Internet Research 10.0 - Internet: Critical
Milwaukee, WI, USA
7-10 October 2009

Automated Discovery of Emerging Online Communities Among Blog Readers: A Case Study of a Canadian Real Estate Blog

Anatoliy Gruzd, PhD
School of Information Management
Dalhousie University, Canada
gruzd@dal.ca

Introduction

According to the latest Technorati's report (http://technorati.com/blogging/state-of-the-blogosphere), more than 184 million people worldwide have started a web blog, and collectively they attracted at least 346 million blog readers. Due to their popularities, web blogs have been the focus of many recent Internet studies. Aside from being a great publishing platform, many of these studies confirmed the fact that modern blogs with commenting-capabilities are also great places for meeting like-minded individuals and forming new social relationships (e.g., Ali-Hasan & Adamic, 2007; Dennen & Pashnyak, 2008). As a result, it is not surprising that there is also a growing interest in discovering and characterizing online communities that tend to naturally form around some web blogs. This interest is shared not only among Internet researchers studying online communities, but also among everyday Internet users seeking to join interesting conversations online. However, not all blogs are capable of fostering and sustaining a community of loyal readers. Why is that?

Until now, most of the research that tried to answer this question has been conducted using very limited case studies that relied heavily on manual content analysis (e.g., Herring et al., 2005) or surveys (e.g., Blanchard, 2004). However, manual content analysis and surveys are very expensive and time consuming. They are perfectly adequate for analyzing blogs with a small number of readers who leave a relatively small number of comments, but they are not practical for the analysis of blogs that attract hundreds or even thousands of daily readers and commentators. Furthermore, this task becomes even more daunting if we want to find out exactly what is going on in more than just one blog.

To address these problems, this paper proposes an automated approach for the discovery of social networks among blog readers just from their comments posted to a blog. This new approach is called "name network" and is an integral part a companion web-based tool called Internet Community Text Analyzer or ICTA for short (http://textanalytics.net). Working together, this new approach and the ICTA tool can automate the processes for discovering and visualizing social networks among a particular group of blog readers. Once the social network is discovered (if any), it can be used to determine to what degree the blog readers are connected to each other and if their connections are strong enough to be classified as a community.

The paper begins with a brief description of the proposed automated approach. Then as a way to validate the proposed approach, it is used to analyze a popular Canadian real estate blog.

The results suggest that the “name network” method is capable of automatically discovering a social network among blog readers that accurately represents group dynamics.

Method

Generally speaking, to discover a social network, we need to (1) identify all members of a community who will become nodes in the network and then (2) find how these nodes are connected to each other. To find all active people involved in the blog network, we have to identify everybody who posted at least one comment to the blog. This can be accomplished by looking through all comments and retrieving names and nicknames from the “From” field of each comment. As for the second step, one of the most common approaches to automatically find social connections in an online community is to identify who talks to whom in a group. In blog data, this can be done by looking for who replies to whose comments (further referred to as the “reply-to” or “chain network” approach). Unfortunately, not all blogs have a “reply-to” capability. And besides, references to previous posters are not always exclusively found in the “reply-to”-type comments. Thus, if the focus is only on “reply-to” comments, some connections are likely to be missed in the resulting network (see, for example, Gruzd and Haythornthwaite, 2008). To address these challenges, it was decided that all comments within a blog must be used to identify addressees. The method used here is a variation of the “name network” method for blog data which was originally developed to mine social networks in threaded discussions (Gruzd, 2009). Specifically, the proposed approach looks for all mentions of the names and nicknames inside the comments. Every time a commentator mentions a fellow commentator by his or her name or nickname, the “name network” method adds a connection between these two individuals in the resulting network. Since the “name network” method does not need to know whether or not a comment is a “reply-to” comment, it can be used as a stand alone procedure for the automated discovery of social networks in blogs without the “reply-to” capability, or it can also be used in conjunction with the “reply-to” approach to discover additional, previously “hidden”, connections.

While useful, the “name network” method does have its own challenges. For example, sometime it is difficult to differentiate between a word/phrase that is used as a nickname and a word/phrase that just looks like it. For example, in the following two sample blog comments, only the second instance of “go green” refers to a person.

> (1) “… keep harvesting forests without developing the technology to **go green**.”
>
> (2) “**go green** - Check out Kitco / goldmoney.com / Mish’s Global (google for exact site).”

