

University of Würzburg  
Institute of Computer Science  
Research Report Series

**Wikipedia and its Network of Human  
and Automated Authors from a Social  
Network Perspective**

Matthias Hirth, Frank Lehrieder, Stephan  
Oberste-Vorth, Tobias Hoßfeld, Phuoc Tran-Gia

Report No. 494

April 2015

University of Wuerzburg, Germany  
Institute of Computer Science  
Chair of Communication Networks  
Am Hubland, D-97074 Würzburg, Germany  
phone: (+49) 931-3186954, fax: (+49) 931-3186632  
matthias.hirth@informatik.uni-wuerzburg.de



# Wikipedia and its Network of Human and Automated Authors from a Social Network Perspective

**Matthias Hirth, Frank Lehrieder, Stephan  
Oberste-Vorth, Tobias Hoßfeld, Phuoc  
Tran-Gia**

University of Wuerzburg, Germany  
Institute of Computer Science  
Chair of Communication Networks  
Am Hubland, D-97074 Würzburg, Germany  
phone: (+49) 931-3186954, fax: (+49) 931-3186632  
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. For studying temporal evolution and time dynamics we investigate the dynamics of the publicly available collaboration network of the Wikipedia authors as an example for an OSN-like network. The authors are distinguished between human authors contributing to the articles and automated scripts, so called bots, which complete specialized tasks like spell-checking. 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.<sup>1</sup>

**Keywords:** Wikipedia, Network Structure, Social Network, Social Network Analysis, Network Dynamics

## 1 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

---

<sup>1</sup>This work is an extension of the results published in 'Wikipedia and its Network of Authors from a Social Network Perspective' Hirth et al. (2012).

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. Unlike to other OSNs these interactions are publicly available.

However, besides human authors, automated scripts also contribute to Wikipedia articles by accomplishing standardized tasks likes, e.g., spell checking. In the following we refer to these automated scripts as *bots*. If not explicitly stated otherwise, they are also included in the collaboration network. In our analysis we focus on the collaboration network of the English Wikipedia and its evolution from 2001 to 2011, as well as the structural analogies to well known social networks, like the presence of the small world phenomenon and the power-law distribution of the node degree. The results presented here are an extend version of the results previously published in Hirth et al. (2012). In addition to the previous work, we analyze the impact of the bot editions on the collaboration network Wikipedia authors.

This paper is structured as follows. After reviewing related work in Section 2, we describe the generation of the collaboration network graph in Section 3. The evaluation of this graph is presented in Section 4. In Section 5, a conclusion of our major results is drawn.

## 2 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. (2006). 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 and Bonato (2005) 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 (2006). 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 (2011) 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. Laniado et al. (2011) also used networks generated from the discussion pages of the English Wikipedia to analyze patterns 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. (2009) 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.

### 3 Collaboration Graph and Data Basis

In the following we describe the definition of the collaboration graph as well as the Wikipedia data used to extract the author collaboration network and the information about the bots.

#### 3.1 Author Collaboration Graph

Unlike other OSNs, there is no explicit social relationship between Wikipedia authors, like *friends* in Facebook or *followers* and *friend* in Twitter. Thus, we define a *collaboration* relation between 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. If not explicitly states, *authors* subsumes both both bot and human authors.

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

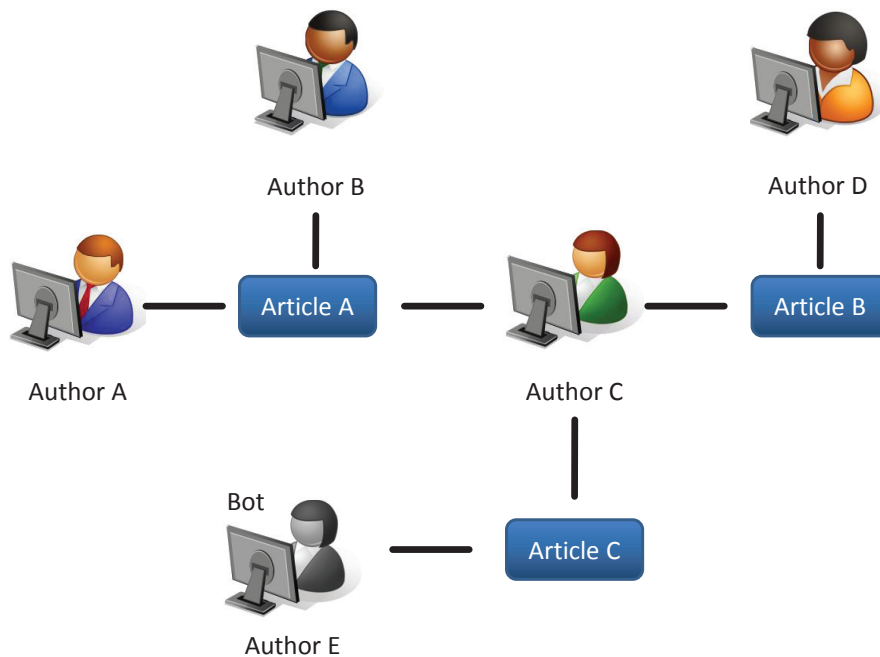


Figure 1: Articles and editing authors

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. Bots are treated alike human authors but the nodes are explicitly flagged to identify them in the later analysis.

Figure 1 and Figure 2 illustrate an example of the generation of a collaboration network. In this example, four different human authors and one bot author contribute to three articles as shown in Figure 1. For the sake of simplicity 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 2. 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.

### 3.2 Wikipedia Data

In the following, we describe the Wikipedia data that we used to create the collaboration network. Wikipedia offers various database snapshots<sup>2</sup>, 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 terabytes the largest snapshots.

