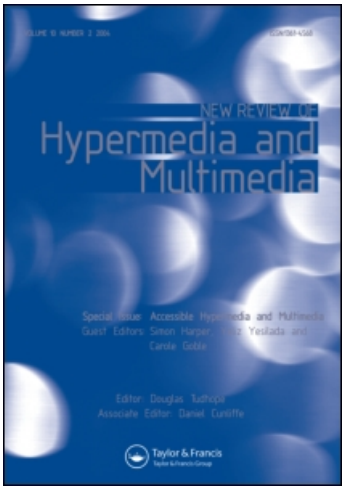


**New Review of Hypermedia and Multimedia**
Publication details, including instructions for authors and subscription information:
http://www.informaworld.com/smpp/title~content=t713599880

# Tracking cohesive subgroups over time in inferred social networks

Alvin Chin[a]; Mark Chignell[b]; Hao Wang[a]
[a] Mobile Social Networking Group, Nokia Research Center, Beijing, China [b] Department of Mechanical and Industrial Engineering, University of Toronto, Toronto, ON, Canada

Online publication date: 04 August 2010

**To cite this Article** Chin, Alvin , Chignell, Mark and Wang, Hao(2010) 'Tracking cohesive subgroups over time in inferred social networks', New Review of Hypermedia and Multimedia, 16: 1, 113 — 139
**To link to this Article: DOI:** 10.1080/13614568.2010.496132
**URL:** http://dx.doi.org/10.1080/13614568.2010.496132

*New Review of Hypermedia and Multimedia,*
Vol. 16, Nos. 1–2, April–August 2010, 113–139

Taylor & Francis
Taylor & Francis Group

# Tracking cohesive subgroups over time in inferred social networks

ALVIN CHIN†*, MARK CHIGNELL‡ and HAO WANG†
†Mobile Social Networking Group, Nokia Research Center, Building No. 2, 5 Donghuan Zhonglu, Beijing 100022, China
‡Department of Mechanical and Industrial Engineering, University of Toronto, 5 King's College Road, Toronto, ON M5S 3G8, Canada

*(Received 12 January 2010; final version received 23 May 2010)*



As a first step in the development of community trackers for large-scale online interaction, this paper shows how cohesive subgroup analysis using the Social Cohesion Analysis of Networks (SCAN; Chin and Chignell 2008) and Data-Intensive Socially Similar Evolving Community Tracker (DISSECT; Chin and Chignell 2010) methods can be applied to the problem of identifying cohesive subgroups and tracking them over time. Three case studies are reported, and the findings are used to evaluate how well the SCAN and DISSECT methods work for different types of data. In the largest of the case studies, variations in temporal cohesiveness are identified across a set of subgroups extracted from the inferred social network. Further modifications to the DISSECT methodology are suggested based on the results obtained. The paper concludes with recommendations concerning further research that would be beneficial in addressing the community tracking problem for online data.

*Keywords:* Cohesive subgroups; Social networks; Community; DISSECT method; Similarity measurement; Centrality; Clustering

## 1. Introduction

With the growth of social networking on the Internet, and Web 2.0 functions such as blogging, social tagging and video sharing, more and more information is becoming available online about how people interact and with whom. The existence of social networks can be inferred from a wide variety of interactions, directly through communications and indirectly through shared interactions on web pages (e.g. making comments). The links that form inferred social networks can be garnered from a wide variety of sources, including feedback and comments and blogs or logged interactions on social networking sites, mobile interactions and various forms of evolving social media.

There are a wide range of powerful social network analysis and data mining tools available to measure and interpret social networks inferred from online

*Corresponding author. Email: alvin.chin@nokia.com

New Review of Hypermedia and Multimedia
ISSN 1361-4568 print/ISSN 1740-7842 online 
http://www.tandf.co.uk/journals
DOI: 10.1080/13614568.2010.496132

interaction. What are the important and interesting questions that can be asked with respect to large social networks representing online interactions? In future community tracking systems, the following questions, among others, may be of interest:

(1) What are the cohesive subgroups of people that exist in the social network?
(2) Which subgroups are ephemeral, and which subgroups persist over time?
(3) How do the persistent subgroups evolve and change over time?
(4) What is the purpose or relationship that makes a subgroup cohesive?
(5) How can subgroups be indexed, and subsequently queried?
(6) How and when are subgroups formed?
(7) Is a subgroup growing or shrinking at a particular point in time?
(8) How strong or cohesive is a particular subgroup?
(9) How strongly are different subgroups related?
(10) To what extent may a related set of subgroups be indicative of a movement or a community?



In this paper we review and demonstrate techniques for addressing the first two of the questions above, using a structure-based approach (i.e. analysis of linking structure without the use of content analysis). Section 2 of this paper provides a literature review of research that can be used for finding subgroups and for tracking changes in subgroups as they evolve. Section 3 briefly describes the Social Cohesion Analysis of Networks (SCAN) method and reports on its use in two case studies (the TorCamp Google group and a set of YouTube vaccination video conversations). Based on analysis of weaknesses in the SCAN method, a revised framework is introduced and referred to as the Data-Intensive Socially Similar Evolving Community Tracker (DISSECT) as described in Section 4, where multiple known subgroups within a social network are tracked in terms of similarity-based cohesiveness over time. In Section 5 the DISSECT method is applied to the Nokia Friend View mobile social network and changes in subgroup membership over the course of the Friend View trial are visualised. Different types of cohesive subgroups are characterised based on their persistence in the network over time. In Section 6, we compare the results for the three case studies in terms of their social network properties, evolution and cohesive subgroups, and areas for further research exploration are suggested. Conclusions are presented in Section 7.

## 2. Finding cohesive subgroups

In this section, we review previous work that can be used for finding and tracking cohesive subgroups. A cohesive subgroup is a group where there is greater interaction and cohesion with members of the group than with members outside the group (cf. Borgatti *et al.* 1990). Previous research has shown that cohesive subgroups form communities of interest (Dixon 1981), have weak ties (Garton *et al.* 1997), and have cohesive bonds that bring

people together (Piper *et al.* 1983). Section 2.1 examines the problem of finding cohesive subgroups, and Section 2.2 then looks at the evolution of subgroups over time.

### *2.1 Identifying subgroups*

Centrality, cohesiveness and clustering and partitioning are properties that can be used for identifying subgroups. Centrality (Freeman 1979) is used to identify the most important or active people that are well connected in the network, where those who are actively involved in one or more subgroups will generally score higher with respect to centrality scores for the corresponding network. In social network analysis, centrality measures how important or central an individual node is in a network. While there are many measures of centrality that have been used to characterise the social behaviour and connectedness of nodes, our focus is on three centrality measures that have been used extensively for identifying cohesive subgroups. These measures are betweenness centrality, degree centrality and closeness centrality.

Betweenness centrality measures the extent to which a node can act as an intermediary or broker to other nodes (Freeman 1979). High betweenness centrality may also indicate stronger subgroup and community membership (Girvan and Newman 2002, Donetti and Munoz 2004, Marlow 2004, Newman and Girvan 2004, Gloor 2005, Tyler *et al.* 2005). In contrast, degree centrality measures the number of direct connections that an individual node has to other nodes within a network (Freeman 1979). Nodes with high-degree centrality have been shown to be more active, more influential and to be associated with a relatively strong sense of community (Fisher 2005, Frivolt and Bielikov 2005, Welser *et al.* 2007, Memon *et al.* 2008). Closeness centrality measures how many steps on average it takes for an individual node to reach every other node in the network. Nodes with high closeness centrality are able to connect more efficiently or easily with other nodes, making them more likely to participate in subgroups. Closeness centrality (Ma and Zeng 2003, Chin and Chignell 2006, Kurdia *et al.* 2007) has been used for characterising influential members.

Cohesive subgroups within social networks can indicate the most active members within a community (Reffay and Chanier 2003, Wellman 1997, Fortunato *et al.* 2004, Sterling 2004). Examples of structural groupings frequently discussed in the literature are cliques and k-plexes. Those structures have been used to characterise groupings in social networks (Wasserman and Faust 1994, Alba 1973, Sterling 2004, Balasundaram *et al.* 2008, Chin and Chignell 2007, Du *et al.* 2007, Reffay and Chanier 2003). Cliques are fully connected subgroups (Wasserman and Faust 1994) where each member has a direct connection to every other member in the subgroup, thus forming a completely connected graph within the subgroup. The criteria for subgroup formation can be relaxed by recognising subgroups where members are not completely connected and where each node in the subgroup

has direct ties to at least *n–k* members. The resulting structure is referred to as a k-plex (Hanneman and Riddle 2005). However, cliques and k-plexes are not well suited as criteria for identifying subgroups in large networks because the required analysis scales exponentially with the number of nodes in the network and their discovery is an NP-complete problem (Balasundaram *et al.* 2008).

