FINDING EVIDENCE OF COMMUNITY FROM BLOGGING CO-CITATIONS: A SOCIAL NETWORK ANALYTIC APPROACH

Alvin Chin
Department of Computer Science
University of Toronto
achin@cs.toronto.edu

Mark Chignell
Department of Mechanical and Industrial Engineering
University of Toronto
chignell@mie.utoronto.ca

ABSTRACT

In this paper, we examine the problem of evaluating communities in blogs. We describe the construction and instrumentation of a research blog (on Canadian independent music) designed as a tool for measuring community effects in blogging. We then identify a number of measures concerning strength and type of communities. Using research results from sociology and psychology concerning how communities grow and function, as well as clustering algorithms from the physical and applied sciences, we demonstrate how these measures can be used in a case study based on the research blog that we developed. In addition to providing these results, the paper also introduces a computational framework based on social network analysis, which can be used to measure and evaluate community in blogs.

KEYWORDS

Blogs, community, social networks, virtual community, sense of community, computer-supported cooperative work

1. INTRODUCTION

Weblogs (blogs) adapt Web technology to allow for instant, updated and frequent communication of information such as events, personal interests, thoughts, and news. Blogs create an online conversation that is recordable on the web, and that is indexed and searchable by blogging search engines. With millions being published, blogs offer a new way for communicating and discovering social metadata, which provide advantages over other forms of content such as e-mail, forums, and wikis.

The conversational nature of blogs can be used to uniquely identify topics and similar areas of interest (Herring et al., 2005). These similarities can then be used to build community, and specifically identify virtual community, in blogs. A common theme that has emerged from past research is that the concepts of sense of community and virtual settlement are prerequisites to finding virtual community (McMillan and Chavis, 1986; Jones, 1997). These prerequisites can then be supplemented with links from blogs, used as indicators of community membership, and by using clustering algorithms to indicate the shape or structures of potentially overlapping communities.

The purpose of this paper is to conceptualize and develop a framework for measuring virtual communities in blogs, and to identify practical measures of community strength and inter-relation from the framework that could be applied to collections of co-citing blogs. Our contributions are the following: first, we propose a framework for measuring and evaluating communities and second, we create a prototype for applying that framework. In this paper, we focus on the first two steps of our framework: motivating the community and identifying the community. Using a specially created indie music blog as a case study for measuring and evaluating community in blogs, we illustrate how the research framework can work in practice.

2. BACKGROUND: BLOGGING AND VIRTUAL COMMUNITIES

According to the Wikipedia, "A *community* is an amalgamation of living things that share an environment… In human communities, intent, belief, resources, preferences, needs…may be present and common…, but the definitive driver of community is that all individual subjects in the mix have something in common." (Wikipedia, 2005). Howard Rheingold describes virtual communities as "social aggregations that emerge from the Net when enough people carry on those public discussions long enough, with sufficient human feeling, to form webs of personal relationships in cyberspace" (Rheingold, 1993). Thus we may infer that a virtual community is a collection of online mediated personal relationships that are motivated or facilitated by a common interest that is shared across the members of that virtual community.

The importance of virtual communities can be explained in at least two ways. First, virtual communities are considered important for social reasons (Wellman and Guilia, 1999). Participation in virtual communities that increase participation in face-to-face communities is a common finding in some researchers' empirical studies (Blanchard, 2003; Wellman et al., 2001). People tend to reap the benefits of social relationships with like-minded others as they become more connected with others through virtual communities.

Second, the existence of a virtual community relates to the sustainability of a virtual group. The term "community" implies an emotionally positive effect, participants are more likely to increase or maintain their participation in virtual communities when they experience feelings of community (Harris, 1999). Conversely, the lack of a community feeling among participants may be key to explaining the frequent demise of many virtual groups.

Based on these two aspects of virtual communities, researchers are examining how blogs facilitate members' social interactions and create a virtual community, a set of blogs linking back and forth to one another's postings, while discussing common topics. Social interactions around blogging have been studied previously by Nardi et al. (2004, pp 222-231), by Blanchard (2004), and by Herring et al. (2005), where they conclude that blogging is used as a means for communicating social activity.