<sup>2</sup>[http://en.wikipedia.org/wiki/Wikipedia:Database\\_download](http://en.wikipedia.org/wiki/Wikipedia:Database_download)

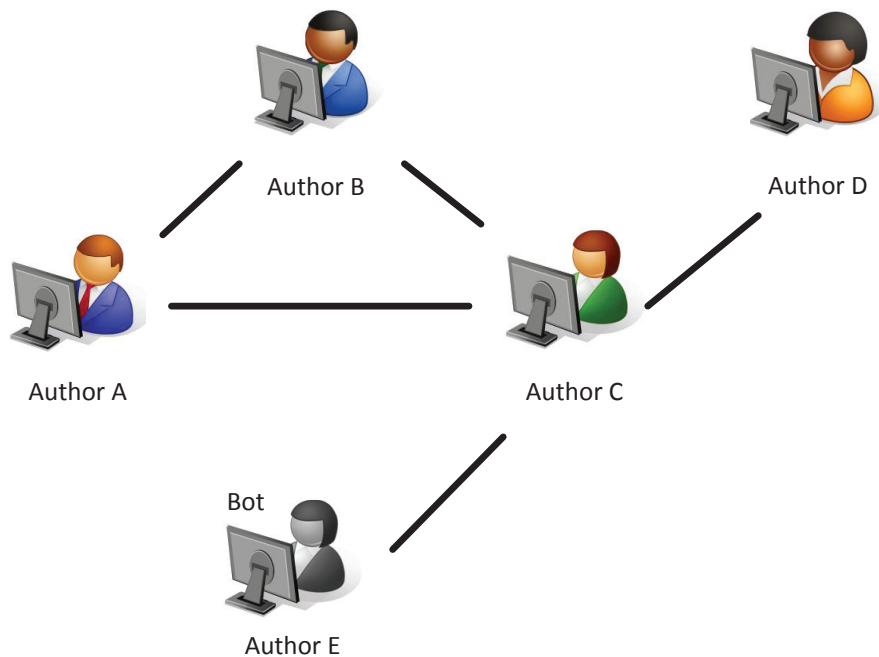


Figure 2: Collaboration network

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 contains among other, the page name, the time of the revision and details about the contributor of the revision. If the contributor is a registered user, 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. Due to this structure, not all registered users of the Wikipedia are included in the stub-meta-history files, but only those who contributed at least one edition.

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.

Bot accounts are not explicitly marked as such and cannot be distinguished from regular user accounts with the information in the stub-meta-history files. Thus, we use the list of the most active bots in terms of editions provided by Wikipedia<sup>3</sup> as of October 31st, 2012. This list contains the names of the 863 most active bots and the corresponding number of editions. 845 bots listed have at least 1 edition, the remaining 18 bots have zero editions. In our analysis, we only consider the bots named on Wikipedia bot list<sup>3</sup>, as all other bots do not contribute to the collaboration network at all.

The calculations and analysis were performed on a desktop PC with a quad-core 3.4 GHz CPU, 16 GB RAM, a 4 TB hard disc and a 254 GB SSD drive. The stub-meta-file was preprocessed with self-developed Java software and the generated collaboration networks were stored using the graph database Neo4j<sup>4</sup>. Depending on the size of the collaboration network snapshot and the analyzed graph metric, the calculations took from several minutes up to several days.

## 4 Temporal Evolution of Wikipedia and its Author Network

In the following we present the results of our analysis of the collaboration network of Wikipedia and the influence of the bots. 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 have a closer look at 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.

### 4.1 Growth over Time and Bot Contribution

First, we study the growth of Wikipedia since its start in 2001. To this end, we consider in Figure 3 the evolution of the cumulative number of editions, the number of articles and the number of active accounts, which accumulate the human authors and the bots who submitted at least one edition.

We observe that the number of accounts grows rapidly during 2001. Afterwards, there is still an exponential growth of the number of accounts but at a lower rate until 2006. After 2006 this rate decreases further until the end of the measurement. The number of articles shows similar trends as the number of accounts, 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 growing number of accounts directly affects the size of the graph, because each bot and 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 accounts. The number of articles and editions affect the number of edges in the graph; however we cannot

---

<sup>3</sup>[http://en.wikipedia.org/wiki/Wikipedia:list\\_of\\_bots\\_by\\_number\\_of\\_edits](http://en.wikipedia.org/wiki/Wikipedia:list_of_bots_by_number_of_edits)