Clustering and techniques such as link analysis (Brin and Page 1998, Kleinberg 1999) and co-citation analysis (Flake *et al.* 2002, Kleinberg 2003, Adar *et al.* 2004, Kumar *et al.* 2004) can be used to detect subgroups. Hierarchical clustering is often used to quantify the structure of community in web networks (e.g. Girvan and Newman 2002, Donetti and Munoz 2004, Clauset 2005, Li *et al.* 2006) where the cluster orderings in the dendrogram form the subgroups. Hierarchical clustering can automate the process of finding subgroups. Agglomerative hierarchical clustering (e.g. Everitt 1974) groups nodes into a cluster if the nodes are similar and then successively merges clusters until all nodes have been merged into a single remaining cluster. Divisive hierarchical clustering, in contrast, builds the hierarchy from the top-down. In contrast to hierarchical clustering, groups formed in partitioning methods are not nested and are non-overlapping. Although fuzzy clustering methods are available (e.g. Sato and Sato 1997), non-overlapping partitioning techniques have been widely used (e.g. the Quickcluster routine in SPSS) because of their computational efficiency. Partitioning methods require that the number of subgroups in the partition be defined prior to the analysis. However, inferring partitioned subgroups from a nested hierarchy (dendrogram) is also potentially problematic, since a criterion must be defined for determining where to partition (cut) the dendrogram (Orford 1976). Criteria and methods aimed at identifying optimal partitions include modularity (Radicchi *et al.* 2004, Danon *et al.* 2005, Duch and Arenas 2005, Ruan and Zhang 2007), vector partitioning (Wang *et al.* 2008) or normalised cut metrics (Zahn 1971, Shi and Malik 2000, Schaeffer 2007, Leskovec *et al.* 2008) for finding subgroups. However, as van Duijn and Vermunt (2006) have noted, it is difficult to determine which measure is the most appropriate to use across a range of applications.

Hierarchical clustering has been shown to produce similar subgroupings as k-plex analysis for some data examples and is less computationally intensive (Chin and Chignell 2008). Modularity has been proposed as an optimising method for partitioning dendrograms. Sometimes clustering and partitioning algorithms are combined in order to identify subgroups (e.g. Lin *et al.* 2008). However, little evaluative research has been carried out for determining which methods of unsupervised subgroup formation work well in subgroup analysis of social networks, and under what conditions.

### *2.2 Modelling evolution*

Online social networks evolve over time and much research has looked into the temporal aspects of social networks changing over time such as Snijders

*et al*. (2007) and Leydesdorff *et al*. (2008). Within social networks, subgroups of people may be found that vary in cohesiveness (Piper *et al*. 1983).

Many researchers have used data from online social networks and have used different methods to model the evolution of a social network, such as: social network analysis; group formation, clustering and partitioning; and mathematical modelling and graph theory.

Using social network analysis, Kumar *et al*. (2006) analysed the structure of Yahoo! 360 and Flickr networks and Barabasi *et al*. (2002) analysed scientific collaborations over time, to create a model of evolution and using simulation to test the model. Hu and Wang (2009) studied the evolutions of degree, network density, clustering coefficient, number of users, modularity and degree assortativity, in order to reveal the properties and evolutionary patterns of the Wealink online social network.

Other researchers have created models of evolution using group formation, clustering and partitioning methods. For example, Backstrom *et al*. (2006) developed a method for measuring movement of individuals between communities, examined the properties of membership of how groups formed and identified which communities grew over time. Tang *et al*. (2008) adopted a spectral clustering framework using temporal information to detect, identify and model community evolution in dynamic multi-mode networks, where both users' membership of community and their interactions were evolving. Cortes *et al*. (2002) proposed a bottom-up data structure to represent all small subgroups based on the "communities of interests" concept on each user in the dynamic network, and then updating all communities.

Mathematical modelling and graph theory can be used for modelling evolution of communities. Lin *et al*. (2008) detected communities using a non-matrix factorisation followed by implementation of an iterative algorithm. Lin *et al*. (2009) extended their earlier work by reformulating the problem in terms of maximum a-posteriori estimation, where they showed that there was a close relationship between their generative probabilistic model and the optimisation framework for solving the evolutionary clustering problem in Chakrabarti *et al*. (2006) and Lin *et al*. (2008). Sun *et al*. (2007) proposed a tool, named GraphScope, based on information theoretical principles, to monitor communities and their membership changes in a stream of graphs efficiently in an online and anytime fashion. Leskovec *et al*. (2007) created a mathematical theoretical model for characterising and describing the densification and shrinking diameter phenomenon in social networks over time.

In our earlier research (Chin and Chignell 2008) we proposed a similarity modelling approach to quantify changes in subgroup structure over time. In this approach cohesiveness over time is quantified in terms of the similarity of the subgroupings that are identified in different time periods.

### *2.3 Summary*

While cliques and k-plexes have some desirable properties, they are generally too computationally complex to be implemented on extremely large networks

that could potentially contain millions of nodes. Clustering and partitioning methods automate the subgrouping process and are relatively computationally efficient. Thus they will be preferred if they can provide adequate subgroupings. However, both clustering and partitioning require some kind of selection process to obtain particular subgroupings from the multiple groups created by the methods (multiple groupings due to nesting in the case of hierarchical cluster analysis and to different numbers of partitions in the case of partitioning methods).

Modularity has been proposed as an optimising method for cutting dendrograms into subgroupings. Researchers (e.g. Lin *et al*. 2008) sometimes use clustering and partitioning algorithms together in order to identify subgroups instead of relying on one alone. However, relatively little evaluative research has been carried out thus far on which methods of unsupervised subgroup formation work well in subgroup analysis of social networks, and under what conditions. For modelling evolution, few researchers have looked into the visualisation and activity of the most influential members in the network over time, concentrating instead on the statistics of the network.



## 3. Social Cohesion Analysis of Networks (SCAN): method, case studies and limitations

In this section, we briefly describe the SCAN method (Chin and Chignell 2008) for identifying cohesive subgroups. The SCAN method assumes that a social graph can be obtained from inferred social networks (e.g. from online community interactions), where the links are untyped (i.e. there are no associated semantics). The SCAN method consists of the following three steps.

### *3.1 Select: selecting potential members of cohesive subgroups*

A measure of network centrality is used to filter out members whose centrality falls below a designated cutoff value. This results in a subgraph of people who may possibly belong to subgroups.

### *3.2 Collect: grouping these potential members into subgroups*

Weighted average hierarchical clustering is used to cluster similar members from the selected subgraph into a set of hierarchically nested, non-overlapping clusters.

### *3.3 Choose: choosing cohesive subgroups that have a similar membership over time*

After finding the cohesive subgroups for a particular time period and a particular cutoff centrality, criteria such as similarity (e.g. Krantz and Tversky 1975) or modularity (Newman 2006) may be used to assess cohesiveness across different time periods (Chin and Chignell 2008).

Figure 1 shows an example of how subgroupings changed over time in the TorCamp Google group from Chin and Chignell (2008).

The movement of the members into clusters in different time periods is indicated by arrows, while the shading of the nodes indicates in which time period the corresponding person first appeared as a member of the subgrouping. This example demonstrates how the membership of subgroups may change markedly between time periods with only some of the subgroupings turning out to be cohesive in nature. In a second example, a set of vaccination videos and their associated comments on YouTube (Keelan *et al*. 2007) spanning a time period of 30 months starting in April 2006 were analysed. The anti-vaccination members in the anti-vaccination video conversations were found to form subgroups that generally contain only anti-vaccination members, whereas for pro-vaccination videos, pro-vaccination and anti-vaccination members were combined together based on the conversations that occurred around the videos as shown in Figure 2.



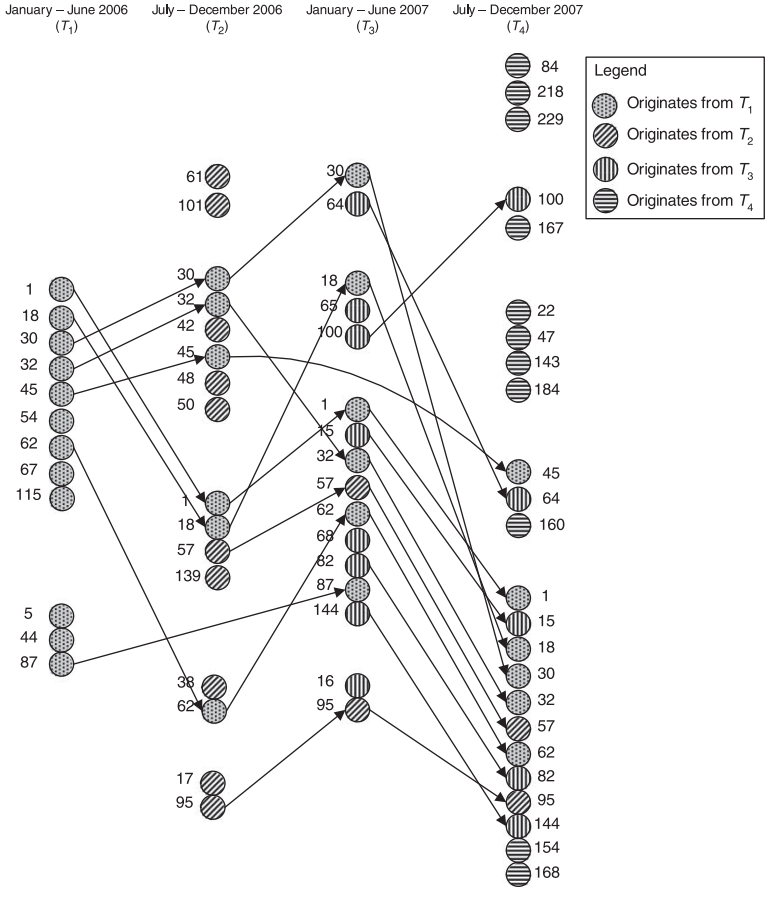


Figure 1. Cohesive subgroupings from the SCAN method in the TorCamp Google group from 2006 to 2007, where the members have been anonymised for privacy.

