Analyzing Reading Behavior by Blog Mining

Tadanobu Furukawa
Mitsuru Ishizuka
University of Tokyo
7-3-1 Hongo, Bunkyo-ku
Tokyo 135-8656, Japan

Yutaka Matsuo
AIST
1-18-13 Sotokanda
Chiyoda-ku, Tokyo
101-0021, Japan

Ikki Ohmukai
National Institute of Informatics
2-1-2 Hitotsubashi
Chiyoda-ku, Tokyo
101-8430, Japan

Koki Uchiyama
Hotto Link Inc.
1-6-1 Otemachi
Chiyoda-ku, Tokyo
100-0004, Japan

Abstract

This paper presents a study of the various aspects of blog reading behavior. The analyzed data are obtained from a Japanese weblog hosting service, *Doblog*. Four kinds of social networks are generated and analyzed: citation, comment, trackback, and blogroll networks. In addition, the user log data are used to identify readership relations among bloggers. After analysis of more than 50,000 users for about two years, we reveal some interactions between social relations and readership relations. We first show that bloggers read other weblogs on a regular basis (50% of weblogs that are read at least three times are read every five times a user logs in). We call this relation a *regular reading relation (RR relation)*. Then, prediction of RR relations is done using features from the four kinds of social networks. Lastly, information diffusion on RR relations is analyzed and characterized. Results of this study show that the blogs in RR relations have an important role in bloggers' activities. We find the features which have a correlation with RR relations.

Introduction

Web logs (blogs, or weblogs) constitute a prominent social medium on the internet that enables users to publish individual experiences and opinions easily. Analyzing these data helps to understand the user's behavioral pattern on the blogosphere, and it supports the development of web services such as recommendation of blogs. Numerous studies have examined weblogs, especially addressing their social aspects: Bloggers read other blogs and leave *comments* and send *trackbacks* as they update their own blogs. Users might mention other blogs in their postings, and express their suggested contacts in *blogrolls* (a sidebar within a particular blog listing the other blogs the blogger frequents). These activities present an interwoven record of multiple relationships among blogs (sometimes called a *multiplex graph* in sociology), which is an interesting source of information to characterize user behavior, community structure, and information diffusion.

To date, various studies have specifically examined social network aspects of weblog authors (Marlow 2004; Adamic & Glance 2005). Among several studies that analyzed social networks on the blogosphere, some have used *citation* (mention of urls in the posting) to show relations among blogs; others have used blogrolls or trackbacks as evidence of relations. At least one study has surveyed comment relations among bloggers (Lento *et al.* 2006).

Another important relation exists among bloggers: readership relations. Readership relations are not observable through publicly available data, but they are an important source of information because bloggers read other blogs and write their own blogs. The importance of readership relations is described in the literature (such as (Efimova, Hendrick, & Anjewierden 2005; Nardi, Schiano, & Gumbrecht 2004)). A recent study has analyzed the structural properties of weblog readership networks (Marlow 2006).

Although those studies reveal various interesting findings related to the social networks of weblogs, no comprehensive study of them has examined all social and behavioral relations simultaneously. A comparison among different relations provides the general overview of each relation and its associated pattern of interaction. For this study, we analyze multiple social networks of weblogs: citation, comment, trackback, and blogroll networks. Subsequently, the user log data (which include which blogs a user browses and when) are used to identify readership relations among bloggers. Therefore, five types of social and behavioral networks are analyzed in this paper. We use the database of a blog-hosting service in Japan called *Doblog*[1].

Analyses of 1.5 million entries made by more than 50,000 users for about two years reveal interesting interactions involving social relations and readerships. This paper describes salient aspects of the following analysis.

- We first illustrate the four kinds of social networks and characterize them.
- Bloggers read some blogs on a regular basis. We can discern these behaviors quantitatively: 50% of weblogs that are read at least three times are read every five times a user logs in. We call this relation a *regular reading relation (RR relation)*.
- *Link prediction* of RR relations is done using features from the four kinds of social networks. Some attributes

Copyright © 2007, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

[1]Doblog (`http://www.doblog.com/`), provided by NTT Data Corp. and Hotto Link, Inc.