A few studies have looked into identifying virtual communities in blogs. One method is to use content analysis by examining the content of the actual blog posts and comments. Wei et al. (2004) recorded the statistics of a knitting blog to find norms that indicate membership rules as an indicator of community. Nardi et al. (2004, pp 222-231) performed text analysis on bloggers' posts and comments in order to elicit feelings of community. A second method involves observing participants of a blog through interviews and surveys. Nardi et al. also audiotaped interviews with bloggers and asked them questions in the form of surveys, then analyzed the responses from bloggers for existence of community interactions.

However, one difficulty with previous methods for identifying community is that they involve opinions from bloggers themselves, which are highly subjective, time consuming, and prone to wrongful interpretation by humans. Fortunately, the valuable qualitative information that is obtained by interviewing bloggers on topics relating to community can be supplemented with more formal methods to generate quantitative measurements of virtual community. One source of measures of virtual community in blogs is the wealth of literature on established methods from psychology and sociology for finding community in face-to-face and online virtual communities, such as e-mail and newsgroups. One common theme that emanates from this literature is the concept of sense of community in which an individual becomes part of this sense of community if she has feelings of membership, feelings of influence, support by others, and shared emotional connection (McMillan and Chavis, 1986).

Blanchard (2003) studied sense of community in the Julie/Julia project blog and discovered that for the most part, a blog is not a virtual community. Blanchard found that in order for a blog to be a virtual community, it has to exhibit the characteristics of a "virtual settlement", that is, a virtual place in which people interact by examining artifacts. Finding evidence for virtual settlements is analogous to digging for artifacts in archaeology (Jones, 1997). For instance, Efimova and Hendrick (2005) applied this technique and used blog reading patterns, linking patterns, blog conversations and blogger directories as artifacts for indicators of blog community.

Recently, researchers have employed clustering algorithms to identify communities in blogs. Kumar et al. (2003) used Kleinberg's bursty algorithm to identify temporary bursts of activity where there is intense linking of blogs to each other based on certain topics like world events. Merelo-Guervos et al. (2004) use a pattern recognition and data mining algorithm called Kohonen's self-organizing map to identify communities,

highlight the topic defined by each community, and permit the assignment of new blogs to a community based on its links.

Social network analysis is another technique that is increasingly being used for analyzing blogs. In this approach, interlinking between blogs is used as a way to indicate relationships and determine conversations that can potentially identify communities (Herring et al., 2005; Efimova and Hendrick, 2005). In the following section, we will discuss how techniques such as social network analysis may be combined into a framework for identifying and measuring community in blogs. We will then illustrate the use of that framework in a case study.

3. FRAMEWORK FOR IDENTIFYING AND MEASURING COMMUNITY IN BLOGS

In order to identify community in existing blogs, tools are needed to detect the interactions, structures, and artifacts that are generated by activity within communities. In this paper, we outline a research framework for measuring the community building effects of different ways of constructing and motivating the use of blogs.

Determining community in blogs requires crawling and analyzing the blogs, along with a methodology to discover community patterns from existing blogs. Looking for patterns of community activity in blogs is somewhat akin to the software design patterns that arise in software engineering. Once the amount and type of community activity patterns can be detected, it should be possible to measure quantities such as the strength of community. These measures can then be used as diagnostic tools for comparing blogs and building better communities in blogs.

Based on this methodological motivation, we propose a framework for measuring and evaluating communities in blogs. This framework includes four main steps: 1) motivating community, 2) identifying community, 3) measuring community, and 4) evaluating community.

3.1 Motivating community

The first step in building and measuring community is to motivate people to blog based on shared topics of interest, coupled with easy to use blogging tools that encourage, rather than hamper, interactions through the blog. Web statistics counters can then be used to track various forms of traffic to the blog site such as length of the visit, pages accessed during the visit, and the path taken to reach the page. These statistical results can then be used to form measures of motivation that are associated with the success of the blog. This process of motivating and measuring motivation is illustrated in Figure 1.

