Identifying Subcommunities Using Cohesive Subgroups in Social Hypertext

Alvin Chin
Interactive Media Lab
Department of Computer Science
University of Toronto
Toronto, ON, Canada
achin@cs.toronto.edu

Mark Chignell
Interactive Media Lab
Department of Mechanical and Industrial Engineering
University of Toronto
Toronto, ON, Canada
chignell@mie.utoronto.ca

ABSTRACT

Web pages can be modeled as nodes in a social network, and hyperlinks between pages form links (relationships) between the nodes. Links may take the form of comments, for example on blogs, creating explicit connections between authors and readers. In this paper, we describe a novel methodology and framework for identifying subcommunities as cohesive subgroups of n-cliques and k-plexes within social hypertext. We apply our methodology to a group of computer technologists in Toronto called TorCamp who communicate using a Google group. K-plex analysis is then used to identify a group of people that forms a subcommunity within the larger community. The results are then validated against the experienced sense of community of people inside and outside the subcommunity. Statistically significant differences in experienced sense of community are found, with people within the subcommunity showing higher levels of perceived influence and emotional connection.

Categories and Subject Descriptors

H.1.2 [**Models and Principles**]: User/Machine Systems – *Human factors, Human information processing.* H.3.5 [**Information Storage and Retrieval**]: Online Information Services – *Web-based services.* H.5.4 [**Information Interfaces and Presentation**]: Hypertext/Hypermedia – *Architectures, Theory.* J.4. [**Social and Behavioural Sciences**]: Sociology.

General Terms

Measurement, Human Factors, Theory, Design, Algorithms.

Keywords

Social networks, social hypertext, virtual community, cohesive subgroups, k-plexes, n-cliques, subcommunities

1. INTRODUCTION

The patterns of interconnections between web pages form a social hypertext, where web pages are nodes in a social network and hyperlinks between pages form links (relationships) between the nodes. Feedback on web pages, such as annotations or comments, create explicit links between authors and readers.

As people communicate with each other through networks of interconnected web pages, common ties may be established and social interactions may develop which can then emerge into virtual community [3]. Structures of virtual community can be discovered through a top-down approach by mapping elements of community to the social hypertext [5] or through a bottom-up approach by finding cohesive structures [9, 10] on the social network extracted from the social hypertext.

In this paper, we describe a novel methodology for identifying subcommunities within social hypertext using a bottom-up approach called *cohesive subgroups analysis*. We propose a method that computes the cohesive subgroups based on n-cliques and k-plexes, and validates the existence of inferred subcommunities by showing that experienced sense of community is greater for members of the subcommunity than it is for members of the broader community. By finding subcommunities, we can identify leaders and connectors from whom others can connect to, and recommend those people to new members to grow their community. We illustrate the use of our approach with the TorCamp community that functions online as a Google group.

2. BACKGROUND AND RELATED WORK

Virtual communities can be identified from social networks arising from conversations in social hypertext [5]. A number of methods have been proposed for identifying structure within communities based on mathematical analysis of the social network. Girvan and Newman [7] used the measure of betweenness centrality to infer community structure, and Newman [8] used eigenvalues of matrices to infer community structure.

In this paper, we chose to define a subcommunity as a cohesive subgroup within a community and then used the traditional sociological approach of clique analysis to identify cohesive subgroups of nodes within the social network. Using the definition of a clique, if a node exchanges at least c messages with every other node, then the nodes form a subgroup called a *clique at level c* [9]. If each node in the group has direct ties to at least n-k other members where n is the total number of nodes, then a *k-plex* is created [10]. Nodes can be grouped according to high cohesion, high connectivity and high reciprocity [1]. The research reported in this paper examines *k-plexes* at different group sizes to infer the existence of a subcommunity. For further review of alternative approaches to inference of subcommunity structure, we refer readers to Bird [2].

HT'07, September 10–12, 2007, Manchester, United Kingdom.
Copyright 2007 ACM 978-1-59593-820-6/07/0009...$5.00.

