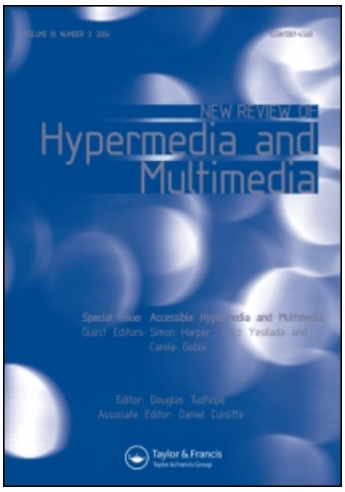

New Review of Hypermedia and Multimedia
Publication details, including instructions for authors and subscription information:
http://www.informaworld.com/smpp/title~content=t713599880

Automatic detection of cohesive subgroups within social hypertext: A heuristic approach

Alvin Chin [a]; Mark Chignell [b]
[a] Department of Computer Science, University of Toronto, Toronto, Ontario, Canada [b] Department of Mechanical and Industrial Engineering, University of Toronto, Toronto, Ontario, Canada

Online Publication Date: 01 January 2008

To cite this Article Chin, Alvin and Chignell, Mark(2008)'Automatic detection of cohesive subgroups within social hypertext: A heuristic approach',New Review of Hypermedia and Multimedia,14:1,121 — 143
To link to this Article: DOI: 10.1080/13614560802357180
URL: http://dx.doi.org/10.1080/13614560802357180

New Review of Hypermedia and Multimedia,
Vol. 14, No. 1, July 2008, 121–143

Taylor & Francis
Taylor & Francis Group

Automatic detection of cohesive subgroups within social hypertext: A heuristic approach

ALVIN CHIN*† and MARK CHIGNELL‡
†Department of Computer Science, University of Toronto, 10 King's College Road, Toronto, Ontario, M5S 3G4, Canada
‡Department of Mechanical and Industrial Engineering, University of Toronto, 5 King's College Road, Toronto, Ontario, M5S 3G8, Canada

The problem of identifying cohesive subgroups in social hypertext is reviewed. A computationally efficient three-step framework for identifying cohesive subgroups is proposed, referred to as the Social Cohesion Analysis of Networks (SCAN) method. In the first step of this method (Select), people within a social network are screened using a level of network centrality to select possible subgroup members. In the second step (Collect), the people selected in the first step are collected into subgroups identified at each point in time using hierarchical cluster analysis. In the third step (Choose), similarity modeling is used to choose cohesive subgroups based on the similarity of subgroups when compared across different points in time. The application of this SCAN method is then demonstrated in a case study where a subgroup is automatically extracted from a social network formed based on the online interactions of a group of about 150 people that occurred over a two-year period. In addition, this paper also demonstrates that similarity-based cohesion can provide a different, and in this case more compelling, subgroup representation than a method based on splitting a hierarchical clustering dendrogram using an optimality criterion.

Keywords: Social Hypertext; Social Networks; Cluster Analysis; Cohesive Subgroups; Similarity Measurement; Online Community; Network Centrality; Subgroup Evolution

1. Introduction

Social hypertexts (Erickson 1996) are networks of documents connected explicitly and implicitly to networks of people. Through crawling and mining of data on blogs, and on sites such as those for video sharing, social bookmarking, and social networking, social networks at a variety of scales may be inferred from online interactions. However, methods are needed to identify meaningful and coherent subgroups within the resulting social hypertexts as a key step toward building enhanced tools for search and other functions where those tools utilize knowledge about groups and communities that people belong to.

*Corresponding author. Email: achin@cs.toronto.edu

New Review of Hypermedia and Multimedia
ISSN 1361-4568 print/ISSN 1740-7842 online
http://www.tandf.co.uk/journals
DOI: 10.1080/13614560802357180

Semantic Web technologies such as Friend-of-a-Friend (FOAF) attempt to add social networking metadata to the Web, however the resulting information is static and not yet available for the majority of Web data. While there has been recent research interest in automatic detection of subgroups in social networks (e.g. Sterling 2004, Balasundaram *et al.* 2007, Du *et al.* 2007, Tantipathananandh *et al.* 2007), that research has generally adopted an optimization approach using techniques such as dynamic programming. The networks analyzed using these techniques have tended to be relatively small (fewer than a hundred people) and the analytic methods employed are computationally intensive. However, in many cases it is of interest to find subgroups in large networks consisting of many thousands of people. Thus heuristic methods for subgroup identification are needed that can scale up to and handle very large social networks.

The research reported in this paper is concerned with identifying subgroups that are cohesive over time. This approach screens out transitory groupings of people that only occur over short time periods. While the analysis of transitory subgroupings is also a potentially interesting subtopic, it is not the focus of this paper. From a methodological perspective, the benefit of using a cohesiveness criterion is that it provides a clear guideline for selecting subgroups that are meaningful and potentially useful.

The research question to be addressed in this paper is: How can cohesive subgroups be identified in large social networks in an efficient manner? This question will be answered in the following sections through the presentation of a new three-step method for identifying cohesive subgroups, referred to as the Social Cohesion Analysis of Networks (SCAN) method. The application of this method in a realistic case study will also be described.

2. Background

In this section previous research relevant to the identification of cohesive subgroups in social networks will be reviewed. First, currently available methods for identifying subgroups will be discussed, followed by review of three general techniques for analysis of social networks. The three techniques considered here are chosen based on their relevance to the three tasks of selecting people who are potential members of subgroups (network centrality), collection of eligible people into subgroups (clustering and partitioning), and assessing cohesion of subgroups over time (similarity measurement).

2.1 Subgroup identification

The problem of defining and evaluating cohesion within groups has been a challenging issue in sociology (Friedkin 2004). However, it is possible to operationalize a construct of subgroup cohesion based on network properties. For instance, Wasserman and Faust (1994) defined a cohesive subgroup as a set of actors (nodes) that are relatively dense and directly connected through reciprocated (bi-directional) relationships (links).

Sociological analyses often examined subgroups over time, but the size of the networks examined tended to be small. For instance, the widely used Southern Women dataset (Freeman 2003) traced interactions between 18 women over a nine-month period.

Kumar *et al.* (1999) addressed the problem of finding emerging subgroups on the Web that had not yet coalesced into larger communities. Their method relied on in-degree (related to degree centrality) as a way of screening potential subgroups. While an early example of subgroup identification and of a heuristic style of research that was designed to be scalable to very large networks, their method was not fully automated and involved some filtering and interpretation by humans.

Research on finding Web communities has tended to utilize content analysis of text and tags associated with Web pages. This approach allows the subgroup identification task to be linked to powerful search engine algorithms. For instance, Flake *et al.* (2002) presented a heuristic community identification algorithm which used Web pages as topic seeds. Searches and link analysis were then used to identify a "community" of pages relating to the topic seeds. In similar vein, Chau *et al.* (2005) developed a method identifying the communities associated with business Web sites by tracking back through the incoming links and carrying out data mining on the resulting network.

There has also been considerable graph-theoretic research on finding densely connected subgraphs within larger graphs. For instance, Gibson *et al.* (2005) presented a method for detecting densely connected groups of servers on the Web. However, these analyses (particularly when applied on a large scale) have typically been focused on static social networks.

Recently, researchers (Tantipathananandh *et al.* 2007) have addressed the issue of finding subgroups in dynamic social networks using an optimization approach. However, their evaluations have tended to involve either synthetic datasets or networks with relatively few members (considerably fewer than 100).

In summary, while there has been considerable work on methods of automated subgroup identification, much of it has been on small or static networks, or has used known topics or Web pages as seeds. In addition, much of the focus has been on finding communities based on links between text content, rather than on links between people as reflected in their online interactions. Thus, the development of a scalable and valid method for discovering cohesive subgroups of people based on large scale online interactions within social networks is still an open research problem.