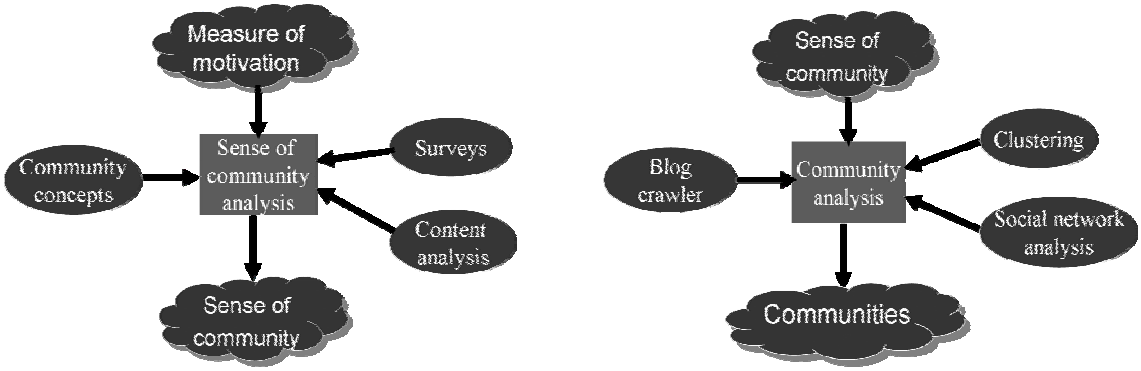
Figure 1. Block diagram for motivating community

3.2 Identifying community

The second main step in the proposed research framework is to determine if the motivation used does indeed create and build community. We follow the methods used by other researchers such as Nardi et al. (2004, pp 41-46), Wei (2004), and Efimova and Hendrick (2005) to identify sense of community. We distribute questionnaires and surveys to users of the blogs, asking them about general blogging habits, behaviour, and

feelings of sense of community. Statistical analyses can then be carried out on the results from the questionnaires and surveys. The resulting interpretations, along with content analysis of posts and comments, are then used to create a sense of community model for blogs as illustrated in Figure 2 (a). Note here that the interpretation of the survey and content analysis results is influenced by the way in which the community construct is defined as indicated by the inclusion of "community concepts" in Figure 2 (a).

It is useful to supplement ratings of community with statistical analyses that are performed on analyses of links and associated graphs constructed using blog crawlers. We have developed a blog crawler that mines the links between blogs and we create a social network that illustrates the relationships between blogs. From this social network, we identify possible communities by performing clustering and social network analysis. The complete process is shown in Figure 2 (b).



(a) Model (b) Analysis

Figure 2. Block diagram for identifying community. The left panel (a) shows the sense of community model which is then used with community analysis in (b) to identify communities

3.3 Measuring community

After the communities are identified, the next step is to measure community. This involves translating the sense of community model that we create in section 3.2 into a computational model where features of the sense of community model can be mapped into specific quantitative measures. Social network analysis can be adapted for use in measuring community. For example, we can use *network centrality* (de Nooy et al., 2005) to measure how important or central an individual blog or node is to the network.

Degree centrality (de Nooy et al., 2005) measures the network activity of a node and gives an indication as to which people are influential and may be useful nodes to focus on in community building. It may also be possible (and useful from a calibration perspective) to develop mappings between questionnaire-based sense of community measures and somewhat analogous quantitative measures such as degree centrality. *Closeness centrality* (de Nooy et al., 2005) measures how many steps on average it takes to reach every other node in the network. A node C with high closeness centrality can most efficiently make contact with other nodes in the network. This suggests that if blog A connects to blog C then blog A can reach other blogs through blog C, thus extending blog A's community to include blog C's community. *Betweenness centrality* (de Nooy et al., 2005) measures the extent to which this node can act as an intermediary or broker to other nodes in the network. We can use betweenness centrality to determine if two disparate blog clusters (at least in principle) can be joined to become one community.

3.4 Evaluating community

After creating the metric for measuring community, a blog may be evaluated in terms of its sense of community. Ideally, it should be possible to come up with a score or rank of blog effectiveness relative to different communities. Thus not only will strength of community for a blog be mentioned but also strength

for different types of community. For example, interest in professional wrestling might be one community, but there could be other communities of interest that a blog could potentially relate to. Therefore, strengths of community for a particular blog could be evaluated for a range of different communities of interest. With respect to particular communities of interest, blogs could then be ranked in terms of their strength and representativeness. The resulting rankings could then be used for a variety of applications such as ranking search results with respect to blogs, prioritizing advertising expenditures on blogs, etc.

4. CASE STUDY: INDIE MUSIC BLOG

We apply our framework from the previous section to independent (indie) music as our shared interest, and created a blog for this on MSN Spaces, a popular blog website hosted by Microsoft. This blog showcases upcoming indie bands from an indie music website and is designed to promote and add value to the indie music virtual community. In this section, we apply the steps of motivating community and identifying community from our framework to the indie music blog, with demonstrations of other steps in our research framework reserved for future work.

