of Practice: Learning, Meaning, and Identity“; Cambridge University Press, New York (1998).

[Zhu et al. 03] Zhu, H., Wang, X., Zhu, J.: “Effect of aging on network structure“; Phys. Rev. E, American Physical Society, 68:056121 (2003).