2009

How Do Bloggers Comment: An Empirical Analysis of the Commenting Network of a Blogging Community

Tingting Jiang
The University of Pittsburgh, tij4@pitt.edu

Xiaoguang Wang
Wuhan University, wxguang@whu.edu.cn

HOW DO BLOGGERS COMMENT: AN EMPIRICAL ANALYSIS OF THE COMMENTING NETWORK OF A BLOGGING COMMUNITY

Research-in-Progress

Tingting Jiang
School of Information Sciences
University of Pittsburgh
135 North Bellefield Avenue
Pittsburgh, PA 15260, USA
tij4@pitt.edu

Xiaoguang Wang
School of Information Management
Wuhan University
Luojiashan, Wuhan, 430072, China
wxguang@whu.edu.cn

Abstract

The blog is one of the most familiar social media to Web users. Its social nature suggests that blogging is not only about writing and publishing bloggers' experience and thoughts, but also about establishing social relationships and networks with others through various links including citations, blogrolls, comments, and trackbacks. We have a particular interest in comments, and develop a commenting model that considers both unidirectional and bidirectional commenting relationships. This paper presents a social network analysis of the commenting network of CSDN blogging community, the largest Chinese language IT blogging community in China. According to our findings, this is a sparse network, and the most noticeable nodes in it are a few central bloggers who possess a lot of incoming relationships. Each of them is surrounded by a number of ordinary bloggers who recognize the central blogger as their exclusive information source in a star topology. Our continuing study will focus on these central bloggers and examine their influences on the information diffusion in CSDN blogging community at the semantic level.

Keywords: Blog, blogging community, commenting network, commenting relationships, social network analysis, information diffusion

Introduction

Blogging, the maintaining of regularly updated entries, is one of the earliest and most prevalent phenomena of Web 2.0. As witnessed by the past decade, the blog has evolved into a powerful information medium. For example, bloggers are enabled to enrich their blogs with photos, audios, videos, etc. while early blogs were primarily textual. Most importantly, however, bloggers are supported to establish social relationships with each other and form social networks through various links (Blood, 2004). Multimedia just change the way information is presented, but links reshape the blog into a social medium for communication and interaction. Citations, blogrolls, comments, and trackbacks are the four types of links most frequently seen in today's blogs. Although linking always happens on a one-to-one basis, i.e. between two bloggers, the links are by default viewable to others (Furukawa et al., 2007). By publishing digital content and connecting to those sharing common ground via one or more types of links, bloggers gather into all kinds of blogging communities. Without clear boundaries, the blogging communities are the outcome of bloggers' self-organization. Bloggers' awareness of their own community responsibilities will determine how they behave in the diffusion of information within the community (Marlow, 2004). It is therefore no longer appropriate to regard blogs as individual information sources without reference to the communities they belong to.

This research study empirically examines an IT blogging community to understand the effectiveness of the blog in disseminating information and ultimately to provide useful guidance for the development of blogging software or services in the future. To achieve these research goals, we need to divide our study into two stages considering the social nature of the blog. First to be investigated is the social network of the blogging community, the underlying structure that represents the interweaving paths in which information travels. What are the roles played by different bloggers in the network? How do they affect the directions of information flow? Only after answering these questions can we proceed to the second stage and dig into the semantic level to see why certain bloggers have stronger influences than others and why information diffuses in the manner we've found.

We acquired abundant longitude data (the entire year of 2007) from CSDN blog (http://blog.csdn.net/), the largest blogging community of IT engineers and computer scientists in China. Although all the four types of link data are available and each can be used to construct a social network that reflects the structure of CSDN blogging community in one way, we are most interested in comments. This is not only because that the majority of previous studies are limited to citation and blogroll links, but also that comments exclusively contain expressed meanings that are indispensible to our semantic analysis. For one thing, a social relationship is implicitly built between the commenting blogger and the one commented. For another, the information transmitted in commenting is explicitly recorded with texts which can be retrieved later. The significance of commenting to blogging can be evidenced by the large quantity of comments that accounts for 30% of the total volume of blog entries (Mishne and Glance, 2006). This paper will only present the first stage of our study that has been completed already. Existing results are based on the social network analysis of CSDN commenting network constructed with our model of blog commenting.