(such as graph distance) and some networks (blogrolls and citations) are demonstrably useful for predicting RR relations.

- Information diffusion through RR relations is analyzed and characterized. Some information is likely to be conveyed through RR relations. Generally, information propagates in a shorter time and with higher probability through RR relations than through non-RR relations.

Our findings provide an overview of social relations and reading behavior. These results support those of existing studies of social network analyses of the blogosphere.

This paper is organized as follows: first, we describe related studies. Next, we explains the definition of social relations among weblogs, and also define the RR relation and characterize it with social relations. We then produce a model to infer the existence of RR relations as a link prediction problem, and analyze information diffusion through RR relations and show the effect of RR relations. Finally, after a discussion of analytical limitations, we conclude the paper.

Related Works

Many studies have specifically undertaken analysis of the blogosphere as a social medium: trend detection, network analysis, user profiling, and splog (spam blog) detection. We introduce several works that are closely related to ours.

Several studies have analyzed social networks that exist in the blogosphere: L. Adamic and N. Glance study the link patterns (citations and blogrolls) and discussion topics of political bloggers (Adamic & Glance 2005). This study detected differences in the behaviors of politically liberal and conservative blogs, with conservative blogs linking to each other more frequently. Lento et al. conduct data analyses regarding the Wallop system and compares users who remain active to those who do not (Lento *et al.* 2006). Similarly to our work, the use of data from the hosting service enables them to detect social ties such as comment relations and invitation relations, which are usually impossible to obtain. G. Mishne and N. Glance analyze blog comments (Mishne & Glance 2006). Those studies extract some relations among blogs and produce a social network for analysis.

Social networks are used for several applications such as blog/entry ranking and community detection: E. Adar et al. proposes a ranking algorithm called iRank, which is based on implicit routes of information transmission as well as explicit links (Adar *et al.* 2004). For community detection, Y. Lin et al. seek interesting aspects of social relations (Lin *et al.* 2006). They develop a computational model for mutual awareness that incorporates specific action types including commenting and changing blogrolls. The mutual awareness feature is used for community extraction. The social network also provides information for blog classification; P. Kolari et al. investigates splogs (Kolari, Java, & Finin 2006) and concludes that although ordinal blogs show a power-law distribution when counting citations, splogs deviate from that pattern.

Several studies have investigated weblog relationships and real-world relationships: J. Cummings et al. discusses online and offline social interactions (Cummings, Butler, & Kraut 2002). Computer-mediated communication (in particular, e-mail) is less valuable for building and sustaining close social relationships than face-to-face contact and telephone conversations. R. Kumar et al. investigates profiles of more than one million livejournal.com bloggers in 2004, and analyzes users' demographic and geographic characteristics (Kumar *et al.* 2004). More recently, Ali-Hasan and L. Adamic find interesting characteristics of bloggers' online and real-life relationships (ALi-Hasan & Adamic 2007). They investigate three blog communities using an online survey, which reveals that few blogging interactions reflect close offline relationships; furthermore, many online relationships were formed through blogging.

Nardi et al. conducted audiotaped ethnographic interviews with 23 bloggers, with analysis of their blog posts (Nardi, Schiano, & Gumbrecht 2004). The motivations of blogging are enumerated. They provide good insights that support the background of our research: bloggers write blogs to (1) update others on activities and whereabouts, (2) express opinions to influence others, (3) seek others' opinions and feedback, (4) "think by writing", and (5) release emotional tension. Nardi et al. remark that

> Our research leads us to speculate that blogging is as much about reading as writing, and as much about listening as talking. We specifically examined the production of blogs, but future research will address blog readers and to assess the relations between blog writers and blog readers precisely.

In the following section, we investigate relations between blog writers and readers from a social network perspective.

Four Social Networks among Weblogs

In this section, we define four kinds of relations among blogs and depict social networks defined by these relations.

We consider four types of relations between two blogs:

Citation We define that there is a *citation* relation from A to B if an entry of blog A includes a hyperlink to blog B.

Blogroll We assert a *blogroll* relation from A to B if a blogroll (a list of weblogs in the front page) of A includes blog B.

