

Dynamic Network Analysis of Wikis

Ralf Klamma, Christian Haasler

(RWTH Aachen University, Information Systems
Ahornstr. 55, 52056 Aachen, Germany
{klamma|haasler}@dbis.rwth-aachen.de)

Abstract: Wikis have their seeds in the easy collaborative editing and maintenance of web pages. This was picked up by tremendously successful public projects such as the online encyclopedia *Wikipedia*. Creating, modifying and maintaining of wiki articles implies social structures and dependencies between wiki authors and wiki articles themselves. The general challenge of this work is to consider these structures as dynamic evolving networks and to point out prominent behaviors in large wiki-based networks. We present an environment capable of handling data management, measurement and visualization issues for the dynamic network analysis of publicly available wiki data.

Key Words: Dynamic network analysis, wiki, Wikipedia, social networks, visualization, measurement, data management

Category: E.1, G.2.2, H.3.1, H.3.3, H.3.7, H.4

1 Introduction

Wikis and blogs are among the most successful social software platforms of the last years [Kumar et al., 2004, Vossen and Hagemann, 2007]. A whole industry has now been established around advancing wiki software, hosting wikis and offering added-value services. A variety of wiki projects are hosted on the *MediaWiki* engine, e.g. the most famous *Wikipedia*, an online encyclopedia with millions of entries in hundreds dozens of languages edited by a countless, non-paid crowd of editors. The enormous number of public and organizational wikis has created a long tail. Besides the very successful and very visible wiki-based knowledge creation and sharing projects, there are many others with much less editors and edits. These phenomena have drawn also a lot of scientific attention, e.g. [Vega-Redondo, 2007, Adler and de Alfaro, 2007, Kittur et al., 2007, Aronsson, 2002]. A lot of studies have already been performed to analyze wikis. One of the first comprehensive researches of Wikipedia was conducted in 2005 by J. Voß[Voss, 2005]. Wikipedia was measured [Barabási et al., 1999] to find out that the distribution of links behaves with respect to *growth* and *preferential attachment*. Wikipedia also revealed a scale-free character in its link structure. D. Wilkinson and B. Huberman figured out that the quality of an article highly depends on the number of its modifications. They demonstrated that the accretion of edits to an article is described by a simple stochastic mechanism, resulting in a heavy tail of highly visible articles with a large number of edits [Wilkinson and Huberman, 2007]. [Kittur et al., 2007] examined the success of

Wikipedia. In particular, they analyzed if it is a great number of contributors where each deals with only a few articles or if it is only a small elite group of contributors that has the lion's share. In spite of Wikipedia's equal treatment of editors, some members get a leading role [Reagle, 2007]. On this qualitative view on Wikipedia the work of [Priedhorsky et al., 2007] dealt with the creation and destruction of Wikipedia articles. The researchers quantified the influence of edits and revisions in relation to the visitors. Most astonishing, vandalized articles are only read by three readers in average before being repaired by the Wikipedia community. In general, those studies can be classified in studies which make use of the publicly available wiki data (dumps) themselves and in studies making use of additional data like access log files [Priedhorsky et al., 2007]. In this paper we concentrate on the analysis of publicly available wiki dumps. In this regards, we can further classify studies concentrating on the static analysis of wiki dumps [Voss, 2005, Hu et al., 2007] and those concentrating on the dynamic aspects. In this paper we concentrate on the dynamic analysis of wikis, especially dynamic network analysis (DNA).

DNA is an emerging area of science enriching traditional social network analysis [Carley, 2003] by the idea that networks evolve over time in terms of changes of nodes in the networks and changes of links between nodes. The kind of changes depends strongly on the network under enquiry. Why is DNA of any relevance for wikis? We argue that first, revealing prominent nodes within the network, "important" and "unimportant" nodes, centrality, hidden dependencies within networks, power laws, heterogeneity, distinct hierarchical structures, small distances etc. are not possible without any network analysis. But within this static view it is not possible to analyze growth, saturation and adjustment of networks. For wiki users, wiki managers, and wiki hosting services it is extremely important to know, if wikis are still going to grow in numbers of authors, edits and wiki articles or if the wiki is going into a phase of stagnation? When a node is "important" will it stay important over the lifetime of the wiki or will its importance change over time? If a network is heterogeneous will it become homogeneous after a while or will it by that way for ever? When it comes to the dynamic network analysis of wikis we investigate two different kind of nodes, authors and articles and two different kind of relations, author–author relations and article–article relations to answer questions like those stated above.