Theoretical Framework and Research Questions

This section firstly provides a review of related work on blogs and blogging communities with a particular emphasis on commenting. We then illustrate a general model of blog commenting, created at the beginning of our study as the foundation for later analysis of the commenting network. The research questions which we intend to answer in this paper are put forward after the model.

Blogs and Blogging communities

The blog, as popularized by such pioneer blogging services as Blogger (http://www.blogger.com/) and LiveJournal (http://www.livejournal.com/) started in 1999, has become one of the primary social media on the Web. Essentially, a blog contains the owner's posts released to the public. Usually arranged reverse-chronologically, these posts may document the blogger's daily activities or express his or her opinions, among other motivations of blogging (Nardi et al., 2004). Blogs are more like personal websites than private diaries, because bloggers intrinsically seek publicity on the Web by attracting and retaining readers, who are either a small circle of family and friends, or a large audience sharing something in common (Kumar et al., 2004).
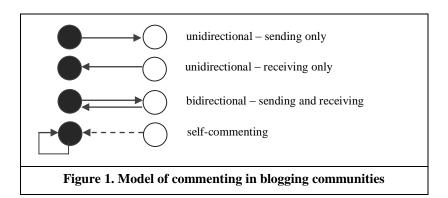
Blog readers are often bloggers themselves (Ali-Hasan and Adamic, 2007). Their attention paid to a specific blog can be traced in several ways. On the one hand, readers can include citation links in their own posts that reference this blog or leave comments on its posts. These are direct means of linking the blogger with readers and involve textual critique of the blog or blog posts. On the other hand, however, there are link-only means, i.e. blogroll and trackback. Readers can add this blog to their blogrolls which are lists of recommended blogs or websites, or send trackbacks to this blog to let the blogger know who have read his or her posts and written relevant posts (Blood, 2004; Furukawa et al., 2007).

However, none of the above ways came with the original form of the blog that emphasized on content instead of links (Blood, 2004). The developments of blogging software during the past decade introduced these links and made linking more convenient than ever, which had greater social significance than a technological impact (Cayzer, 2004). With the increasing connectedness between bloggers, they cluster in the blogspace to shape various virtual communities underlain by social networks, also known as "blogosphere" (Herring et al., 2005). These blogging communities may be characterized with locations, ages, genders, occupations, themes and so on. Bloggers are generally conscious of their membership to communities, probably because their interpersonal relationships in the form of any of the four social links are publicly observable (Furukawa et al., 2007). Within the blogging communities, different bloggers assume different roles, e.g. leader, gatekeeper, or observer, which determines the community structures. And this in turn determines how information diffuses among the bloggers (Marlow, 2004).

Model of Commenting and Research Questions

A considerable number of empirical studies have been carried out to reveal the structural characteristics of blogging communities, such as Adamic and Glance (2005), Herring et al. (2005), Furukawa et al. (2007), Ali-Hasan and Adamic (2007), just to name a few. In spite of the variety of the blogging communities examined, these studies relied heavily on social network analysis and were mainly interested in blog networks that focused on blogroll and citation links. Network analysis is a method well suited for blog research thanks to the visibility and availability of the social relationships between bloggers (Marlow, 2004). Neglecting commenting networks, nevertheless, is the major limitation or insufficiency of the existing studies.