Comment A *comment* relation from A to B pertains if the blogger of blog A comments on blog B.

Trackback A *trackback* relation from A to B exists if an entry of blog B contains a back-reference by the trackback function to blog A.

These are mentioned in (Marlow 2004) and other literature. We call these four relations *social relations* because the relations are publicly observable and therefore involve some degree of social consciousness and manifestation.

All four relations are directed. We can use some thresholds (minimum number of times) to define citation, comment, and trackback relations; however, in this paper we set the threshold as 1. Therefore, if at least one citation, comment, or trackback relation is apparent, we regard the two blogs as having a correspondent relationship. Throughout the paper, use blog A and blogger (or a user) A interchangeably; usually a blog is owned and maintained by a blogger.

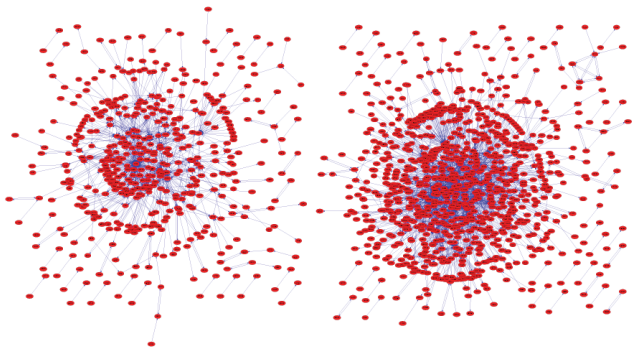

Figure 1: Citation (left)/Blogroll (right) social network.

Table 1: Network indices on social networks.

Relation	n (GCC*)	k	L	C
citation	472 (404)	981	4.01	0.179
blogroll	918 (833)	2312	4.13	0.162
comment	947 (875)	3301	3.63	0.163
trackback	420 (364)	697	4.02	0.166

*Number of nodes in the giant connected component (GCC).

We can illustrate the networks consisting of either relation. Figure 1 depicts social networks among users that have been detected by citations/blogrolls. (The network of trackback/comment has a similar structure.) For illustration, we use data of the 2,648 bloggers who are selected randomly. The entire dataset consists of 1,540,077 entries by 52,976 users from October 2003 to June 2005. All four networks have a dense core in the middle, with isolated groups in the periphery.

The network indices are shown in Table 1. In the table, n is the number of nodes that are involved in each relationship among 2,648 bloggers, and k is the number of edges. Following (Watts 2003), L is the characteristic path length, which is defined as the average distance between two nodes on the network, and C is the clustering coefficient, defined as the chance that two friends are themselves friends. We can see that C is almost the same for the four networks, and L is slightly small for the comment network.

Table 2 shows the QAP correlation of the social networks (Wasserman & Faust 1994): if the QAP correlation is higher (up to 1.0), then the two networks are more similar. The comment network and the blogroll network are similar; in addition, the citation network and the comment network are similar. Because blogroll and comment relations are sometimes created among friends and acquaintances, the networks have high correlations each other.

Readership Network

In this section, we define the readership relations and analyze the user log.

Behavioral Relation

We define *behavioral relations* (in contrast to social relations) as relations that are observable only from the user log. Behavioral relations are not recognizable from the public data. They include the readership relations between two weblogs, direct messaging, invitation, and so on. In this paper, we specifically examine the readership relation because reading other blogs is the most dominant identifiable activity of bloggers other than writing their own blogs; it well reflects users' interests and information-seeking behavior.

Examining the user log, we can recognize which blog a user browses and when; various criteria can be used to define the readership. For instance, we can define a readership relation as a user browsing another blog at least once.

Table 2: QAP correlations.

	citation	blogroll	comment	trackback
citation	–	–	–	–
blogroll	0.283	–	–	–
comment	0.364	0.432	–	–
trackback	0.338	0.182	0.302	–
RR network	0.194	0.350	0.347	0.149

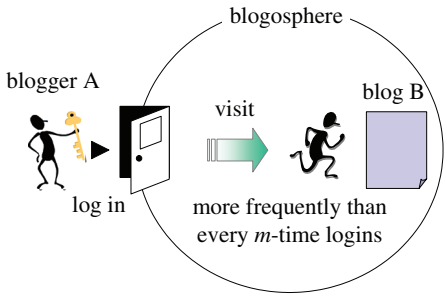

Figure 2: Definition of Regularly Reading from A to B.