4.1 Motivating community

To stimulate the interest of readers and foster the formation of online community, we incorporate numerous features into the blog (blog tools) in addition to regular posts and comments on particular topics. A media player automatically loads the featured song of the day and allows users to stop and re-play the song as desired. A photo album displays new and exciting pictures of artists, tours, and concerts. Custom lists categorize and index reviewed music by genre (ie. rock, pop, folk, etc.) with links to download the reviewed songs. A rating scheme rates reviewed songs on a five-point scale and encourages readers to post their impressions of the music in the comments. One particular song is selected as the 'song of the week' and allows new visitors to sample the different types of indie music. A screenshot of the indie music blog is illustrated in Figure 3.

In addition, we regularly visit and comment on other blogs, which encourage readers to visit our blog in return. Active participation also includes responding to comments and participating in activities featured in the community. A statistics counter was used to measure the statistics that record the number of visits, the visitors, length of visit and the path taken.

4.2 Identifying community

As an initial starting point for representing a model for sense of community in blogs (Figure 2 (a)), we use the definition for sense of community by McMillan and Chavis (1986) as a community concept and ignore surveys and content analysis. Then, we perform community analysis as shown in Figure 2 (b) by analyzing the links that connect the blogs in the indie music blog from the posts and comments.

We first find nodes and links in the blog network by creating a blog crawler that crawled the indie music blog up to a link depth of 2 or two degrees of separation for inlinks (blogs that link to the target blog) and outlinks (blogs that link from the target blog), and recorded the frequency of links. In addition, we restrict crawling to MSN Spaces and use the RSS feed for crawling the blogs. To obtain the inlinks, we used the Google search engine along with the results from Technorati, a popular blogging search engine. Merging the inlinks and outlinks together and then anonymizing the blogs, resulted in the complete blog trace needed for blog data analysis. After crawling the blog, we use the UCINET social network analysis software by Borgatti et al. (2002) to create the nodes and links in the network.

The network surrounding the indie music blog, for up to 2 degrees of separation including inlinks and outlinks from the blog crawl on November 8, 2005, is visualized using Pajek (a social network visualization software by de Nooy et al. (2005)) in Figure 4. The blog network consists of 604 blogs, it is relatively sparse, and thus it is difficult to identify clusters that could possibly represent communities. To find possible communities in this network, we need to filter out those blogs which are clearly not part of any community in order to simplify the network. We use clustering and social network analysis from Pajek to then identify possible communities and explain the results in Section 5.