We raise this problem based on the fact that commenting is inherently integrated with blogging and vital to the interactive nature of blogs (Efimova and de Moor, 2005). Comments are attached to the bloggers' posts, link to the commenters, and reflect their thoughts in text, i.e. written information. In contrast, citations and trackbacks are not as straightforward as comments. Readers of current blogs need to be redirected to the other blogs, where the citations point to or the trackbacks come from, to see all the content of the conversations between the bloggers. The blogroll, even more differently, is about managing contacts, without any textual interpretation of relationships. Comparatively speaking, comments are the best indications for the status of blogs and the most detectable information exchange in the blogging communities (Mishne and Glance, 2006).

Figure 1. Model of commenting in blogging communities

Given the importance of comments, we created the following commenting model applicable to most blogging communities (Figure 1). This model explains commenting relationships rather than comments themselves. A commenting relationship between two bloggers is established when either side leaves the first comment. Subsequent comments only add to the weight of the relationship. To one of the bloggers, the commenting relationship can be unidirectional (exclusively sending comments to or receiving comments from the other) or bidirectional (both

sending and receiving). The comments in opposite directions involved in bidirectional commenting need not respond to each other. But self-commenting takes into account the situations that bloggers attach comments to their own posts in response to, sometimes immediately after, the comments by other bloggers.

One fundamental implication of the commenting relationship is that the blogger at the receiving end is the original information disseminator. Comments are the reactions of the blogger at the other end. That is to say, we can infer that the main information flow in the community is in the opposite direction of commenting relationships. In our definition, the commenting network of a blogging community excludes those isolated bloggers who never send or receive comments, and consists of the remaining bloggers and the commenting relationships between all the connected pairs. In order to fill the research gap of understanding how information flows in the blogging community from the perspective of commenting, we made social network analysis of the comment data from a typical blogging service. The two general research questions addressed at current stage of research are:

1) What are the characteristics of the overall commenting pattern of the blogging community? Is it random or regular?

2) What are the characteristics of different bloggers' behavior of sending and receiving comments?

Methods

The blogging service providing data for our study is CSDN blog (http://blog.csdn.net/), an influential platform for IT engineers and computer scientists in China to exchange their expertise, experience, and ideas. Started in early 2004, it has brought up China's largest blogging community of computer science and technology. With simple functionalities, CSDN blog drives the users to concentrate on the two elementary activities – blogging and commenting, which makes it an ideal candidate for studying commenting network. Thanks to a mature anti-spam system, CSDN bloggers are seldom infected by spam comments of any kind.

The original dataset contains all the comments submitted between Jan 1 and Dec 31, 2007. As exported from CSDN log database, these comments add up to 90,175 and are distributed over 33,568 posts that belong to 11,463 bloggers. After a quick scan of who submitted these comments, we filtered out 31,547 comments left by anonymous readers whose identities were not recognizable. The remaining dataset was further downsized by sampling for easier manipulation. In that the nodes of commenting network are bloggers, sampling bloggers should better guarantee the similarity between the sample network and the original than sampling comments or posts. The result of our systematic sampling with an interval of 4 was 2,452 bloggers, and the dataset reduced to 15,615 comments which they received both from other bloggers and themselves.

Considering the difficulties in identifying the implicit and subtle relationships established in self-commenting, we decided to eliminate the 2,934 self-comments from the sampled dataset for our current study. The resulting dataset included 12,681 comments by others. It was then studied with whole network analysis by using Ucinet, a popular application for analyzing social networks. Whole network analysis is one of the two major branches of social network analysis (SNA), and the other is ego network analysis. Investigating actors as nodes and their ties as edges, SNA has been widely used by information scientists (Evelien and Ronald, 2002; Kretschmer, 2004; Björneborn, 2004; Naoki, 2007), as well as in research on blogs (Herring et al., 2005; Java et al., 2006; Takama et al., 2006).