Regularly Reading Weblogs

We analyze the frequency of a bloggers reading other weblogs. Some users write weblogs everyday, but others write less. Therefore, we normalize the time interval by the average interval of login to Doblog. In this way, we can see how often a blogger reads other blogs when logged in.

Figure 3 shows the normalized interval of users reading other weblogs. The value of 1.0 indicates that the interval is equivalent to the average interval of the user login; in other words, the user always reads the weblog when logged in. About 50% of the blogs (that are read at least three times) are read by a user every five times that the user logs in.

We define a weblog *regularly reading* relation as follows:

Regularly Reading (RR) We say that blogger A has an RR relation with blog B if blogger A reads blog B more frequently than every m times the blogger logs in (Fig. 2).

In this paper we set m=5. It corresponds to about the half of blogs that are read more than three times [2]. We designate the blogs that show an RR relation with blogger A as RR blogs of A.

[2]In our analyses, the trend does not change if we use a different value for m.

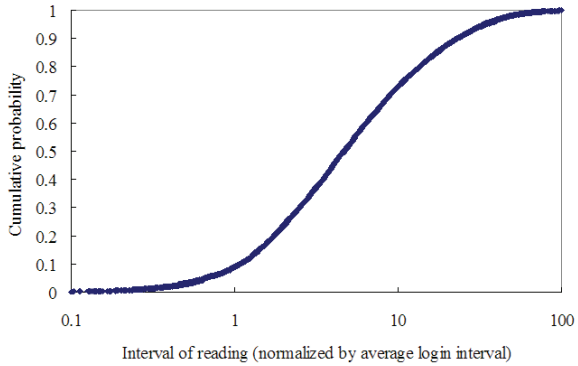

Figure 3: Interval of reading.

The RR relation reveals intimate communication; bloggers often share interests and exchange comments or trackbacks. These RR relations sometimes hold for very popular blogs, typically known as "A-list" blogs. On the other hand, among all of the 5829 RR relations, 3757 relations (64.5%) do not correspond to any of the four relations. Therefore, it is difficult to find RR relations using direct social links.

Link Prediction of Regularly Reading

What causes users to regularly read other blogs? We can build a recommendation of blogs for each user if we can create a model and predict whether a user will regularly read a blog or not. In this section, we describe the algorithm and results of the prediction using a machine-learning approach.

Liben-Nowell and Kleinberg propose a link prediction problem (Liben-Nowell & Kleinberg 2003): Given the network, the task is to predict whether a link exists or not (or a link will be generated or not). If we can predict whether RR relations hold between two weblogs from their social relations, then we can generate a social network of RR relations using publicly available data.

We can generate a number of attributes between two users based on their network topology. Although several studies have examined link prediction (Getoor & Diehl 2005), Liben-Nowell and Kleinberg test various attributes; we use those that are reported to be effective:

- Adamic/Adar: $\sum_{z \in \Gamma(x) \cap \Gamma(y)} 1/\log|\Gamma(z)|$
- graph distance: length of shortest path between x and y
- common neighbors: $|\Gamma(x) \cap \Gamma(y)|$
- Jaccard's coefficient: $|\Gamma(x) \cap \Gamma(y)|/|\Gamma(x) \cup \Gamma(y)|$
- preferential attachment: $|\Gamma(x)| \cdot |\Gamma(y)|$

We denote a set of neighbors of x in the network as $\Gamma(x)$. We prepare these attributes for four social networks.

We construct a decision tree using C4.5 (Quinlan 1993) using these attributes. Our dataset consists of 6000 pairs of RR blogs as positive examples. We also use the same number of negative examples, for which two weblogs do not show an RR relation. The performance is measured using three-time leave-one-out cross-validation. It is noteworthy that we use no information of a direct relation between blog A and blog B when predicting the RR relation from A to B because we seek to predict a possible RR relation when no recognizable relation exists between A and B.

