Wikipedia and its Network of Authors from a Social Network Perspective

Matthias Hirth, Frank Lehrieder, Stephan Oberste-Vorth, Tobias Hoßfeld, Phuoc Tran-Gia
University of Würzburg, Institute of Computer Science, 97074 Würzburg, Germany
Email: matthias.hirth@informatik.uni-wuerzburg.de

***Abstract*—Online social networks (OSNs) become more and more important in today's social and business life. Therefore, considerable effort is put in research to gain a deeper knowledge of the development of these networks and their dynamics. However, most of the existing literature is based on very limited subsets of the network data, which is often filtered by the OSN operator providing the data or biased by the crawling mechanisms used to obtain the data. This makes it difficult to analyze the temporal evolution of OSNs based on complete data. To overcome this issue, we investigate the dynamics of the publicly available collaboration network of the Wikipedia authors as an example for an OSN-like network. In particular, we study the temporal evolution of this network since its beginning and demonstrate that it exhibits prominent similarities to well known social networks such as the small-world phenomenon. This indicates that the insights gained from the analysis of Wikipedia's collaboration network might be transferable to social networks in general.**

I. Introduction

At least with the rise of Facebook, online social networks (OSNs) have become one of the most important developments in the recent years and subject to many research efforts. Especially, a detailed analysis of the temporal development of these networks is of great interest because it gives insights about the evolution of mass movements or the success of viral marketing campaigns.

However, most of the current research results are based on a crawled subset of the network data, which do not allow an unbiased view of the networks. Crawled data suffers often from biases introduced by the crawling algorithms, like the underestimation of less connected nodes when using breath-first or depth-first crawling algorithms. Furthermore, automated crawling of social network data is usually against the terms of service of the OSN provider and thus only small data sets are available if at all. The analysis of the temporal changes within these networks is even harder since the temporal information in the available data sets is limited. A solution to this problem is the use of publicly available data from social network like systems for the analysis of the temporal changes. Results gained from this analysis can afterwards be adapted to the closed OSNs using the available snapshot data.

In this study we pursue this approach and analyze the collaboration network of Wikipedia authors as an example of a social network like structure. Wikipedia has developed to one of the most important sources of information nowadays. This success is rooted in the contribution of thousands of volunteer authors contributing to the Wikipedia articles. These authors interact in various ways with each other during the edition of the articles. Consequently, we argue that the collaborations can be seen as an example of a social interaction and define the collaboration network of Wikipedia authors in the following way. We consider all registered Wikipedia authors as vertices and an edge between two authors exists if there is a Wikipedia article that both authors have edited. Therefore, this definition of collaboration replaces the friendship relation in other OSNs. Furthermore – unlike to other OSNs – these interactions are publicly available.

The contribution of our work is twofold. First, we provide an analysis of the evolution of the collaboration network of the English Wikipedia from 2001 to 2011. Our analyzed collaboration network is based on public meta-information about the edits of the Wikipedia articles, which are provided by the Wikipedia Foundation. Second, we show that there are structural analogies between the collaboration network and social networks, like the presence of the small world phenomenon and the power-law distribution of the node degree.

This paper is structured as follows. After reviewing related work in Section II we describe the generation of the collaboration network graph in Section III. The evaluation of this graph is presented in Section IV. In Section V, a conclusion of our major results is drawn.

II. Related Work

Before analyzing the collaboration network of Wikipedia and its temporal changes, we briefly review related work. The Wikipedia and its authors have been subject to various studies before. However, we focus in the following on related work dealing with networks generated from Wikipedia content and the Wikipedia authors.

An extensive analysis of the network structure of the articles of the 30 largest Wikipedias (in different languages) was performed by Zlatić et al. in [1]. The authors compared several well-known metrics like degree distributions, growth, topology, reciprocity, clustering, assortativity and path lengths of the resulting networks and showed that many network characteristics are common among all studied Wikipedias. Bellomi et al. [2] used a snapshot of the English Wikipedia in 2005 to generate a network of linked articles. Using different ranking algorithms they were able to retrieve information about social biases in the Wikipedia. An approach to visualize

the relationships between articles was presented by Biuk-Aghai [3]. The relationship of the articles were determined using the link structure of the articles, as well as information about the co-authorship of articles' editors.While Zlatić et al., Bellomi et al., and Biuk-Aghai focused on the network structure of the Wikipedia articles, we investigate the collaboration network of the Wikipedia authors in this study.