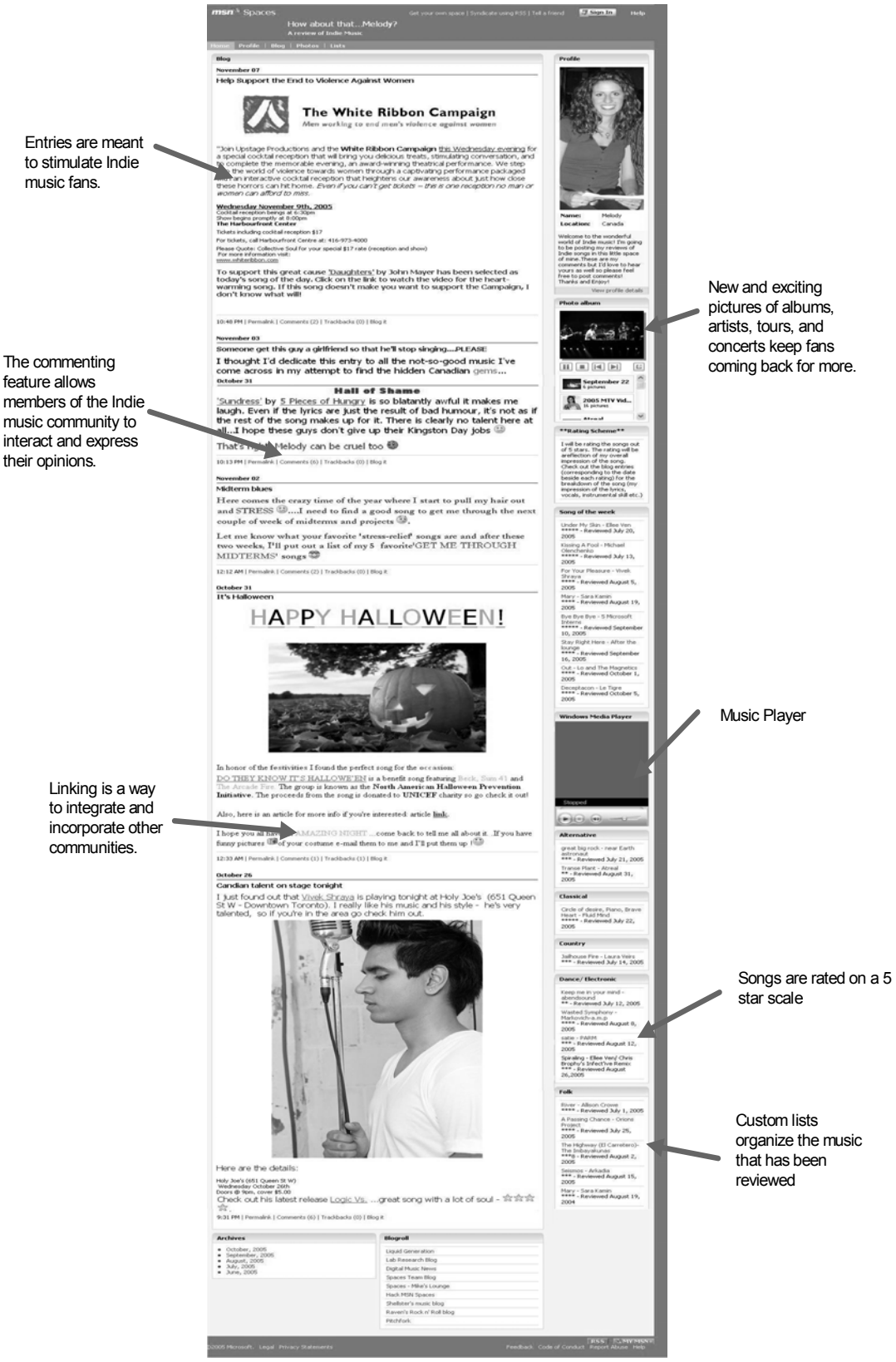


Figure 3. Screenshot of the indie music blog from http://spaces.msn.com/members/howaboutthatmelody

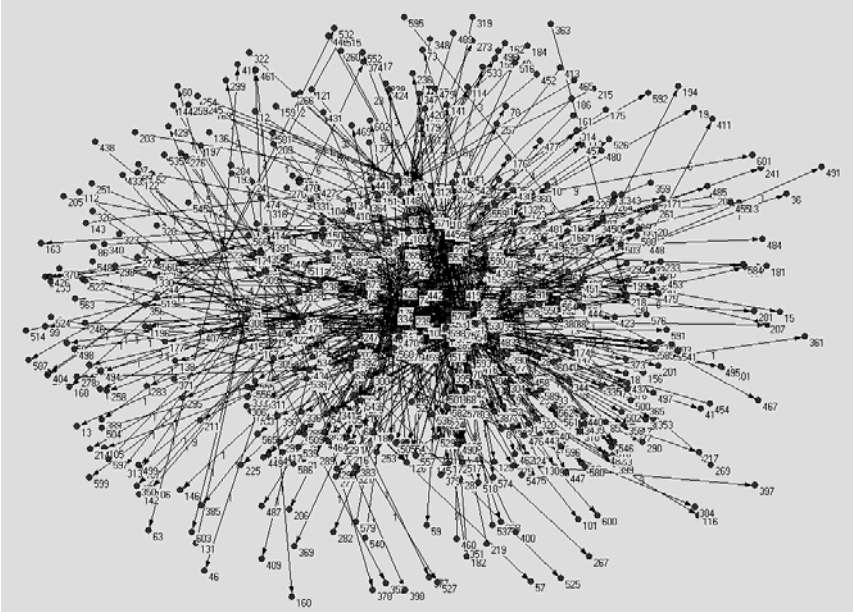


Figure 4. Network for the indie music blog up to 2 degrees of separation for inlinks and outlinks

5. DATA ANALYSIS AND PRELIMINARY RESULTS

In this section, we analyze the blog trace from the indie music blog and discuss the properties of the network around the music blog, in order to identify patterns of interaction that could possibly be communities. We use the Pajek social network visualization software by de Nooy et al. (2005) to visualize the network, then perform clustering and social network analysis as explained below.

