Social Knowledge Dynamics: A Case Study on Modeling Wikipedia

Benyun Shi

Abstract

We show the current research progresses and focuses on Wikipedia to try to understand principles behind social knowledge dynamics. Wikipedia is a popular web-based encyclopedia, which can be edited freely and collaboratively by its users. This kind of truly collaborative and social platform provides a good opportunity for sociologist, physicists, computer scientists, and other researchers on traditional social knowledge dynamics to better understand the evolution of social cognition - that is, the ability of a group of people to remember, think, and reason. We survey recent research work on how to characterize global features of Wikipedia and how to model the growth of Wikipedia. While previous work mainly focus on statistical analysis of a mass of date from Wikipedia, it is extremely hard to understand the nature behind the Wikipedia phenomena. In this work, we try to study how to build bottom-up models based on autonomy-oriented computing (AOC) to better understand the nature and fundamental principles of social knowledge dynamics through a case study of Wikipedia.

1 Introduction

Wikipedia is a freely available online encyclopedia, that anyone can create, edit, as well as delete articles. The unique character of the free editing policy and the large number of participants make the success of Wikipedia. Each article of Wikipedia can be treated as a collective knowledge of a group of users who have made updates on it. Sociologists define knowledge as follows. "*Knowledge is embodied in people gathered in communities and networks. The road to knowledge is via people, conversations, connections and relationships. Knowledge surfaces through dialog, all knowledge is socially mediated and access to knowledge is by connecting to people that know or know who to contact.*" In terms of this definition, the evolution of Wikipedia can be treated as a type of social knowledge dynamics for the following reasons: i) the formation of each article of Wikipedia is contributed by a collective of users that gathered together on the page of the article, ii) users on the article page can exchange their knowledge through "talk" page (i.e., user interaction), iii) users with similar opinions or users who act frequently on a specific article may form community on the article, and iv) the underlying structure of some articles may inversely influence uses' knowledge on some other articles. In this case, finding principles behind about the evolution of Wikipedia may help to have a better understanding about social knowledge dynamics.

Social knowledge dynamic is a research branch of social dynamics. Social dynamics mainly focuses on the study of a society of individuals to react to inner and/or outer changes and tries to find ways to explain some social phenomena. Researches [7] on social dynamics have shown that interesting global patterns, such as phase transition and criticality in culture dynamics [8], can emerge from even a group of simple individuals (in term of their behavior rules and/or relationships). Understanding the driven force behind such emergence is the first step for us to realize the complex real world. Previous studies show that individual's social behaviors play important roles in this process. For example, Surowiecki's "The Wisdom of Crowds" [20] discusses the question about why the many are smarter than the few and how collective wisdom shapes business, economies, societies and nations.

The physics community [7] has long aimed at the discovery of fundamental principles behind emergent properties, such as phase transition, of the social systems by local dynamic models, for example, the simple sand pile model [6]. However, for any investigation of local dynamic models of social dynamics, there are two levels of difficulty: i) the definition of sensible and realistic microscopic models, and ii) the usual problem of inferring the macroscopic phenomenology out of the microscopic dynamics of such models. However, without intact data of social activities, it is extremely hard i) to build a realistic microscopic model, and ii) to observe explicit macroscopic phenomena. With a millions of articles, hundreds of thousands of contributors, and tens of millions of fully recorded article revisions, Wikipedia's freely available database [2] made it possible to study how human knowledge is recorded and organized through an open collaborative process.

In this work, we survey recent research work on how to characterize microscopic features of Wikipedia and how

to model the growth of Wikipedia. These previous work mainly focus on statistically analyzing a mass of data from Wikipedia, and building statistical models to explain some macro-level patterns. We argue that the macro-level analysis and modeling is not enough to understand the nature of the evolution of Wikipedia, as well as social knowledge dynamics, for the reason that the real Wikipedia is driven by the social dynamics, including user-to-user interactions, use-to-group interactions, and group-to-group interactions, rather than simple stochastic processes.

We then discuss how to build bottom-up models based on autonomy-oriented computing (AOC) [12] to better understand the nature and fundamental principles of social knowledge dynamics. AOC defines and deploys a system of local autonomy-oriented entities, which spontaneously interact with their environment and operate based on their behavior rules. Since the features of AOC match the formation of Wikipedia very well, we believe the AOC-based models can help us to understand the whole Wikipedia system, the topic evolution, and the dynamic of user communities on Wikipedia.