Massa et al. [4] focus on generating a social network of authors. However in contrast to our collaboration network based on the article edits, they used the information from the discussion pages of Wikipedia. To this end, they developed two different algorithms to automatically extract the social network from discussion pages and compare them to a manually extracted social network from the Venetian Wikipedia discussion pages. Laniadro et al. [5] also used networks generated from the discussion pages of the English Wikipedia to analyze patters of interactions of the authors and found structural differences among the discussions about articles from different semantic areas.

Similar to our approach, Brandes et al. [6] analyze the collaboration of the Wikipedia authors using networks based on the edits of the articles. Brandes et al. focus on the editing networks of single articles and identify different roles of the authors, by tracking their different editing activities like adding, deleting or revising parts of single articles. Furthermore, they present visualization techniques for these local editing networks to gain a quick overview of the different roles of the authors and visualize the collaboration structure of different articles. In contrast to Brandes et al., our work deals with the global collaboration graph of the Wikipedia and its temporal changes.

III. Collaboration Graph and Data Basis

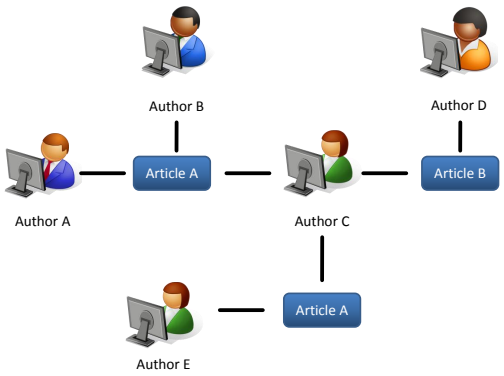
In the following we describe the definition of the collaboration graph and the Wikipedia data used to extract the author collaboration network.

A. Author Collaboration Graph

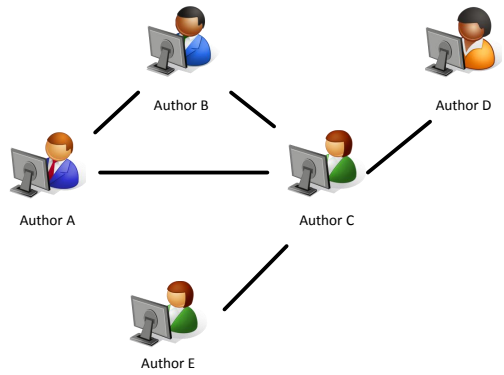
Unlike other OSNs, there is no explicit social relationship between Wikipedia authors, like *fiends* in Facebook or *followers* and *friend* in Twitter. Thus, we define a *collaboration* relation between authors in order to connect authors. According to our definition, two authors collaborate if there is at least one article that has been edited by both authors. This is a very broad view of collaboration since we do not distinguish whether the authors subsequently add new content to an article or if they change others' contributions.

Using this definition of collaboration, we can represent the collaboration network of the authors as an undirected, loop-free graph $G(V, E)$. In this graph, the nodes V correspond to the authors, the bidirectional edges E to the collaboration relations between the authors. An edge E_{ab} exists, if the two authors a and b edited at least one article in common.

Figure 1 depicts an example of the generation of a collaboration network. In this example, five different authors contribute to three articles as shown in Figure 1a. For sake of simplicity



(a) Articles and editing authors



(b) Collaboration network

Fig. 1: Generation of the author collaboration network

multiple article editions of the same author are represented by a single edge since we focus on the fact whether collaboration exists and not on the intensity of the collaboration. Applying our definition of collaboration to the given example results in the collaboration network shown in Figure 1b. All authors working on *Article A* are connected with each other, *Author D* and *E* are only connected with *Author C* who worked on all articles.

B. Wikipedia Data

In the following, we describe the Wikipedia data that we used to create the collaboration network. Wikipedia offers various database snapshots [7], which contain different subsets of Wikipedia content. The most comprehensive snapshots include all articles and all their revisions; others comprise only the current version of the articles or only the abstracts of the articles. The size of these snapshots varies from a few gigabytes to more than 5 terrabytes the largest snapshots.