2.2 Network centrality

Network centrality (Freeman 1978) can be used to identify the most important people that are at the center of the network and are the most well connected in subgroups. Network centrality measures may be calculated using software such as Pajek (de Nooy *et al.* 2005) and UCINET (Borgatti

et al. 2002). Numerous centrality measures have been used for characterizing the social behaviour and connectedness of nodes within networks (Mizruchi *et al.* 1986, Dwyer *et al.* 2006). Three centrality measures frequently used in analyzing social networks are degree centrality, closeness centrality and betweenness centrality.

Degree centrality measures the number of direct connections that an individual node has to other nodes within the social network (Freeman 1978). Thus in a social network it measures the number of friends one has, that is, how many people a person is directly connected to. It has been shown that nodes with higher degree centrality than others in the network are more active and important as members within the community (Frivolt and Bielikova 2005) and are more influential (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 (Freeman 1978). Thus closeness centrality measures how close, on average, a person is to other people in the network. Closeness centrality has been used to identify important nodes within social networks (e.g., Ma and Zeng 2003, Kurdia *et al.* 2007, Chin and Chignell 2007a).

Betweenness centrality measures the extent to which a node can act as an intermediary or broker to other nodes (Freeman 1978). The more times that a particular node lies on paths that exist between other pairs of nodes in the network, the higher the betweenness centrality is for that node. Nodes that have a high betweenness centrality may act as brokers between subgroups and they may have stronger membership in surrounding communities (Girvan and Newman 2002, Donetti and Munoz 2004). Betweenness centrality has been used to reveal the hierarchical structure of organizations in e-mail (Tyler *et al.* 2005), mailing lists (Gloor *et al.* 2003) and blogs (Tremayne *et al.* 2006), and to identify opinion leaders in blogs (Marlow 2004). Betweenness centrality is one of the most frequently used centrality measures in research on social networks (Gloor *et al.* 2003, Marlow 2004, Newman and Girvan 2004, Tyler *et al.* 2005, Tremayne *et al.* 2006).

2.3 Clustering and partitioning

Finding subgroups within social networks is a problem that has attracted considerable interest (e.g., Reffay and Chanier 2003, Wellman 2003 and Sterling 2004). Clique analysis and related methods look directly at the links that occur in a network and identify specific patterns of connectivity (e.g., subgroups where everyone in the subgroup has a direct connection to everyone else). Clustering and partitioning methods are less direct (but more computationally efficient) in that they base their groupings (clusters) on proximity measures (similarities or distances) derived from the connection patterns between network nodes.

Cluster analysis (Sokal and Sneath 1963) and related methods have frequently been used to analyze networks. For instance, link analysis has been used to identify topics within the clusters of web pages (Kleinberg 1999).

Co-citation analysis has been used to rank search engine results (Yaltaghian and Chignell 2003) and to find groupings in Web pages (Kumar *et al.* 1999, Flake *et al.* 2002), blogs (Kleinberg 2002, Adar *et al.* 2004, Chin and Chignell 2007a), and tags associated with web pages (Golder and Huberman 2005, Brooks and Montanez 2006, Marlow *et al.* 2006).

Hierarchical clustering has been used to quantify the structure of community in documents (Brooks and Montanez 2006), web pages (Girvan and Newman 2002, Donetti and Munoz 2004, Clauset 2005), blogs (Paolillo and Wright 2005), and discussion groups (Chin and Chignell 2007b, 2007c, Gomez 2008). Hierarchical clustering results in a hierarchy (tree) being formed where the leaves of the tree are the nodes that are clustered. The resulting trees can be visualized as dendrograms, examples of which are shown later in this paper.

In contrast to hierarchical cluster analysis, the groups formed in partitioning methods are not nested. The k-means algorithm (Hartigan 1975) is a popular method for partitioning that is available in widely used statistical packages such as SPSS. K-means analysis has been used to detect clusters in blogs (Adar *et al.* 2004, Breuer and Ratkiewicz 2005).

Partitioning methods are relatively efficient, but they require that the number of subgroups in the partition be defined prior to the analysis. On the other hand, hierarchical cluster analysis does not yield a partition and the hierarchy (dendrogram) that is output needs to be cut in order to identify a particular set of subgroups. In practice, for both partitioning analysis and hierarchical cluster analysis, the method needs to be supplemented with an additional selection criterion. For partition analysis, the method is run using a number of different values of k (i.e., number of groups in the partition) and the selection criterion is used to define which of the possible partitions should be chosen as the best subgrouping. For hierarchical clustering, the selection criterion is used to decide at which point the dendrogram should be cut in order to obtain a non-nested set of subgroups.

Two general approaches to implementing the required criterion (as noted in the preceding paragraph) will now be considered. In the approach considered first (Section 2.3.1), the criterion is based on a mathematical model of optimality. In the second approach (Section 2.4), the criterion is based on maximizing the cohesion of subgroups identified at different points in time, using a similarity measure.

2.3.1 Optimality criteria. Orford (1976) described a range of criteria for determining where to partition a dendrogram. Orford made the important point that the best criterion to use will generally vary with the problem context. In contrast to Orford's eclectic approach, recent research has tended to assess specific measures for obtaining an optimal partition (e.g., Jung *et al.* 2003). The modularity (designated as Q) discussed by Newman and Girvan (2004) has been proposed as a definitive measure of the quality of clustering. Newman (2006) claimed that maximizing modularity results in a set of clusters that best represents optimum subgroup structure. Modularity has

been used for finding community structure and subgroups (Bird 2006, Ruan and Zhang 2007). The computational performance of different algorithms based on modularity was evaluated in Danon *et al.* (2005) and algorithms have been created to improve over Newman's original method (Radicchi *et al.* 2004, Duch and Arenas 2005). Other approaches for partitioning based on optimality include vector partitioning (Wang *et al.* 2007) and normalized cut metrics (Leskovec *et al.* 2008).

Despite much work being done to create more efficient algorithms for modularity, there has been relatively little research on evaluating its effectiveness in finding meaningful partitions and cohesive subgroups. As noted by Radicchi, Castellano, Cecconi *et al.* (2004), it is not clear whether the "optimal" partitions that are discovered using the modularity criterion are representative of real collaborations in the corresponding online communities. Van Duijn and Vermunt (2005) noted that it is difficult to determine which measure is the most appropriate to use across a range of applications. Thus, Orford's (1976) original insight still seems relevant, that is, the best criterion for splitting a dendrogram may depend on factors such as the type of data being collected and compared.

2.4 Similarity measurement

How can cohesion in subgroups over time be assessed? Intuitively, cohesive subgroups should have a core of people that remain in them over different time periods. The situation is complicated by the fact that subgroups may split or merge, so that cohesiveness is not necessarily a property of a single subgroup, but may sometimes relate to a family of one or more related subgroups. Cohesive families of subgroups at one time period should be similar to corresponding subgroups at a different time period.

Similarity is a topic that has received attention in a wide variety of scientific fields. Mathematically, similarity is often viewed as a geometric property involving the scaling or transformation necessary to make objects equivalent to each other. Similarity can be defined as the inverse of distance, with a well-known distance measure being Euclidean distance (Elmore and Richman 2001), which itself is a special case of a family of distance measures known as Minkowski metrics (Santini and Jain 1999). However, distance measures typically require a vector (spatial) model of the entities being compared, which is often not appropriate for comparing aggregations of nodes in a network.