The rest of the paper is organized as follow. In Section 2 we characterize wikis as social networks where DNA is applicable. In Section 3 we are presenting the main results of our analysis of different wikis. We conclude our paper with a discussion and an outlook on further research.

2 Wikis as Social Networks

Social network analysis is concerned with patterns of relationships between social actors [Breiger, 2004]. In this manner, mediated relations between actors as in Wikis can be regarded as social networks. Authors as well as articles can be seen as actors in a social network which helps to achieve the aim of establishing the wiki. Relations between node pairs emerge while two authors are writing an article in common. These networks of different actors and relations naturally evolves through the wiki editing process. The most significant restriction of classical social network analysis (SNA) is the lack of dynamic components such as the growth and adjustment of the social network. Dynamic network analysis (DNA) considers acting and behaviour of the network objects during the evolution process. Thus, a time component will be added to the networks. We use now classical centrality measures as well as the Small World Phenomenon and the power law distribution in social networks, for details cf. to e.g. [Brandes and Erlebach, 2005].

As aforementioned two network models are set up. Thus the (undirected) author graph is defined as $G_{author} = (V_{author}, E_{author})$. A label consisting of the author name and type (*registered/anonymous*) is allocated to every node $v \in V_{author}$. Because the node labels are unique there won't be an explicit distinction between a node and its label below. Each edge $e \in E_{author}$ represents a relation between two authors, i.e. the collaboration of them on a wiki article. So the connective edge can be understood semantically as that article on which both authors have cooperated the first time. In the (directed) article graph $G_{article} = (V_{article}, E_{article})$ each node $v \in V_{article}$ corresponds to a wiki article and its name space. E.g. the node for the Wikipedia discussion site of the article *Chemistry* is denoted by *Talk:Chemistry*. The directed edges of article graphs represent references/links from one article to another article or to external resources like websites. The references are contained in the text body of a wiki article. The evolution process of a wiki is journalized by means of revisions and special revision pages. Hence each modification of an article is reproducible via the corresponding revision. Each wiki possesses a sequence of timestamps TS . The smallest ("oldest") element of TS is that time stamp of the first article modification. It corresponds to the wiki generation. The greatest ("youngest") element corresponds to the last modification, i.e. the moment of the wiki dump creation. Author graphs as well as article graphs depend on timestamps $t \in TS$. By using timestamps it is possible to map the state of a wiki graph according to t . Consequently $G_{article}(t)$ and $G_{author}(t)$ represent the graph at time stamp t .

3 Dynamic Analysis of Wikis

Due to the space restrictions, we can only give some examples of hypotheses for the dynamic analysis of wikis we tested already with our system. If readers are interested in further analysis they can contact us or visit our system online, e.g. at <http://www.prolearn-academy.org>. We have developed a two-stage system prototype to handle all data management issues like data extraction from different wiki hosting platforms, transcription of wiki dumps into dynamic social network data sets reflecting dynamic author-author article-article dependencies, the dynamic analysis of network data, and last but not least the aesthetically appealing visualization of network data. In order to generate networks according to this goal a two-staged system was implemented. It consists of a software prototype that takes care of data extraction, transferring them into a system database, preparing them for generating and visualizing networks as well as applying measurement methods. Stage 1 is able to handle with XML dumps of arbitrary file size. Parsing is done in linear time and constant space using SAX. Stage 2 uses the advantages of graph tool kits and their network analysis methods and algorithms. The details of the design and the implementation are described in [Klamma and Haasler, 2008]. The developed system allows for analyzing the gained network data on different levels. On the database interface information can be filtered and prepared for further processing. There are two main aspects in the dynamic analysis of network data. Data can be classified in *network dimension data* and in *network structure data*. In general the dimension aspect refers to the size of the wiki networks and its changing during the evolution process. In author networks the considered characteristics may be the number of authors, the number of edits per author, and the corresponding rate of growth. In the article case network dimensions may refer to the number of links and its rate of change. The structural matter applies network measurements like centrality, shortest paths or clustering. Both the dimension and the structural aspect can be analyzed via the gained network information based on public data. Wiki log files or direct measurements at the wiki database are abandoned. In the following a couple of hypothesis concerning network characteristics are introduced. They will give at least some hints how wiki networks evolve and behave.