The rest of this work is organized as follows. In section 2, we briefly introduce the roles of Wikipedia in the development of Web 2.0. In section 3, we summarize the microscopic phenomena analysis of Wikipedia data dump. In section 4, we survey the recent studies of modeling of Wikipedia growth. In section 5, we discuss the possibility of designing AOC-based model to study the nature and fundamental principle of social knowledge. Finally, we conclude our work in section 6.

2 Roles of Wikipedia

Wikipedia relies on server-side technology that allow both registered and anonymous users to make instant update (i.e., create, edit, delete) to a page or link via a web interface. It has archiving systems that record all previous edits of a page and make it simple to revert to an earlier version. This archiving system ensures that no permanent harm can be caused by cankered editing. The key technology behind Wikipedia is its online heavyweight Web-based collaboration. Ed H. Chi's work [9] on Web 2.0 technologies provides the following social web collaboration spectrum (Figure 1).

At the lightweight end of the social collaboration spectrum, researchers are focusing on information-foraging and behavioral models. In the middle the spectrum, mathematicians and social scientists are developing new theories and algorithms to model, mine, and understand social structure. While at the heavy end of the spectrum, researchers should focus on studying what hinders and fosters coordination on large group projects, which is especially important for understanding collaborative co-creation system such as

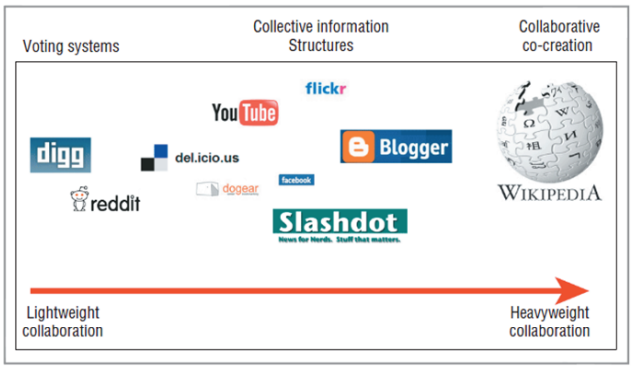

Figure 1. Social Web collaboration spectrum. (adopted from Fig. 2 in [9].)

Wikipedia.

For the richness of data from Wikipedia, it is important to understand features behind these data so that further studies can be carried out. In the next section, we summarize the recent data analysis about Wikipedia.

3 Related Work on Wikipedia

Since Wikipedia provides rich data sources to download, including i) articles, ii) categories, iii) images and multimedia, iv) user, help, and talk pages, and v) redirects, templates and broken links, there are many work on analyzing these data with different purposes. For example, D. Spinellis and P. Louridas [19] analyzed the evolution of the ratio between incomplete and complete articles, and the relation of references and definitions of articles. They found that "the addition of new articles in not a purely random process following the whims of its contributors but that references to nonexistent articles trigger the eventual creation of a corresponding article". R. Almeida et al. [3] studied the updates of articles in Wikipedia, and found that the evolution of updates is governed by a self-similarity process. A. Kittur et al. [11] proposed a mapping approach to study the distribution of topics in Wikipedia. Also, there are also other researches [21] focusing on visualization of history flow of Wikipedia.

In this work, we mainly focus on the analysis where Wikipedia is treated as complex networks, where the articles of Wikipedia represent the nodes of the network while the hyperlinks pointing from one article to another are treated as links. Many characteristics [17] of complex networks can be used to analyze the structure of Wikipedia, for instance, degree distribution, clustering, path length, network topology, reciprocity, motifs, and so on. Here, we briefly introduce analysis results for three characteristics that may relate to bottom-up model designing.

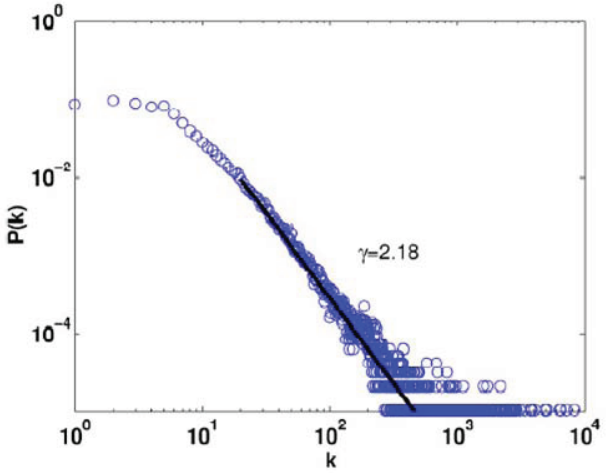

Figure 2. The in-degree distribution of Japan Wikipedia. (adopted from Fig. 3 in [23].)