In developing methods to assess the similarity between different species, numerical taxonomists have developed and utilized a number of similarity measures (Sokal and Sneath 1963). Many of these measures involve some sort of correlation, a construct that is conceptually related to similarity. One correlation measure is the cosine distance or dot product that measures the angle between two objects represented as vectors of numerical features (Wang *et al.* 2002). However, since features cannot always be expressed on a well-defined numerical scale, researchers (e.g., psychologists)

have developed feature models of similarity that assess similarity based on a comparison of matching and mismatching features, using a set-theoretic approach.

Tversky's feature contrast model (Tversky 1977) expressed the degree of similarity of two stimuli to a linear combination of their common and distinctive features. Gregson (1975) recommended a content similarity model where similarity was expressed as the ratio of the intersection of the features for the objects being compared to the union of their features. A simplified version of the content similarity model is the Jaccard coefficient (first proposed in 1901) which is defined as the size of the intersection divided by the size of the union of the objects being compared (Jaccard 1901).

In the research problem addressed in this paper, the number of nodes is not fixed, since members may become more or less active or may enter or leave the community, so that the people under consideration as subgroup members will tend to change over time. Thus a custom-built similarity measure will be introduced later in this paper to meet the demands of identifying subgroups in a dynamic social network, based on the content model of similarity (Gregson 1975).

2.5 Summary

A number of methods already exist for finding subgroups and clusters, but there is as of this writing no method that will scale up to handle very large social networks and that can identify subgroups that remain cohesive over time. The research reported in this paper fills this gap by introducing a heuristic and scalable method for automatically identifying subgroups of people in social networks that are cohesive over time. In the next section, the lessons learned from the preceding review of relevant research literature are utilized in formulating this methodology.

3. The Social Cohesion Analysis of Networks (SCAN) method

The SCAN method is a three-step process that could be used to analyze social networks in general, but that is intended for use in identifying cohesive subgroups on the basis of social networks inferred from online interactions. The SCAN method assumes that the social network has been previously inferred using one of the number of data mining and crawling techniques that are available.

From a review of relevant research literature in Section 2, and a logical analysis of the problem, the following three steps for efficient and effective identification of cohesive subgroups have been identified: (1) *selecting* potential members of cohesive subgroups from the social network; (2) *collecting* these potential members into subgroups; and (3) *choosing* cohesive subgroups that have a similar membership over time.

3.1 Step 1: Select

In the first step of the SCAN method, the objective is to recognize potential members of cohesive subgroups from the online environment. This is achieved by finding members from the social network that are relatively well connected and thus more likely to be part of subgroups. As noted in the preceding literature review, centrality measures have been recommended by a number of researchers as a means to identify influential members of communities who are also more likely to be members of subgroups. Betweenness centrality in particular has been found to be a good indicator of community in past research, although in very large networks it may be necessary to supplement or replace it with centrality measures that are less expensive to compute.

In the initial version of the SCAN method proposed here, betweenness centrality is used to determine potential candidates for inclusion in subgroups. People having an amount of betweenness centrality that is below a certain threshold are filtered out of the subsequent subgroup analysis due to their overall low connectivity to, and cohesion with, others in the social network. In cases where very large parent networks need to be scanned for subgroups, the problem can be kept manageable by adopting one of the following strategies. The first strategy is to either switch to degree centrality since it is relatively easy to compute, or to employ a multiple screen where degree centrality is used as an initial filter with further rounds of filtering then employing closeness centrality and/or betweenness centrality being carried out as needed. A second approach to implementing the Select step in a very large network involves randomly selecting subgraphs of the social network and then calculating the centrality measure for each person as the average betweenness centrality obtained for that person across the set of subgraphs that he or she was sampled in.

Whichever of the methods in the preceding paragraph is used, once the centralities are calculated, a subgraph from the original social network is constructed where all the nodes have a centrality measure above a chosen cutoff. It should be noted that the appropriate threshold of centrality for selecting people probably varies with context and that knowledge of the community of interest might help in assessing how many members in the overall community of interest are likely to be part of active subgroups. A rough starting point for determining the starting cutoff betweenness centrality can be obtained by creating a frequency distribution graph of centralities of all members in the social network and then selecting the cutoff as a break in the distribution that separates a smaller group of higher centrality members from the larger group of people with lower centralities.

3.2 Step 2: Collect

Once the potential members for inclusion in subgroups have been identified in the preceding step, the second step collects the potential members into subgroups. The Collect step performs a hierarchical cluster analysis on the

members selected in the Select step. Average weighted hierarchical clustering is used in this step as it has previously been found to work well for large-scale clustering (Cutting *et al.* 1992).

The output of hierarchical clustering is a set of nested clusters (or tree) in a dendrogram. As shown in figure 1, this dendrogram then has to be cut at some point (corresponding to a level of similarity) in order to create a partition that contains specific (non-nested) subgroups. In order to determine the cutoff point the initial approach that we tried was to maximize modularity Q (Newman 2006). Since Q has not received a lot of empirical testing in this context, one of the goals of the case study reported in Section 4 was to determine how good the subgroups formed using the Q cutoff criterion are in a realistic context.

3.3 Step 3: Choose

The previous two steps (Select and Collect) can be repeated to discover candidate subgroups at different points in time. Similarity analysis is then used to see which of the subgroups are cohesive over time. The initial version of the SCAN method uses a set-theoretic approach (Gregson 1975), where the similarity is implemented as the ratio of the set intersection of subgroups to the corresponding set union.

The formula for calculating similarity between subgroupings identified at different times is shown in Equation 1. From the subgroups identified in the Collect step, cohesion is examined for the largest subgroup in the first of the two time periods being compared. The largest group in the first time period is chosen as larger groups are less likely to arise by chance and are more likely to retain a quorum of members between time periods. The similarity measure defined in Equation 1 tracks how well the subgroup observed in the first time

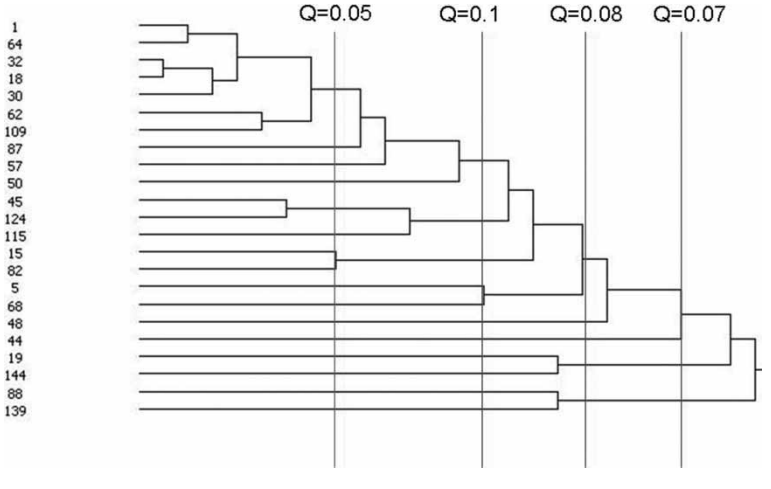

Figure 1. Example of dendrogram showing modularity and possible cutpoints for creating partitions at different levels of similarity.

period (T_1) stays together when subgroups are recalculated in the second period (T_2). The similarity measure looks at all the possible pairwise relationships between members of the largest subgroup in the first time period, and sees how many of the pairs still exist in the second time period. The similarity can then be calculated using the following formula:

$$Sim_{T_1-T_2} = \frac{\text{N(common pairs in subgroups from } T_1 \text{ in } T_2)}{\text{N(possible pairs in largest subgroup in } T_1)} \quad (1)$$

where N() is a cardinality operator that counts the pairs that meet the criteria defined in the numerator and denominator, respectively of Equation 1.

As an example of how Equation 1 works, if there were a subgroup of five people which then split into a group of two and a group of three, then there would be a combination of five choose two or 10 pairs from the first time period and after that there would be three choose two or three pairs plus 1 pair together in the second time period, so the similarity would be 4/10 $=0.4$. This measure of similarity can then be used to assess cohesiveness over time.