The rate of new authors/articles into a wiki network falls off after a period of time. The idea is to come from a “foundation fever” of a wiki. Figure 1 shows both the growth rate of the number of authors and articles. A few wikis are treated in the diagrams. In general, the assumption can not be verified. It couldn't determined a fall off in the rate of growth in both cases. The growth's characteristics may be up to semantic aspects of a wiki, e.g. up-to-date incidences that animate new users to write new articles. It has to be proved individually. In the case of *Wikia Search* it seems to be clear. In January 2008 it went online for public – observably in the sharp bend in both network types. The mea-

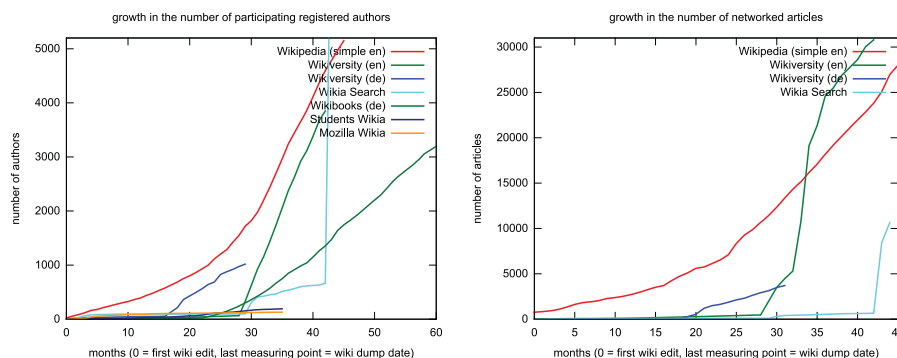


Figure 1: Rate of growth (author/article networks).

measurements of *Wikipedia (Simple English)* show a progressive growth rate, in the case of *Wikiversity* it fluctuates sometimes or it may have leaps and bounds in other wikis. Because Wikiversity’s articles are strongly categorised, further name spaces are included. There is a remarkable observation that wasn’t intended when considering both growth rates separately. Joining new authors to a wiki mostly means new articles – it does not mean working on already existing articles.

Wiki networks are heterogeneous during the whole evolution process. In homogeneous networks the number of k links per node is about the average $\langle k \rangle$ [Albert et al., 2000]. Such a uniform distribution couldn’t be verified in (social) wiki networks. Applying and measuring the degree centrality showed an imbalance between the network nodes in terms of their links. According to a lot of situations in social structures a small portion of actors have above-average links and do most of the work, i.e. editing articles and establishing new relations. This is shown in figure 2 where two author networks are given (left, center). To make contact to other users, one needs to edit a lot of articles. But, this kind of users are the minority. This distinctive heterogeneity not only occurs in author networks, but also in article networks (see figure 2, right). For article networks this is proven in figure 3 by using the degree centrality. Incoming as well as outgoing article edges and links respectively are observed over a certain time period. The measurements showed in all considered wikis a continuous strong standard deviation of edges to nodes. Depending on semantic issues there may be a very high standard deviation of outgoing links. This is given in *Aachen Wiki* which serves as a information wiki for the city of Aachen and as an index which naturally has many outgoing references.

Central nodes hold their important role during the evolution process. As described, the “importance” of a node can be determined by using the betweenness

