Identifying Active Subgroups in Online Communities

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

Mark Chignell
Interactive Media Lab
Dept. of Mech. & Industrial Engineering
University of Toronto
chignell@mie.utoronto.ca

Abstract

As online communities proliferate, methods are needed to explore and capture patterns of activity within them. This paper focuses on the problem of identifying active subgroups within online communities. k-plex analysis and hierarchical clustering are used to identify and contrast subgroups, and the methodology is demonstrated in a case study involving the TorCamp Google group community. We assessed the validity of the subgroups obtained in the case study by comparing them with the members' experienced sense of community, and their self-reported acquaintanceships. Results suggest that active subgroups of people not only interact with each other at a higher rate, but also have a greater experienced sense of community. It is concluded that detection of active subgroups in online communities can be implemented widely using automated tools for analyzing the social networks implied by online interactions.

1 Introduction

Discovering members who are the leaders and connectors in the community can lead to better methods for building community. Community building is an important task in a number of settings, including startup companies and new open source projects [1]. Standard tools are not yet available for identifying leaders and followers within these communities. Our focus on finding active subgroups is motivated by the expectation that leaders are more likely to be actively involved within communities and thus subgroups are likely to form around active, and like-minded, people. These subgroups may then play a disproportionate role in driving the goals and activities of the community as a whole.

How can active subgroups within communities be efficiently recognized? What quantitative methods could be used to assess if active subgroups exist in the community and if so which members are they comprised of? In this paper, we examine two quantitative approaches (k-plex analysis and hierarchical cluster analysis) to address these questions and then contrast their effectiveness in a case study. Preliminary results show that hierarchical clustering produces similar subgroups to k-plex analysis, and that network centrality and number of acquaintances may also provide supplementary evidence concerning community activity and subgroup formation.

2 Background and Related Work

Several quantitative methods have been proposed [2, 5] for identifying structure within communities based on analysis of the associated social network. Hierarchical clustering has been used to quantify the structure of community in citation networks [5] and open-source community projects [1]. An alternative method, however, is to classify members into a community based on a behavioural model. Our previous work [2] used McMillan and Chavis' sense of community instrument [7] to

create a social hypertext model to assess community membership.

In this paper, we use k-plex analysis [3] to identify subcommunities as cohesive subgroups within a community, and then compare this approach with hierarchical cluster analysis. We then use sense of community and other measures to validate the subcommunities found.

3 Structural Model of Online Community

A k-plex is a structure where each node has direct ties to at least n-k other members and n is the size (number of nodes) of the k-plex. To identify an active subgroup, we compute all the k-plexes for various sizes of the k-plex, where the size ranges from the minimum of 2k-1 [3] to the maximum size for which k-plexes are found, and k ranges from 2 to the maximum geodesic distance from which k-plexes are found. Community members that appear in many overlapping k-plexes are then identified as forming a possible subgroup.

As an alternative to the k-plex approach, we also used weighted average hierarchical clustering [6] on the same dataset that merges the pair of clusters in each iteration with the highest cohesion. Clusters found may then be interpreted as subgroups.

3.1 Validating subgroups as subcommunities

People within an active subgroup may show a greater sense of community than other members of the surrounding community. This suggests that subgroups identified through quantitative analysis, such as frequency of interaction and network centrality, may then be validated by assessing whether or not subgroup members have higher experienced sense of community than other community members.

4 Case Study: TorCamp Group

In order to study online community, an online group was needed that had high connectivity and cohesion, a sense of place, common ties, and social interaction [4]. The Toronto-based TorCamp group met the above criteria and was chosen for the case study. The self-described goals of TorCamp are to build an open community of individuals and companies, and to provide open events to inspire and instill a sense of community in the Toronto technology scene.

TorCamp holds face-to-face meetings often to discuss and share ideas about computer technology. This helps to build a physical sense of community [7] which is extended online through the TorCamp Google group. TorCamp members are highly opinionated and passionate. They post an average of more than five messages per day, with an average of more than five messages posted per thread. In carrying out the case study we examined the links between messages using a crawler and then analyzed the resulting social network. To validate whether those members were part of a subcommunity, we then administered two questionnaires. The first questionnaire evaluated participants' sense of community and their personality. Participants were then asked to list the people that they personally knew in TorCamp along with their frequency of communication with those people, in the second social networking questionnaire.