For our analysis of the author network, we require information about the editions of the individual authors. Thus, we use the *stub-meta-history* snapshots. These XML-files include all meta information of every revision of any Wikipedia page but not the page content itself. The meta information contain among other, the page name, the time of the revision and details about the contributor of the revision. If the contributor is a registered uses, the meta information

contains his unique user name and the user's id. If the contribution was submitted by an anonymous user, it contains the IP address of the editing device.

Wikipedia pages are grouped in namespaces which reflect the main purpose of the page. Pages containing content for the encyclopedia belong to the *main namespace*. In addition, namespaces exist also for discussions or home pages of the users. In this work we only consider pages in the main namespace, which we denote as *articles* in the following, even if the page contains a redirection, stub or disambiguation.

We also limit our analysis to registered authors only since it is not possible to use a pure IP-based identification of the anonymous authors. On the one hand, the same author might edit articles from different devices and thus use various IPs. On the other hand, one device can be used by multiple authors.

The following results are based on the stub-meta-history files of the English Wikipedia from May 26th, 2011, which includes every revision of every page of the English Wikipedia from its start in 2001 until the creation date of the stub-meta-history file. Applying our limitation on the articles and the authors, this data set contains 3.6 million authors, who contributed to 8.5 million articles and are connected by 2.7 billion edges. In order to limit the computational efforts for the temporal analysis, we create snapshots of the collaboration network in intervals of six month. These snapshots are also based on the stub-meta-file from May 26th, 2011, but all revisions after the time of the snapshot are neglected.

The calculations and analysis were performed on a desktop PC with a quad-core 3.4 GHz CPU, 16 GB RAM and a 4 TB hard disc. The stub-meta-file was preprocessed with self-developed Java software and the generated collaboration networks were stored using the graph database Neo4j [8]. Depending on the size of the collaboration network snapshot and the analyzed graph metric, the calculations took from several minutes up to several days.

IV. Temporal Evolution of Wikipedia and its Author Network

In the following we present the results of our analysis of the collaboration network of Wikipedia. In order to get a better understanding of the collaboration network, we first study basic statistics like the development of the number of articles and authors. Afterwards, we study the collaboration of the authors and whether the author network is split in several unconnected components. Finally, we show that the collaboration network is a typical small-world network like other OSNs.

A. General Wikipedia Statistics

First, we study the growth of Wikipedia since its start in 2001. To this end, we consider the evolution of the number of registered authors and articles, and of the cumulative number of editions shown in Figure 2. Note that the y-axis is in logarithmic scale. We observe that the number of authors grows rapidly during 2001. Afterwards, there is still an exponential growth of the number of authors but at a lower rate until 2006. After 2006 the registration rate of the authors decreases further until the end of the measurement. The number of articles shows similar trends as the number of authors, beginning with a rapid growth until 2003 and followed by a slightly lower growth rate until approximately 2006. The growth rate decreases even more after 2006. The same applies also for the number of editions in the graph.

The growth of the number of authors directly affects the size of the graph because each author is represented by a node in the collaboration graph. As a result, the growth of the number of nodes in the graph is identical to the growth of the number of authors. The number of articles and editions affect the number of edges in the graph; however we cannot derive the number of edges directly from these two values. If an article is edited by authors that have already cooperated before, the structure of the graph is not changed since no new edges are generated. On the opposite side, an article or an edition changes the structure if the contributing authors have not interacted before. Thus, the number of editions and articles might be used to estimate the number of connections within the collaboration graph, but it does suffice to determine it exactly.

In the next step, we compare the number of editions in the whole Wikipedia to the ones in the main namespace and to the ones which we consider for generation of the collaboration graph, i.e., editions in the main namespace done by registered authors. This gives an estimate of which fraction of all editions are represented in the collaboration graph. Figure 3 shows the number of edits per half-year performed in the entire Wikipedia, the main namespace and our dataset. Again, the y-axis is in logarithmic scale.