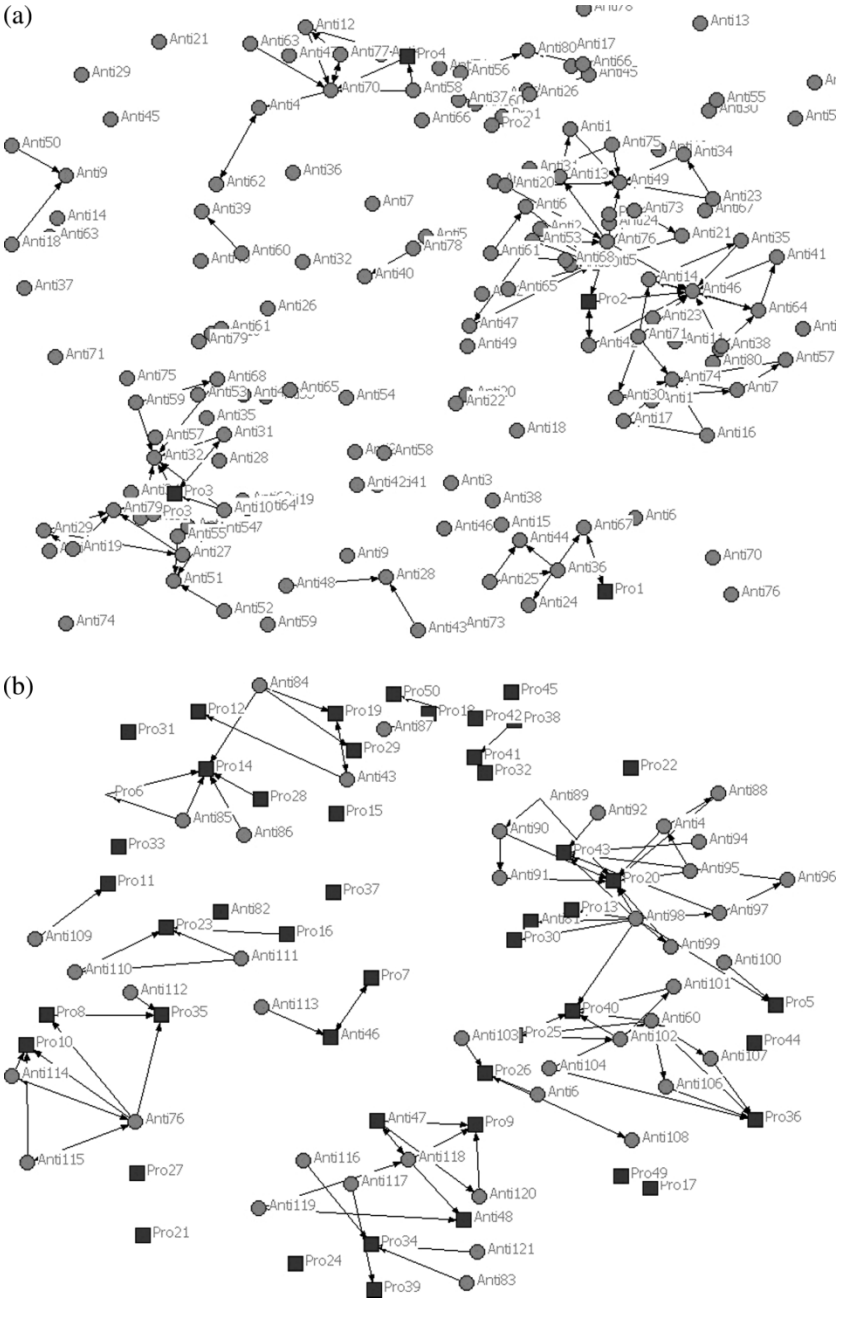


Figure 2. Comment discussions network from (a) YouTube anti-vaccination videos and from (b) YouTube pro-vaccination videos (Chin *et al*. 2010, published by Wiley and used with permission). The squares indicate members who have a pro-vaccination view and the circles show those that have an anti-vaccination view.

It was impossible in this case to identify cohesive subgroups of only anti-vaccination, or only pro-vaccination people based on linking structure alone (content analysis was also needed).

## 4. Data-Intensive Socially Similar Evolving Community Tracker (DISSECT)

Chin and Chignell (2010) proposed the DISSECT method as an enhanced method of identifying cohesive subgroups that addressed the following limitations of the SCAN method:

(1) The SCAN method only focuses on betweenness centrality; other centrality measures may be useful.
(2) The time periods used in the SCAN method are defined ad hoc as a matter of convenience, without any systematic evaluation.
(3) The SCAN method fails if semantic properties determine subgroup membership.

The DISSECT method consists of the following steps.

### *4.1 D1: find the initial time periods for analysis*

We divide the dataset into time periods (which may be equal or unequal in duration) in order to track subgroups in the network over time. Time periods should be long enough that there is enough data to distinguish potential subgroups, and there should be a sufficient number of them to estimate cohesion over time. In the absence of well-defined methodologies for selecting time periods, in the version of the DISSECT method used by the authors as of this writing, time periods are chosen through subjective judgement, based on the timeframe of the data collected, the number of phases of evolution to be examined, and the properties of the network.

### *4.2 D2: label subgroups of people from the network dataset using content analysis and semantic properties (optional)*

For each time period defined from the previous step, content analysis may be performed to label the links and/or individuals within the network. Techniques such as noun-phrase analysis (Gruzd and Haythornthwaite 2008) and other natural language-processing techniques that analyse content of the posts (Gruzd and Haythornthwaite 2008) can be used for classifying the links and the nodes. Content analysis is useful here to eliminate those links and individuals which have no relation to the subgroup (e.g. comments that are considered as spam or have no relation to the topic of the conversation). While content analysis is a useful supplement to the structural methods described in this paper, it is not always feasible (e.g. for data where there is little, if any, associated text).

### *4.3 D3: select the possible members of known subgroups that you want to track (from previous step) using Select from the Social Cohesion Analysis of Networks (SCAN) method*

While betweenness centrality appears to be a useful filter for screening potential subgroup members (Chin and Chignell 2008), other centrality measures such as degree and closeness centrality may also be used (Chin and Chignell 2010). Degree centrality may be a good default measure with which to screen potential subgroup members because it deals with direct interactions where the ties have stronger bonds that indicate stronger cohesion, and also because it has the lowest computational complexity compared to other centrality measures (Chin 2009).

### *4.4 D4: perform clustering of snapshots in time of known subgroups of people using the Collect step from the Social Cohesion Analysis of Networks (SCAN) method*

This step is identical to the Collect step from the SCAN method, with the provision that other cluster methods may be used in addition to, or instead of, weighted average hierarchical clustering.



### *4.5 D5: repeat steps D3 and D4 for different values of centrality*

Since there is as yet no known "best" or most appropriate centrality cutoff value for selecting potential subgroup members (and even if such ideal cutoff values exist they will differ for different centrality measures and may also differ within the same measure for different types of social networks/ applications) a search process may be used to identify particular cutoff values that lead to identification of the most cohesive groups. Different types of search strategy may be employed (some more exhaustive than others) but they would involve repetition of steps D3 and D4 over a range of values of centrality, with the goal of maximising the cohesiveness (self-similarity over time) of the obtained subgroupings. In this case, the similarity measures suggested by Chin and Chignell (2008) are recommended, although other similar measures proposed in the literature (e.g. Sokal and Sneath 1973) may also be used.

### *4.6 D6: select and characterise the obtained subgroupings*

Select and characterise the obtained subgroupings in terms of their cohesiveness and their behaviour over time (e.g. growth, shrinkage, aggregation, etc.). If the members of the network were labelled through content analysis or other means, then the purpose or relationship implied in cohesive subgroups may be interpreted based on the pattern of labels observed in their members.

The search process briefly described above may also be expanded by search over different definitions of time periods, as well as different centrality measures and centrality cutoff values. For instance, both the starting points

and durations of time periods could be varied. However, it remains to be seen how beneficial the search process envisioned here will be in improving the identification of cohesive subgroups. It seems likely that strongly cohesive subgroups that remain intact over a sustained period of time should be "easy to find" with a range of time period definitions and centrality measurement and filtering strategies. In contrast the search process envisioned above might be useful in finding more ephemeral subgroups that exist for only short periods of time and for tracking, in more detail, evolution in subgroupings.

## 5. Application of the Data-Intensive Socially Similar Evolving Community Tracker (DISSECT) method: Nokia Friend View

In this section, we apply the DISSECT method to a large mobile social networking application called Nokia Friend View. This dataset differs from readily available public datasets of online social media like Flickr, Twitter and YouTube in that it is a corporate dataset involving data collected from a mobile social network application. Another feature of this dataset is that it captured the complete lifecycle of the service from beginning to end (in contrast to most other datasets that include data from only a portion of the time that the target service is running).



### *5.1 Overview of Nokia Friend View*

Friend View is a location-enhanced microblogging application and service from Nokia Research Center that was launched in Nokia Beta Labs in November 2008, and was discontinued at the end of September 2009. It allowed users to post messages about their status with other friends from GPS-enabled Nokia S60 phones or from the web. Two users become friends if one user sends an e-mail to the other, or if one sends a friend request which is accepted by the other user. Users can send two types of messages, an "I am here" message which posts the current location coordinates of where the user is from the phone (without writing a message), or a message which includes a text note. Users can see their friends' locations on a map which is navigable and zoomable, along with their messages, and comments from others. Figure 3 (a) shows the interface of Nokia Friend View with the friends' status messages and location, while figure 3 (b) shows example comments.