Different from ego network analysis that focuses on individual nodes, whole network analysis treats the network as a whole and emphasizes the collective characteristics of all the nodes. One of its major variables is centralization which is the dispersion of centrality. There are *indegree centralization*, *outdegree centralization* and *betweenness centralization*. A network with high degree centralization is likely to have a few nodes with many ties and many nodes with few ties, and high betweenness centralization means a few nodes working as bridges who control how information flows in the network. Other variables of whole network analysis calculated in this research include *density* – the proportion of ties in the network relative to the total number of ties possible, *average path length* – the average of the distances of all pairs of nodes, and *clustering coefficient* – the likelihood that two neighbors of a node are neighbors themselves. Density indicates the intensity of connections, while the other two can be used to determine whether small-world phenomenon exists in a social network. If a network has a very short path length and a large clustering coefficient, it is a small-world network (Watts, 1999).

Results and Analysis

In CSDN commenting network, nodes are the bloggers who have sent or received at least one of the 12,681 comments, and edges are their commenting relationships. In practice, we removed the weight of all the commenting relationships since the absolute majority (88.6%) of them carry just one comment. The simplified network is a directed and unweighted one. The variables of whole network analysis mentioned above have been calculated for this network, as in Table 1 followed by detailed explanations of each one.

Table 1. Whole network characteristics of CSDN commenting network								
Variable	Nodes	Edges	Density	Indegree Centralization	Outdegree Centralization	Betweenness Centralization	Average Length	Clustering Coefficient
Value	**8,480**	**9,016**	**.0001**	**.0767**	**.0028**	**.0009**	**4.638**	**.038**

There are 8,480 nodes and 9,016 edges in CSDN commenting network, meaning that on average a blogger only has 2.13 neighbors connected to him or her in either way, by receiving or sending comments. And the network density (.0001) is correspondingly low. The values of indegree centralization (.0767) and outdegree centralization (.0028) suggest detectable aggregation of edges at certain nodes which are in the minority. That's to say, a few bloggers have a lot of neighbors while many bloggers have very few. This network, therefore, is rather sparse on the whole, but some parts look denser. At the same time, the higher indegree centralization (than outdegree) tells us that the aggregation of incoming ties is even more evident than that of outgoing ones. This means that, in the blogging community, while the differences between the numbers of other bloggers whom the bloggers have ever commented are not that distinct, there are a few popular bloggers who draw unusual attention, as implied by their significantly more commenters than others.

The betweenness centralization of the network is as low as .0009, which indicates the lack of bridge nodes, i.e. intermediary bloggers with bidirectional relationships. Thanks to the great visibility of peer bloggers and their posts, bloggers are able to comment straightforwardly, rather than relying on others to convey their opinions in relay comments. The small world phenomenon is common in social networks and is also found here. This is due to the short average path length (4.638) and large clustering coefficient (.038) (these two variables of the random network with the same parameters are 6.222 and .0004).
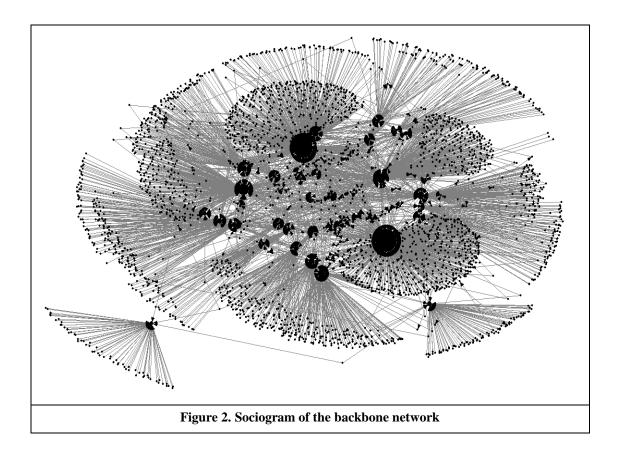
So far we have identified that the most noticeable nodes in CSDN commenting network are characteristic of many incoming ties. They are in very small quantity, but directly reachable from a considerable portion of the entire network. In the interest of finding out their impact on the network structure, we chose 27 bloggers with very large indegree centralities (i.e. more than 40 incoming ties) and all their neighbors to construct a "backbone network". They are respectively called "central bloggers" and "ordinary bloggers". The threshold of 40 was selected intentionally to balance the size and the representativeness of the resulting sub-network. Whole network analysis has also been conducted for the backbone network (Table 2).