Our previous work [5] used McMillan and Chavis' sense of community in order to classify members of community, then quantitatively identified community using network centrality and co-citations of the underlying blog network. In the present research, we use sense of community measures to validate a subcommunity that is identified through cohesive subgroups analysis.

3. SUBCOMMUNITIES WITHIN SOCIAL HYPERTEXT

This section describes a method for identifying highly connected and cohesive structures representing subcommunities within a larger community. We use cohesive subgroups analysis to enumerate the possible subgroups in the social network that are indicative of community, with each subgroup structure forming a subcommunity and the nodes forming its members. We then validate the existence of an inferred subcommunity using sense of community measures.

3.1 Procedure for Finding Subgroups as Possible Subcommunities

To find possible subcommunities, we first find subgroups. We compute all the n-cliques at a frequency c and k-plexes at a frequency c for varying sizes of n-cliques and k-plexes, where the size is the minimum number of members in the n-clique or k-plex and the size ranges from the specified starting size to the maximum size of the n-clique or k-plex. For n-cliques, the starting size is 2, whereas for k-plexes, it is 2k-1 [6]. This repeats for all values of n and all values of k ranging from the starting size to the maximum geodesic distance (shortest path between any two nodes in a graph) from which an n-clique at level c or k-plex at level c is found. The collection of all n-cliques at level c and k-plexes at level c from this procedure form subgroups and subcommunities are identified as groups of people who consistently appear together in various cliques.

3.2 Validating Subcommunities

To determine whether the discovered subgroups are subcommunities, we survey the participants for their experienced sense of community using a standard instrument called the *Sense of Community Index* [4]. In addition to the total sense of community score, the subscales of membership, influence, reinforcement of needs and shared emotional connection are also examined. Our hypothesis is that people within a subcommunity should show a greater sense of community than members of a community who are not associated with a subcommunity. In addition to the experienced sense of community as a validating criterion, we also look at the frequency of interaction, personality factors, and centrality measures as further predictors of subcommunity membership. Once the scientific theory behind subcommunity formation is well established, it may then be possible in some instances to automatically find subcommunities within community through the structural analysis of the associated social networks.

4. CASE STUDY: TORCAMP GROUP

In the following discussion, a case study is used to demonstrate and validate the proposed method for identifying subcommunities. The TorCamp group that was used in the case study is a community of designers, developers, and entrepreneurs who work with technology in Toronto. Conversations occur through the TorCamp Google group. TorCamp holds physical meetings often (eg. DemoCamp) and this face-to-face interaction helps to build a physical sense of community which is extended online through the TorCamp Google group.

We crawled the TorCamp Google group for a two-year period up to May 2007 during which time 381 topics were discussed by the group. We generated a directed graph *G(V,E)*, where *V* is a non-empty finite set of nodes, each node *u* represents a person that posted to the TorCamp Google group, and *E* is a finite set of edges between nodes. A directed edge *e* from node *u* to node *v* with edge weight *w* exists in *G*, if user *u* has directly replied to a comment by user *v* or to the original post by user *v*, where *w* is the number of replies between *u* and *v*. The result of this analysis was a densely connected graph (social network) with 146 nodes.

4.1 Finding Subgroups in the Social Network

We used k-plexes to search for cohesive subgroups indicative of possible subcommunity in TorCamp because n-cliques nearly produced the entire network as a subgroup. As expected, from Figure 1, the number of k-plexes decreased with increasing minimum size of the k-plex (which we denote as s). As k increases, the distribution shifts to the right and up, showing more k-plexes for the same value of s.