Although not common, this problem can significantly reduce the accuracy of the resulting network. To prevent such problem from negatively affecting the results, a semi-automated approach was developed to allow researchers to examine, and if necessary, override the resulting network data manually. Future work will include adding automated procedures for word disambiguation based on additional syntactic and semantic clues to ensure that only those instances where a word/phrase actually refers to a real person are included into the resulting network.

The “name network” method has been incorporated into an online tool called Internet Community Text Analyzer (ICTA) and is available at http://textanalytics.net. Once a social network is automatically discovered using ICTA, researchers can also use ICTA’s built-in interactive network visualization feature to visualize and explore connections among group members (see Figure 1). For example, ICTA's user can find out how or why group members are connected by clicking on an edge that connects a pair of two individuals in the social network. This action will open a new window showing all comments that were used to establish the connection between the blog readers (see Figure 2).

Case Study: Evaluating the "Name Network" Method

In order to validate the proposed automated approach for the discovery of social networks, the approach is used in the analysis of a popular Canadian real estate blog, http://www.greaterfool.ca. The analysis attempts to answer a single question: whether blog readers who post comments to this blog form a virtual community or not?

The blog was started in March of 2008 and covers a wide variety of topics related to the Canadian real estate markets and the ongoing financial meltdown in the world economy. This blog was chosen for this study because it generates a large number of blog comments, thus making it an ideal candidate for automated analysis. The owner of the blog, Garth Turner, a best selling Canadian author and former Member of Parliament, posts a new entry almost every day, and each new entry in turn generates on average about 133 readers' comments. Due to the large number of daily comments, it would have been very time consuming to analyze this blog manually.

After a quick review of some of the comments posted on the blog, the initial assumption was that this blog is not likely to contain any social community to speak of. This assumption was based on two observations. First, all of the comments on the blog were posted anonymously under user-created pseudonyms. And second, many of the comments on this blog express strong disagreement and sometimes even verbal attacks against the blogger, fellow blog readers and commentators. To confirm or reject this initial assumption, ICTA was used to discover and describe the social network that might or might not exist among the readers who left comments on this blog.

The dataset for this study consisted of two separate samples: all comments posted by blog readers in October, 2008 (1526 comments) and in January, 2009 (3217 comments). Dapper.net software was used to retrieve all comments and generate RSS feeds for subsequent import into ICTA. After the data was collected and imported into ICTA, the "name network" method was used to build one social network for each of the two periods in the study. For the purpose of the study, all non-reciprocal connections were removed from the resulting networks. Non-reciprocal connections appear when one person replies to somebody's comment, but there is no follow up interaction between the two. And since non-reciprocal connections are less likely to contribute to the formation of social connections among participates (Jones, 1997), they were safely removed to ensure that connections discovered by ICTA represent only actual social connections in the blog. The two resulting networks are shown in Figure 4 and 5 correspondently.

After the social network for each of the two sample datasets was discovered using the "name network" algorithm, ICTA's interactive network visualization feature was used to characterize the discovered social networks. From a network perspective, the social network for the period during October of 2008 cannot be characterized as a community. This is because the network is not very dense suggesting that the number of existing connections is much lower than the actual number of all possible connections among blog commentators. Furthermore, there is only one central actor among all of the commentators without whom the whole network would break up into isolated nodes and a few small components. However, the network for the January 2009 period demonstrates a significant growth in the size and density of this group. This suggests that the blog has become more popular, and the comments have become more interactive. This fact can be confirmed by looking at the changes in a number of standard measures that are commonly used in social network analysis. For example, the decrease in network fragmentation (from 0.57 to 0.31) and the increase in total degree centrality (from 6.9 to 18.36) and the betweenness centrality (from 3.06 to 31.1) suggest that the commentators became more connected and more of them took a stand in the group.

The next step was to validate the social networks that were built automatically and determine if these networks are indeed accurate representations of online communities among blog readers. With the help of ICTA, a manual content analysis was conducted to verify and explain the nature of the connections between readers who left comments. The process consisted of manually reading all comments that were discovered and used by ICTA to establish these connections.