Table 3: Performance of link prediction for RR relation.

Features	Precision	Recall	F-measure
all features	0.607	0.569	0.588
Adamic/Adar	0.580	0.705	0.637
graph distance	0.557	0.686	0.615
common neighbors	0.609	0.278	0.381
Jaccard's coefficient	0.617	0.252	0.357
preferential attachment	0.572	0.760	0.653
citation	0.585	0.257	0.358
blogroll	0.569	0.597	0.583
comment	0.595	0.589	0.592
trackback	0.569	0.239	0.336

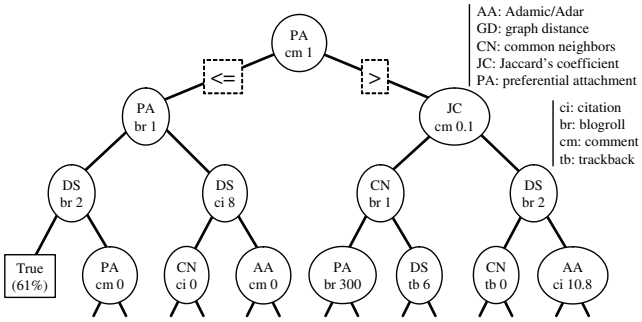

Figure 4: Decision tree of RR blogs.

The results are shown respectively in Table 3 for predicting RR blog. If we use all features, the precision is about 0.6 and the recall is 0.57, yielding the F-measure of 0.6. These figures are not high because there are equal numbers of positive and negative examples. However, some interesting findings are available to illustrate the performance by selection of attributes.

Among the five types of features, Jaccard's coefficient shows the highest precision: if weblogs have similar neighbors to a user, they have a high probability of being read regularly by a user. The implications gained using this measure correspond to the social behavior of ordinary users, as described in (Herring *et al.* 2005) and (Nardi, Schiano, & Gumbrecht 2004): The ordinary blogs (compared to A-list blogs) are densely interconnected within the community and are linked sparsely to other blogs. On the other hand, preferential attachment brings the highest recall. This results indicates that blogs with many incoming relations have a large audience, as shown partly in (Marlow 2004). Therefore, if we want to make a precise recommendation, we can show a user blogs with a proximity of social relations. If a user seeks an exhaustive list of potentially interesting blogs, we can present blogs with numerous incoming links in the community.

Among the four social networks, the comment network shows the highest precision/recall; comments are the explicit signal of user interests, so the comment relation works

well in a transitive manner when detecting a user's interest. As already shown in Table 2, the blogroll network and the comment network have a higher correlation than other pairs. Although comment relations are usually unidentifiable, blogroll relations might serve as a good proxy of comment relations. The trackback network functions worst when predicting RR relations, perhaps because users often do not use trackbacks (as apparent from the sparse nature of the trackback network), and also because of the prevalence of trackback spam messages.

Figure 4 shows the obtained decision tree using all features to classify RR blog. In those decision trees, the highly influential relation occupies the upper position as a node: We can see that the preferential attachment on blogrolls and comments has a high impact. In addition, other features on blogrolls are also important. Generally, blogrolls and comments have high predictive power for RR blogs.

Information Diffusion on Regular Reading Channels

Several studies have examined how information diffuses through online social networks (such as (Leskovec, Adamic, & Huberman 2005)). Adar et al. focuses on urls mentioned in the posting, and uses a machine-learning approach to detect implicit relations between two blogs (Adar & Adamic 2005). In our study, we have useful relations for detecting diffusion: readership relations. Using those relations, we then investigate information diffusion on the readership network. Concretely, inspired by Adar's method, the diffusion of urls on RR relations is analyzed in this section.

Our analysis is twofold. We seek to answer two questions: (i) How likely is information to diffuse between two blogs with and without RR relations? (ii) What kind of information is likely to diffuse through RR relations? In both cases, inclusion of a url in an entry is analyzed. We conjecture that the url information is propagated from blog A to blog B if blog A mentions a url and, subsequently blog B (which is in some close proximity to blog A) mentions the same url. We define the distance as two for all of the four social networks. Although this is only a rough approximation, the general trend is apparent from the analysis.