<sup>4</sup><http://neo4j.org>



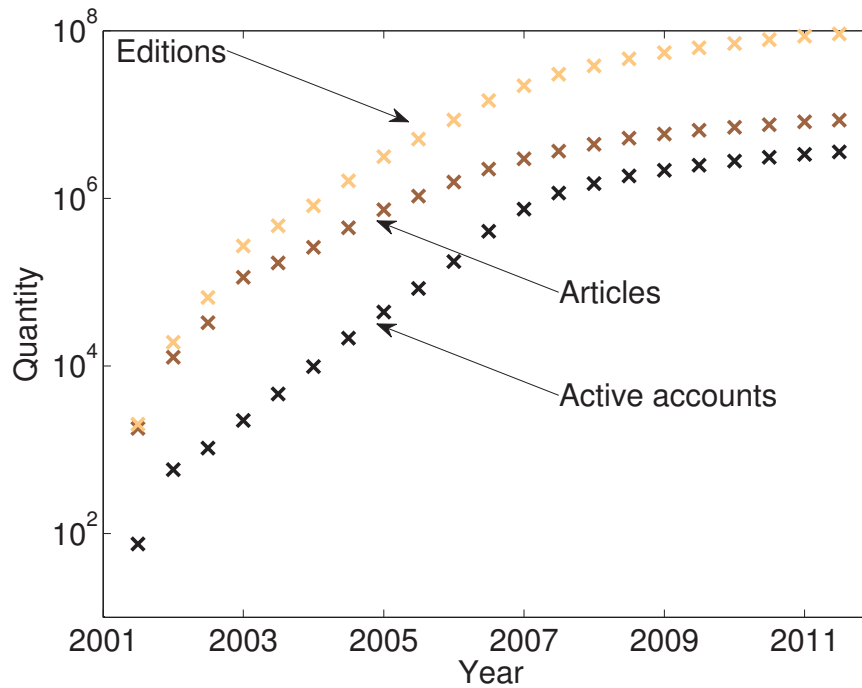


Figure 3: Number of active accounts, articles, and editions

derive the number of edges directly from these two values. If an article is edited by authors or bots 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 and bot 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 analyze the share of editions, articles and accounts created by bots. The results are depicted in Figure 4. Our dataset does not contain any bots before 2003. Afterwards, the share of bot accounts remains at a very low level between 0.016% in 2011 to 0.047% in 2003. This share remains almost constant even though the total number of authors increased significantly. Despite the rather small share of the bot accounts, they generate between 3% and 21% of all editions within our dataset in 2004, respectively 2011. Furthermore, there seem to be some articles which are generated exclusively by bots. This occurs, if all other editions are made by unregistered authors which are neglected in our analysis.

Authors as well as bots are represented as nodes in the collaboration network. The bots only create a small number of additional nodes, however, the large amount of editions does not imply that they also generate the same amount of edges within the collaboration graph, because of the above mentioned reasons.

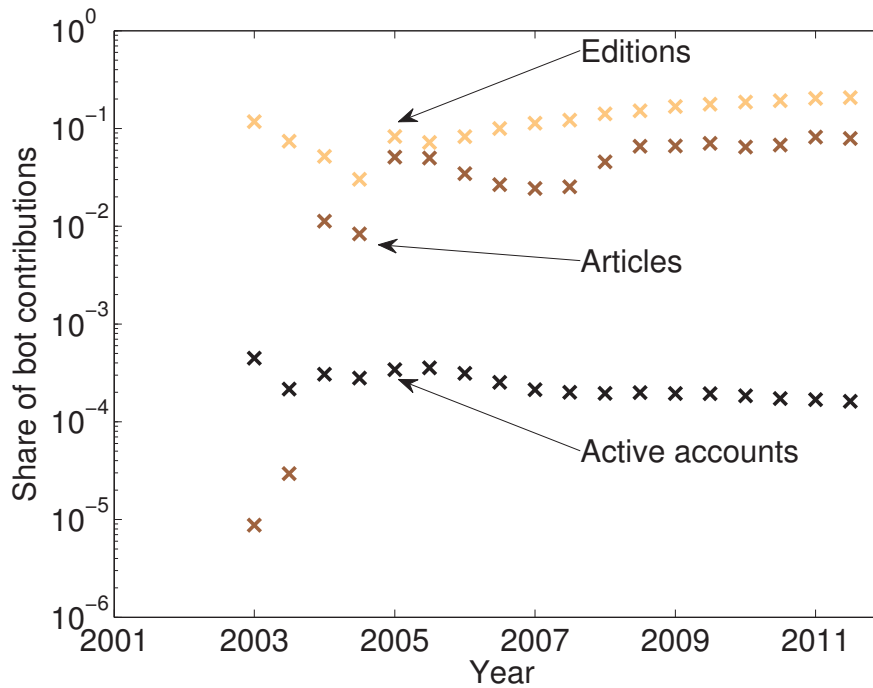


Figure 4: Contribution of bots

## 4.2 Collaboration of Authors

After studying the growth process of the Wikipedia, we now focus on the structure of the collaboration network and how this human generated structure is affected by the bots. 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 human or bot author. The degree distribution of the collaboration network, including the bots, for three different snapshots at the beginning of our available data (mid 2002), at an intermediate point of time (mid 2005), and at the end (beginning of 2011) is shown in Figure 5. The continuous lines show power-law approximations, which we detail on later.

In all three snapshots, the graph shows many of authors 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 authors with a very large number of collaborations, at most over 1 million in 2011.

In order to identify if there are differences between the human authors and bots in terms of their number of collaborations we analyze the maximum and the median of the node degree for both groups. The results are depicted in Figure 6. The diamond-shaped markers depict the maximum node degree, the median of the node degrees are marked with a cross, the whiskers indicate the 25%, respectively the 75% quantile of the node degrees. The dark markers indicate values of bot authors, the light markers of human authors. Before 2003, there have not been any bots in our data. In 2003, there was only one active bot, resulting in the maximum, the