5.1 Simplifying the network using reciprocal links

We use McMillan and Chavis' definition of sense of community, just like other researchers (Blanchard, 2003; Efimova and Hendrick, 2005), as the community concept for simplifying the network in Figure 4. Recall that sense of community involves having feelings of membership, feelings of influence, shared emotional connection and support by others (McMillan and Chavis, 1986). Feelings of membership arise from having people voice their own opinions and receiving some positive response back in return. In blogs, feelings of membership or membership ties can be represented by reciprocal links in which a post by, for example, blogger A is commented by blogger B (which is denoted by the link A->B), and blogger B writes a post on blog B which is commented by blogger A (which is denoted by the link B->A). Then the combination of inlink (A->B) and outlink (B->A) forms a reciprocal link. Efimova and Hendrick (2005) use reciprocal links in their blog analysis to identify communities, and we follow their method here as well.

In our approach, we remove all unidirectional arcs (which are links in one direction which is an inlink or an outlink) from the network and convert input and output links in each direction (also known as bidirectional arcs) into undirected edges, which results in leaving only blogs in the network that have reciprocal links. We apply recursive degree reduction to eliminate all nodes that have degree less than 2 because these nodes only have one tie, which hardly is any indication of a community.

Employing this method of reciprocal links on the network from Figure 4 reduces the number of nodes from 604 to 54, and simplifies the network as shown in Figure 5. We can see from Figure 5 that there are 8 clusters as marked in Figure 6 (a). These clusters could form possible communities, but we need to conduct further analysis in order to confirm their role as either separate communities or as parts of a larger community.

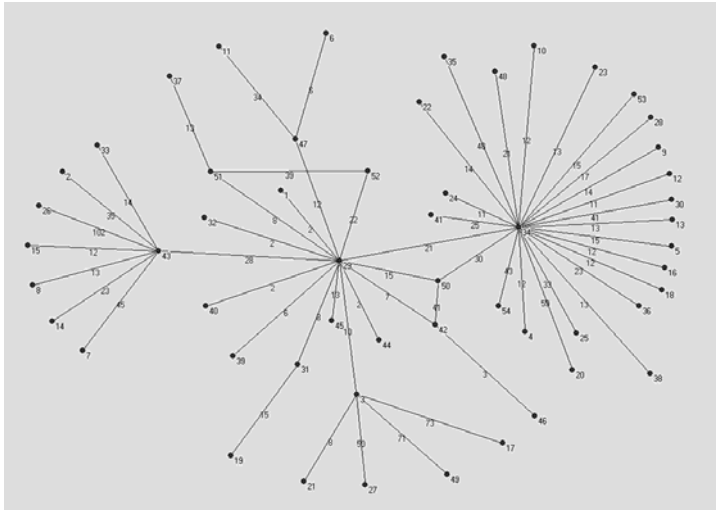
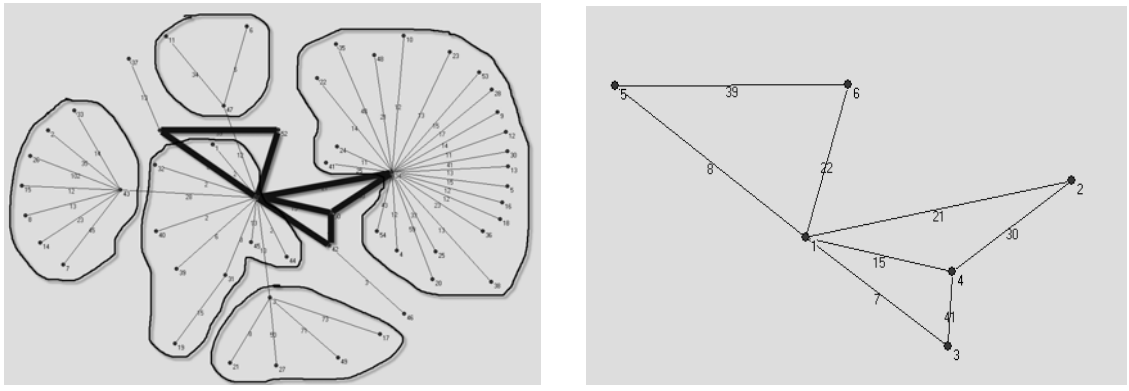


Figure 5. Network for the indie music blog for reciprocal links only

5.2 Using social network analysis to identify possible communities

We use social network analysis to validate our initial intuition of identified clusters, as indicated in Figure 6 (a), and further refine them to create groups. In social network analysis, researchers use measures of cohesion to identify components that indicate the intense clusters of interaction. These intense clusters in blogs represent conversations and if we compare these conversations to face-to-face conversations, then there is the possibility of finding shared emotional connection and inferring support by others. We can use these components for reducing the initial clusters and create dense clusters of blogs known as cohesive subgroups.