Cohesiveness, as implemented by similarity analysis, can be used as a criterion for choosing a threshold for the centrality measure used in Step 1 of the SCAN method, and it can also be used to choose where to cut the dendrogram in Step 2 of the method. For example, in the case of the centrality cutoff, the cutoff would be chosen that maximizes the observed cohesion of subgroups and measure of similarity in the sample. While we know of no formal mathematical justification for this approach, we observe that if subgroups are cohesive they should tend to change less over time and that high levels of cohesion are unlikely to arise by chance. Thus capitalizing on the best betweenness cutoff to create the most cohesive subgroups is unlikely to distort the patterns that are inherent in the data. This approach is similar in spirit to the Box and Cox (1964) argument that the transformation that maximizes the value of the F-ratio is de facto the most appropriate transformation since random effects will tend to decrease rather than increase the F-ratio. The "best" cohesive subgroups will then be the ones that are the most similar, therefore it is appropriate to choose cutoffs and criteria that maximize the similarity of subgroups obtained over different time periods.

In the next section, we provide an initial test of the SCAN method by applying it to a case study.

4. Applying the Social Cohesion Analysis of Networks (SCAN) method: TorCamp Google group

The TorCamp group is a community of designers, developers, marketers, and anyone who works with and is passionate about technology in Toronto, Canada. In this case study, we describe how the SCAN method works in practice for automatically finding cohesive subgroups and their members in TorCamp that communicate using the TorCamp Google group.

4.1 Data gathering and social network creation

We first crawled the TorCamp Google group, recording all posts and comments. To analyze and track the subgroup growth in the TorCamp Google group, we decided to use the first two calendar years (January through December, for both 2006 and 2007) after the inception of the group in October 2005. We then divided those two years into six-month intervals for further analysis. Figure 2 illustrates the characteristics of each time period studied. The number of topics, messages and members all increased over the four time periods, except for July–December 2007 where the number of messages decreased slightly relative to the corresponding number for the preceding time period.

The inferred social network was constructed from the crawled data in the following way. A directed graph $G(V,E)$ was generated for each time period, 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 had directly replied to a comment by user v if it existed within the threaded discussion corresponding to the post or to the original post by user v, where w is the number of times that user u made a comment to user v's comment or post. The collection of all the people, links and weights formed the inferred (parent) social network. We repeated this method for all four time periods. One problem that arose in gathering this data was that some threads were carried over from one of the time periods into the next. For these cases, we included the entire thread in each of the time periods that it appeared in.

Figure 3 illustrates the social networks for each of the time periods studied and shows how the density of the network increased from 2006 to 2007.

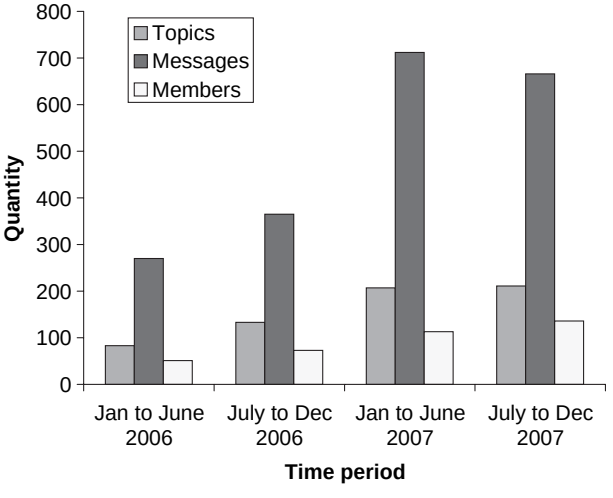

Figure 2. TorCamp Google group statistics from 2006 to 2007.

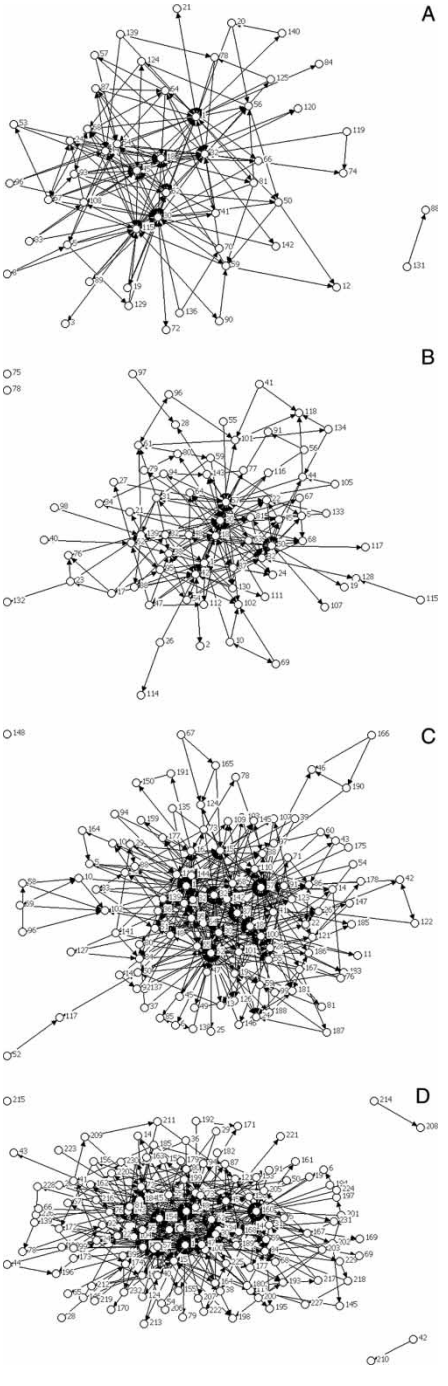

Figure 3. Social networks for each of the time periods studied during 2006 to 2007 for the TorCamp Google group (January to June 2006 (a), July to December 2006 (b), January to June 2007 (c), and July to December 2007 (d)).

4.2 Step 1: Select

To select possible members of cohesive subgroups, we calculated the betweenness centrality for each member in each of the time periods using Freeman's betweenness centrality measure. We selected people as subgroup candidates at four levels of inclusiveness, corresponding to betweenness centrality cutoffs that captured the top 10%, 15%, 20% and 25%, respectively of the nodes in the network. We chose these percentages from looking at the frequency distribution of betweenness centrality for the TorCamp Google group. Table 1 summarizes the corresponding cutoff betweenness centralities that corresponded to the four percentile thresholds for each time period.

It can be seen that cutoff betweenness centrality decreases from a high in the first half of 2006 to a low in the first half of 2007, but increases again in the second half of 2007. A subgraph of members that have betweenness centralities greater than a cutoff of 20% (0.0214) in the July–December 2007 time period is shown in figure 4. After potential subgroup members were selected based on the cutoff betweenness centralities, the analysis proceeded to the second step in the SCAN method.

4.3 Step 2: Collect

In the second step (Collect) of the SCAN method, members that survived the screening in the Select step were collected into subgroups for each time period using hierarchical cluster analysis. The dendrogram produced by the hierarchical clustering was then cut into a subgroup partition using a criterion that yielded the largest value of modularity Q. Table 2 shows the modularity Q that was obtained after clustering the subgraphs using each of the four betweenness centrality cutoffs.

Figure 5 illustrates an example of the dendrogram generated by weighted average hierarchical clustering for the top 20% of members with highest betweenness centrality for the July–December 2007 time period. The vertical line indicates the cutoff where maximum modularity was found which was $Q = 0.109$. This results in the subgroups of (1, 15, 18, 30, 32, 57, 62, 82, 95, 144, 154, 168), (84, 218, 229), (100, 167), (22, 47, 143, 184), and (45, 64, 160) as indicated by the overlaid rectangles in figure 5. We repeated this same procedure of using weighted average hierarchical clustering and maximum modularity as calculated in table 2 for all the cutoff betweenness centralities and time periods.