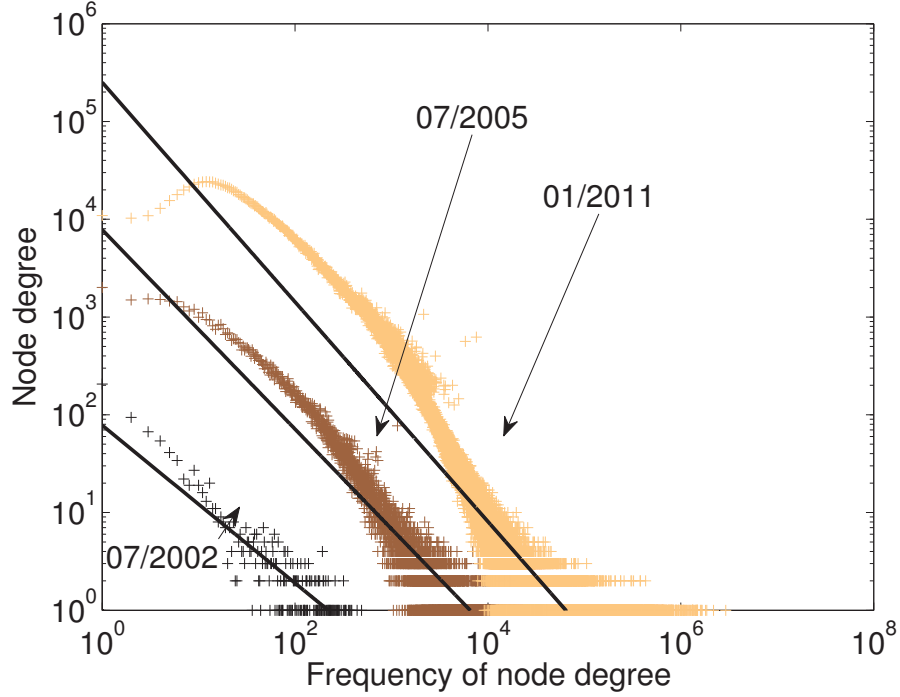


Figure 5: Node degree of the human and bot authors

average node degree, as well as the quantiles showing the same value.

From 2001 to 2011 the maximum node degree of the human authors constantly increases due to the growing number of authors and editions creating connections among them. The same trend applies to the maximum node degree of the bot. Until 2006, the maximum node degree of the human authors was larger or approximately even to the maximum node degree of the bots. Afterwards, the most connected author is a bot with a node degree about 1.6 times larger than the one for the most connected human. During the whole analyzed period, the average node degree of the bot authors was larger than the one of the human authors. This corresponds to the observations made in Figure 4, even if the number of bots is relatively small in comparison to the total number of authors, they are responsible for a relative large amount of editions. With more editions per author, bots are more likely to be higher connected, i.e. have a higher node degree, than human authors. The analysis of the maximum and average node degree lead to the conclusion that nodes with high node degrees are more likely to be bots. However, due to they relatively small number and the relatively small difference in the maximum node degree, bots do not change the “natural” human generated degree distribution.

A further analysis of the distribution 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 be approximated by  $P(k) \propto ck^{-\gamma}$ , with a constant factor  $c$  and the power-law exponent  $\gamma$ . The fitting of a power-law to the measured node degrees is shown with the continuous lines in Figure 5.

The accuracy of the power-law fitting decreases for nodes with lower node degrees in 2005 and 2011. This is typical for other OSNs Mislove et al. (2007). However, the reasons for that phenomenon are different in our case. In OSN analysis the data is usually collected using

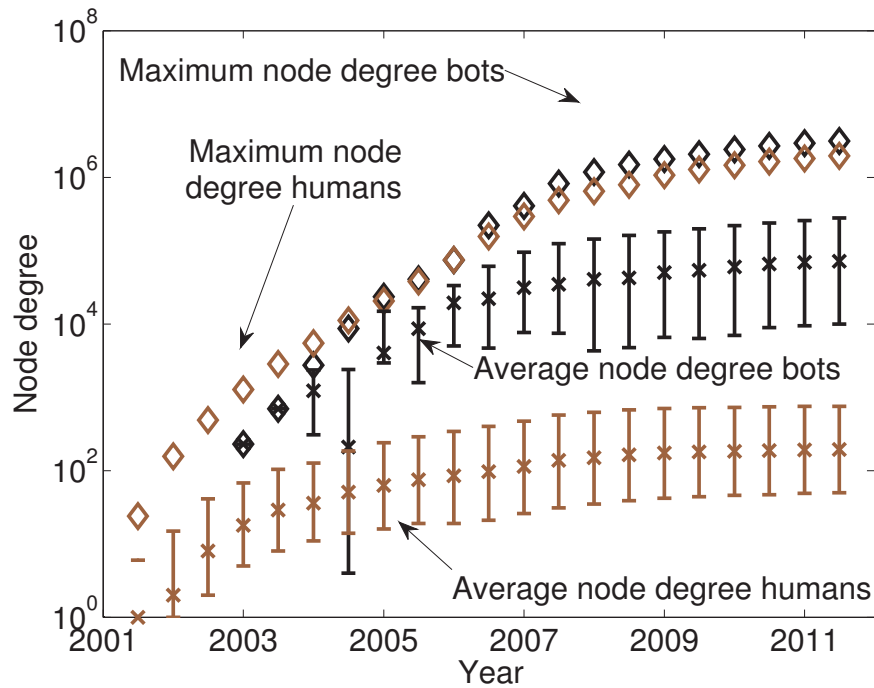


Figure 6: Maximum and average node degree of human and bot authors

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  $\gamma$ . Figure 7 shows the variation of  $\gamma$  from 2001 to 2012, both with and without including the bots. We first focus on the common slope of both curves. Until 2003 we observe some larger changes in the power-law exponent. This is caused by the fast changes in the graph structure at the beginning of the collaboration process. Between 2003 and 2007 we see an increase of the power-law exponent, which means the slope of the node degree distribution steepens. This can be explained with the results of the statistical analysis of the Wikipedia shown in Figure 3. During this phase, the number of newly joining authors grows faster than the number of editions. Consequently, the fraction of highly connected nodes decreases. Since 2007 the power-law exponent  $\gamma$  does not show any significant changes, because the number of authors and editions increases at approximately the same speed. When comparing the different values