### *5.2 Dataset*

We obtained a dataset from Nokia Friend View from 1 November 2008 to 30 September 2009. From this dataset, we extracted two social networks, the interaction network (or comment network) and the friend network. The interaction network is a directed graph $G(V,E)$, where $V$ represents the set of Friend View users and $E$ represents the set of comments to status messages posted by users, where a directed edge or tie exists between users A and B if user A posted a comment in response to user B's status message and the

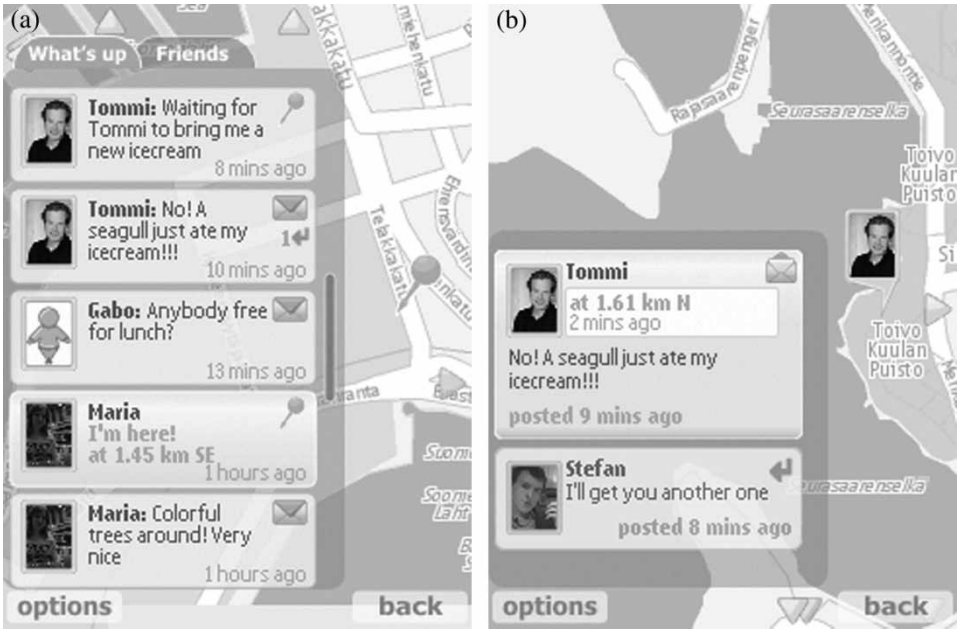


Figure 3. (a) Status updates from friends and their locations and (b) comments to a status message from Nokia Friend View.

edge weight $w$ indicates the number of comments that A made in response to all of B's status messages. The friend network is an undirected graph $G(V,E)$, where $V$ represents the set of Friend View users and $E$ represents the set of friend relations similar to Java *et al.* (2007) for the Twitter friend network, except that Friend View friend relations are undirected (since when a person accepts a friend request, the two people automatically become friends), compared to Twitter where the friend relations (i.e. followers in Twitter) are directed (user u can be a follower of user v, but not necessarily vice versa). For both networks nodes without edges were removed.

A total of 34,980 users were registered to Friend View, with 8,443 users that had at least one friend, and a total of 20,873 friendship links were created. In Friend View, a total of 62,736 status messages were posted by 16,176 users, providing an average of 3.88 status messages per user and 22,251 comments from 2,283 users, providing an average of 2.38 comments per status message and 9.75 comments per user. The Friend View comment network consisted of 2,898 users who made at least one comment.

### *5.3 Initial time periods for analysis*

Figure 4 (a) shows the number of new users and new friends for every month over the course of the trial, and figure 4 (b) shows the total number of users and friends.

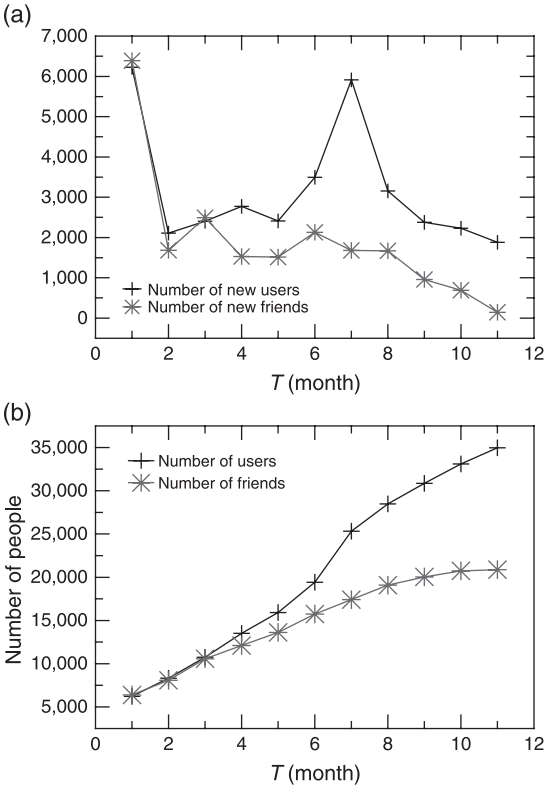


Figure 4. Time evolution of (a) the number of new users and new friends and (b) the total number of users and friends in the Friend View network.

Based on the pattern of data shown in figure 4, five time periods were identified, as shown in Table 1.

### *5.4 Finding subgroups in the evolution of Nokia Friend View*

Since we did not obtain the content of the status messages and comments, no content analysis was performed. Due to the anonymised nature of the data there were also no labelling information attached to the people in the network. The DISSECT method was applied as follows. For each time period in the Select step, we first selected the cutoff points for normalised betweenness and degree centrality (as suggested by Chin and Chignell 2010) in order to select the possible subgroup members. In the Collect step, we used weighted average hierarchical clustering to find non-overlapping subgroups.

**5.4.1 Centrality cutoff points.** To determine the betweenness and degree centrality cutoff points, the frequency distributions for betweenness and degree centrality were inspected for each time period. Betweenness centrality

Table 1. Time periods chosen for analysis in the evolution of Nokia Friend View.

| Time period | Time range | Growth/decline within time period |
|---|---|---|
| $T_1$ | 1–30 November 2008 | Beginning, initial growth, mostly early technology adopters |
| $T_2$ | 1 December 2008 to 28 February 2009 | Early growth |
| $T_3$ | 1 March to 31 May, 2009 | Rapid growth, many new users trying the service |
| $T_4$ | 1 June to 31 July, 2009 | Rapid slowing of growth |
| $T_5$ | 1 August to 30 September, 2009 | Further, but more gradual slowing of growth |

for each member was computed using Freeman's (1979) betweenness centrality measure as shown in Equation (1).

$$C_B(n_i) = \sum_{\substack{j \\ j < k}} \sum_{k} \frac{g_{jk}(n_i)}{g_{jk}} \quad i \neq j \neq k \tag{1}$$

where $C_B(n_i)$ is the betweenness centrality, and $g_{jk}(n_i)$ is the number of geodesics linking the two nodes $j$ and $k$ that contain node $i$. A geodesic is the number of links in the shortest possible walk from one node to another. Normalising the result between 0 and 1 yields:

$$C'_B(n_i) = \frac{C_B(n_i)}{[(g-1)(g-2)]/2} \tag{2}$$

where $C'_B(n_i)$ is the normalised betweenness centrality, $C_B(n_i)$ is the betweenness centrality, and $g$ is the total number of nodes in the network. The degree centrality can be calculated by the following formula (normalised between 0 and 1) as

$$C'_D(n_i) = \frac{d_i(n_i)}{g-1} \tag{3}$$

where $C'_D(n_i)$ is the normalised degree centrality for the $i$th node $n_i$, $d_i(n_i)$ is the degree for the $i$th node $n_i$, and $g$ is the total number of nodes in the network.

Figure 5 shows the betweenness centrality frequency distribution for the first time period $T_1$. Note that log scales are used in Figures 5 and 6 to represent the frequencies, in order to make it easier to visualise the tail of the distribution. This tends to result in a cutoff value that focuses subsequent analysis on a relatively small set of active people.

From this distribution, the cutoff was chosen to be the highest value (an educated guess) in the distribution that appeared to separate a smaller group of higher centrality members from a larger group of people with lower centralities (Chin and Chignell 2008). In this case a betweenness centrality cutoff of 0.2 was chosen for $T_1$. Based on inspection of the corresponding

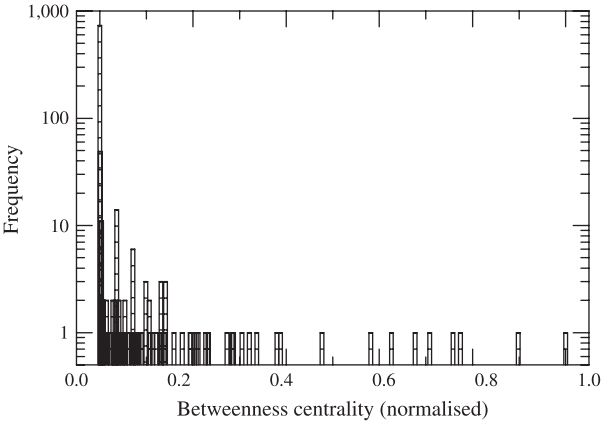


Figure 5. Betweenness centrality distribution for the first time period in Nokia Friend View.



distributions for $T_2$ through $T_5$, a betweenness centrality cutoff of 0.2 was also chosen for those time periods. For degree centrality, the same process was also used and a cutoff of 0.005 was chosen (the corresponding frequency distribution for $T_1$ is shown in Figure 6).

Other values of betweenness and degree centrality selected for the search process (step D5 listed above) were a betweenness centrality cutoff of 0.1 and a degree centrality cutoff of 0.03. Higher cutoff values were not considered in this case as they would have resulted in a very small number of people being considered for possible subgroup membership.