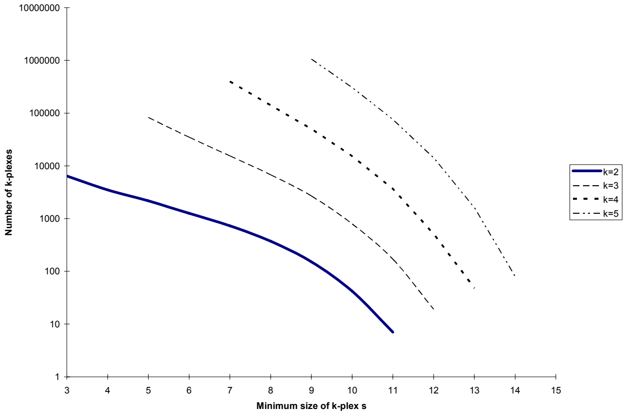

Figure 1. k-plex distribution for various k and minimum size of k-plex s

We discovered seven 2-plexes of minimum size 11 and nineteen 3-plexes of minimum size 12, with the different 2-plexes being generally similar in composition but differing from each other in terms of a couple of members. This pattern of data from Figure 1 shows that roughly the same number of people remain in the subgroup even after the parameter k in the k-plexes is relaxed several times. Thus there is evidence of a group of between 11 and 14 individuals that have formed a subcommunity. This expectation was then confirmed by viewing and comparing the composition of the various k-plexes. For k=3, for each member in the 3-plex, we computed the number of 3-plexes that each member was found in and plotted this as the social network shown in Figure 2, where the size of each node is proportional to the number of 3-plexes found for that member. Using the k-plex criterion it can be seen that potential members of the subcommunity vary in terms of how strongly they are associated with the subcommunity. Thus it seems that subcommunity membership is a somewhat fuzzy criterion.

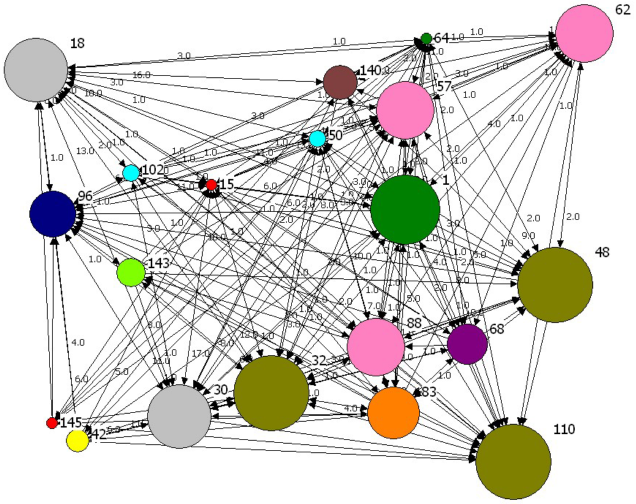

Figure 2. Network of 3-plexes with minimum size 12

4.2 Validating Subcommunity Structure

The k-plex analysis of the TorCamp data suggested the existence of a subcommunity of k-plexes consisting of somewhere between 11 and 14 members. We then administered a survey to collect further data about TorCamp community members and compare them to the k-plexes with k=3. This survey consisted of a ten-item version of the Big Five personality inventory as well as the Sense of Community Questionnaire [4]. There were a total of 25 responses, of which 18 were in the TorCamp social network.

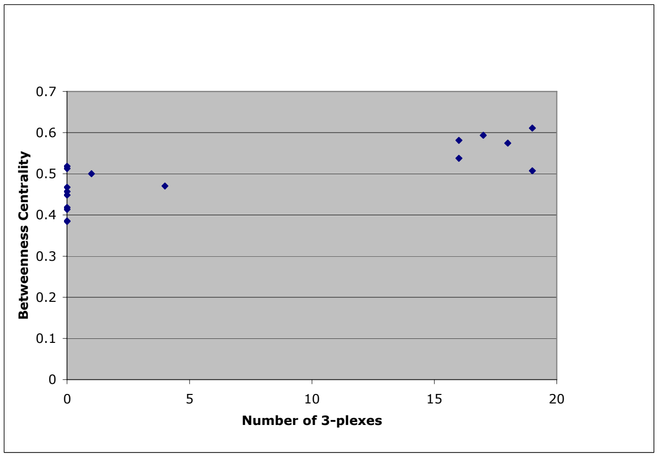

Figure 3. Betweenness centrality and number of 3-plexes

Figure 3 shows the relationship of betweenness centrality and k-plex involvement for k=3 as a scattergram. It can be seen that people involved in a large number of 3-plexes generally have a betweenness centrality above 0.5 and those with few messages have centrality scores below 0.5, but with a small amount of overlap between the two groups.