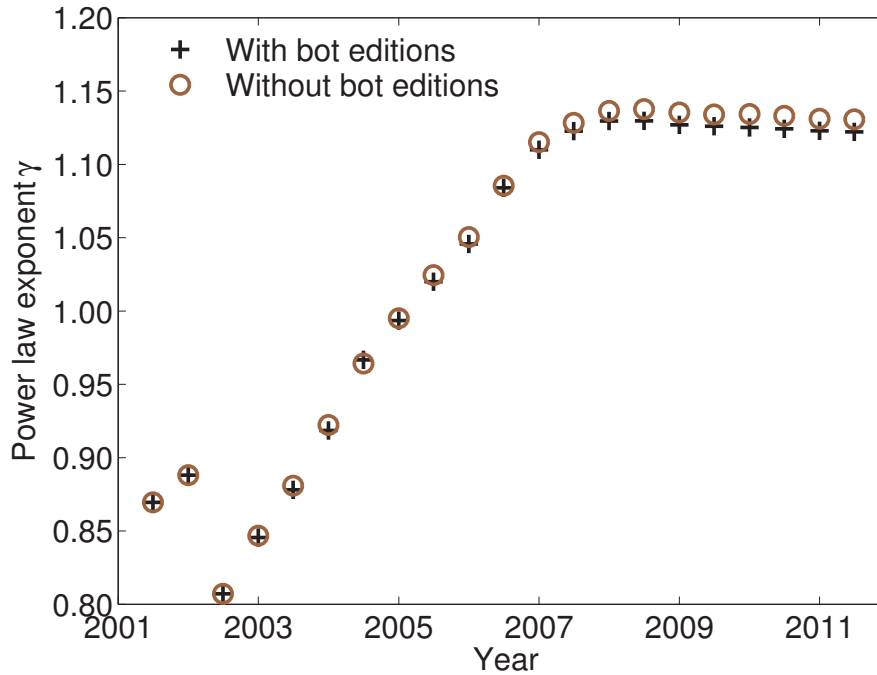


Figure 7: Power-law exponent of the author node degree

of  $\gamma$  for bot and human authors, we observe that they are similar until 2006. Since mid 2006,  $\gamma$  is larger if the bots are not included in the snapshots. This corresponds with the development of the maximum node degree in Figure 6. Since mid 2006, the nodes with the highest degree are bots. If these nodes are removed, the fitted power-law curve flattens, resulting in a decrease of  $\gamma$ . One of the major insight grains from this analyse is that even if the network is growing, its structure does not significantly change.

In order to achieve a better understanding of the connections among the authors, we have a look at the density of the collaboration network. Wasserman and Faust (1994) define the density  $d$  of a network as the ratio of present edges to the maximum number of possible edges and can be calculated by  $d = \frac{|E|}{V \cdot (V-1)/2}$ . The density of the collaboration graph from 2001 to 2012 is shown in Figure 8.

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 snapshots, the development of the network density corresponds to the development of the power-law exponent. Starting with 2003, the number of authors increases

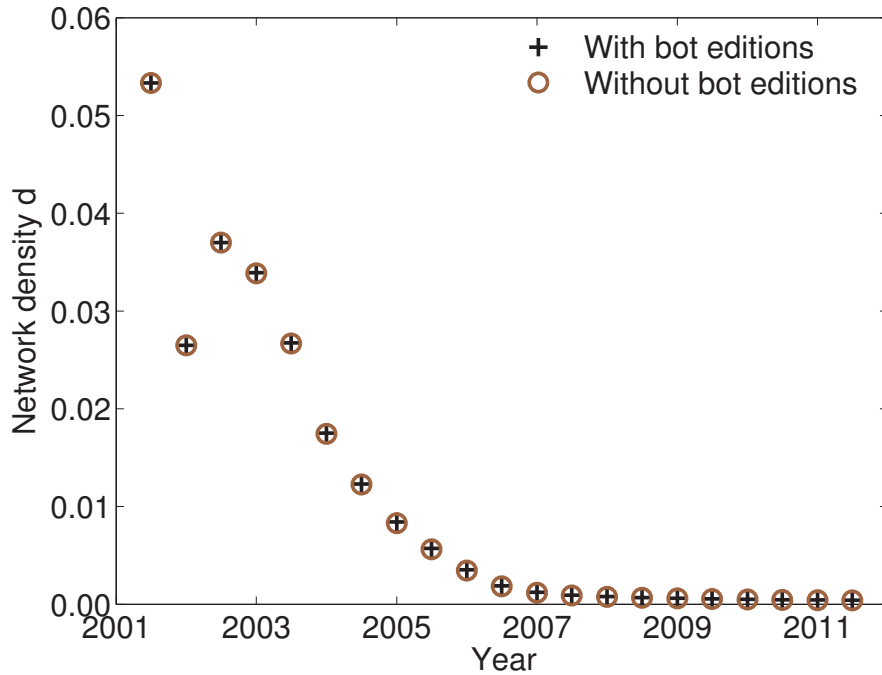


Figure 8: Network density

faster than the number of editions. As a result the number of author with a small node degree increases, causing the density of the network to decreasing and the exponent of the power-law fitting to raise. Since 2007 the density of the collaboration network remains constant and thus also the power-law exponent  $\gamma$ . Our analysis shows that the density of the collaboration network is almost independent of the bots.

### 4.3 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 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 9 depicts the number and size of the connected components of the author network, including bots and humans, 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.