The number of editions in the main namespace is always smaller than the number of editions in the whole Wikipedia since the main namespace is just a subset of all Wikipedia pages. For our analysis we use all pages of the main namespace, however, we only take edits from registered authors into account. Thus, the number of editions in our studied data is smaller than the number of edition in the whole main namespace. Nevertheless, our analyzed subset covers almost ⅔ of the edits in the main namespace and ¼ of all Wikipedia edits. Since we focus on the collaboration network retrieved from the Wikipedia articles, ⅔ of all main namespace editions can be assumed to be a representative subset of the overall data.

B. Collaboration of Authors

After the study of the growth process of the Wikipedia, we now focus on the structure of the collaboration network of the authors. First, we investigate the node degree distribution. The node degree is defined as the number of edges that the node is connected to. In our case, the node degree represents the number of collaborations of the author. The degree distribution of the collaboration network for three different snapshots at mid 2002, 2005, and 2011 is shown in Figure 4.

In all three snapshots, the graph shows a lot of nodes with a small node degree. The number of authors with high degree drops rapidly with an increasing number of collaborations. However, there are a few nodes with a very large number

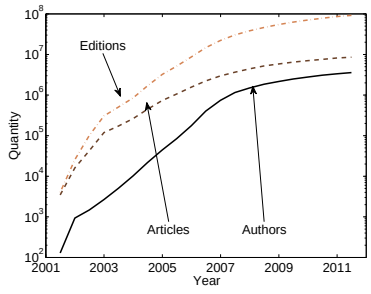


Fig. 2: Number of authors, articles, and editions

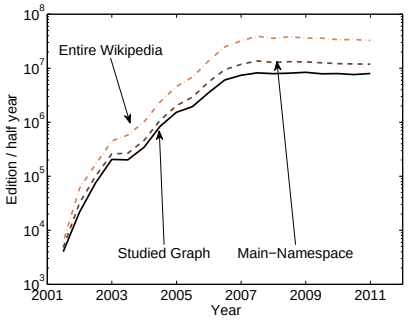


Fig. 3: Editions per half-year

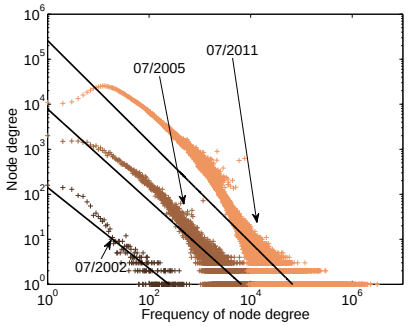


Fig. 4: Node degree of the authors

of collaborations, at most over 10^6 in 2011. These highly connected authors are mainly automated scripts, which perform tasks like spell correction on a huge number of articles, or semi-automatized accounts of power-users who contribute content and run self-designed automated scripts. In our further analyses we also treat these accounts as regular authors.

A further analysis of the node degrees shows that it follows roughly a power-law distribution, i.e., the probability P of a node having k connections can by approximated by $P(k) \propto ck^{-\gamma}$, with a constant factor c and the power-law exponent γ. The fitting of a power-law to the measured node degrees is shown with the continuous lines in Figure 4.

The accuracy of the power-law fitting decreases for nodes with lower node degrees in 2005 and 2011. This is typical for the node degree of social networks where the node degree also follows a power-law distribution. However, the reasons for that phenomenon are different in our case. In OSN analysis the data is usually collected using breath-first or depth-first crawling algorithms. It is well known that low degree nodes are underrepresented in these crawled samples, as the probability to reach a node decreases with its number of edges. The underrepresentation is hence caused by incomplete data in OSN analysis. In the case of the collaboration network of the authors, however, we use a complete snapshot of the collaboration network and crawling biases are consequently not present. The underrepresentation of the low degree nodes (compared to the power-law fitting) results here from the fact that there are more authors working on popular articles and authors working on specialized articles are rare. If an author edits only specialized articles with only a few other contributors, his node degree is smaller than the degree of an author who edits a popular article with hundreds of other authors.

Even if the power-law fitting is not perfect, it approximates the node degree reasonably well. Thus, we proceed with an investigation of the temporal change of the power law exponent γ. Figure 5 shows the variation of γ from 2001 to 2012. In the first phase until 2003 we observe a decrease of the power-law exponent, which means the slope of the node degree distribution, flattens. This can be explained with the results of the statistical analysis of the Wikipedia shown in Figure 2. From 2001 to 2003 the number of editions grows faster than the number of authors. As a result, the number of edges in the graph increases faster than the number of nodes, which leads to an increase of higher connected nodes. Between 2003 and 2007 we see an increase of the power-law exponent. During this phase, the number of newly joining authors grows faster than the number of editions. Consequently, we observe the opposite development of that until 2003 and the fraction of highly connected nodes decreases. Since 2007 the power-law exponent γ does not show any significant changes.