Table 2. Whole network characteristics of the backbone network								
Variable	Nodes	Edges	Density	Indegree Centralization	Outdegree Centralization	Betweenness Centralization	Average Length	Clustering Coefficient
Value	**2,928**	**4,007**	**.0004**	**.2224**	**.0047**	**.0003**	**4.241**	**.153**

Central bloggers and the ordinary bloggers surrounding them amount to 34.5% of all the nodes (2,928/8,480), but account for 44.4% of all the edges (4,007/9,016). It can be inferred that the denser parts in the original network are mostly occupied by them because the density (.0004) is slightly higher after removing other noise bloggers. With a great increase in indegree centralization, the incoming ties are more aggregative around the central bloggers, meaning that they become even more noticeable. Relatively, the changes in outdegree and betweenness

centralizations are not so obvious. However, the backbone network shows stronger small world traits as reflected by the even shorter average length and larger clustering coefficient.

To complement these results of whole network analysis, we created a sociogram for the backbone network (Figure 2). With node size indicating the indegree centrality, the huge nodes are central bloggers encircled by a number of small nodes which are ordinary bloggers. This visualization clearly depicts the commenting patterns in the backbone network: ordinary bloggers directly commenting on central bloggers is dominating while the opposite is rarely seen; ordinary bloggers generally keep commenting relationship with only one central blogger, but not vice versa; ordinary bloggers infrequently comment on each other, even if they are interested in the same central blogger.

Figure 2. Sociogram of the backbone network

A ubiquitous structure that can be easily detected in Figure 2 is the star topology. It represents the star and fan relations between a central blogger and his or her subordinate ordinary bloggers. Below we only include the star topology based on the central blogger with the most commenting relationships (Figure 3), because those based on other central bloggers are very similar to it and just smaller in size. Star bloggers broadcast interesting content to fan bloggers though blog posts and obtains their feedbacks in the form of comments. In face of the latter's enthusiasm in them, the former however seldom express their thoughts or feelings about it by initiating outgoing comments. With most ordinary bloggers unambiguously attaching themselves to only one central blogger, different clusters with distinguishable boundaries emerge, each of which is led by a central blogger. But they are not true "cliques" because the members generally neglect each other.
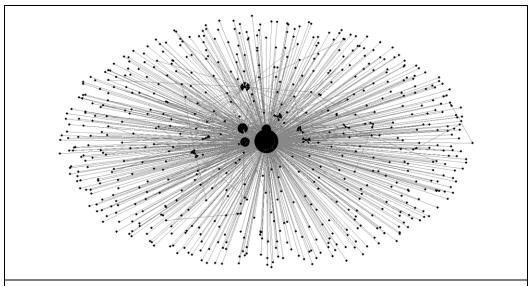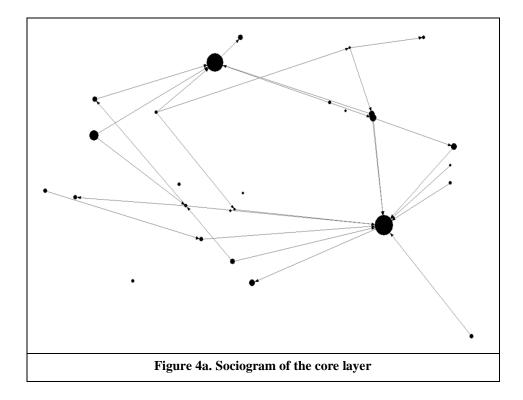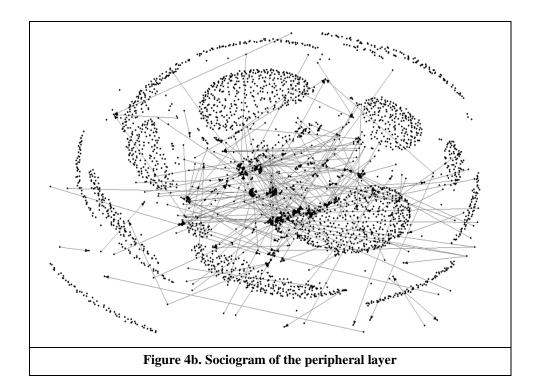
Figure 3. Star topology of the central bloggers with the most commenting relationships