Figure 5 shows the time for a url to be propagated among two blogs that mention the same url. When the RR relation holds, in about 60% of cases, the url diffuses within 200 h. In contrast, with no RR relation, the url diffuses in less than 30% of cases, and does not reach greater than 50% diffusion in 1000 h.This result seems straightforward; we can understand that if a user is regularly reads other blogs, the information on the blog diffuses in a shorter time and with higher probability. Still, we can assess the effectiveness of RR relations quantitatively from this result.

Table 4 shows some examples of urls that are diffused more on RR relations and more on non-RR relations. In RR relations, we can find urls for entertainment web pages such as horoscopes, and some technology sites on the web. On the other hand, information propagated less on RR relations includes news articles, a web page of TV programs, and products. These are effects of mass media and the large number of advertisements propagated online.

Table 4: Examples of urls that propagate through RR/non-RR blogs.

more on non-RR
http://headlines.yahoo.co.jp/hl?a=20050201-00000111-yom-soci - A city news article in Yahoo! Japan.
http://www.nhk.or.jp/asadora/ - A web page of the well-known TV drama, updated weekly.
http://www.apple.com/jp/macmini/ - An introduction of Mac min on Apple's website.
more on RR
http://u-maker.com/38267.html - A website for predicting one's personality.
http://www.geocities.co.jp/Milkyway-Aquarius/7075/trainman1.html - A complete survey of *Train man*, a love romance story of an otaku and a beautiful lady, which actually occurred online.
http://www.seo-association.com/ - A web page of SEO competition.

It is apparent that RR relations tend to convey information that is interesting to particular users or to a particular community. Some information propagates well through RR relations, and ultimately diffuses. Depending on the type of information, social relations on the blogosphere can exert a great effect on information diffusion.

Discussion

In this study, we used log data on Doblog. This research has several limitations that are intrinsic to the data: the sample users are not representative of all blog users; we did not crawl the internet to obtain weblog data from the entire web. It is important to analyze the data integrated with publicly available data in the future. The nature of communities depends on system architectures and user characteristics. Therefore, comparative research would be promising, such as Ali-Hasan's work on US, Kuwait, and UAE blogs (ALi-Hasan & Adamic 2007).

In other studies (such as (Adar *et al.* 2004), (Kumar *et al.* 2004) and (Marlow 2004)), social relations among weblogs are analyzed through examination of blogrolls and citations. Although other relations exist, such as comments and trackbacks, the usage of two relations is a feasible approach: Trackbacks are less numerous, and have little predictive power of RR relations. Comments give useful information on predicting RR relations, but show higher correlation with blogrolls. However, it is also important to conduct analyses considering the weight of links in the future.

In previous section, we showed analysis of information diffusion on RR relations. However, we note the following possibilities: we use urls because they are easily identifiable, but information is rarely represented in the form of urls; users do not mention the url in their entries even if the information of the url is propagated; the information might propagate from other blogs over a long distance, or media

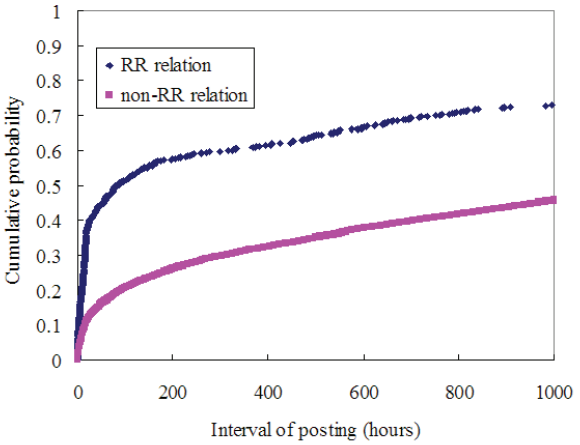

Figure 5: Interval of posting between RR blogs. We have respectively investigated 699 and 9682 pairs of RR blogs and non-RR blogs.

other than weblogs, which causes coincidental detection of information propagation.