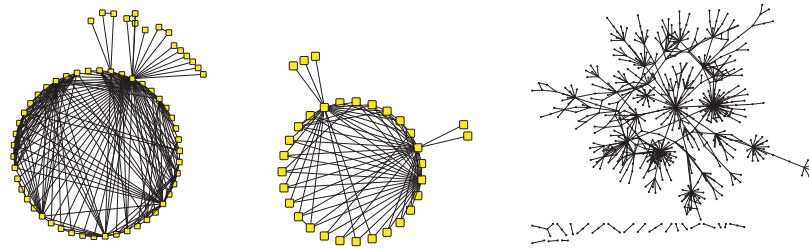


Figure 2: Heterogeneous author/article networks

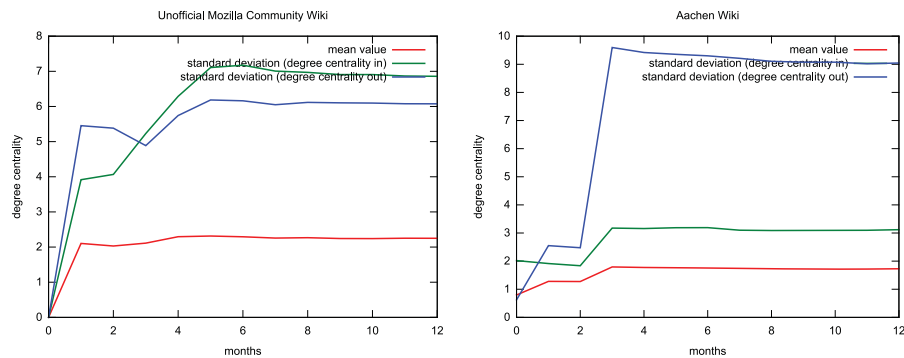


Figure 3: Article networks: degree centrality and standard deviation

centrality. This means that most shortest paths in the network go through these nodes. This measurement is done for *Wikia Search* for the time period August 2004 to August 2005. The left side of figure 4 gives for every registered author its betweenness centrality depending on time (unnormalized for a better view). Like the degree centrality there is only a small part of authors that have a high betweenness centrality. In general they hold or increase their high value during the evolution process. The survey can be found in article networks as well. The right side of figure 4 shows the betweenness centrality for *Jabber Wiki*, a wiki as the name suggests.

Important nodes can be found, too. Most central nodes keep their role during the process. The measure betweenness centrality also supports the assumption of heterogeneity of wiki networks. It shows both central authors and articles controlling the information flow in a wiki network. Author nodes with high centrality serve as intermediary of author relationships as well as they have a connection function of different author groups. Article nodes with a high index are visited often (above average) by “clicking” through the wiki.

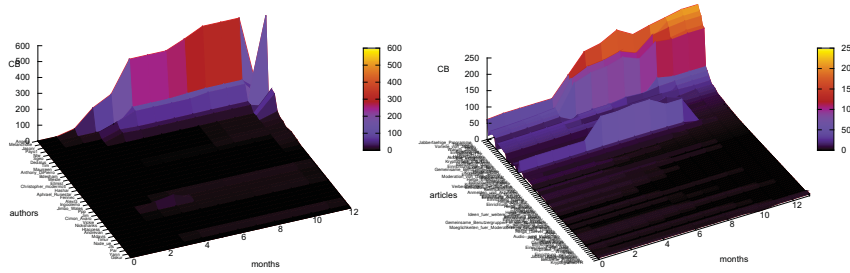


Figure 4: Betweenness centrality of author and article networks

4 Conclusions and Outlook

Primary goal of this work was to establish a new view on wikis towards social networks and furthermore as continuous mutating and growing network structures. A deeper understanding of the network evolution and its dynamics was given by considering different kinds of quantitative and qualitative characteristics. Therefore, formal network models were established, regarding both network types author–author and article–article. To come up with the steady changing of networks a time component was added. The applied DNA refers to structural as well as to dimension aspects of wiki networks. One of the prominent characteristics gives the topology of both network types. Measurements of centrality indices revealed a growing heterogeneity in both cases. Like other social networks it could be determined a strong hierarchical structure of important and unimportant nodes. Furthermore, it could built a bridge to the Small World Phenomenon [Milgram, 1967, Watts and Strogatz, 1998] that can be found in social science frequently. It was shown a continuous growth in the number of authors and articles with a remarkable correlation. But there could not made a general assertion about the kind of growth. This has to be checked in any particular case. But, it offers interesting starting points for further research in cross-medial network types like author–article that contains author as well as article nodes. Beyond that research towards more semantic matters is interesting. What effect does weighting of edges have? What is the influence of *minor edits*? In addition, semantic analyses of corresponding discussions, talk or user pages in terms of growth and changing may be interesting, too.

Acknowledgements

This work was supported by the German National Science Foundation (DFG) within the collaborative research center SFB/FK 427 “Media and Cultural Communication”, within the research cluster established under the excellence initiative of the German government “Ultra High-Speed Mobile Information and Communication (UMIC)” and within the cluster project CONTICI. We thank our colleagues for the inspiring discussions.