In order to expose more patterns in the above sociogram of the backbone network, we further divided it into two layers, the core layer composed of 27 central bloggers (Figure 4a) and the peripheral layer 2,901 ordinary bloggers (Figure 4b). These two layers respectively have 52 and 372 internal links that never go beyond the corresponding layers. The density of the central layer (.0741) is dramatically higher than that of the peripheral ($4.4e^{-5}$). These statistics imply much closer between-cluster connection between central bloggers than between ordinary bloggers. The former are responsible for exchanging information with the outside and then serve as the exclusive information sources of the latter within the clusters. As we can see in Figure 4b, most ordinary bloggers become isolated due to the absence of their central bloggers from the peripheral layer.

Figure 4a. Sociogram of the core layer

Figure 4b. Sociogram of the peripheral layer

Conclusions and Future Work

Commenting is a natural way that bloggers and their readers communicate with each other. It lays the ground for an effective system which enables them to discuss common topics and develop online social connections. Aiming at revealing the structural characteristics of a blogging community, this stage of the study stage examines the commenting network that embodies all the directed commenting relationships. Based on the comment data from CSDN blog, our social network analysis produces a series of findings related to the two research questions.

For the first research question, we detect random commenting patterns from CSDN commenting network through whole network analysis. The commenting relationships distribute highly unevenly, with a handful of bloggers possessing a lot of relationships (either incoming or outgoing) and the rest possessing very few relationships. That's to say, there are not many typical information providers (who receive comments from a lot of bloggers) or information consumers (who send comments to a lot of bloggers) in CSDN blogging community. Since bridge nodes are rare, CSDN blogging community also lacks intermediary bloggers who act as information hubs. Information generally flows flatly, rather than hierarchically. The entire CSDN commenting network is a small-world network.

For the second research question, we identify a small number of central bloggers and their surrounding ordinary bloggers while other bloggers' commenting behavior is difficult to characterize. Central bloggers are the typical information providers by attracting considerable readers to their blogs, reflected in the large numbers of incoming relationships they have. However ordinary bloggers are not typical information consumers because they usually comment exclusively on their central bloggers. An ordinary blogger is loyal to only one central blogger once recognizing him or her as their reliable information source, rather than observing multiple. Each central blogger and his or her associated ordinary bloggers constitute a cluster that assumes the star topology with the former in the middle disseminating information to the latter. Ordinary bloggers tend to be neglected by the counterparts in their own clusters, let alone those in other clusters. As a result, information exchange between clusters is often accomplished by central bloggers.

The dominant one-way commenting from ordinary bloggers to central bloggers in CSDN blogging community has great practical significance to the development of blogging services. Such knowledge can be utilized in blogger rating and discovery of hot topics, and especially, central blogs are the perfect places to put advertisements. During our next stage of research, we will focus on these central bloggers and explore how they become prominent. The reasons they gain prominence may be found in bloggers' personal background or the interface and functionality

design of the blogging service. For example, every blogger has a profile introducing their interests and expertise, and CSDN blog categorizes them according to domain. Hotter domains such as Java programming and Web 2.0 will naturally catch more attention. Also, CSDN blog promotes a certain number of blog posts by pre-defined "experts" on the homepage every day. Readers should be willing to take advantage of such convenience, thus following the recommendations. Taking such factors into account, we can extract and analyze related semantic information from CSDN website to see if they contribute to the popularity of central bloggers. Whether they are beneficial to the information diffusion in the blogging community is yet to be determined by the analysis of the comment content that can be positive or negative.