**5.4.2 Similarity analysis and visualisation for evolution of cohesive subgroups.** From the subgroups identified in the Collect step of the DISSECT method, cohesion is examined within subgroups of three or more people across the two time periods being compared. The similarity measure defined in Equation (4) tracks how well the subgroups in one of the time periods are

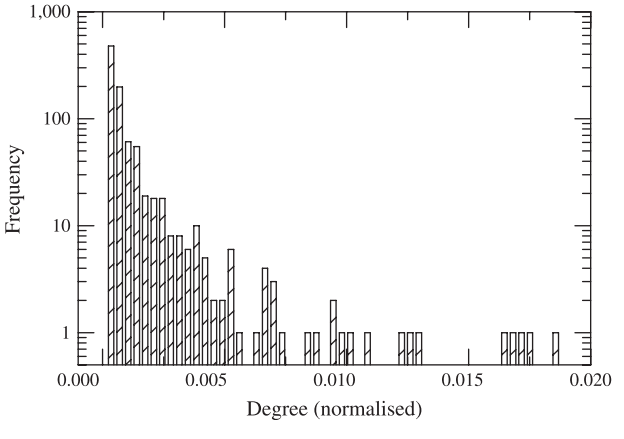


Figure 6. Degree centrality distribution for the first time period in Nokia Friend View.

matched by the subgroups in the other time period. The similarity measure looks at all the possible pairwise relationships between members of the subgroups in each time period, and compares that value with the number of pairs within subgroups at the second time period (note that for both time periods, only pairs within subgroups are considered). The similarity can then be calculated using the following formula:

$$\mathrm{SIM}_{T_1-T_2} = \frac{N(\text{common pairs in subgroups from } T_1 \text{ in } T_2)}{N(\text{pairs that exist in subgroups in } T_1)} \tag{4}$$

where $N()$ is a cardinality operator that counts the number of pairs within subgroups. Typically this measure uses the earlier time interval as the baseline for comparison. Essentially the measure divides the number of pairings in one time period by the number of subgroup pairings in the other. However, the identity of subgroup members is also important. Thus, subgroupings are cohesive only to the extent that the same members are involved in the pairings. The denominator of the equation is the base against which the similarity comparison is made, and only pairings that involve members of subgroups in the base time period are included in calculating the number of matching pairings in the other time period (in the numerator of the similarity equation). To see how the equation works, consider the case where there are two subgroups (a, b and c) and (d, e and f) in $T_1$, and one subgroup (a, b, c, d, e and f) in $T_2$. There are six pairings within the two subgroups in $T_1$ ((a, b), (b, c), (a, c), (d, e), (e, f) and (d, f)). In $T_2$, there are 15 pairings within the merged subgroup, but only six of them match the pairings seen in $T_1$. Thus the similarity in this case is $6/6 = 1$. Note that if $T_2$ had been used as the baseline in this case, then the similarity would have been 6/15.

Table 2 shows the results of the similarity analysis using the designated betweenness and degree centrality cutoff points.

Table 2 shows that cohesiveness, as measured by similarity (using the similarity measure in Equation (4), was generally higher in the middle three time periods ($T_2$ through $T_4$). It can also be seen that the answer to the question of which cutoff values to use depends on which time periods are being compared. In comparing $T_2$ and $T_3$ a liberal cutoff that included more people led to the greatest cohesiveness in subgroups identified, whereas for $T_3$ vs. $T_4$ a more conservative (restrictive) cutoff produced the highest similarity (cohesiveness) measure. The cutoff centralities chosen for all the time periods based on highest similarity are indicated in Table 3.

Table 2. Cutoff values for betweenness and degree centrality, and similarities between consecutive time periods.

| Cutoff point | SIM($T_1$, $T_2$) | SIM($T_2$, $T_3$) | SIM($T_3$, $T_4$) | SIM($T_4$, $T_5$) |
|---|---|---|---|---|
| Betweenness = 0.1, degree = 0.003 | 0.11 | 0.63 | 0.62 | 0.47 |
| Betweenness = 0.1, degree = 0.005 | 0.09 | 0.6 | 0.6 | 0.47 |
| Betweenness = 0.2, degree = 0.005 | 0.2 | 0.47 | 0.77 | 0.47 |

Table 3. Centrality cutoffs for each time period based from DISSECT.

| Time period | Betweenness centrality cutoff | Degree centrality cutoff |
|---|---|---|
| $T_1$ | 0.2 | 0.005 |
| $T_2$ | 0.1 | 0.003 |
| $T_3$ | 0.1 | 0.003 |
| $T_4$ | 0.2 | 0.005 |
| $T_5$ | 0.1 | 0.003 |

Figure 7 shows the subgroups, their members and the connectivity between members for each of the time periods that come from the result of the similarity analysis above, using NetDraw and the spring embedding layout algorithm (Borgatti 2002). We can see that subgroups members tend to be tightly connected with each other, suggesting strong cohesion.

Figure 8 shows the movement of people between the Friend View subgroups based on the similarity analysis. The members have been grouped together based on the subgroups found from the DISSECT method, with each member having a shape that corresponds to the time period where that member is first found as indicated in the legend. This differs from figure 1 where we indicate at which period the member originally was found in the cohesive subgroup. The arrows indicate how the member or subgroup moved from one subgroup in the previous period to the next period. If there is a space between two adjacent members in the same time period, this indicates that the two members are not part of a cohesive subgroup. From this figure, it can be seen that there are two subgroups at $T_1$ (3, 32, 37, 7, 9, 34 and 16) and (5, 21, 17, 24, 28, 11, 32, 33 and 13). The second of these groups started splitting up in $T_2$ and had completely dispersed by $T_3$. Across the entire trial only three people stayed together across all the time periods (7, 9 and 37) and a further two people paired up in $T_2$ and then stayed together for the remainder of the trial. Other people moved around between the time periods, sometimes moving to a different subgroup and sometimes appearing to drop out and become singletons. Since the data were anonymised, it is not possible to identify who participated in the cohesive subgroups. However, ID numbers were assigned in the order of adoption (beginning with the number 1) so the low ID numbers shown in figure 8 (all below 50) indicated that the people who participated in subgroups were the early adopters in the trial. In the Friend View trial the early adopters could be characterised as either external people who always keep track of what Nokia is doing, or employees or researchers at Nokia.

The subgroups shown in figure 8 can be characterised in terms of how persistent they were. The most persistent group (37, 7 and 9) is indicated as a bold rectangle around the subgroup and remained together through all the time periods. Semi-persistent groups may be defined as groups that stayed together for some portion of the trial. They included the following groupings (6 and 26) that persist from $T_2$ to $T_5$, and (5 and 21) from $T_1$ to $T_2$ as shown

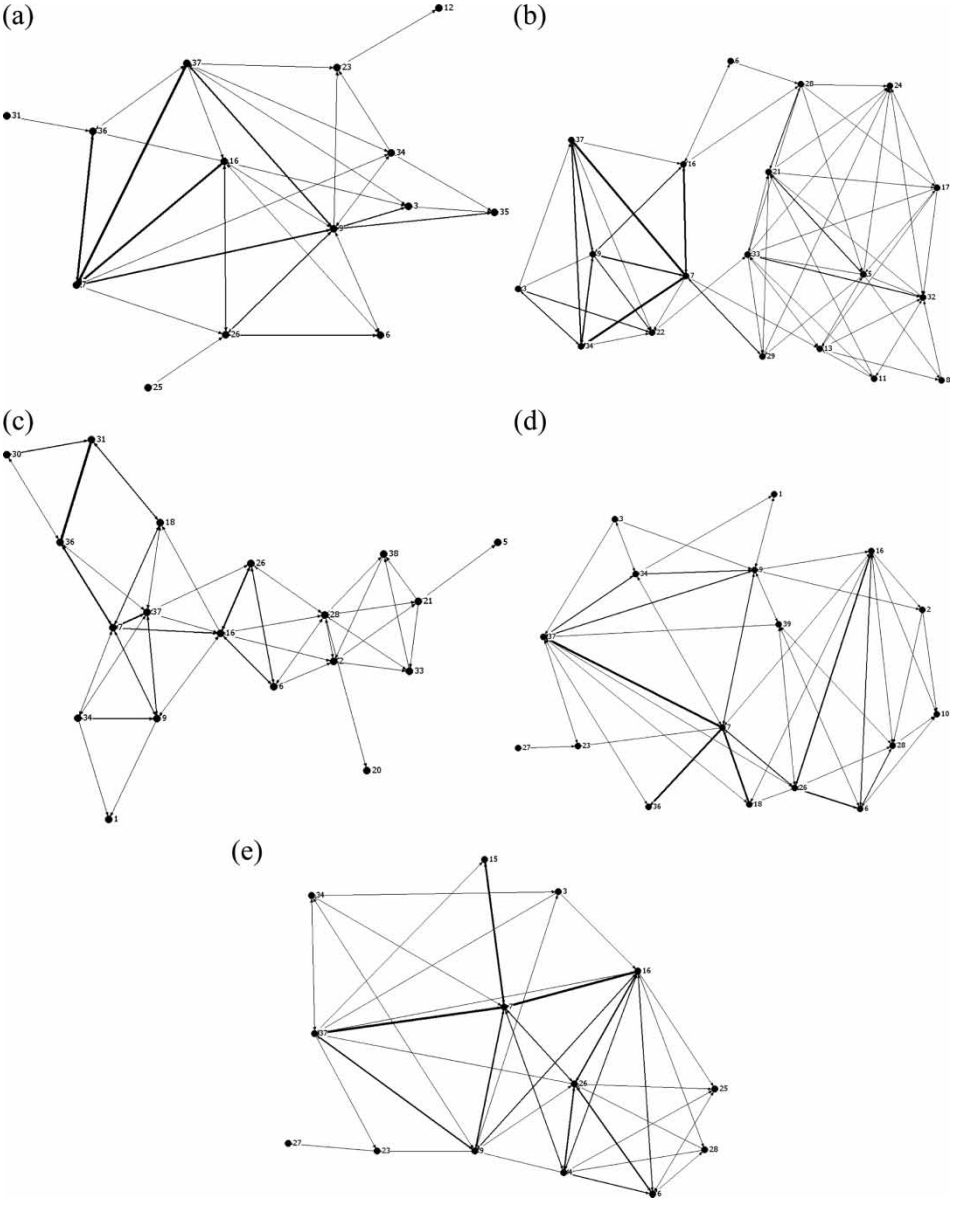


Figure 7. Subgroups and their members identified for time periods $T_1$ (a), $T_2$ (b), $T_3$ (c), $T_4$ (d) and $T_5$ (e) for the Friend View data.