The validation process is a very important part of this study because the discovery of social networks by itself does not automatically constitute the presence of an online community. However, it is expected that the social network detected by the "name network" method is likely to represent actual social connections among blog readers and thus suggests the presence of an online community. This is because the "name network" method, as described in the previous section, relies on the use of names and nicknames to determine connections among people, and names and nicknames are "one of the few textual carriers of identity" in discussions on the web (Doherty, 2004, p. 3). Furthermore, their use is crucial to create and maintain a sense of community (Ubon, 2005) and social presence (Rourke, 2001). As Ubon (2005) put it, by addressing each other by name, participates "build and sustain a sense of belonging and commitment to the community" (p.122).

Following McMillan and Chavis' (1986) notion of "sense of community" and Jones' (1997) notion of "virtual settlement", each comment was checked for presence of indicators known to characterize an online community. For example, some of the characteristics of an online community that were examined included interactivity, sustained membership, feelings of membership and influence, and shared emotional connection. By manually exploring the content and types of comments posted to the blog, it was confirmed that the social networks automatically discovered by ICTA included many of the characteristics that one would normally associate with an actual community. This was especially true with the network found in the January 2009 dataset. The characteristics include the presence of highly interactive discussions where 1 in 3 comments directly addresses or references another poster and the creation of shared norms as shown through examples of the self-moderation on the blog. For instance,

- "Whatever you call the matter between your two. Please refrain from bad language" or
- "Try to hear what people are saying, and not twist their words".

And, interestingly, it was revealed that over time the anonymous posters learned to recognize and acknowledge each other and could even predict how a specific poster might respond to a particular issue, as demonstrated in the following sample comments:

- "Watch out for our resident *Real Estate expert*" or
- "What happened to *brazer*?", followed by "Hey welcome back *brazer*, funny what happens when you disappear for a few days and your links seem missed."

There were also many instances of interactions that are important in developing stronger connections between people such as comments that provide help, support, express humor, and share information that other readers of the blog might find useful including comments like:

- “Sorry to hear about your particular financial situation,” or
- “Thanks for the insight, it is much appreciated! I have bookmarked the links, will make for some good reading tomorrow.”

Finally, the posters were shown to be particularly loyal to the blog. They voluntarily came back to the blog repeatedly over a long period of time. For example, almost half of people who commented in October also posted comments in January.

Conclusion

These discoveries about the nature of the online community found within this blog are significant in two respects. First, it demonstrates that the social networks discovered automatically by the “name network” method are good and accurate approximations of the actual social networks among online blog readers and commentators. The method was able to accurately show the emergence of a new online community that started out as a small disjointed group of individual readers with only one central actor in October 2008 and evolved into a highly interactive and very connected group with multiple central actors by January of 2009. Second, despite the initial expectation of little or no community among the readers of this particular blog, the name network and ICTA have shown that even in a blog dominated by mostly anonymous and argumentative commentators, a community can still be formed and strengthened. More interestingly, it should be noted that many of the key actors in the two social networks discovered by ICTA were actually those who posted the most argumentative or controversial comments. But despite the high level of disagreement among some posters (or maybe even because of it), during the period between October and January the community has thrived and grew from a mere 50 active posters to 88. This observation is in line with the previous research on the so-called “flaming phenomena”. While studying responses to antagonism on YouTube, Lange et al. (2004) found that “not all participants view certain critical comments as a problem that necessitates regulatory mechanisms that threaten to limit participation” and “[p]articipants often wish to preserve an aura of free speech and promote self-expression” (pp.2-3). So perhaps, aside from self-interest in the primary topic of the blog, readers are coming back to this blog because it contains elements of self-moderation that many of them value. The present of such norms provided readers with a “safe” environment to debate different opinions with fellow blog readers. This is something that other bloggers might consider if they want to establish a more sustainable online community around their blogs.

References

Ali-Hasan, N. and Adamic, L. (2007). Expressing Social Relationships on the Blog through Links and Comments. In the *Proceedings of International Conference on Weblogs and Social Media*, Boulder, Colorado, USA.

Blanchard, A. (2004). Blogs as Virtual Communities: Identifying a Sense of Community in the Julie/Julia Project. In S. A. L.Gurak, L. Johnson, C. Ratliff, & J. Reyman (Eds). *Into the Blogosphere: Rhetoric, Community, and Culture of Weblogs*. University of Minnesota.

Dennen, V.P. and Pashnyak, T.G. (2008). Finding Community in the Comments: The Role of Reader and Blogger Responses in a Weblog Community of Practice. *International Journal of Web Based Communities* 4(3): 272 – 283.