Bird (2006) and Newman and Girvan (2006) claimed that values of Q above 0.3 indicate strong community structure. According to this criterion, the TorCamp Google group had a relatively weak community structure, with July–December 2006 exhibiting the strongest presence of community within the two-year dataset.

Table 1. Cutoff betweenness centrality for each time period according to top 10%, 15%, 20% and 25% of all nodes that have highest betweenness centrality.

	Cutoff betweenness centrality			
Time period	Top 10% of nodes with highest betweenness centrality	Top 15% of nodes with highest betweenness centrality	Top 20% of nodes with highest betweenness centrality	Top 25% of nodes with highest betweenness centrality
January–June 2006	0.13	0.04	0.027	0.0175
July–December 2006	0.05	0.0282	0.0161	0.0133
January–June 2007	0.0362	0.025	0.014	0.01
July–December 2007	0.044	0.027	0.0214	0.0168

4.4 Step 3: Choose

After carrying out the previous analyses, four sets of subgroups were obtained for each time period (one subgroup for each of the four betweenness centrality cutoffs). In the Choose step, cohesiveness was then measured between time periods in order to determine which betweenness centrality cutoff was the most appropriate prior to choosing a final subgroup representation. The similarity in subgroups was calculated between the first and second halves of 2006, and separately between the first and second halves of 2007, for each of the Q modularities (from table 2) obtained from clustering the subgraphs corresponding to the four betweenness centrality cutoffs, using Equation 1 to measure the similarity. The resulting similarity scores are shown in table 3.

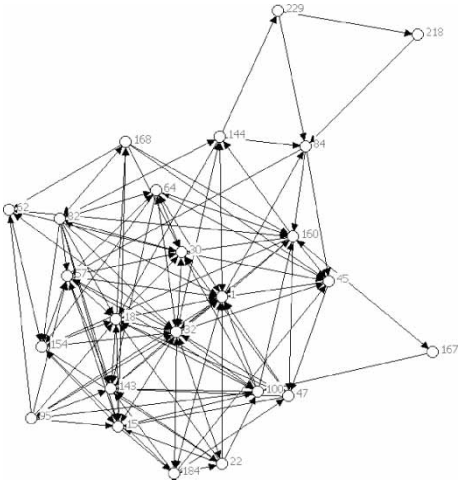

Figure 4. Subgraph of top 20% of members with highest betweenness centrality in TorCamp from July to December 2007.

Table 2. Q modularity found for each time period according to clustering of top 10%, 15%, 20% and 25% of all nodes that have highest betweenness centrality.

	Q Modularity			
Time period	Top 10% of nodes with highest betweenness centrality	Top 15% of nodes with highest betweenness centrality	Top 20% of nodes with highest betweenness centrality	Top 25% of nodes with highest betweenness centrality
January–June 2006	−0.018	−0.006	0.020	0.042
July–December 2006	0.147	0.144	0.184	0.191
January–June 2007	0.006	0.067	0.069	0.106
July–December 2007	0.030	0.100	0.109	0.119

Overall, similarity (cohesiveness) was higher in the second year (2007) than it was in the first year (2006). The highest similarity was found using the top 20% cutoff betweenness centrality for both 2006 (0.167) and 2007 (0.318). Therefore, using this criterion, the cohesive subgroups in the TorCamp Google group are the subgroupings at the top 20% betweenness centrality for each time period. Table 4 enumerates the non-overlapping cohesive subgroups found from similarity analysis.

4.5 Visualizing subgroup growth and member migration

Figure 6 visualizes the cohesive subgroups from table 4 and shows how members in the TorCamp Google group move in and out of subgroups at

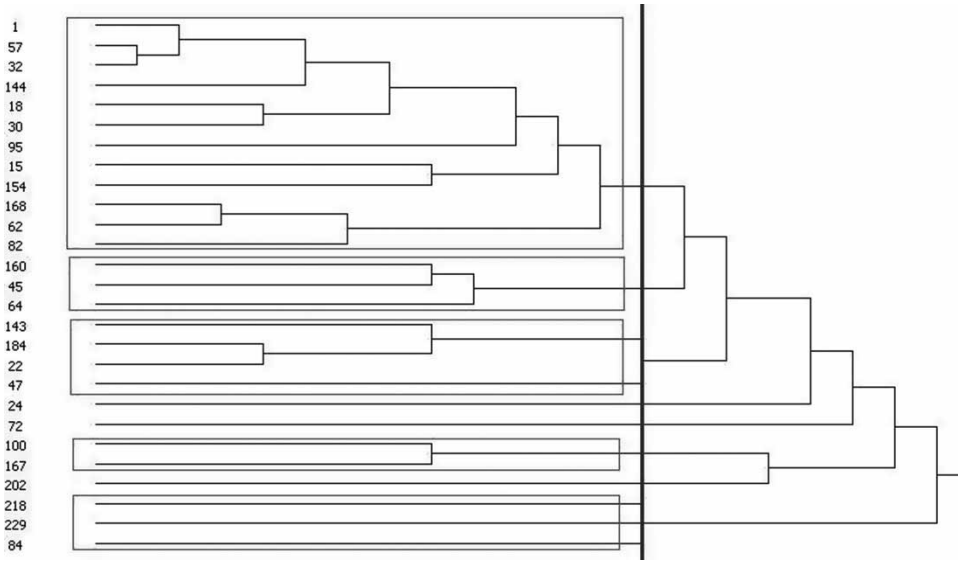

Figure 5. Dendrogram from performing hierarchical clustering for top 20% of TorCamp members with highest betweenness centrality in July to December 2007.

Table 3. Similarity measures for possible cohesive subgroups in the TorCamp Google group.

	Similarity measures			
Consecutive time periods	Top 10% of nodes with highest betweenness centrality	Top 15% of nodes with highest betweenness centrality	Top 20% of nodes with highest betweenness centrality	Top 25% of nodes with highest betweenness centrality
January–June 2006 and July–December 2006	0.33	0.095	0.167	0.11
January–June 2007 and July–December 2007	0.2	0.194	0.318	0.306

different time periods. The movement of the members into clusters in different time periods is indicated by the arrows whereas the shades of the nodes in the legend indicate in which time period the member first appeared as a member of the subgrouping.

As indicated by the low similarity values for 2006 in table 3, the first two columns of figure 6 (representing the first and second half of 2006, respectively) show little evidence of cohesiveness, with members from the larger of the clusters in the first time period of January–June 2006 splitting into three clusters in the second half of that year. In contrast, there is good overlap between the largest clusters in the two halves of 2007 (in the rightmost two columns of figure 6), with seven of the nine people from the largest cluster in the first half of the year also being found in the largest cluster for the second half of the year. Overall, the information summarized in figure 6 shows the emergence of a core group of members in 2007 which remain in the same subgroup over time.

Figure 7 illustrates the dendrogram for the calendar year 2007 and shows that the top nodes (1, 32, 57, 15, 82, 62, 144) are first clustered together into a subgroup, followed by 18 and 100, 95, then 30 and 64. Comparing this with figure 6, it can be seen that the top nodes indicated in the rectangle in the

Table 4. Cohesive subgroups for the TorCamp Google group for 2006 to 2007 from similarity analysis (using betweenness centrality cutoffs of 20% for the 2006 and 2007 data).