by the dashed line around the subgroup. Temporal groups have persistent members in one period and then divide into different groups in the next time period. They are indicated in figure 8 with a dotted line around the subgroup. An example temporal group is (5, 21, 17, 24, 28, 11, 32, 33 and 13) which formed in $T_1$ but then divided into different groups in later time

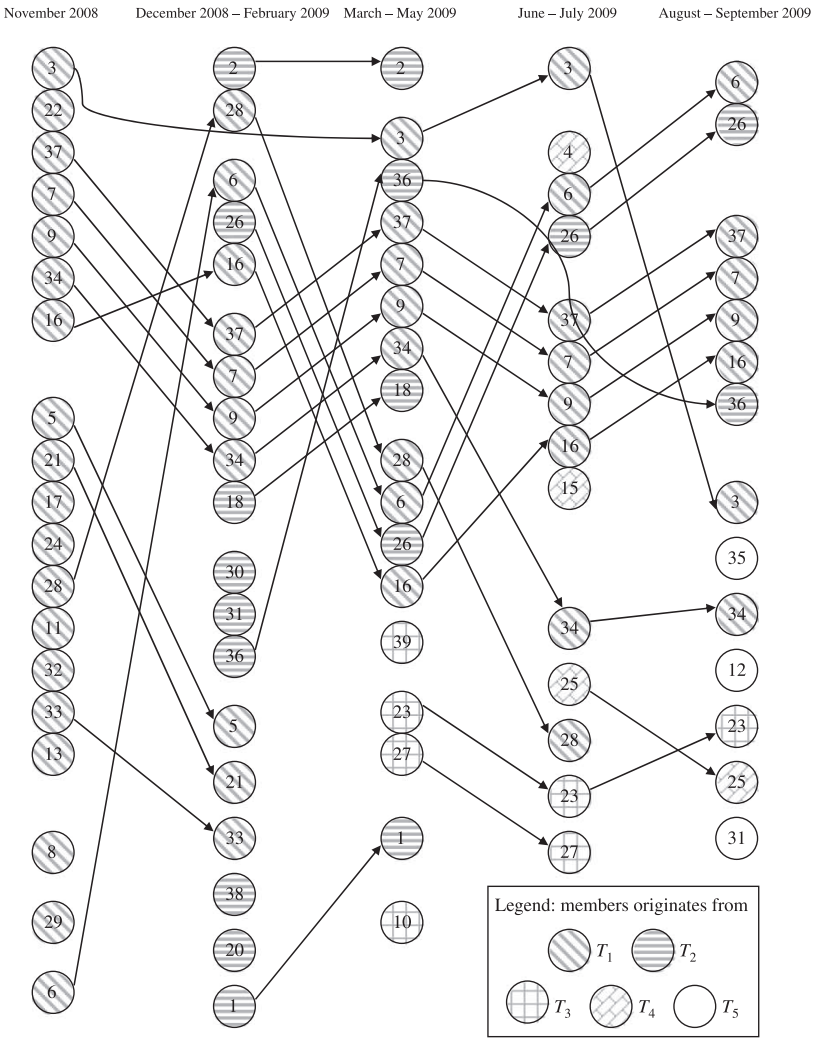


Figure 8. Visualisation of the evolution of members and subgroups in different time periods in Friend View.

periods. Ephemeral groups may be defined as having members that are together in only one period such as (3 and 22) in $T_1$, and (30 and 31) in $T_2$. Based on this analysis, the different types of subgroups observed in the Friend View trial are listed in Table 4 according to their persistence. Note that since this analysis extends over five time periods, a particular person may participate in more than one type of subgroup at non-overlapping time durations.

There was very little evidence of new subgroups (involving new members) being formed after $T_2$. Thus cohesive subgrouping activity was driven by the early adopters, and with most of the cohesive subgroup formation being done

Table 4. Enumeration of all the different type of subgroups and their members discovered from the DISSECT method in Nokia Friend View.

| Type of subgroup | Subgroup members | Persistence |
|---|---|---|
| Persistent core subgroup | (37, 7, 9) | $T_1$–$T_5$ |
| Semi-persistent group | (6, 26) | $T_2$–$T_5$ |
| | (5, 21) | $T_1$–$T_2$ |
| Temporal group | (5, 21, 17, 24, 28, 11, 32, 33, 13) | $T_1$ |
| | (2, 28) | $T_2$ |
| Ephemeral group | (3, 22) | $T_1$ |
| | (30, 31) | $T_2$ |

in the early part of the trial (although with some movement of members between and out of subgroups in later stages of the trial).



### *5.5 Cohesion and messaging activity*

The next analysis sought to examine properties of the subgroups in terms of message activity and centrality of their members. Table 5 reports the correlations (calculated separately for each of the five time periods) between the number of status messages posted for each time period, and various measures.

As expected, there were significant correlations between the number of status messages posted by the members and the centrality measures. In addition, there were correlations of similar size between number of messages and both the number of unique users that the active users made comments to (outdegree), and the number of unique users that made comments to the active users (indegree). The rightmost column of table 5 (subgroup membership) shows that, of those people selected using the centrality cutoff, people in the identified subgroups sent more messages. In this sample the correlations between betweenness centrality and message activity were of the same order as the correlations between degree centrality and message activity.

Table 5. Correlations between the number of status messages posted in each period on the one hand, with degree centrality, indegree and outdegree, betweenness centrality and subgroup membership, respectively.

| Time period | Degree centrality | Indegree | Outdegree | Betweenness centrality | Subgroup membership |
|---|---|---|---|---|---|
| $T_1$ | 0.548 | 0.483 | 0.559 | 0.496 | 0.380 |
| $T_2$ | 0.509 | 0.461 | 0.516 | 0.422 | 0.478 |
| $T_3$ | 0.551 | 0.463 | 0.565 | 0.630 | 0.519 |
| $T_4$ | 0.435 | 0.411 | 0.415 | 0.386 | 0.346 |
| $T_5$ | 0.681 | 0.599 | 0.683 | 0.562 | 0.599 |

## 6. Discussion

In this section, we discuss about what we have learned from applying the DISSECT method to Nokia Friend View, compared with the earlier SCAN method which was applied to the TorCamp Google group and YouTube vaccination videos.

### *6.1 Comparison of results of Social Cohesion Analysis of Networks (SCAN) and Data-Intensive Socially Similar Evolving Community Tracker (DISSECT) methods*

Table 6 shows the comparison between all three networks that we analysed (TorCamp Google group, YouTube vaccination videos and Nokia Friend View) in terms of their size, network properties, network evolution and cohesive subgroups found. Overall, Friend View had the most number of network members (2,898) but had the shortest time duration analysed (11 months), whereas YouTube vaccination videos had the least number of network members (177 for anti-vaccination network and 222 for pro-vaccination network) but the longest time duration analysed (2.5 years). The TorCamp and YouTube datasets were snapshots of the original networks, whereas the Friend View data were a complete network that contained all the data from the trial.

Network density was the highest in the TorCamp Google group network and the lowest in the Friend View network. The TorCamp Google group was a close-knit community that had many technology-centred discussions where the users in the group are all TorCamp members that meet in person in monthly meetings, resulting in higher network density. On the other hand, people in the YouTube vaccination video conversations (even though they have many heated debates through the conversations) and Friend View users may not have been as closely knit (it might be speculated that many did not meet other people in person, resulting in lower network density). The TorCamp Google group was created as an online extension of the TorCamp group that meets regularly every month, whereas with the YouTube vaccination and Friend View users, the service was not created as a result of a physical meetup. The YouTube vaccination video networks had the smallest average shortest path (less than 2) compared to the TorCamp Google group (around 3) and Friend View (around 5). Most people in the YouTube dataset had a closer connection with each other as compared to people in the TorCamp and Friend View networks, which was also reflected by the network diameter where YouTube was the smallest, followed by TorCamp, with Friend View having the largest network diameter.