Doherty, C. (2004). Naming the Trouble with Default Settings. In the *Proceedings of "SFL Ripples in the 21st Century" Australian Systemic Functional Linguistics Association Conference*, Brisbane, Australia.

Gruzd, A. (2009). Studying Collaborative Learning Using Name Networks. *Journal for Education in Library and Information Science* 50(4).

Gruzd, A. and Haythornthwaite, C. (2008). Automated Discovery and Analysis of Social Networks from Threaded Discussions. *Paper presented at the International Network of Social Network Analysts*, St.Pete, Florida, USA.

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

Jones, Q. (1997). Virtual Communities, Virtual Settlements and Cyber-Archaeology. *Journal of Computer Mediated Communication* 3(3).

Lange, P. G. (2007). Commenting on Comments: Investigating Responses to Antagonism on YouTube. *Society for Applied Anthropology Conference*. Tampa, Florida.

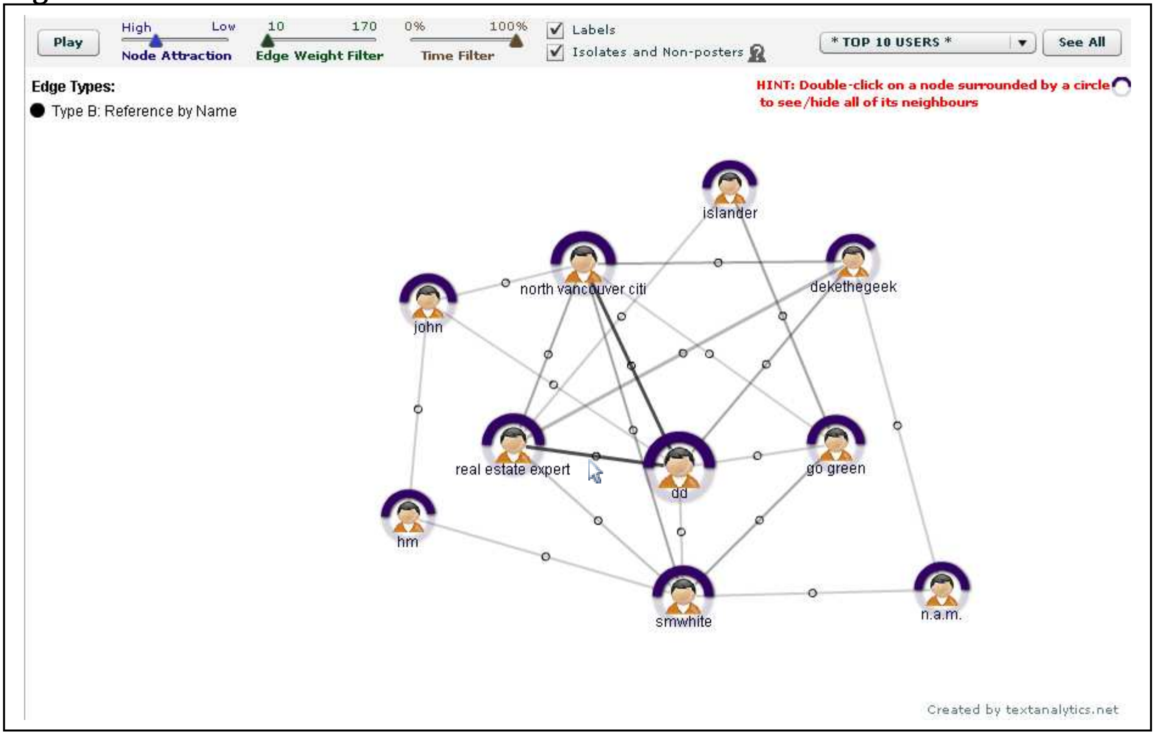
McMillan, D. W. and D. M. Chavis (1986). Sense of Community: A Definition and Theory. *Journal of Community Psychology* 14(1): 6-23.

Rourke, L., Anderson, T., Garrison, D. R., and Archer, W. (2001). Methodological Issues in the Content Analysis of Computer Conference Transcripts. *International Journal of Artificial Intelligence in Education* 12: 8-22.

Ubon, A.N. (2005). Social Presence in Asynchronous Text-Based Online Learning Communities: A Longitudinal Case Study Using Content Analysis. Department of Computer Science, The University of York. Doctor of Philosophy.