Acknowledgements

References

Adamic, L. A., and Glance, N. "The Political Blogosphere and the 2004 U.S. Election: Divided They Blog," in *Proceedings of the 3rd International Workshop on Link Discovery, a workshop at the International Conference on Knowledge Discovery and Data Mining*, 2005.

Ali-Hasan, N., and Adamic, L. A. "Expressing Social Relationships on the Blog through Links and Comments," in *Proceedings of the 1st International Conference on Weblogs and Social Media*, 2007.

Björneborn, L. *Small-World Link Structures Across an Academic Web Space: A Library and Information Science Approach*. PhD Thesis, Department of Information Studies, Copenhagen, Royal School of Library and Information Science, 2004.

Blood, R. "How Blogging Software Reshapes the Online Community," *Communication of the ACM* (47:12), 2004, pp. 53-55.

Borgatti, S. P., Everett, M. G., and Freeman, L. C. *Ucinet for Windows: Software for Social Network Analysis*. Analytic Technologies, Harvard, MA, 2002.

Cayzer, S. "Semantic Blogging and Decentralized Knowledge Management," *Communication of the ACM* (47:12), 2004, pp. 47-52.

Efimova, L., and de Moor, A. "Beyond Personal Webpublishing: An exploratory Study of Conversational Blogging Practices," in *Proceedings of the 38th Hawaii International Conference on System Sciences*, 2005.

Evelien, O., and Ronald, R. "Social Network Analysis: A Powerful Strategy, Also for the Information Sciences," *Journal of Information Science* (28:6), 2002, pp. 441-453.

Furukawa, T., Matsuo, Y., Ohmukai, I., Uchiyama, K., and Ishiuka, M. "Analyzing Reading Behavior by Blog Mining," in *Proceedings of the 22nd Conference on Artificial Intelligence*, 2007.

Herring, S. C., Kouper, I., Paolillo, J. C., Scheidt, L. A., Tyworth, M., Welsch, P., Wright, E., and Yu, N. "Conversations in the Blogosphere: An Analysis 'From the Bottom Up'," in *Proceedings of the 38th Hawaii International Conference on System Sciences*, 2005.

Java, A., Kolari, P., Finin, T., and Oates, T. "Modeling the Spread of Influence on the Blogosphere," in *Proceedings of the 15th International World Wide Web Conference*, 2006.

Kretschmer, H. "Author Productivity and Geodesic Distance in Bibliographic Co-authorship Networks and Visibility on the Web," *Scientometrics* (60:3), 2004, pp. 409-420.

Kumar, R., Novak, J., Raghavan, P., and Tomkins, A. "Structure and Evolution of Blogspace," *Communication of the ACM* (47:12), 2004, pp. 35-39.

Marlow, C. "Audience, Structure and Authority in the Weblog Community," in *Proceedings of the Communication Association Conference*, 2004.

Mishne, G., and Glance, N. "Leave a Reply: An Analysis of Weblog Comments," In *Proceedings of the 3rd Annual Workshop on the Weblogging Ecosystem: Aggregation, Analysis and Dynamics, a workshop at the 15th International World Wide Web Conference*, 2006.

Naoki, S., Yuya, K., and Katsumori, M. "Topological Analysis of Citation Networks to Discover the Future Core Articles," *Journal of the American Society for Information Science and Technology*, (58:6), 2007, pp. 872-882.

Nardi, B., Schiano, D., Gumbrecht, M., and Swartz L. "Why We Blog," *Communications of the ACM*, (47:12), 2004, pp. 41-46.

Takama, Y., Matsumura, A., and Kajinami, T. "Visualization of News Distribution in Blog Space," in *Proceedings of the IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology*, 2006.

Watts, D. J. *Small Worlds: The Dynamics of Networks between Order and Randomness*. Princeton University Press, Princeton, 1999.