One measure of cohesion that we use is k-cores (de Nooy et al., 2005). A k-core is a cluster where the minimum degree of each node in the cluster is at least k. The degree of a node is the number of neighbours in an undirected network (which we consider for the analysis of communities). In blogs, this means that each blog in a k-core has at least k conversations. From the Pajek visualization software, we find three 2-cores (minimum degree of each node is at least 2) where each 2-core connects 3 blogs together, otherwise known as a triad. The triangles, highlighted in Figure 6 (a) from the network and extracted in Figure 6 (b), illustrate the 2-cores. Every node in the 2-core or triad is completely connected with a connection to each of the other nodes.



(a) Clusters found manually

(b) 2-cores (or triads) found from Pajek

Figure 6. The left panel (a) shows k-cores (indicated by the thick triangles) where k = 2 and other clusters found from Figure 5, whereas the right panel (b) zooms in on those k-cores

Since we have identified a 2-core as a cluster, then we calculate the fraction of blogs in the network that belong to those clusters to determine how well blogs in the indie music blog network belong to possible communities. Taking the three 2-cores and finding those from the entire network in Figure 4, we find that 0.66%, 0.17%, and 0.17% of the 604 nodes were found in each of the three clusters respectively. The majority (99% of the nodes) were not identified as comprising communities in this analysis. This shows that this initial methodology that we are using does not find strong evidence of community overall on the indie music blog, since less than 1% of the blogs belong to one of the 2-cores. However on closer analysis, we do find that there are only a few bloggers (2, 3, 4, 5, 6) that form some type of community, which would hardly be evident at first glance.

5.4 Discussion

Our initial analysis of the indie music blog network does show existence of possible communities. Further interpretation of these results can then be carried out using social network analysis on the links. To determine if the clusters that we found are indeed communities and verify their existence, we need to take into account the bloggers' behaviour to determine if they feel like they are a part of community. By having bloggers complete surveys, which we have started to do, we can identify behavioural patterns and features and combine that with sense of community.

We have not done any detailed content analysis of the posts and comments to determine whether the threads can indicate that there may be a sense of community. However, by briefly looking at the list of relevant bloggers (1,2,3,4,5,6) and examining their blog conversations, there is evidence to show that they are the prominent bloggers that engage in frequent dialogue with the author of the indie music blog (1), and that they should be included in the community. Combining content analysis along with statistical analysis from the surveys will allow us to create a complete and comprehensive model for sense of community in blogs.

Our preliminary analysis neglects the frequency of the undirected or bidirectional links (or edge counts). We can infer that the higher the edge count, then there is more conversation, which indicates higher membership ties (from sense of community), and can lead to a higher probability of community. It would be interesting to distinguish different types of community based on the strength and intensity of interactions.

Furthermore, we need to develop a computational model that maps the features of sense of community in blogs into quantitative social network analytic measures such as degree centrality. From there, we can create an algorithm for the computational model that takes into account the social network analytic measures and the bloggers' behaviour, to create a community rating that differentiates a well-developed community from an under-developed one.

6. CONCLUSIONS

The research reported in this paper sought to address the gap that exists between what users perceive as community compared to what can be measured quantitatively based on online interactions. Consequently, a framework was developed for identifying practical measures of community strength and inter-relation that can be applied to collections of co-citing blogs.

We began by reviewing prior research on virtual communities and blogs. We then proposed a sense of community model for blogs based on interpretation of content analysis and survey data. Next, a blog crawler was developed to mine the blogs for links and infer a social network structure for the blogs based on those links. We applied this analytic approach to the network of blogs surrounding an indie music blog that we developed. After creating the network (using Google and Technorati to identifying incoming links), we then used clustering and social network analysis techniques to identify clusters in the network that could be possible communities.