Examining the network evolution, we see that in the TorCamp Google group, there is a gradual increase in membership over time, whereas in Friend View there was slow growth, followed by rapid growth, rapid decline in growth and then gradual decline in growth, as was also found in the Wealink online social network (Hu and Wang 2009). Comparing the cohesive subgroups found, Friend View had persistent core subgroups that stay within all the time periods, whereas the TorCamp Google group only had a core

Table 6. Comparisons between the various social networks of TorCamp, YouTube vaccination videos and Nokia Friend View.

| Network | Content/type | Size | Period | Network density | Average shortest path | Network diameter | Network evolution | Method used to find cohesive subgroups | Cohesive subgroups found |
|---|---|---|---|---|---|---|---|---|---|
| TorCamp | Tech/Google group | 228 | 2 years | 0.042 | 2.956 | 9 | Gradual increase in members over time | SCAN | Core group of members in 2007 which remain in the same subgroup over time |
| YouTube | Vaccination and health/video | 177 (anti), 222 (pro) | 2.5 years | 0.019 (anti), 0.015 (pro) | 1.647 (anti), 1.344 (pro) | 5 (anti), 4 (pro) | N/A | SCAN (omitting Choose step) | Anti-vaccination members are in the subgroups of anti-vaccination network, pro- and anti-vaccination members are in the subgroups of pro-vaccination network |
| Friend View | Mobile location-based social network/ microblog | 2,898 | 11 months | 0.00032 | 5.136 | 16 | S-curve (initial growth, then rapid growth, rapid decline then gradual decline) | DISSECT | Persistent core groups, semi-persistent, temporal and ephemeral subgroups. Persistent core groups remain the most cohesive over time and have highest comment indegree and outdegree |

subgroup in the second half of the time analysed (2007). Friend View also had semi-persistent, temporal and ephemeral subgroups. In contrast to the TorCamp and Friend View case studies, it was not possible to identify homogeneous cohesive subgroups of purely pro-vaccination members only in the pro-vaccination YouTube video conversation network, due to the mixing of people with very different affiliations and beliefs in the same conversational interaction (from which the social network was inferred).

### *6.2 Limitations*

Since the Friend View dataset did not contain the actual content of the status messages and the comments, we could not perform a content analysis in order to more accurately filter out other Friend View users that would not be part of cohesive subgroups. Neither was information available to label people and thereby interpret the relationships between subgroup members, nor the overall purpose of the subgroup. No analysis was done concerning different sized durations of time periods. In future work it may be beneficial to use similarity measures that take into account variable time windows and all possible combinations of time windows (rather than pairwise consecutive time windows only). Probabilistic stochastic models (Snijders *et al*. 2007), sliding time windows (Moody *et al*. 2005, Falkowski *et al*. 2006) and time graphs and burst analysis (Kumar *et al*. 2006, Backstrom *et al*. 2006) may be used to select possible time periods.

Relatively few subgroups were discovered in the Friend View data. Based on our experience in dealing with a number of datasets we have found that betweenness centrality seems to be a better indicator of activity in a relative small and densely connected network (such as in the TorCamp Google group). In very large networks (such as the one formed in the Friend View trial) where there are only isolated pockets of people who know each other, degree centrality (with its emphasis on who people are directly connected to) may be a better reflection of activity and a better filter for selecting people who are likely to belong to subgroups that might be cohesive. Thus for large networks of thousands or more nodes it is recommended that degree centrality be used in preference to betweenness centrality.

Another feature of the Friend View analysis was that only a few relatively small subgroups were discovered in spite of the thousands of people who participated in the trial. This is most likely because a relatively stringent centrality cutoff criterion was used. It would be useful to have algorithmic methods of centrality cutoff selection that supplement the visual inspection of the frequency distribution used in the research reported here.

## 7. Conclusions and future work

In this paper, we presented a method for finding and tracking community in social media using similarity-based cohesive subgroups. We used this method to propose a framework called DISSECT for tracking community evolution

in an online community. This revised framework was a response to the limitations of our previous method, the SCAN method, for finding cohesive subgroups in online interactions. The new DISSECT framework is designed to be a step-by-step guide on how to track the evolution of community members. We tested our DISSECT framework with the Nokia Friend View mobile social network and discovered that DISSECT found different types of cohesive subgroups based on persistence in time. Those types were: persistent core groups, semi-persistent groups, temporal groups and ephemeral groups. We compared the persistent core groups with the original post and comment statistics and discovered that the most active members were the ones who were part of the persistent core groups found from the DISSECT method.

In future work, it would be beneficial to examine the value of the DISSECT approach with a variety of different datasets. Relevant research issues include the choice of time periods using various statistical and time-based models, developing other measures for similarity assessment and using content analysis to determine semantic properties that govern or explain subgroup membership. The DISSECT framework may also be applied to other online social networks such as Twitter and Facebook to determine whether the similar cohesive subgroups can be used for improving friend recommendations, an important component in growing social networks.

Perhaps most importantly, the research in this paper is a step towards community trackers, a new class of online tools that may in future index and query cohesiveness and community based on online collaboration and communication. Like search engines and recommender systems before them, it is envisioned that community tracking systems will have a revolutionary impact. In the case of community tracking, this impact will be on inferring detailed community structure from diffuse patterns of online collaboration and communication. Just as search engines use text retrieval and relevance ranking algorithms to identify documents and Web pages relevant to a search query, community trackers will be able to answer queries about community structures using methods such as social network analysis and content analysis.



**Acknowledgements**

The authors would like to thank James Reilly and the Nokia Friend View team for providing the Friend View anonymised dataset for our analysis.

**References**

E. Adar, L. Zhang, L. Admic and R. Lukose, "Implicit structure and the dynamics of blogspace", in *Workshop on the Weblogging Ecosystem, 13th International World Wide Web Conference*, 18 May 2004, New York, pp. 44–54, 2004.

R. Alba, "A graph-theoretic definition of a sociometric clique", *Journal of Mathematical Sociology*, 3(1), pp. 113–126, 1973.

L. Backstrom, D. Huttenlocher, J. Kleinberg and X. Lan, "Group formation in large social networks: Membership, growth, and evolution", *Proceedings of the 12th ACM SIGKDD International Conference on*

*Knowledge Discovery and Data Mining*, Philadelphia, PA, USA, 20–23 August 2006. KDD '06. New York: ACM, pp. 44–54, 2006.

B. Balasundaram, S. Butenko, I.V. Hicks and S. Sachdeva, 2008. Clique relaxations in social network analysis: The maximum k-plex problem. Available online at: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.60.4294&rep=rep1&type=pdf (accessed 15 June 2010).

A. Barabasi, H. Jeong, Z. Neda, E. Ravasz, A. Schubert and T. Vicsek, "Evolution of the social network of scientific collaborations", *Physica A: Statistical Mechanics and its Applications*, 311(3–4), pp. 590–614, 2002.

S. Borgatti, 2002. Netdraw. Available online at: http://www.analytictech.com/downloadnd.htm (accessed 12 January 2010).

S.P. Borgatti, M.G. Everett and P.R. Shirey, "LS sets, lambda sets and other cohesive subsets", *Social Networks*, 12, pp. 337–357, 1990.

S. Brin and L. Page, "The anatomy of a large-scale hypertextual Web search engine", *Computer Networks and ISDN Systems*, 30(1–7), pp. 107–117, 1998.

D. Chakrabarti, R. Kumar and A. Tomkins, "Evolutionary clustering", in *Proceedings of the 12th ACM SIGKDD international Conference on Knowledge Discovery and Data Mining*, Philadelphia, PA, USA, New York: ACM Press, pp. 554–560, 2006.

A. Chin, *Social cohesion analysis of networks: A method for finding cohesive subgroups in social hypertext*, PhD Dissertation, University of Toronto, Toronto, 2009.

A. Chin and M. Chignell, "A social hypertext model for finding community in blogs", in *Proceedings of the Seventeenth Conference on Hypertext and Hypermedia*, Odense, Denmark, 22–25 August 2006, HYPERTEXT '06. New York: ACM, pp. 11–22, 2006.

A. Chin and M. Chignell, "Identifying subcommunities using cohesive subgroups in social hypertext", in *Proceedings of the Eighteenth Conference on Hypertext and Hypermedia*, Manchester, UK, 10–12 September 2007. HT '07. New York: ACM, pp. 175–178, 2007.

A. Chin and M. Chignell, "Automatic detection of cohesive subgroups within social hypertext: A heuristic approach", *New Review in Hypermedia and Multimedia*, 14(1), pp. 121–143, 2008.

A. Chin and M. Chignell, "DISSECT: Data-intensive socially similar evolving community tracker", in *Computational Social Network Analysis: Trends, Tools and Research Advances, Series: Computer Communications and Networks*, A. Abraham, A.-E. Hassanien, and V. Snášel (Eds), London: Springer, pp. 81–105, 2010.

A. Chin, J. Keelan, V. Pavri-Garcia, G. Tomlinson, K. Wilson and M. Chignell, "Automated delineation of subgroups in web video: A medical activism case study", *Journal of Computer-Mediated Communication*, 15(3), pp. 447–464, 2010.

A. Clauset, "Finding local community structure in networks", *Physical Review E*, 72(2), pp. 26132–26139, 2005.

C. Cortes, D. Pregibon and C. Volinsky, "Communities of interest", *Intelligent Data Analysis*, 6(3), pp. 211–219, 2002.

L. Danon, A. Díaz-Guilera, J. Duch and A. Arenas, "Comparing community structure identification", *Journal of Statistical Mechanics: Theory and Experiment*, 2005, pp. P09008–P09018, 2005.

J. Dixon, *Towards an understanding of the implications of boundary changes -with emphasis on community of interest*, Technical report, Armidale: University of New England, 1981.

L. Donetti and M. Munoz, "Detecting network communities: A new systematic and efficient algorithm", *Journal of Statistical Mechanics: Theory and Experiment*, 2004, pp. P10012–P10020, 2004.

N. Du, B. Wu, X. Pei, B. Wang and L. Xu, "Community detection in large-scale social networks", in *Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 Workshop on Web Mining and Social Network Analysis*, San Jose, California, 12–17 August 2007. WebKDD/SNA-KDD '07. New York: ACM, pp. 16–25, 2007.

J. Duch and A. Arenas, "Community detection in complex networks using extremal optimization", *Physical Review E*, 72(2), pp. 27104–27108, 2005.

B.S. Everitt, *Cluster Analysis*, London: Halsted Press, 1974.