4.1 Subgroups within the social network

We crawled the TorCamp Google group from 2005 up to May 2007, for a total of 381 topics, each topic defined as a post followed by a list of replies. A social network was then constructed where, for each post, links were recorded from the post to each person who replied to that post. We also inferred links from each reply to the immediately previous reply. Following this procedure for all topics led to an inferred TorCamp Google group social network which was a high densely connected graph with 146 nodes.

We then applied k-plex analysis to the inferred TorCamp Google group social network. By varying k from 2 to 5, we discovered a subgroup consisting of between 11 and 14 members. Since the 2-plexes were generally similar to the 3-plexes except differing in one or two members, we decided to choose 3-plexes for subsequent analysis. For each member in the 3-plexes, we computed the number of 3-plexes in which that member participated. We then visualized this membership function in the social network illustrated in Figure 1, where the size of each node

was proportional to the number of 3-plexes that a person was involved in.

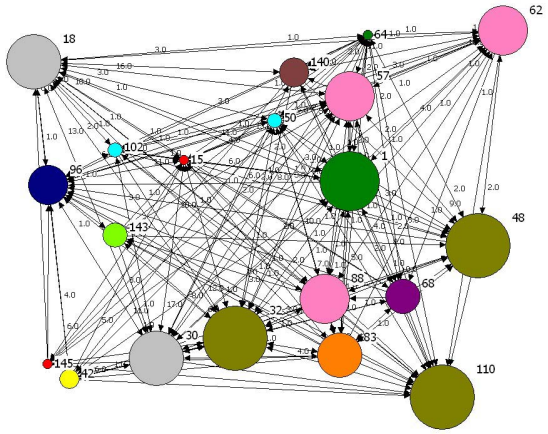

Figure 1. Social network of 3-plexes with minimum size 12 with size of nodes (shown as circles) being proportional to the number of 3-plexes found in which that node appeared

As an alternative approach to finding subgroups, we also used weighted average hierarchical clustering [6] as implemented in UCINET on the inferred TorCamp Google group social network. Figure 2 shows a portion of the resulting dendrogram. The TorCamp members shown in the cluster at the top of the figure have strong overlap with the subgroup identified earlier using k-plex analysis, with 15 members of the identified cluster also appearing in a 3-plex containing 20 members.

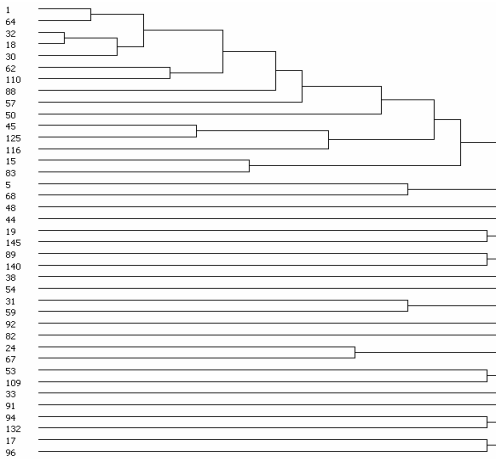

Figure 2. Partial dendrogram from hierarchical clustering (weighted average) of TorCamp Google group social network

4.2 Validating the sub-community structure

The k-plex analysis from the previous section, indicated that there is a particularly active sub-community of between 11 and 14 members. People involved in a large number of 3-plexes generally had higher betweenness centrality (above 0.5) and sent more messages.

There were a total of 25 responses to the sense of community questionnaire, of which 18 were in the inferred TorCamp Google group social network. We also asked participants in the social network questionnaire to list others that they knew in TorCamp based on their level of acquaintance (do not know, have met, somewhat close, very close), to compare the recorded social interactions with people's perceptions. We obtained a total of 14 responses, of which 10 were in the TorCamp Google group social network. Table 1 summarizes the Pearson correlation (two-tailed test) between the sense of community subscales with level of acquaintanceship with other members, centrality and k-plex involvement for k=3, size=12.