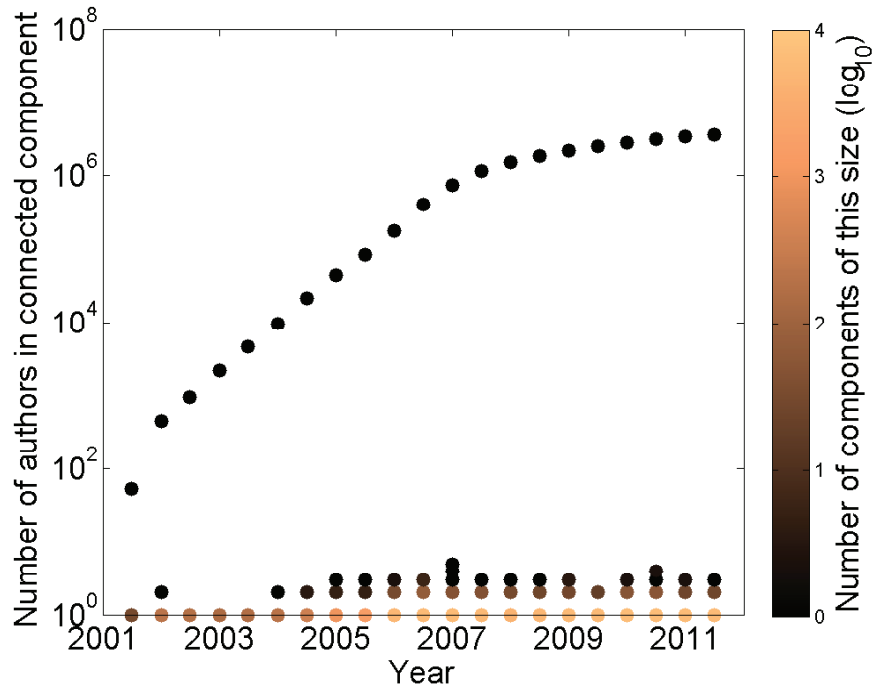


Figure 9: Size and number of connection components with bot editions

To analyze the impact of the bots on the group forming of the authors, we first focus on the number of connected components with and without bots, depicted in Figure 10. If we include the bot authors in our collaboration network, the number of connected components increases slowly up to approximately 1 000 in 2005. Afterwards we have a large increase to about 2 500 in 2006 and then the number of connected component fluctuates between 2 000 and 3 000. Without the bot authors, we observe the same slope until 2006, but afterwards the number of connected components increases significantly higher. This indicates that the collaboration decomposes if the bots are removed leaving some separated human authors, respectively author groups.

As second step we analyze the impact of the bots on the size of the largest connected component. The results are depicted in Figure 11. The left y-axis and the up facing triangles show the relative size  $\epsilon$  of the largest connected component with bots authors normalized to the size of the largest connected component when the bots are removed. The down facing triangles and the right y-axis shows the absolute difference  $\delta$  between the sizes of largest connected component with and without bots. We again observe that the graph decomposes as the bots are removed, because  $\delta > 0$  and  $\epsilon > 1$ . Furthermore,  $\delta$  approximately equals the difference in the number of connected components, cf. Figure 10. This indicates that the newly generated connected components, when removing the bots, are human authors separated from the main component. However, even if some human authors are disconnected as the bots are removed, the size of the largest connected component remains almost constant.

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

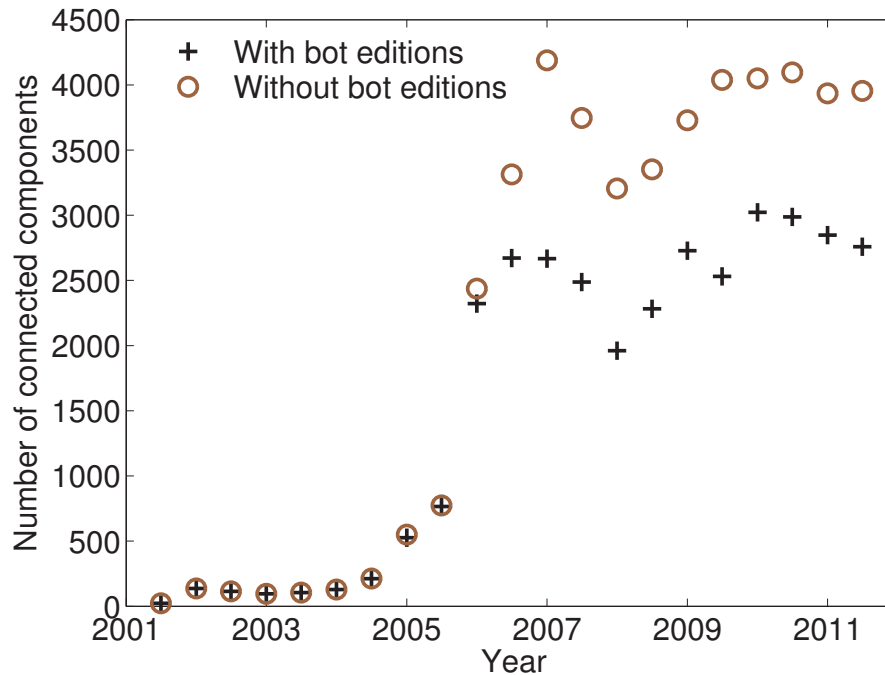


Figure 10: Impact of bots on the number of connected components

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. Moreover the results show that bots have an notable impact on the number of connected components in the collaboration network, but they do not significantly affect the size of the largest connected component.