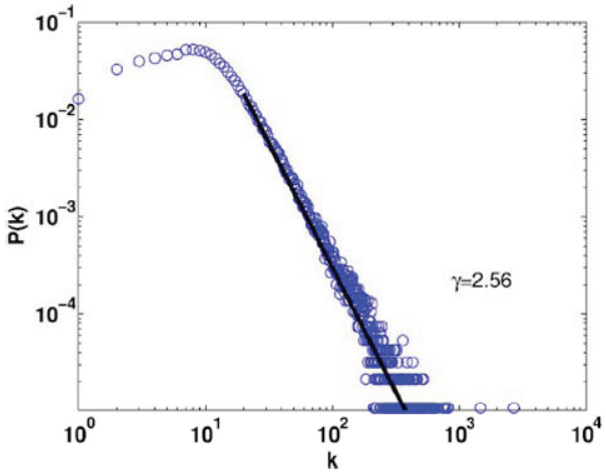

Figure 3. The out-degree distribution of Japan Wikipedia. (adopted from Fig. 4 in [23].)

3.1 Degree distribution

The in-degree (respectively, out-degree) of a node in Wikipedia network measures the number of articles that link into the article (respectively, the number of articles that the article links to as references). In most cases, two articles sharing a link reflect some kinds of relations in term of their contents. Hence, analyzing degree distributions (both in- and out-degree distributions) of Wikipedia may help to understand the article structure of Wikipedia, and may future help to understand the intrinsic relation of different kinds of knowledge.

V. Zlatic et al. [23] presented an analysis of Wikipedia in several languages as complex networks. They found that both the in-degree and out-degree of distributions of most Wikipedia follow power-law distributions. Figures 2-3 are examples of in-degree and out-degree distributions of the Japan Wikipedia.

The similar degree distributions for different kinds of languages supports the assumption that the Wikipedia in different languages represent realizations of the same process of network growth, which in turn shows that there must be some fundamental principles behind the social knowledge dynamics. Hence, various statistical models can be proposed to study the Wikipedia growth. We will introduce these models in section 4.

However, most of recent work analyze Wikipedia network as a whole. We argue that to have a better understanding of knowledge dynamic, it is necessary to do analysis about Wikipedia on different granularity (e.g., analysis on topic level to measure the topic distribution; analysis within different topics to measure the distribution of articles in a specific topic). By doing so, we can observe whether or not the observed scale-free phenomena exist in the Wikipedia structure. In other words, does self-similarity exist?

3.2 Reciprocity and feedback loops

Another interesting features observed in Wikipedia network is their reciprocity [23]. Reciprocity quantifies mutual "exchange" between two nodes. Reciprocal links are just the links pointing from the node i to the node j for which exists a link pointing from node j to the node i. The reciprocity is then defined as

$$\rho = \frac{L_{bd}/L - \bar{a}}{1 - \bar{a}}$$

where L_{bd} represents the number of bidirectional links, i.e., links for which a reciprocal link exists. L is the total number of directed links and $\bar{a} = \frac{L}{N(N-1)}$ is the density of links.

Another feature that similar with reciprocity is feedback loops in the networks. A feedback loop in Wikipedia network can be defined as a loop with directed links that start from and end with the same node. The ecological study in [15] observed that the number of feedback loops in the species network is correlated with system lifetime. Though there have not many papers dealing with the origin of reciprocity or network evolution models that capture this quantity, we believe that reciprocity and feedback loops play important roles in the evolution of Wikipedia. In this case, analysis about reciprocity and feedback loops on Wikipedia is necessary and important work before designing models on Wikipedia to understand social knowledge dynamics.

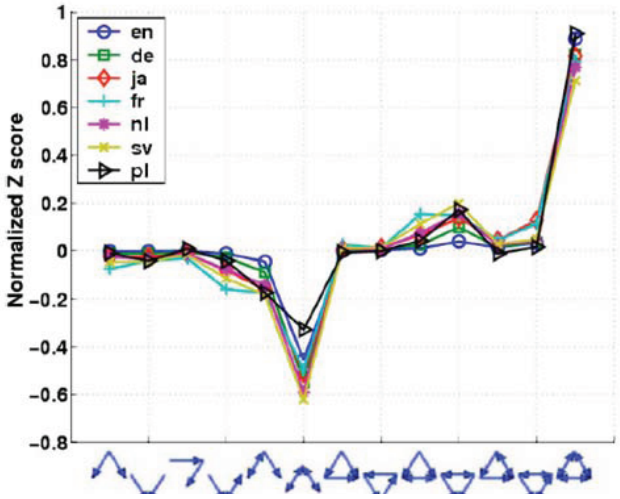

Figure 4. The triadic motif profiles of Wikipedias. (adopted from Fig. 10 in [23].)