Time period	Non-overlapping cohesive subgroups
January–June 2006	(1,18,30,32,45,54,62,67,115), (5,44,87)
July–December 2006	(61,101), (30,32,42,45,48,50), (1,18,57,139), (38,62), (17,95)
January–June 2007	(30,64), (18,65,100), (1,15,32,57,62,68,82,87,144), (16,95)
July–December 2007	(84,218,229), (100,167), (22,47,143,184), (45,64,160), (1,15,18,30,32,57,62,82,95,144,154,168)

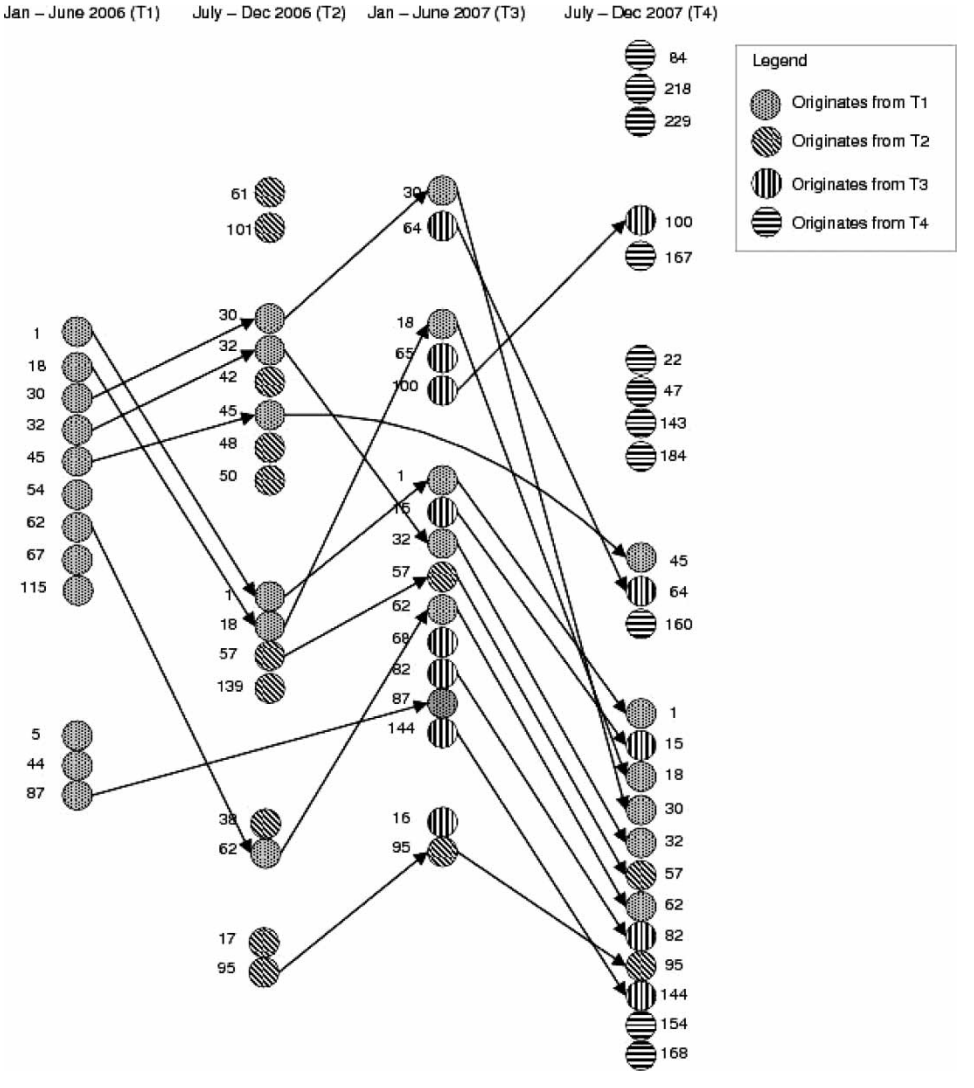

Figure 6. Visualizing and tracking cohesive subgroupings in the TorCamp Google group from 2006 to 2007.

dendrogram indeed correspond to the large cluster of nodes in both the first half of 2007 and second half of 2007.

4.6 Discussion

The similarity measure worked well in capturing the increased evidence of subgroup cohesion in 2007. However, there was a negative correlation between similarity as a measure of cohesion and the modularity criterion as a measure of strength of community that seems counterintuitive. If Q

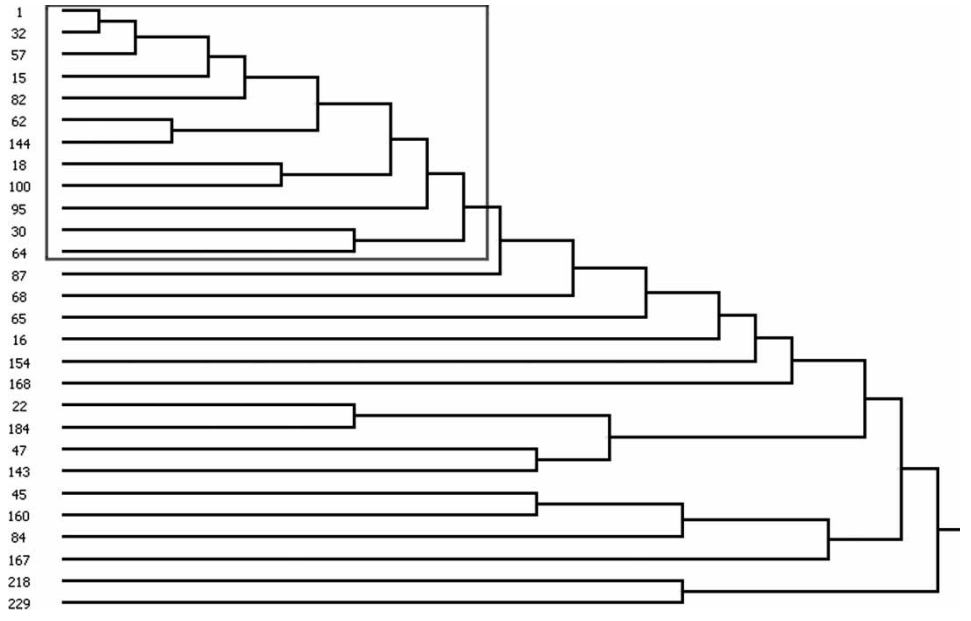

Figure 7. Dendrogram for the calendar year 2007 using weighted average hierarchical clustering for the TorCamp Google group.

predicts cohesiveness, then based on this data the subgroups should be most cohesive in 2006 (where modularity is the highest), whereas similarity predicts the opposite in that subgroups were the most cohesive (similar) in 2007. In our view similarity is a more direct measure of cohesiveness and the present results cast doubt on the value of modularity as a criterion for choosing cohesive subgroups.

Another issue relating to the application of the SCAN method is how to determine the appropriate duration of each time period. For the TorCamp case study, it was decided to choose four equally spaced time periods within the two years as a matter of convenience. Other time periods might conceivably have been used. In general, selecting an appropriate time period for analysis seems to require balancing two sides of a tradeoff which is characterized in the following sentences. Selecting too large a duration for time periods will result in a smaller number of periods for tracking subgroups, making it difficult to visualize the movement of members in and out of subgroups. On the other hand, selecting too small a duration for time periods will result in a larger number of periods, producing greater complexity in the detailed movements of members in and out of subgroups. Perhaps the best solution may be to select the duration of time period which results in the highest average similarity throughout the time in the case study using the same reasoning as was applied earlier with respect to cutting the dendrogram and selecting a betweenness centrality cutoff threshold.

The results from the TorCamp Google group case study suggest that it may not be possible to set a single cutoff value for betweenness centrality when

screening potential subgroup members. Further research is needed to determine what is the best centrality measure and cutoff measure to use for different datasets. Prior to such research being conducted it is suggested that the 20% betweenness centrality cutoff may be a reasonable place to start in implementing the Select step of the SCAN method for other datasets.