Figure 4 shows the relationship between k-plex involvement and number of messages for k=3. It can be seen that while high message use is associated with being in a larger number of 3-plexes, there are a few exceptions to this general rule.

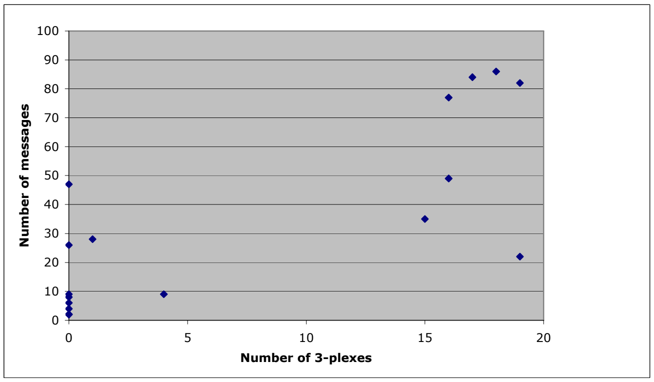

Figure 4. Number of 3-plexes and messages

The results shown in Figures 3 and 4 suggest that people involved in only a few 3-plexes (N<5) might be better placed in the same group as those with no 3-plexes. This led to a comparison of the 7 people involved in many 3-plexes versus the 11 others who were involved in either few 3-plexes or none.

The 7 people involved in many 3-plexes were contrasted with the 11 others who were involved in either few 3-plexes or none. The results of the t-tests carried out are summarized in Table 1.

Table 1. Summary of t-test results for the 3-plex effect

	t(16)	p	3-plex effect
Extraversion		ns	
Agreeableness		ns	
Conscientiousness	3.51	0.003	- (4.5 vs. 6)
Emotional Stability		ns	
Openness	-1.99	0.07	+ (6.6 vs. 5.9)
Total SOC	-1.95	0.08	+ (48.4 vs. 42.6)
Membership		ns	
Influence	-2.84	0.01	+ (12.7 vs. 9.5)
Reinforcement of Needs		ns	
Emotional Connection	-2.25	0.04	+ (13 vs. 11.6)
Number of Messages	-4.51	0.002	+ (62 vs. 13)
Degree Centrality		p < .01	higher
Betweenness Centrality		p < .01	higher
Closeness Centrality		p < .05	higher

In terms of personality, people involved in 3-plexes had lower conscientiousness but tended to have greater openness. The total sense of community tended to be higher for those involved in 3-plexes, and this was attributable to significantly higher levels of perceived influence and emotional connection. People involved in 3-plexes typically sent many more messages and had higher centrality scores than the others.

Table 2 examines the correlations between three sense of community subscales and the other measures collected in this study. None of the correlations involving reinforcement of needs were found to be significant so it was not included. In addition, the total sense of community score was also not included in the table because it had only one borderline significant effect and that was attributable to the effects of betweenness centrality on the influence and emotional connection subscales. As summarized in Table 2, both the influence and emotional connection scales appear to be related to conscientiousness, number of messages, and betweenness centrality. In addition, people who claimed higher levels of influence tended to have higher levels of agreeableness, while higher emotional connection tended to be associated with higher degree centrality.

Table 2. Correlations between sense of community subscales and the other measures used (ns indicates p>.10)

	Membership	Influence	Emotional Connection
Extraversion	ns		ns
Agreeableness	ns	r=.49, p<.05	ns
Conscientiousness	ns	r=.45, p<.10	r=.48, p<.05
Emotional Stability	ns		ns
Openness	r=.44, p<.10		ns
Number of Messages	ns	r=.41, p<.10	r=.47, p<.05
Degree Centrality	ns		r=.45, p<.10
Betweenness Centrality	ns	r=.43, p<.10	r=.57, p<.05
Closeness Centrality	ns		ns