References

- [Kumar et al., 2004] Kumar, R., Novak, J., Raghavan, P., Tomkins, A.: Structure and Evolution of Blogspace. *Communications of the ACM* 47(12) (2004) 35–39
- [Vossen and Hagemann, 2007] Vossen, G., Hagemann, S.: *Unleashing Web 2.0. - From Concepts to Creativity*. Morgan Kaufman, Burlington, MA (2007)
- [Vega-Redondo, 2007] Vega-Redondo, F.: *Complex Social Networks*. Econometric Society Monographs. Cambridge University Press, Cambridge (2007).
- [Adler and de Alfaro, 2007] Adler, B.T., de Alfaro, L.: A content-driven reputation system for the Wikipedias. In: *WWW (2007)* 261–270
- [Kittur et al., 2007] Kittur, A., Chi, E.H., Pendleton, B.A., Suh, B., Mytkowicz, T.: Power of the few vs. wisdom of the crowd: Wikipedia and the rise of the bourgeoisie. In: *25th Annual ACM Conference on Human Factors in Computing Systems (CHI 2007)*; 2007 April 28 - May 3; San Jose, CA. (2007)
- [Aronsson, 2002] Aronsson, L.: Operation of a large scale, general purpose wiki website: Experience from susning.nu's first nine months in service. In: *Proceedings of the 6th International ICC/IFIP Conference on Electronic Publishing*, Karlovy Vary, Czech Republic, (2002), 27–37
- [Priedhorsky et al., 2007] Priedhorsky, R., Chen, J., Lam, S.T.K., Panciera, K., Terveen, L., Riedl, J.: Creating, destroying, and restoring value in wikipedia. In: *GROUP '07: Proceedings of the 2007 international ACM conference on Supporting group work*, New York, NY, USA, ACM (2007) 259–268
- [Voss, 2005] Voss, J.: Measuring wikipedia. In: *Proceedings of the 10th International Conference of the International Society for Scientometrics and Informetrics*. (2005)
- [Hu et al., 2007] Hu, M., Lim, E.P., Sun, A., Lauw, H.W., Vuong, B.Q.: Measuring article quality in wikipedia: models and evaluation. In: *CIKM '07: Proceedings of the sixteenth ACM conference on Conference on Information and Knowledge Management*, New York, NY, USA, ACM, 243-252
- [Carley, 2003] Carley, K.M.: Dynamic network analysis. In Breiger, R., Carley, K.M., eds.: *Summary of the NRC workshop on Social Network Modeling and Analysis*, National Research Council (2003)
- [Barabási et al., 1999] Barabási, A.L., Albert, R., Jeong, H.: Mean-field theory for scale-free random networks (1999)
- [Wilkinson and Huberman, 2007] Wilkinson, D.M., Huberman, B.A.: Assessing the value of cooperation in wikipedia (Feb 2007)
- [Reagle, 2007] Joseph M. Reagle, J.: Do as i do:: authorial leadership in wikipedia. In: *WikiSym '07: Proceedings of the 2007 international symposium on Wikis*, New York, NY, USA, ACM (2007) 143–156
- [Breiger, 2004] Breiger, R.L.: The analysis of social networks. In Hardy, M., Bryman, A., eds.: *Handbook of Data Analysis*. London, SAGE Publications (2004) 505–526
- [Brandes and Erlebach, 2005] Brandes, U., Erlebach, T.: Fundamentals. In Brandes, U., Erlebach, T., eds.: *Network Analysis: Methodological Foundations*. Springer (2005)
- [Klamma and Haasler, 2008] Klamma, R., Haasler, C.: Wikis as social networks: Evolution and dynamics. In: *Proceedings of 2nd ACM SNA-KDD Workshop at KDD 2008*, August 24, 2008, Las Vegas, NV. (2008)
- [Albert et al., 2000] Albert, R., Jeong, H., Barabási, A.L.: Error and attack tolerance of complex networks. *Nature* 406 (2000) 378–382
- [Milgram, 1967] Milgram, S.: The small-world problem. *Psychology Today* 1(1) (1967) 60–67
- [Watts and Strogatz, 1998] Watts, D.J., Strogatz, S.H.: Collective dynamics of 'small-world' networks. *Nature* 393 (1998) 440–442