#### 4.4 The Collaboration Graph as a Small-World Network

Next we analyze if the collaboration networks exhibits the small-world property according to the definition by Watts and Strogatz (1998). 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 12 shows the distribution of the shortest paths between two authors in the largest connected component including both author types of the collaboration network from 2001 to 2005. For snapshots of the author networks later than 2005, the calculation of the shortest path



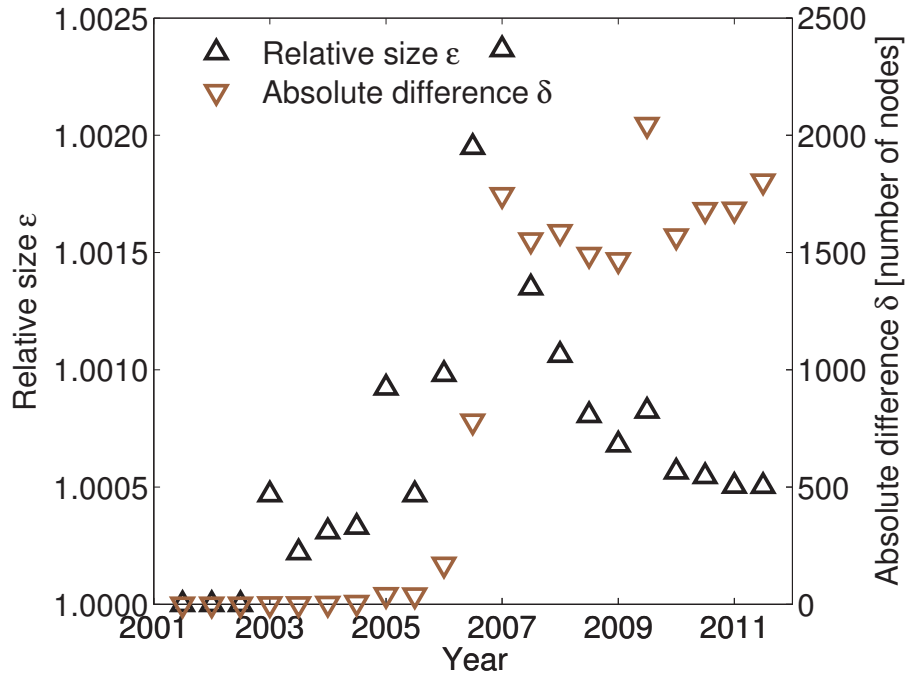


Figure 11: Impact of bots on the size of the largest component

distribution was not possible due to computational limitations. During the whole period, the shortest path is always below 6 hops and most of the authors are connected via a 2-hop path.

In order to investigate the influence of the growing network, we show the shortest path length over time in Figure 13. We see that the maximum path length is 6 in 2002 and 5 or less in all the other snapshots. Since 2003 the average path length remains almost at a length of 2. 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 Watts and Strogatz (1998). Similar to the graph density, the bot authors have only a neglect able influence on the average path length.

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 14. For larger snapshots an analysis was not possible due to the computational limitations. During the analyzed phase, the clustering coefficient was not affected by the bot authors and for both snapshot types the clustering coefficient increases from 0.46 to 0.85

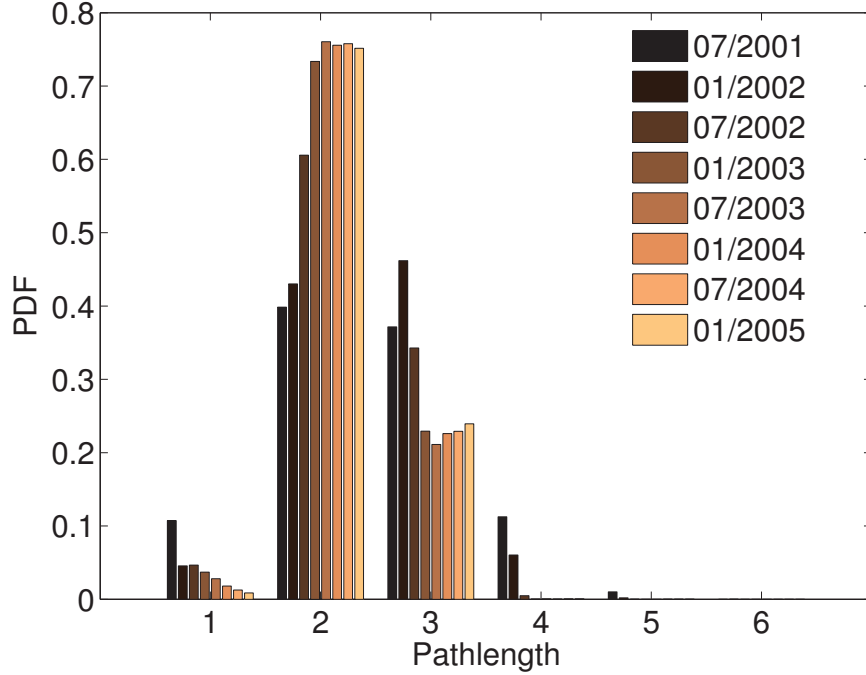


Figure 12: Shortest path length

indicating an increasing cliquishness of the collaboration network.

In order to compare the measured clustering coefficient to the clustering coefficient of a random network, we used the formula given by Dorogovtsev (2003) 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 Newman et al. (2001), 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 also shown in Figure 14. 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 even if automated authors also contribute to the collaboration network.