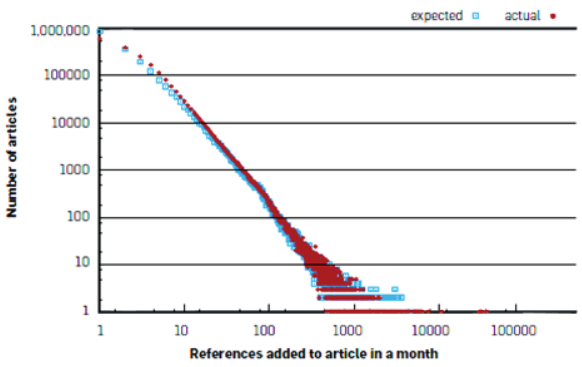

Figure 5. The frequency distribution of the expected and actual number of references added each month to each article (adopted from Fig. 3b in [19]).

3.3 Motifs

Motifs introduced in [16] are small subgraphs (in [16], triadic subgraphs are concerned) of networks, which are used to systematically study similarity in the local structure of networks. By comparing significance profiles of the original network with that of randomly generated networks, networks with similar local structure can be grouped together. Interesting results shows that networks with similar functions may have similar motif profiles (see Fig. 1 in [16] for more details).

By analyzing the motifs in Wikipedia network, V. Zlatic et al. [23] show that the triadic motif significance profiles of Wikipedia networks with different languages are very similar, though Wikipedia networks with different have different size. In Figure 4, the x axis depicts all possible triadic motifs of a directed network, while the y axis represents the normalized Z score [16] for a given motif.

The similar motif significance profiles for different languages indicate that there maybe exist common fundamental principles that drive the growth of Wikipedia. Then, what are the principles behind this? Do these principles related to principles that drive the social knowledge dynamics? We should keep these questions in mind when we design models to discover the essential principles.

4 Modeling Wikipedia's Growth

In this section, we will mainly focus on the scale-free phenomena in Wikipedia. The models for scale-free can be divided into two groups: i) scale-free as the result of an optimization or phase transition process [2], and ii) scale-free as the results of a growth model, such as preferential attachment. Recent study [18] also shows that scale-free can also be formed by deliberate removal.

It is impossible to examine the emergence of scale-free in other big real-world networks, as there is no full record of their evolution. Wikipedia provides a platform to allow us to witness, and validate preferential attachment. Two models based on preferential attachment will be introduced in this section to explain two different power law phenomena. We believe that lots of research should be down in the future to find more fundamental principles.

4.1 A model about reference

The following model is from the study of D. Spinellis and P. Louridas in [19]. Consider a model where at each time step t a month, a variable number of articles and r_t references are added. The references are distributed among all entries following a probability $p(k_{i,t}) = \frac{k_{i,t}}{\sum_{j,t} k_{j,t}}$, with the sums and the connectivity calculated at the start of t. The expected number of references added to entry i at month t is then $E(k_{i,t}) = r_t p(k_{i,t})$. The authors find a close match between the expected and the actual numbers in Wikipedia data. Figure 5 shows a log-log plot of frequency distribution of the expected and actual number of articles gaining a number of references in a month.

4.2 A model about degree distribution

The following model is from the work of V. Zlatić et al. [22]. The model consists of two steps. In the first one a new node, introduced in the network at time t and therefore labeled as t, attaches to the network with m outgoing links.

The probability that the given link, from these m links, will attach itself to some node $s < t$ is proportional to the in-degree $k_i(s)$ of the node s. In the second step for every new link with the probability r a new reciprocal link is formed between nodes s and t. The results also show a perfect fit between the in-degree distribution of Wikipedia and the expected distribution of this model. See figures 2-3 in [22] for more details.