In order to achieve a better understanding of the connections among the authors, we have a look at the density of the collaboration network. The density d of a network is defined as the ratio of present edges to the maximum number of possible edges and can be calculated by $d = \frac{|E|}{V \cdot (V-1)/2}$ [9]. The density of the collaboration graph from 2001 to 2012 is shown in Figure 6. The first snapshot in 2001 exhibits a very high network density in comparison to all other snapshots. After the density drops considerably in the second half of 2001, it increases again until 2003. From 2003 to 2007 we see a constant decrease and the density remains constant at a very low level from 2007 on to the end of the studied data. This means that of the large number of possible collaborations comparably few of them are present in the actual graph.

The very high network density in the first snapshots might root from the intensive interactions of the early adopters of the Wikipedia idea. In this snapshot there are only very few authors who contributed to a relative small number of articles. This results in a highly connected network with a high density.

Except the first snapshot, the development of the network density corresponds to the development of the power-law exponent. From the second half of 2001 until 2003, the number of edges increases faster than the number of nodes, and thus the network density increases and with it the γ decreases. Afterwards, the density of the network is decreasing and the relative number of nodes with a small degree increases and with this the exponent of the power-law fitting. Since 2007 the density of the collaboration network remains constant and thus also the power-law exponent γ.

C. Author Groups in the Collaboration Graph

In the next step we analyze the connectivity within the collaboration network to determine whether the majority of the

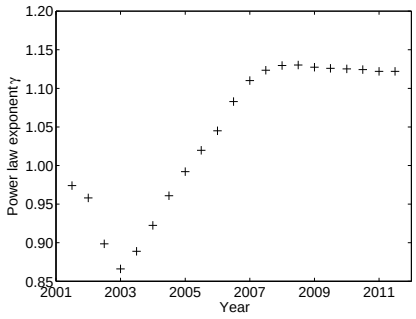


Fig. 5: Power-law exponent of the author node degree

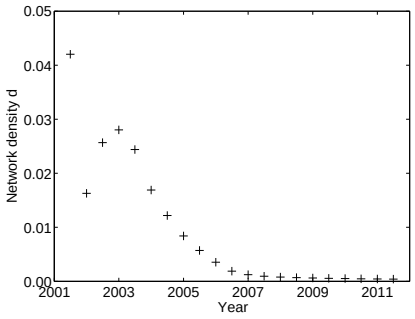


Fig. 6: Network density

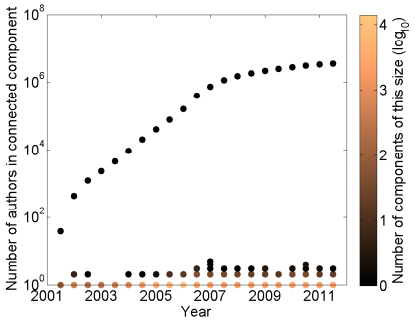


Fig. 7: Size and number of connection components

authors is connected or if the graph decomposes in numerous components. For that purpose, we calculate the number and the size of the connected components in the collaboration network.

Figure 7 depicts the number and size of the connected components of the author network from 2001 until 2012. The size of each connected component is shown on the logarithmic y-axis; the number of connected components with the size given on the y-axis is encoded by the color of the markers. During the whole period the majority of the authors are included in one large connected component which constantly grows. Besides this large component there exist several small components, which include up to 4 workers. At most there are 52 small components besides the largest one in 2006 if we exclude all connected components that contain only a single author, i.e., of size 1. These are registered authors who only edited articles on their own without any collaboration. As the color in Figure 7 indicates, their number ranges in the order of 10^2 to 10^4.

The analysis shows that most of the authors are connected with each other and there are only very few authors in isolated groups. In particular, the analysis reveals that there are no groups of authors of considerable size that have no interactions with other authors at all. The small isolated groups result from authors who only collaborated on specialized topics or on single isolated pages like re-directions that were not edited after their creation.