SOC subscale	# Acq	Deg Cen	Betw Cen	Close Cen	# 3-plexes
Membership	.737	.589	-.408	.542	-.186
Emotional connection	.634	.292	-.552	.453	.127
Influence	.211	.096	-.398	.182	-.078
Needs	.089	.173	-.563	.196	-.430

Table 1. Correlations between sense of community subscales and acquaintances

Each of the individual sense of community subscales was correlated with betweenness centrality, as shown in Table 1. In addition, the membership and emotional connection subscales of sense of community were both strongly correlated with acquaintanceship.

We discovered that the number of acquaintances directly affected the degree centrality in the inferred social network. This is illustrated in Figure 3. This suggests an alignment between the social network that can be inferred from online interaction, and the social network that can be constructed from reported acquaintanceships. There were six people who belonged to the sub-community identified and who also provided information about their acquaintanceships.

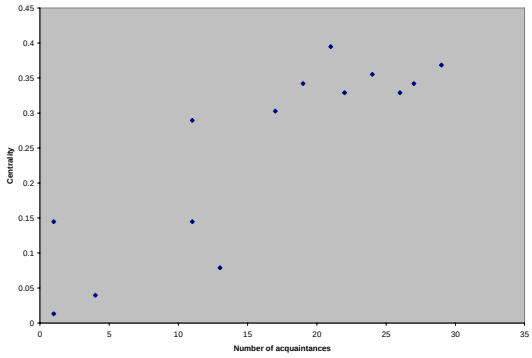

Figure 3. Effect of number of acquaintances on degree centrality

Table 2 shows the relationship between level of acquaintanceship and number of messages sent across the possible 30 pairwise asymmetric relations. It can be seen that 19 of the 30 relations can be classified as weak ties (people who one has met but do not know well) and that mean number of messages was highest for the weak tie relations. This is consistent with earlier social networking research demonstrating that communication activity on the internet is often highest between people who are related by weak ties, with synchronous communication (e.g., face to face or by telephone) being preferred in maintaining strong ties.

Acquaintance	Count	Mean	SD
None	2.00	2.00	0.00
Have Met	19.00	3.32	3.73
Somewhat Close	7.00	1.86	1.77
Very Close	2.00	3.00	0.00
Total	30.00	2.87	3.12

Table 2. Relationship between level of acquaintanceship and number of messages sent

5 Conclusions

In this paper, we showed how social network analysis and clustering can be used to identify subgroups within an online community and demonstrated this approach on the TorCamp group. We found evidence for a subcommunity of between 11 and 14 members using both k-plex analysis and hierarchical cluster analysis, with a strong overlap between the memberships of the subgroups identified by each of the two methods. Subgroup membership was also found to be related to betweenness centrality in this study.

The research results suggest that social network analysis and cluster analysis can identify active subgroups of people who not only interact with each other at a higher rate, but who also have a greater experienced sense of community. While further research is needed to confirm the present results and interpretations, it seems that active subgroups may be a useful way to identify emerging or defacto leaders in otherwise poorly structured communities. Since active subgroups may be inferred from social networks implied by patterns of online interaction, it seems likely that active subgroup detection in online communities can be implemented on a large scale using automated methods, without the need for surveys.

References

[1] C. Bird. Community Structure in OSS Projects. Downloaded from http://wwwcsif.cs.ucdavis.edu/~bird/, July 13, 2007.

[2] A. Chin and M. Chignell. A social hypertext model for finding community in blogs. In Pro-c. of the 17th International ACM Conference on Hypertext and Hypermedia, ACM, 2006, pages 11–22.

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

[4] G. A. Hillery. Definitions of community: Areas of agreement. Rural Sociology, Vol. 20, (1955), pages 779-791.

[5] J. Hopcroft et al. Tracking evolving communities in large linked networks. PNAS 101: 5249-5253; published online before print as 10.1073/pnas.0307750100.

[6] S. C. Johnson. Hierarchical Clustering Schemes. Psychometrika, Vol. 2, (1967), pages 241-254.

[7] D.W. McMillan and D.M. Chavis. Sense of community: A definition and theory. Journal of Community Psychology, Vol. 14, No. 1, (1986), pages 6-23.