Though these two models could be used to reflect some principles behind network growth, such as preferential attachment, we should realize that the real Wikipedia is driven by the social dynamics, including user-user interactions, use-group interactions, and group-group interactions, rather than the simple stochastic processes. Autonomy-oriented computing provides a bottom-up way to study the global emergence of Wikipedia. In the next section, we will discuss the possibility of adopting AOC-based model [12] to design the social knowledge dynamics mechanisms.

5 AOC-Based Models

AOC defines and deploys a system of local autonomy-oriented entities, which spontaneously interact with their environment and operate based on their behavior rules. Self-organization as a soul of AOC allows entities to self-organize both their relationships and their local dynamics, with respect to some predefined settings, so that problems with various complicated properties can be adaptively solved. Feedbacks (both positive and negative) play important roles in self-organization process. The fundamental difference of positive feedback and negative feedback is that i) positive feedbacks aim at accelerate aggregations with non-linear amplification (e.g., reproduction), while ii) negative feedback (e.g., collective regulation) aim at self-correction/self-tuning.

The features of AOC match the formation of Wikipedia very well in terms of i) Wikipedia is formed by editing of spontaneous users, ii) users may interact/discuss with each other in "talk" page of each article, iii) contents of other pages may give feedbacks to regulate or aggregate users decisions, and iv) communities can be formed during the evolution of Wikipedia, which inversely play important roles for Wikipedia's evolution.

In this case, what are the fundamental behavior rules of entities to form global patterns of Wikipedia? How do entities self-organize themselves during the evolution of Wikipedia? Do these rules and self-organization reflect the formation rule of social knowledge and social organization? To answer these questions, we should carefully define entities behavior rules and relation structures during model designing.

In the rest of this section, we propose three possible research directions on Wikipedia for our future research.

5.1 Wikipedia as a system

Any natural systems have processes of birth, boom, and death. As a collaborative system based solely on users' spontaneous actions, what's the driven of its birth, boom, and death?

Robert M. May [13] [14] has studied the impact of interaction strength, connectance, on stability of large complex ecosystems. His results and subsequent work indicate that large randomly assembled ecosystems tend to be less stable as they increase in complexity, where the complexity is measured by the connectance and the average interaction strength between species. R. Mehrotra et al. [15] have studied an evolutionary model that exhibits spontaneous growth, stasis, and then a collapse of its structure. They find that the typical lifetime of the system increase with the diversity of its components. They also find that the number of feedback loops play important roles in the process of collapse of the system. There are also other researchers trying to understand nature behind some phenomena of complex system, such as catastrophe [10], punctuated equilibrium [5], and so on.

We argue that as a system, Wikipedia may also have similar phenomena as the ecosystems do. However, up to now, most existing models are based on statistical analysis or stochastic simulations, which are not enough to reflect real-world phenomena as we argued in section 1. Hence, in the future, we would like to build AOC-based models to simulate and analyze the Wikipedia system.

5.2 Topic evolution on Wikipedia

Previous work on topic mining focus on mining specific topics from a large volume of data, where for most cases, i) the data are static (some researches on dynamic topic mining adopt time windows to reflect the topic evolution), and ii)the mining processes are based on sematic/content analysis.

However, in Wikipedia, topics are dynamically evolving as a result of users dynamics. Different with traditional topic evolution problem, we can treat the topic evolution on Wikipedia as a results of user-to-user interactions, or even the interaction among groups of users. In this case, we can further observe the evolution of social knowledge.

While traditional data mining strategies cannot explain the driven force behind the evolving the topics, we argue that AOC-based model on Wikipedia topics may reveal some principles by focusing on entities/topics local interactions and collective regulations. The studies on culture evolution [8] [4] may provide some ideas on this research aspect.

5.3 User community dynamics on Wikipedia

Traditional community discovery algorithms [1] are mostly based on the quantitative measure of modularity Q. The modularity is defined as $Q = \sum_i (e_{ii} - a_i^2)$, where e_{ii} measures the fraction of edges that have both ends pointing at nodes in group i, and a_i measures the fraction of edges whose end points belong to at least one of vertices in group i. Some researchers [1] argue that the linkage-based measurement cannot reflect multiple relationships.

In Wikipedia, each user may associate into multiple articles. While for each article, there will be multiple users acting on it. This kind of bi-party network may provide novel definitions of community so that new approaches about community evolution can be proposed. Also, by AOC-based models, communities may emerge from entities local interactions, and may further dynamically evolve as entities activities change over time.

For above mentioned research aspects, many work need to be done in the future to get more deep understanding about Wikipedia and its dynamics.