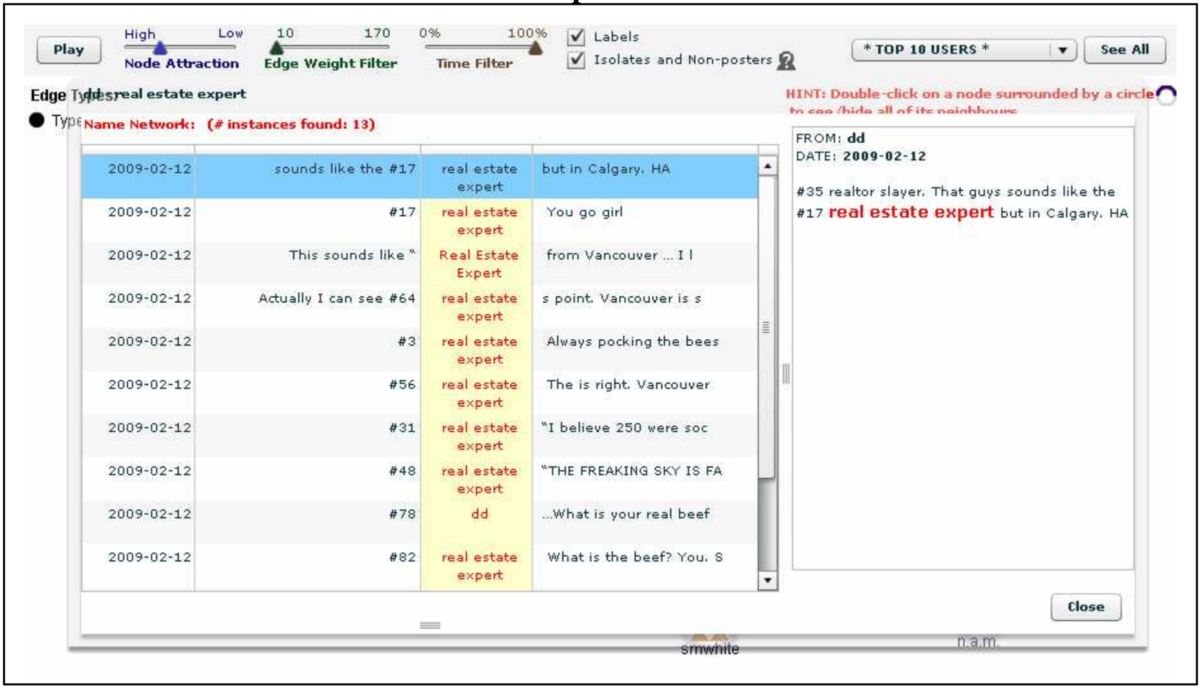
3,100 words
August 30, 2009

Figure 1: ICTA's interactive network visualization feature.



Note: By clicking on an edge that connects any two blog commentators, one can see all references/comments exchanged between these two individuals.

Figure 2: A pop-up window showing all references/comments that were used by ICTA to establish the tie between "real estate expert" and "dd".



dd - real estate expert

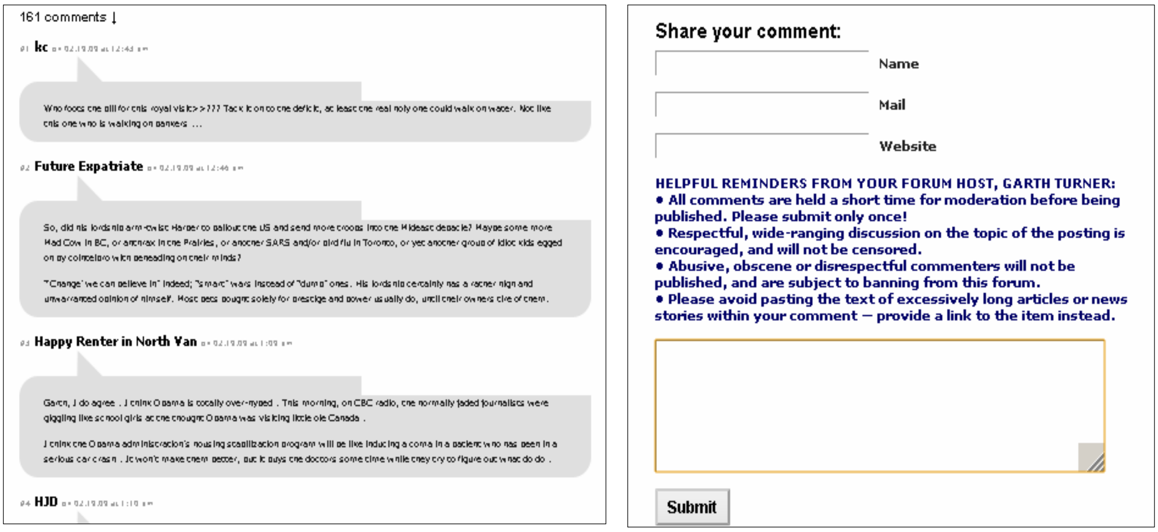
Name Network: (# instances found: 13)