T. Falkowski, J. Bartelheimer and M. Spiliopoulou, "Mining and Visualizing the Evolution of Subgroups in Social Networks", in *Proceedings of the 2006 IEEE/WIC/ACM International Conference on Web Intelligence*, Washington, DC, USA, pp. 52–58, 2006.

D. Fisher, "Using egocentric networks to understand communication", *IEEE Internet Computing*, 9(5), pp. 20–28, 2005.

G. Flake, S. Lawrence, C.L. Giles and F.M. Coetzee, "Self-organization and identification of web communities", *Computer*, pp. 66–71, 2002.
S. Fortunato, V. Latora, and M. Marchiori, "Method to find community structures based on information centrality", *Physical Review E*, 70(5), pp. 56104–56117, 2004.
G. Frivolt and M. Bielikov, "An approach for community cutting", *Proceedings of the 1st Int. Workshop on Representation and Analysis of Web Space*, Prague, Czech Republic, pp. 49–54, 2005.
L. Garton, C. Haythonthwaite and B. Wellman, "Studying online social networks". *Journal of Computer Mediated Communication*, 3(1), 1997. http://jcmc.indiana.edu/vol3/issue1/garton.html
M. Girvan and M. Newman, "Community structure in social and biological networks", *Proceedings of the National Academy of Sciences*, 99(12), pp. 7821–7829, 2002.
P. Gloor, "Capturing team dynamics through temporal social surfaces", in *Proceedings of the Ninth International Conference on Information Visualisation*, London, pp. 939–944, 2005.
A. Gruzd and C. Haythornthwaite, "Automated discovery and analysis of social networks from threaded discussions", Presented at *International Network of Social Network Analysis (INSNA) Conference*, St. Pete Beach, Florida, 22–27 January, 2008.
R.A. Hanneman and M. Riddle, *Introduction to social network methods (online textbook)*, Riverside, CA: University of California, 2005.
A. Java, X. Song, T. Finin and B. Tseng, "Why we twitter: understanding microblogging usage and communities", in *Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 Workshop on Web Mining and Social Network Analysis*, San Jose, California, New York: ACM, pp. 56–65, 2007.
J. Keelan, V. Pavri-Garcia, G. Tomlinson and K. Wilson, "YouTube as a source of information on immunization: A content analysis", *JAMA*, 298(21), pp. 2482–2484, 2007.
J. Kleinberg, "Authoritative sources in a hyperlinked environment", *Journal of the ACM (JACM)*, 46(5), pp. 604–632, 1999.
J. Kleinberg, "Bursty and hierarchical structure in streams", *Data Mining and Knowledge Discovery*, 7(4), pp. 373–397, 2003.
D.H. Krantz and A. Tversky, "Similarity of rectangles: An analysis of subjective dimensions", *Journal of Mathematical Psychology*, 12(1), pp. 4–34, 1975.
R. Kumar, J. Novak, P. Raghavan and A. Tomkins, "Structure and evolution of blogspace", *Communications of the ACM*, 47(12), pp. 35–39, 2004.
R. Kumar, J. Novak and A. Tomkins, "Structure and evolution of online social networks", *Proceedings of the 12th ACM SIGKDD international Conference on Knowledge Discovery and Data Ming*, Philadelphia, PA, USA, New York: ACM, pp. 611–617, 2006.
A. Kurdia, O. Daescu, L. Ammann, D. Kakhniashvili and S.R. Goodman, "Centrality measures for the human red blood cell interactome", *Engineering in Medicine and Biology Workshop, 2007 IEEE Dallas*, IEEE, pp. 98–101, 2007.
J. Leskovec, J. Kleinberg and C. Faloutsos, "Graph evolution: Densification and shrinking diameters", *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 1(1), pp. 1–41, 2007.
J. Leskovec, K. Lang, A. Dasgupta and M. W. Mahoney, "Statistical properties of community structure in large social and information networks", in *Proceeding of the 17th international Conference on World Wide Web*, Beijing, China, 21–25 April 2008. WWW '08. New York: ACM, pp. 695–704, 2008.
L. Leydesdorff, T. Schank, A. Scharnhorst and W. De Nooy, 2008. Animating the development of social networks over time using a dynamic extension of multidimensional scaling. Available online at: http://arxiv.org/pdf/0809.4655 (accessed 15 June 2010).
X. Li, B. Liu and P.S. Yu, "Mining community structure of named entities from web pages and blogs", in *AAAI Spring Symposium*, 2006. Available online at http://www.aaai.org/Papers/Symposia/Spring/2006/SS-06-03/SS06-03-021.pdf (accessed 15 June 2010).
Y. Lin, Y. Chi, H. Sundaram and B.L. Tseng, "Facetnet: A framework for analyzing communities and their evolutions in dynamic networks", in *Proceeding of the 17th International Conference on World Wide Web*, Beijing, China, New York: ACM, pp. 685–694, 2008.
Y. Lin, Y. Chi, H. Sundaram and B.L. Tseng, "Analyzing communities and their evolutions in dynamic social networks", *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 3(2), pp. 1–31, 2009.
H. Ma and A. Zeng, "The connectivity structure, giant strong component and centrality of metabolic networks", *Bioinformatics*, 19(11), pp. 1423–1430, 2003.

C. Marlow, "Audience, structure and authority in the weblog community", in *International Communication Association Conference*, New Orleans, LA, 2004. Available online at http://alumni.media.mit.edu/~cameron/cv/pubs/04-01.pdf (accessed 15 June 2010).

N. Memon, H. Larsen, D.L. Hicks and N. Harkiolakis, "Detecting hidden hierarchy in terrorist networks: Some case studies", *Lecture Notes in Computer Science*, 5075, pp. 477–489, 2008.

J. Moody, D. McFarland and S. Bender-deMoll, "Dynamic network visualization 1", *American Journal of Sociology*, 110(4), pp. 1206–1241, 2005.

M. Newman, "Modularity and community structure in networks", *Proceedings of the National Academy of Sciences*, 103(23), pp. 8577–8582, 2006.

M. Newman and M. Girvan, "Finding and evaluating community structure in networks", *Physical Review E*, 69(2), 26113, pp. 1–16, 2004.

J. Orford, "Implementation of criteria for partitioning a dendrogram", *Mathematical Geology*, 8(1), pp. 75–84, 1976.

W. Piper, M. Marrache, R. Lacroix and B.D. Jones, "Cohesion as a basic bond in groups", *Human Relations*, 36(2), pp. 93–108, 1983.

F. Radicchi, C. Castellano, F. Cecconi, V. Loreto and D. Parisi, "Defining and identifying communities in networks", *Proceedings of the National Academy of Sciences*, 101(9), pp. 2658–2663, 2004.

C. Reffay and T. Chanier, "How social network analysis can help to measure cohesion in collaborative distance learning", 2003. Available online at http://edutice.archives-ouvertes.fr/edutice-00000422 (accessed 15 June 2010).

J. Ruan and W. Zhang, "An efficient spectral algorithm for network community discovery and its applications to biological and social networks", in *Seventh IEEE International Conference on Data Mining*, Omaha, Nebraska, USA, IEEE, pp. 643–648, 2007.

M. Sato, Y. Sato, and L. Jain, *Fuzzy clustering models and applications (studies in fuzziness and soft computing vol. 9)*, Amsterdam: Springer-Verlag, 1997.

S. Schaeffer, "Graph clustering", *Computer Science Review*, 1(1), pp. 27–64, 2007.

J. Shi and J. Malik, "Normalized cuts and image segmentation", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(8), pp. 888–905, 2000.

T. Snijders, C. Steglich and M. Schweinberger, "Modeling the co-evolution of networks and behavior", in *Longitudinal Models in the Behavioral and Related Sciences*, London: Routledge, pp. 41–71, 2007.

S. Sterling, Aggregation techniques to characterize social networks, *Storming Media*, 2004. Available online at http://www.au.af.mil/au/awc/awcgate/afit/sterling_socnet.pdf (accessed 15 June 2010).

J. Sun, C. Faloutsos, S. Papadimitriou and P.S. Yu, "Graphscope: Parameter-free mining of large time-evolving graphs", in *Proceedings of the 13th ACM SIGKDD international Conference on Knowledge Discovery and Data Mining*, San Jose, California, USA, New York: ACM, pp. 687–696, 2007.

L. Tang, H. Liu, J. Zhang and Z. Nazeri, "Community evolution in dynamic multi-mode networks", in *Proceeding of the 14th ACM SIGKDD international Conference on Knowledge Discovery and Data Mining*, Las Vegas, Nevada, USA, New York: ACM, pp. 677–685, 2008.

J. Tyler, D. Wilkinson and B.A. Huberman, "Email as spectroscopy: Automated discovery of community structure within organizations", in *Communities and Technologies*, M. Huysman, E. Wenger and V. Wulf, (Eds.). Deventer, The Netherlands: Kluwer, pp. 81–96, 2003.

M. van Duijn and J. Vermunt, "What is special about social network analysis", *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences*, 2(1), pp. 2–6, 2006.

G. Wang, Y. Shen and M. Ouyang, "A vector partitioning approach to detecting community structure in complex networks", *Computers and Mathematics with Applications*, 55(12), pp. 2746–2752, 2008.

S. Wasserman and K. Faust, *Social Network Analysis: Methods and Applications*, New York: Cambridge Univ Pr, 1994.

H. Welser, E. Gleave, D. Fisher and M. Smith, "Visualizing the signatures of social roles in online discussion groups", *The Journal of Social Structure*, 8(2), 2007.

C. Zahn, "Graph-theoretical methods for detecting and describing gestalt clusters", *IEEE Transactions on Computers*, 20(1), pp. 68–86, 1971.