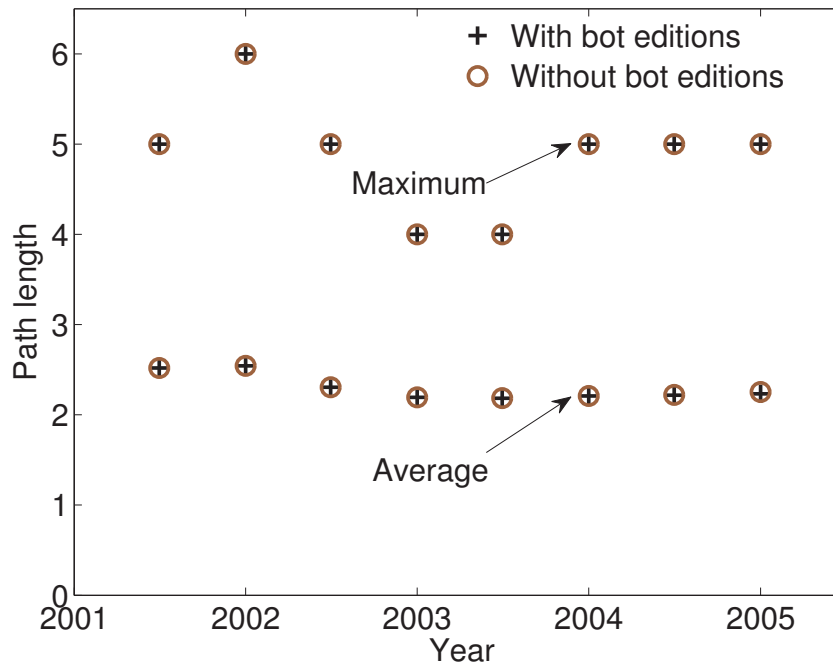


Figure 13: Maximum and mean path length

## 5 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 human and bot 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 even if automated scripts are also included. 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 structure does not change, although the size of the network still increases. In particular the power-law exponent of the degree distribution, density of the network, and the clustering coefficient remain 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. Even though bots have an impact on the number of connected components, they do affect the size of the largest component only marginally. 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

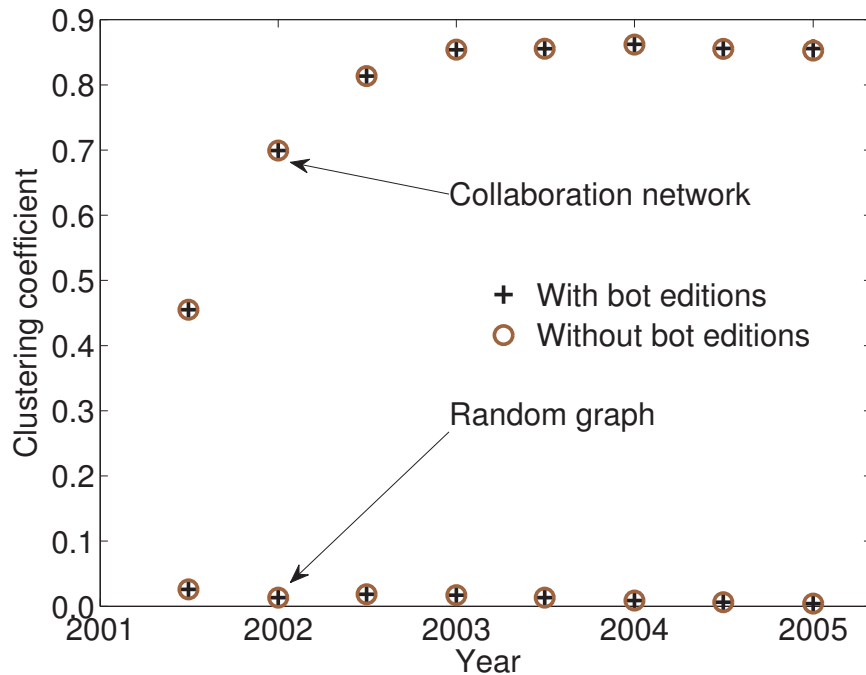


Figure 14: Mean clustering coefficient

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. We are convinced that the insights gained in this work can be used for analyzing networks dynamics in general and developing new analysis methods, e.g. to determine if a network has reached its steady state. This will be subject to future work.

## References

- Bellomi, F., Bonato, R.: “Network analysis for wikipedia”; Proc. Wikimania; 2005.
- Biuk-Aghai, R.: “Visualizing co-authorship networks in online wikipedia”; Proc. Symposium on Communications and Information Technologies; 2006.
- Brandes, U., Kenis, P., Lerner, J., Van Raaij, D.: “Network analysis of collaboration structure in wikipedia”; Proc. International Conference on World Wide Web; 2009.
- Dorogovtsev, S.: “Clustering of correlated networks”; Arxiv preprint cond-mat/0308444; (2003).
- Hirth, M., Lehrieder, F., Oberste-Vorth, S., Hoßfeld, T., Tran-Gia, P.: “Wikipedia and its Network of Authors from a Social Network Perspective”; Proc. International Conference on Communications and Electronics (ICCE); 2012.

- Laniado, D., Tasso, R., Volkovich, Y., Kaltenbrunner, A.: “When the wikipedians talk: Network and tree structure of wikipedia discussion pages”; Proc. International Conference on Weblogs and Social Media; 2011.
- Massa, P.: “Social networks of wikipedia”; Proc. Conference on Hypertext and Hypermedia; 2011.
- Mislove, A., Marcon, M., Gummadi, K. P., Druschel, P., Bhattacharjee, B.: “Measurement and analysis of online social networks”; Proc. Conference on Internet Measurement; 2007.
- Newman, M., Strogatz, S., Watts, D.: “Random graphs with arbitrary degree distributions and their applications”; Physical Review E; 64 (2001), 2.
- Wasserman, S., Faust, K.: Social Network Analysis: Methods and Applications; Cambridge University Press, 1994.
- Watts, D., Strogatz, S.: “Collective dynamics of small-world networks”; Nature; 393 (1998), 6684.
- Zlatic, V., Božičević, M., Štefančić, H., Domazet, M.: “Wikipedias: Collaborative web-based encyclopedias as complex networks”; Physical Review E; 74 (2006), 1.