The modularity Q was used as a dendrogram partitioning tool in this case study. While a cohesive subgroup was found, it was only one subgroup and it was clearly identifiable across a wide range of similarity cutoffs. Thus its successful identification in this case is likely not attributable to the use of the Q criterion in selecting the partitioning cutoff value. In the TorCamp dataset, the value of Q tended to increase with the number of members in the group, indicating that it should probably not be compared across dendrograms containing different numbers of members. Another problem was that low values of Q were obtained in spite of evidence to suggest the presence of a cohesive subgroup. One possible explanation for this is the following. Modularity assumes that most if not all the objects or people being clustered should belong in one subgroup or another. There may also be the implicit assumption that communities should elicit fairly active involvement from most of their members. However, for an online group this will often not be the case.

As observed in the case study, there may be a core of active people in one subgroup with a few other (floating) subgroups that change over time (e.g. members 18 and 62 are part of the same subgroup in the first half of 2006 but are not in the second half of 2007) and do not lead to more long-lived interactions. On the other hand, many people who are interested in the group and actively read the information available to group members may not actively participate in online interactions. Trying to put inactive people into subgroups in cases such as this will be problematic at best. Thus, based on the present results, it is recommended that Q should be treated with caution when considering its application to the assessment of cohesion in online groups.

One further indicator of problems with the use of Q in this context was that a significant negative (Pearson) correlation between cutoff betweenness centrality and Q ($r = -0.509$, $N = 16$, $p < 0.05$) was obtained in the present case study. This may be explained by the fact that modularity is related to the proportion of people who are part of well-defined subgroups. Therefore, if the clustering is restricted to only a few people with the highest betweenness centralities, then they are likely to be in the core subgroup and there will be few people who are outliers or who are in small floating groups. In contrast, when the cutoff is increased, more people are added to the analysis, a smaller portion of which will be part of the core subgroup.

5. Conclusions

The ability to find cohesive subgroups in large social networks derived from online interactions promises to revolutionize many activities, including online search. Prior to the development of the SCAN method, research efforts had

not yet discovered a reliable system for identification of cohesive subgroups on a large scale. In formulating the SCAN method we combined the most promising of previous techniques (screening with a centrality method, and subgroup formation using clustering and partitioning) with an additional step that uses similarity measurement to assess subgroup cohesion over time. While there are many remaining research issues concerning when and how to use different centrality measures, different cutoff values, different partitioning criteria, and different similarity approaches, the general approach utilized in the SCAN method seems to be very promising. In the realistic case study reported in this research, the SCAN method was able to automatically identify a cohesive subgroup. More research is now needed to determine how the SCAN method can be adapted for different types of social networks and online interaction types.

The contributions made in this paper are as follows. First, a novel and computationally efficient three-step method for identifying cohesive subgroups was developed. Second the application of this SCAN method was demonstrated in a case study where a subgroup was automatically extracted from a social network formed based on the online interactions of a group of about 150 people that occurred over a two year period. Third, it was also found that the modularity criterion (Q) for discovering subgroups was negatively correlated with a measure of cohesiveness based on measured similarity over time. It is argued that similarity over time is a more direct measure of cohesiveness and that these results raise questions about the applicability of the Q criterion that need to be addressed in future research.

While the more complex topic of tracing the structure of online communities is outside the scope of the research, we expect that the scalable methods of subgroup formation proposed here will also form a useful starting point for research on automatic identification of broader communities based on patterns of online interaction.

Acknowledgements

We would like to thank the members of TorCamp for their generous participation in this research and for their many helpful comments and suggestions. We would also like to thank Bell University Laboratories for their generous support of this research, and the reviewers who helped to improve this paper.

References

E. Adar, L. Zhang, L.A. Adamic, and R.M. Lukose, "Implicit structure and the dynamics of blogspace. Workshop on the Weblogging Ecosystem", in *13th International World Wide Web Conference*, 2004.

B. Balasundaram, S. Butenko, I. Hicks, and S. Sachdeva, *Clique relaxations in social network analysis: The maximum k-plex problem*. Technical report, Texas, A and M Engineering, 2007.

C. Bird, *Community Structure in OSS Projects* [online], 2006. University of California, Davis. Available online at http://wwwcsif.cs.ucdavis.edu/ ~bird/papers/community-structure.pdf (accessed 8 January 2008).

S. Borgatti, M. Everett and L.C. Freeman, *Ucinet for Windows: Software for Social Network Analysis*. Lexington, KY: Analytic Technologies, 2002.

G.E. P. Box and D.R. Cox, "An analysis of transformations", *Journal of Royal Statistical Society*, 26 (Series B), pp. 211–246, 1964.

A. Breuer and J. Ratkiewicz, *Blogs in brief experiments in query result presentation using mead,* Technical report, Indiana University, 2005.

C.H. Brooks and N. Montanez, "Improved annotation of the blogosphere via autotagging and hierarchical clustering", in *WWW '06: Proceedings of the 15th international conference on World Wide Web*, ACM Press, 2006, pp. 625–632.

M. Chau, B. Shiu, I. Chan and H. Chen "Automated identification of Web Communities for business intelligence analysis", in *Proceedings of the Fourth Workshop on E-Business* (*WEB*), Las Vegas, 2005.

A. Chin and M. Chignell, "Identifying communities in blogs: Roles for social network analysis and survey instruments", *International Journal of Web Based Communities*, 3(3), pp. 345–363, 2007a.

A. Chin and M. Chignell, "Identifying subcommunities using cohesive subgroups in social hypertext", in *Proceedings of the 18th International ACM Conference on Hypertext and Hypermedia*, ACM Press, 2007b, pp. 175–178.

A. Chin and M. Chignell, "Identifying active subgroups within online communities", in *Proceedings of the 17th IBM Centre for Advanced Studies Annual International Conference on Computer Science and Software Engineering* (CASCON 2007), 2007c, p. 280–283.

A. Clauset, "Finding local community structure in networks", *Physical Review E*, 72, p. 026132, 2005.

D.R. Cutting, D.R. Karger, J.O. Pedersen and J.W. Tukey, "Scatter/gather: A cluster-based approach to browsing large document collections", in *Proceedings of the 15th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, ACM Press, 1992, pp. 318–329.

L. Danon, A. Díaz-Guilera, J. Duch and A. Arenas, "Comparing community structure identification", *Journal of Statistical Mechanics: Theory and Experiment*, 29(9), p. P09008, 2005.

W. de Nooy, A. Mrvar and V. Batagelj, *Exploratory Social Network Analysis with Pajek*, New York: Cambridge University Press, 2005.

L. Donetti and M.A. Munoz, "Detecting network communities: A new systematic and efficient algorithm", *Journal of Statistical Mechanics: Theory and Experiment*, 10, p. P10012, 2004.

N. Du, B. Wu, X. Pei, B. Wang and L. Xu, "Community detection in large-scale social networks", in *Proceedings of the Joint 9th WEBKDD and 1st SNA-KDD Workshop '07*, ACM Press, 2007, pp. 1–10.

J. Duch and A. Arenas, "Community detection in complex networks using extremal optimization", *Physical Review E* (*Statistical, Nonlinear and Soft Matter Physics*), 72(2), pp. 027104, 2005.

T. Dwyer, S.H. Hong, D. Koschutzki, F. Schreiber and K. Xu, "Visual analysis of network centralities", in *APVis '06: Proceedings of the 2006 Asia-Pacific Symposium on Information Visualisation* (*Darlinghurst, Australia*), Australian Computer Society Inc., 2006, pp. 189–197.

K.L. Elmore and M.B. Richman, "Euclidean Distance as a Similarity Metric for Principal Component Analysis", *Monthly Weather Review*, 129(3), pp. 540–549, 2001.