2009-02-12	sounds like the #17	real estate expert	but in Calgary. HA
2009-02-12	#17	real estate expert	You go girl
2009-02-12	This sounds like "	Real Estate Expert	from Vancouver ... I l
2009-02-12	Actually I can see #64	real estate expert	s point. Vancouver is s
2009-02-12	#3	real estate expert	Always pocking the bees
2009-02-12	#56	real estate expert	The is right. Vancouver
2009-02-12	#31	real estate expert	"I believe 250 were soc
2009-02-12	#48	real estate expert	"THE FREAKING SKY IS FA
2009-02-12	#78	dd	...What is your real beef
2009-02-12	#82	real estate expert	What is the beef? You. S

FROM: **dd**
DATE: **2009-02-12**

#35 realtor slayer. That guys sounds like the #17 **real estate expert** but in Calgary. HA

Figure 3: Examples of how comments are posted on the blog (to the left) and the blog comment submission form (to the right).



161 comments ↓

01 **kc** on 02.19.09 at 12:43 am

Who foots the bill for this royal visit>>??? Tack it on to the deficit, at least the real holy one could walk on water. Not like this one who is walking on bankers …

02 **Future Expatriate** on 02.19.09 at 12:46 am

So, did his lordship arm-twist Harper to bailout the US and send more troops into the Mideast debacle? Maybe some more Mad Cow in BC, or anthrax in the Prairies, or another SARS and/or bird flu in Toronto, or yet another group of idiot kids egged on by cointelpro with beheading on their minds?

"Change" we can believe in" indeed; "smart" wars instead of "dumb" ones. His lordship certainly has a rather high and unwarranted opinion of himself. Most pets bought solely for prestige and power usually do, until their owners tire of them.

03 **Happy Renter in North Van** on 02.19.09 at 1:09 am

Garth, I do agree. I think Obama is totally over-hyped. This morning, on CBC radio, the normally jaded journalists were giggling like school girls at the thought Obama was visiting little ole Canada.

I think the Obama administration's housing stabilization program will be like inducing a coma in a patient who has been in a serious car crash. It won't make them better, but it buys the doctors some time while they try to figure out what to do.

04 **HJD** on 02.19.09 at 1:10 am

Share your comment:

Name

Mail

Website

HELPFUL REMINDERS FROM YOUR FORUM HOST, GARTH TURNER:
- **All comments are held a short time for moderation before being published. Please submit only once!**
- **Respectful, wide-ranging discussion on the topic of the posting is encouraged, and will not be censored.**
- **Abusive, obscene or disrespectful commenters will not be published, and are subject to banning from this forum.**
- **Please avoid pasting the text of excessively long articles or news stories within your comment – provide a link to the item instead.**

Figure 4: Social network of blog readers/commentators for October 2008.

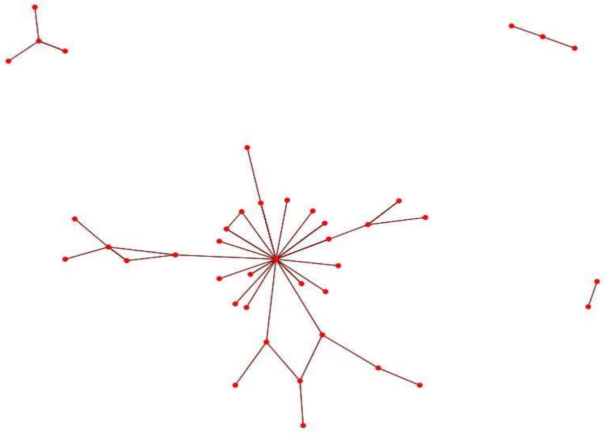


Figure 5: Social network of blog readers/commentators for January 2009.