4.3 Discussion

The interpretation of the survey results is limited by the number of responses that we were able to obtain, with 25 respondents overall of whom 18 were in the social network that we constructed. Nevertheless a number of significant and borderline significant results were obtained, with the results generally being in line with prior expectations. A subcommunity was identified using k-plex analysis (k=3), and that subcommunity was also consistent with the betweenness centrality scores, as expected from past research literature. In addition, the influence and emotional connection sense of community subscale scores tended to be higher for those in the subcommunity. One intriguing new result was that conscientiousness scores were significantly lower for people in the subcommunity while openness scores tended to be higher. This raises the possibility that the personalities of group members may influence how subcommunities form within social hypertexts and who joins them. It is also interesting to note that extraversion was not associated with subcommunity membership in this case, so the creation of the subcommunity could not be attributed to an effect of outgoing people communicating more often with each other. Aside from betweenness centrality, we also measured distance centrality and closeness centrality. While the other centrality measures also exhibited some relationships, in every case that we examined, betweenness centrality showed the strongest relationship, and thus we agree with past researchers that betweenness centrality is a useful measure in identifying and assessing subcommunities.

5. CONCLUSIONS

As the internet moves from being a network of servers and documents to a network of documents and people, online community interaction and social computing become increasingly important. Methods are needed to evaluate community activity and structure. In this paper we report on research using cohesive subgroups to identify subcommunities within an online community. In the TorCamp case study we found strong evidence for a subcommunity of between 11 and 14 members. K-plex analysis with k=3 indicated the presence of only one subcommunity. As expected, people within the subcommunity showed a higher sense of community than others with the effect being mainly due to higher scores on the influence and emotional connection subscales of sense of community. Further validation was provided by the strong differentiation in betweenness centrality scores between those who were in and those who were outside the subcommunity (since past research has suggested that betweenness centrality is also a good indicator of subcommunity). K-plex analysis is recommended as a useful supplement to other methods of discovering subcommunities because it provides a considerable amount of diagnostic information that helps the analyst get a better understanding of the strength of the subgroup and how sharp the boundaries of the group are (i.e., the distinction between who is in the group and who is out of the group). While this type of subgroup research is still at an early stage, the present results are promising, demonstrating the validity of these approaches in the context of a realistic case study.

6. ACKNOWLEDGMENTS

We would like to thank the members of TorCamp for their generous participation in this research and for their many helpful comments and suggestions, as well as the many reviewers that provided valuable feedback and suggestions.

7. REFERENCES

[1] Ali-Hasan, N. and Adamic, E. Expressing Social Relationships on the Blog through links and comments. In *Proc. of the Intl. Conference on Weblogs and Social Media*, Boulder, CO, 2007.

[2] Bird, C. Community Structure in OSS Projects. Downloaded from http://wwwcsif.cs.ucdavis.edu/~bird/, July 1, 2007.

[3] Blanchard, A. and Markus, M. The experienced sense of a virtual community: Characteristics and processes. *The DATA BASE for Advances in Information Systems*, Vol. 35, No. 1, 2004.

[4] Chavis, D. *Sense of Community Index*. Association for the Study and Development of Community, http://www.capablecommunity.com/pubs/SCIndex.PDF.

[5] Chin, A. and Chignell, M. A social hypertext model for finding community in blogs. In *Proc. of the 17th Intl. ACM Conference on Hypertext and Hypermedia*, Odense, Denmark, 2006, 11-22.

[6] Everett, M. Cohesive subgroups. Analytic Technologies, http://www.analytictech.com/networks/EverettSubgroups.doc

[7] Girvan, M. and Newman, M.E.J. Community structure in social and biological networks. In *Proc. National Academy of Sciences USA,* 99:7821, 2002.

[8] Newman, M.E.J. Finding community structure in networks using the eigenvectors of matrices. Physical Review E, 74:036104, 2006.

[9] Reffay, C. and Chanier, T. How social network analysis can help to measure cohesion in collaborative distance learning. In *Proc. of Computer Supported Collaborative Learning*, Kluwer Academic Publishers, 2003, 343–352.

[10] Sterling, S. Aggregation techniques to characterize social networks. *Master's thesis*, Air Force Institute of Technology, 2004.