D. The Collaboration Graph as a Small-World Network

In our last step we analyze if the collaboration networks exhibits the small-world property according to the definition by Watts and Strogatz [10]. The small-world phenomenon describes the fact that even in a network with a large number of nodes and comparably few edges the average distance between two nodes is small. This is a typical property of social networks and we hence investigate it in the following for the author network of Wikipedia. According to the aforementioned definition, small-world networks are characterized by a very short characteristic path length and a high clustering coefficient. The characteristic path length is the minimum distance of two nodes in the network averaged over all pairs of nodes. The clustering coefficient measures the cliquishness of the networks. We start with the investigation of the shortest paths and consider the clustering coefficient afterwards.

Figure 8 shows the distribution of the shortest paths between two authors in the largest connected component of the collaboration network from 2001 to 2005. For snapshots of the author networks later than 2005, the calculation of the shortest path distribution was not possible due to computational limitations. During the whole period, the shortest path is always below 6 hops. Except for the first two snapshots, most of the authors are connected via a 2-hop path. Furthermore, from 2002 to 2005 the probability for a 2-hop path decreases while the one for a 3-hop path increases.

In order to investigate the influence of the growing network, we show the shortest path length in dependency of the number of authors in the network in Figure 9. We see that the maximum path length is 6 in 2002 and 5 or less in all the other snapshots. Since 2/2002 the average path length constantly increases slightly from 2.05 to 2.25. The analysis shows that even if the networks grows from a few hundred to over 3 million nodes, the characteristic path length remains surprisingly short, similar to random networks [10].

Another important measure to identify small-world networks according to Watts and Strogatz is the clustering coefficient, which is a measure for the cliquishness of the networks. In small-world networks, the clustering coefficient is significantly higher than in random networks. The local clustering coefficient c_i of a node i with n neighbors in an undirected network is defined as the ratio of present edges e between its neighbors and the maximum possible number of edges $\frac{n\cdot(n-1)}{2}$ between its neighbors, which leads to $c_i = \frac{2e}{n\cdot(n-1)}$. It is not unambiguously defined for nodes with only one or without a neighbor, thus we do not consider these nodes in our analysis. Using the previous definition of the local cluster coefficient, the clustering coefficient of a graph is calculated as the mean clustering coefficient of all its nodes.

The mean clustering coefficient of the collaboration network from 2001 to 2005 is shown in Figure 10 by the cross markers. For larger snapshots an analysis was not possible due to the computational limitations. During the analyzed phase, the clustering coefficient increases from 0.6 to over 0.9 indicating an increasing cliquishness of the collaboration network.

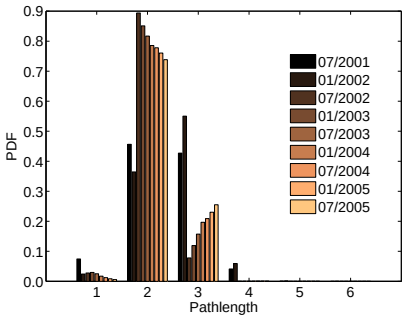


Fig. 8: Shortest path length

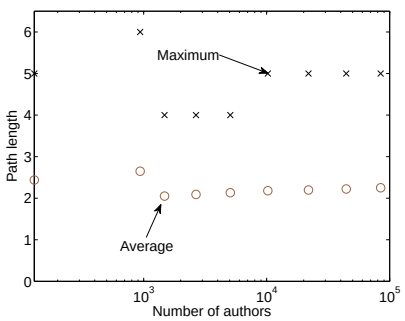


Fig. 9: Maximum and mean path length

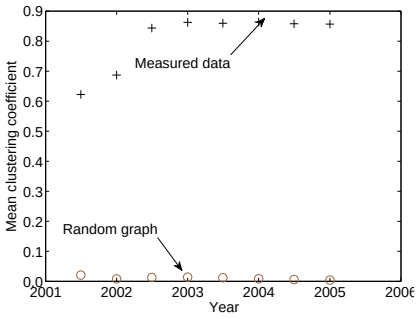


Fig. 10: Mean clustering coefficient

In order to compare the measured clustering coefficient to the clustering coefficient of a random network, we used the formula given by Dorogovtsev [11] for the clustering coefficient of an uncorrelated network

$$C = \frac{(\overline{k^2} - \overline{k})^2}{N\overline{k}^3},$$

with the average node degree $\overline{k}$ and N nodes. According to [12], the probability of a node having degree k in a random network is given by a Binomial Distribution $P(k) = \binom{N}{k} p^k (1-p)^{N-k}$. The probability p can be calculated with the average node degree $\overline{k}$ and the number of nodes N in the graph: $p = \overline{k}/(N-1)$.