Future research will be undertaken: (i) to propose measures that characterize types of information propagated on RR relations, and (ii) to detect key persons on the multiple social networks.

Conclusions

This study has analyzed the properties of behavioral relations of Doblog users. We first described that about 50% of the blogs (visited at least three times) are read by users every five times that a user logs in. Then we showed a predictive method of such RR relations and described the important attribute types and social relations for the prediction. Finally, we explained how RR relations contribute to information diffusion and characterized the urls that are likely to be conveyed in RR relations.

Several issues require further analysis, but we believe that we have shown a useful overview of readership relations that can be associated with social relations. Our study provides insight into multiple social networks among weblogs and supports the usage of publicly available relations, i.e., blogrolls and citations. Future studies will include a comparative study of different weblog hosting services for investigation of the generality of our findings.

References

Adamic, L., and Glance, N. 2005. The political blogosphere and the 2004 u.s. election: Divided they blog. In *LinkKDD-2005*.

Adar, E., and Adamic, L. A. 2005. Tracking information epidemics in blogspace. In *Web Intelligence 2005*.

Adar, E.; Zhang, L.; Adamic, L.; and Lukose, R. 2004. Implicit structure and the dynamics of blogspace. In *Workshop on the Weblogging Ecosystem*.

ALi-Hasan, N., and Adamic, L. 2007. Expressing social relationships on the blog through links and comments. to appear.

Cummings, J.; Butler, B.; and Kraut, R. 2002. The quality of online social relationships. *Communications of the ACM* 45(7).

Efimova, L.; Hendrick, S.; and Anjewierden, A. 2005. Finding 'the life between buildings': An approach for defining a weblog community. In *Proc. Internet Research 6.0*.

Getoor, L., and Diehl, C. P. 2005. Link mining: A survey. *SIGKDD Explorations* 2(7).

Herring, S.; Kouper, I.; Paolillo, J.; Scheidt, L.; Tyworth, M.; Welsch, P.; Wright, E.; and Yu, N. 2005. Conversations in the blogosphere: An analysis "from the bottom up". In *Proc. HICSS-38*.

Kolari, P.; Java, A.; and Finin, T. 2006. Characterizing the splogosphere. In *Proc. 3rd Annual Workshop on Weblogging Ecosystem*.

Kumar, R.; Novak, J.; Raghavan, P.; and Tomkins, A. 2004. Structure and evolution of blogspace. *Communications of the ACM* 47(12).

Lento, T.; Welser, H.; Gu, L.; and Smith, M. 2006. The ties that blog: Examining the relathionship between social ties and continued participation in the wallop weblogging system. In *3rd Annual Workshop on the Weblogging Ecosystem*.

Leskovec, J.; Adamic, L. A.; and Huberman, B. A. 2005. The dynamics of viral marketing. http://www.hpl.hp.com/research/idl/papers/viral/viral.pdf.

Liben-Nowell, D., and Kleinberg, J. 2003. The link prediction problem for social networks. In *Proc. CIKM*, 556–559.

Lin, Y.; Sundaram, H.; Chi, Y.; Tatemura, J.; and Tseng, B. 2006. Discovery of blog communities based on mutual awareness. In *WWW2006 Workshop on Weblogging Ecosystem*.

Marlow, C. 2004. Audience, structure and authority in the weblog community. In *Proc. Communication Association Conference*.

Marlow, C. 2006. Investment and attention in the weblog community. In *Proc. HyperText 2006*.

Mishne, G., and Glance, N. 2006. Leave a reply: An analysis of weblog comments.

Nardi, B.; Schiano, D.; and Gumbrecht, M. 2004. Blogging as a social activity, or would you let 900 million people read your diary? In *Proc CHI 2004*.

Quinlan, J. R. 1993. *C4.5: Programs for Machine Learning*. California: Morgan Kaufmann.

Wasserman, S., and Faust, K. 1994. *Social network analysis. Methods and Applications*. Cambridge: Cambridge University Press.

Watts, D. 2003. *Six Degrees: The Science of a Connected Age*. W. W. Norton & Company.