T. Erickson, "The world-wide-web as social hypertext", *Communications of the ACM*, 39(1), pp. 15–17, 1996.

G.W. Flake, S. Lawrence, C.L. Giles and F.M. Coetzee, "Self-organization and identification of web communities", *IEEE Computer*, 35(3), pp. 66–71, 2002.

L.C. Freeman, "Centrality in social networks: Conceptual clarification", *Social Networks*, 1, pp. 215–239, 1978.

L.C. Freeman, "Finding social groups: A meta-analysis of the Southern Women data", in *Dynamic Social Network Modeling and Analysis*, R. Breiger, K. Carley and P. Pattison, Eds, Washington: The National Academies Press, 2003, pp. 39–77.

N.E. Friedkin, "Social cohesion", *Annual Review of Sociology*, 30, pp. 409–25, 2004.

G. Frivolt and M. Bielikova, "The anatomy of a large-scale hypertextual web search engine", in *RAWS 2005 C Proceedings of the 1st International Workshop on Representation and Analysis of Web Space*, 2005, pp. 49–54.

D. Gibson, R. Kumar and A. Tomkins, "Discovering large dense subgraphs in massive graphs", in *Proceedings of the 31st International Conference on Very Large Data Bases* (*Trondheim, Norway, August 30–September 02, 2005*), 2005, pp. 721–732.

M. Girvan and M.E.J. Newman, "Community structure in social and biological networks", *Proceedings of National Academic of Science of USA*, 99, p. 7821, 2002.

P.A. Gloor, R. Laubacher, S.B.C. Dynes and Y. Zhao, "Visualization of communication patterns in collaborative innovation networks-analysis of some w3c working groups", in *CIKM '03: Proceedings of the*

twelfth international conference on Information and knowledge management, New York, NY, USA: ACM Press, 2003, pp. 56–60.
S.A. Golder and B.A. Huberman, *The structure of collaborative tagging systems.* Technical report, HP Labs, 2005.
V. Gomez, A. Kaltenbrunner and V. Lopez, "Statistical analysis of the social network and discussion threads in slashdot," in *WWW '08: Proceedings of the 17th international conference on World Wide Web*, New York, NY, USA: ACM Press, 2008, pp. 645–654.
R.A.M. Gregson, *Psychometrics of Similarity*, New York: Academic Press, 1975.
J.A. Hartigan, *Clustering Algorithms*, New York: John Wiley & Sons Inc., 1975.
P. Jaccard, "Distribution de la flore alpine dans le bassin des Dranses et dans quelques régions voisines", *Bulletin del la Société Vaudoise des Sciences Naturellese*, 37, pp. 241–272, 1901.
Y. Jung, H. Park, D. Du and B.L. Drake, "A decision criterion for the optimal number of clusters in hierarchical clustering", *Journal of Global Optimization*, 25, pp. 91–111, 2003.
J. Kleinberg, "Authoritative sources in a hyperlinked environment", *Journal of the ACM*, 46(5), pp. 604–632, 1999.
J. Kleinberg, "Bursty and hierarchical structure in streams", in *KDD '02: Proceedings of the eighth ACM SIGKDD international conference on knowledge discovery and data mining*, ACM Press, 2002, pp. 91–101.
R. Kumar, P. Raghavan, S. Rajagopalan and A. Tomkins, Trawling the web for emerging cyber-communities. *Computer Networks*, 1999.
A. Kurdia, O. Daescu, L. Ammann, D. Kakhniashvili and S.R. Goodman, "Centrality measures for the human red blood cell interactome", in *Engineering in Medicine and Biology Workshop*, IEEE Dallas, 2007, pp. 98–101.
J. Leskovec, K.J. Lang, A. Dasgupta and M.W. Mahoney, "Statistical properties of community structure in large social and information networks", in *WWW '08: Proceedings of the 17th international conference on World Wide Web*, New York, NY, USA: ACM Press, 2008, pp. 695–704.
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* [online], 2004. Available online at http://alumni.media.mit.edu/ ∼cameron/cv/pubs/04-01.pdf (accessed 16 June 2008).
C. Marlow, M. Naaman, D. Boyd and M. Davis, "Ht06, tagging paper, taxonomy, flickr, academic article, to read", in *HYPERTEXT '06: Proceedings of the seventeenth conference on Hypertext and hypermedia*, ACM Press, 2006, pp. 31–40.
N. Memon, D. Hicks, N. Harkiolakis and H.L. Larsen, "Detecting hidden hierarchy in terrorist networks: Some case studies", *Lecture Notes in Computer Science*, 5075, pp. 477–489, 2008.
M.S. Mizruchi, P. Mariolis, M. Schwartz and B. Mintz, "Techniques for disaggregating centrality scores in social networks", *Sociological Methodology*, 16, pp. 26–48, 1986.
M.E.J. Newman and M. Girvan, "Finding and evaluating community structure in networks", *Physical Review E*, 69, p. 026113, 2004.
M.E.J. Newman, "Modularity and community structure in networks", *Proceedings of the National Academy of Sciences*, 103(23), pp. 8577–8582, 2006.
J. Orford, "Implementation of criteria for partitioning a dendrogram", *Mathematical Geology*, 8(1), pp. 75–84, 1976.
J.C. Paolillo and E. Wright, *Social network analysis on the semantic web: Techniques and challenges for visualizing foaf* [online], 2005. Available online at http://www.blogninja.com/vsw-draftpaolillo-wright-foaf.pdf (accessed 8 January 2008).
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", in *Proceedings of Computer Supported Collaborative Learning 2003*, ACM Press, 2003, pp. 343–352.
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* (*ICDM 2007*), IEEE Press, 2007, pp. 643–648.
S. Santini and R. Jain, "Similarity measures", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 21(9), pp. 871–883, 1999.

R.R. Sokal and P.H.A. Sneath, *Principles of Numerical Taxonomy*, San Francisco, CA: W.H. Freeman, 1963.

S. Sterling, *Aggregation techniques to characterize social networks*. Thesis (Masters), Air Force Institute of Technology, 2004.

C. Tantipathananandh, T. Berger-Wolf and D. Kempe, "A framework for community identification in dynamic social networks", in *Proceedings of the 13th ACM SIGKDD international Conference on Knowledge Discovery and Data Mining* (*KDD'07*), ACM Press, 2007, pp. 717–726.

M. Tremayne, N. Zheng, J.K. Lee and J. Jeong, "Issue publics on the web: Applying network theory to the war blogosphere", *Journal of Computer-Mediated Communication*, 12(1), 2006.

A. Tversky, "Features of similarity", *Psychological Review*, 84(4), pp. 327–352, 1977.

J.R. Tyler, D.M. Wilkinson and B.A. Huberman, "E-mail as spectroscopy: Automated discovery of community structure within organizations", *The Information Society*, 21(2), pp. 143–153, 2005.

M.A.J. Van Duijn and J.K. Vermunt, "What is special about social network analysis?", *Methodology*, 2, pp. 2–6, 2005.

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.

H. Wang, W. Wang, J. Yang, and P.S. Yu, "Clustering by pattern similarity in large data sets", in *Proceedings of the 2002 ACM SIGMOD international Conference on Management of Data* (*Madison, Wisconsin, June 03–06, 2002*), New York, NY: ACM Press, 2002, pp. 394–405.

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

B. Wellman, "Structural analysis: From method and metaphor to theory and substance", in *Social Structures: A Network Approach*, B. Wellman and S.D. Berkowitz, Eds., Edinburgh, UK: Elsevier Limited, 2003, pp. 19–61.

B. Yaltaghian and M. Chignell, "Searching the hypermedia web: Improved topic distillation through network analytic relevance ranking", *New Review of Hypermedia and Multimedia*, 8(1), pp. 171–197, 2003.