Using the same number of nodes and the same average node degree as observed in our snapshots, we can calculate the clustering coefficient of a random network with the same properties. The clustering coefficients of the random networks are shown in Figure 10 by the circled markers. Comparing these values to the measured clustering coefficient, we see that the clustering coefficient of the collaboration network is significantly higher than in a random network.

In conjunction with the short characteristic path length, the collaboration network is a small world network according to the definition of Watts and Strogatz and thus can be used to analyze small world phenomena in OSNs.

V. Conclusion

In this paper we analyzed the temporal evolution of the network of Wikipedia authors. To this end, we defined a graph of all registered authors and connected two authors in the graph if they collaborated, i.e., edited the same article. Furthermore, we showed that this collaboration network exhibits prominent similarities to other social networks. Hence, it can serve as an example network where all information is publicly available, in contrast to most other social networks.

Our analysis has shown that at the launch of Wikipedia and shortly afterwards, the early adopters formed a highly connected and dense network due to their small number and high activity on the articles. With the growth of Wikipedia and the increasing number of authors, the density of the network decreases and the difference of the degree of highly connected and low connected nodes increases. Since about 2007 the network seems to have reached a steady state, where the power-law exponent of the degree distribution and density of the network remains constant.

Our results further indicate that even if the articles of the Wikipedia cover such a huge range of topics, the major part of the authors is part of a single big connected component. This might either be caused by overlapping interests of individual authors, or by authors, respectively automated scripts, which perform routine tasks, like spell correction on a large number of articles. Analyzing this major connected part of the collaboration network, we found out that it shows small-world properties like social networks. Even if there are several thousands of authors within this connected component, the average path length between those authors is rather small. However, the network has still a very high clustering coefficient, which indicates that the authors work together in groups. These findings are in line with results for other social networks and underpin their similarity to the Wikipedia network, what motivated our study.

References

[1] V. Zlatić, M. Božičević, H. Štefančić, and M. Domazet, "Wikipedias: Collaborative web-based encyclopedias as complex networks," *Physical Review E*, vol. 74, no. 1, 2006.

[2] F. Bellomi and R. Bonato, "Network analysis for wikipedia," in *Proceedings of the Wikimania*, 2005.

[3] R. Biuk-Aghai, "Visualizing co-authorship networks in online wikipedia," in *Proceedings of the Symposium on Communications and Information Technologies*, 2006.

[4] P. Massa, "Social networks of wikipedia," in *Proceedings of the Conference on Hypertext and Hypermedia*, 2011.

[5] D. Laniado, R. Tasso, Y. Volkovich, and A. Kaltenbrunner, "When the wikipedians talk: Network and tree structure of wikipedia discussion pages," in *Proceedings of the International Conference on Weblogs and Social Media*, 2011.

[6] U. Brandes, P. Kenis, J. Lerner, and D. Van Raaij, "Network analysis of collaboration structure in wikipedia," in *Proceedings of the International Conference on World Wide Web*, 2009.

[7] Wikipedia Snapshots, "http://en.wikipedia.org/wiki/Wikipedia:Database_download."

[8] Neo4j, "http://neo4j.org."

[9] S. Wasserman and K. Faust, *Social Network Analysis: Methods and Applications*. Cambridge University Press, 1994.

[10] D. Watts and S. Strogatz, "Collective dynamics of small-world networks," *Nature*, vol. 393, no. 6684, 1998.

[11] S. Dorogovtsev, "Clustering of correlated networks," *Arxiv preprint cond-mat/0308444*, 2003.

[12] M. Newman, S. Strogatz, and D. Watts, "Random graphs with arbitrary degree distributions and their applications," *Physical Review E*, vol. 64, no. 2, 2001.