6 Conclusion

In this work, we survey both the macro- and microscopic studies about Wikipedia. For the macroscopic study, we focus on the data analysis of Wikipedia, especially on the Wikipedia network analysis, and find many interesting phenomena. For the microscopic study, we introduce two statistical models based on preferential attachment to fit the scale-free phenomena in Wikipedia. Since the real Wikipedia is driven by the user dynamics rather than simple stochastic processes, it is necessary to design AOC-based models to study the essential rules behind the Wikipedia dynamics. Several research aspects are proposed for future study at the end of this paper.

References

[1] Community Discovery. http://www.cscs.umich.edu/ crshalizi/notebooks/community-discovery.html.

[2] Download Wikipedia. http://download.wikimedia.org.

[3] R. Almeida, B. Mozafari, and J. Cho. On the evolution of wikipedia. *In Proceeding of International Conference on Weblogs and Social Media*, 2007.

[4] R. Axelrod. The dissemination of culture: A model with local convergence and global polarization. *Journal of Conflict Resolution*, 41(2):203–226, 1997.

[5] P. Bak and K. Sneppen. Punctuated equilibrium and criticality in a simple model of evolution. *Physical Review Letters*, 71(24):4083–4086, 1993.

[6] P. Bak, C. Tang, and K. Wiesenfeld. Self-organized criticality. *Physical Review A*, 38(1):364–374, 1988.

[7] C. Castellano, S. Fortunato, and V. Loreto. Statistical physics of social dynamics. *Reviews of Modern Physics*, 81:591–646, 2009.

[8] D. Centola, J. C. Gonzalez-Avella, V. M. Eguiluz, and M. S. Miguel. Homophily, cultural drift, and the co-evolution of cultural groups. *Journal of Conflict Resolution*, 51(6):905–929, 2007.

[9] E. H. Chi. The social web: Research and opportunities. *Computer in Computer*, 41(9):88–91, 2008.

[10] S. A. Kauffman. *The Origins of Order: Self-Organization and Selection in Evolution*. Oxford University Press, 1993.

[11] A. Kittur, E. H. Chi, and B. Suh. What's in wikipedia? mapping topics and conflict using socially annotated category structure. *In Proceeding of the* 27^{th} *Annual CHI Conference on Human Factors in Computing Systems*, pages 4–9, 2009.

[12] J. Liu. Autonomy-oriented computing: The nature and implications of a paradigm for self-organized computing. *Keynote Talk at The* 4^{th} *International Conference on Natural Computation, and the* 5^{th} *International Conference on Fuzzy Systems and Knowledge Discovery*, pages 3–11, 2008.

[13] R. M. May. Will a large complex system be stable? *Nature*, 238:413–414, 1972.

[14] R. M. May. Qualitative stability in model ecosystems. *Ecology*, 54(3):638–641, 1973.

[15] R. Mehrotra, V. Soni, and S. Jain. Diversity sustains an evolving network. *Journal of the Royal Society Interface*, 6(38):793–799, 2009.

[16] R. Milo, S. Itzkovitz, N. Kashtan, R. Levitt, S. Shen-Orr, I. Ayzenshtat, M. Sheffer, and U. Alon. Superfamilies of evolved and designed networks. *Science*, 303(5663):1538–1542, 2004.

[17] M. E. J. Newman. The structure and function of complex networks. *SIAM Review*, 45(2):167–256, 2003.

[18] M. Salathé, R. M. May, and S. Bonhoeffer. The evolution of network topology by selective removal. *Journal of Royal Society, Interface*, 2(5):533–536, 2005.

[19] D. Spinellis and P. Louridas. The collaborative organization of knowledge. *Communications of the ACM*, 51(8):68–73, 2008.

[20] J. Surowiecki. *The Wisdom of Crowds: Why the Many are Smarter than the Few and How Collective Wisdom Shapes Business, Economies, Societies, and Nations*. Doubleday, 2004.

[21] F. B. Viégas, M. Wattenberg, and K. Dave. Studying cooperation and conflict between authors with history flow visualizations. *In Proceedings of the Conference on Human Factors in Computing Systems*, pages 575–582, 2004.

[22] Vinko and H. Štefančić. Model of wikipedia growth based on information exchange via reciprocal arcs. *Physics and Society*, 2009.

[23] V. Zlatic, M. Bozicevic, H. Stefancic, and M. Domazet. Wikipedias: Collaborative web-based encyclopedias as complex networks. *Physical Review E*, 74(1), 2006.