Our preliminary research results show that communities can be found from a target blog such as the indie music blog that we developed. We still need to further refine the tools and methods by creating a sense of community model for blogs. The tools and methods developed in our research should now be applied to a range of blogs and blog communities for purposes such as developing typologies of blog communities, and for identifying the importance of various blogs and their positions in respective communities.

ACKNOWLEDGEMENT

We acknowledge Bell Canada University Labs and Interactive Media Lab for their support of this research, Annie Xu for assistance with the creation of the surveys, Melody Gilanpour for her creation and maintenance of the indie music blog, and the conference reviewers for their suggestions in improving this paper.

REFERENCES

Blanchard, A., 2003. The effects of dispersed virtual communities on face-to-face social capital. *In M. Huysman & V. Wulf (Eds), Social capital and information technology*. MIT Press, Cambridge, USA.

Blanchard, A., 2004. Blogs as Virtual Communities: Identifying a Sense of Community in the Julie/Julia Project. *Into the Blogosphere: Rhetoric, Community and Culture*. Available from: <http://blog.lib.umn.edu/blogosphere/> [Accessed 11 November 2005].

Blanchard, A. and Markus, M., 2004. The Experienced "Sense" of a Virtual Community: Characteristics and Processes. *The DATA BASE for Advances in Information Systems*, Vol. 35, No. 1.

Borgatti, S.P. et al., 2002. *Ucinet for Windows: Software for Social Network Analysis*. Analytic Technologies, Harvard, USA.

de Nooy, W. et al., 2005. *Exploratory Social Network Analysis with Pajek*. Cambridge University Press, New York, USA.

Efimova, L. and Hendrick, S., 2005. *In Search for a Virtual Settlement: An Exploration of Weblog Community Boundaries*. Available from <https://doc.telin.nl/dscgi/ds.py/Get/File-46041/weblog_community_boundaries.pdf> [Accessed 16 December 2005].

Glance, N.S. et al., 2004. BlogPulse: Automated Trend Discovery for Weblogs. *Proceedings of the World Wide Web*. New York, USA, pp 1-8.

Harris, J., 1999. The idea of community in the study of writing. *In L. Ede (Ed.), On writing research: The Braddock essays*. Bedford, Boston, USA.

Herring, S.C. et al., 2005. Conversations in the Blogosphere: An Analysis "From the Bottom Up". *Proceedings of the Thirty-Eighth Hawai'i International Conference on System Sciences*. Los Alamitos, USA, pp 107-118.

Jones, Q., 1997. Virtual-communities, virtual settlements and cyber-archaeology: A theoretical outline. *Journal of Computer Supported Cooperative Work*, Vol. 3, No. 3.

Kumar, R. et al., 2003. On the Bursty Evolution of Blogspace. *Proceedings of the 12th International Conference on World Wide* Web, New York, USA, pp 568-576.

McMillan, D. W. and Chavis, D. M., 1986. Sense of community: A definition and theory. *Journal of Community Psychology*, Vol. 14, No. 1, pp 6-23.

Merelo-Guervos et al., 2004. *Mapping weblog communities*. Available from <http://arxiv.org/pdf/cs.NE/0312047> [Accessed 16 December 2005].

Nardi, B. A. et al., 2004. Blogging as social activity, or, would you let 900 million people read your diary? *Proceedings of the 2004 ACM Conference on Computer Supported Cooperative Work,* New York, USA, pp 222–231.

Nardi, B.A. et al., 2004. Why we blog. *In Communications of the ACM*, Vol. 47, No. 12, pp 41–46.

Rheingold, H., 1993. *The virtual community: Homesteading on the electronic frontier*. Addison-Wesley, Toronto, Canada.

Wei, C., 2004. Formation of norms in a blog community. *Into the Blogosphere: Rhetoric, Community and Culture of Weblogs*, University of Minnesota, Minnesota, USA.

Wellman, B. and Guilia, M., 1999. Net Surfers don't ride alone: Virtual communities as communities. *In B. Wellman (Ed.), Networks in the global village: Life in contemporary communities*, Westview Press, Boulder, USA.

Wellman, B. et al., 2001. Does the Internet increase, decrease, or supplement social capital? Social networks, participation and community commitment. *American Behavioural Scientist*, Vol. 45, No. 3, pp 436-455

Wikipedia, 2005. *Community* [online]. Available from <http://en.wikipedia.org/wiki/Community> [Accessed 11 November 2005].