Chapter 4
DISSECT: Data-Intensive Socially Similar Evolving Community Tracker

Alvin Chin and Mark Chignell

Abstract This chapter examines the problem of tracking community in social networks inferred from online interactions by tracking evolution of known subgroups over time. Finding subgroups within social networks is important for understanding and possibly influencing the formation and evolution of online communities. A variety of approaches have been suggested to address this problem and the corresponding research literature on centrality, clustering, and optimization methods for finding subgroupings is reviewed. This review will include a critical analysis of the limitations of past approaches. The focus of the chapter will then turn to novel methods for tracking online community interaction. First, the method proposed by Chin and Chignell called SCAN will be briefly introduced, where a combination of heuristic methods is used to identify subgroups in a manner that can potentially scale up to very large social networks. Then, we present the DISSECT method where multiple known subgroups within a social network are tracked in terms of similarity-based cohesiveness over time. The DISSECT method relies on cluster analysis of snapshots of network activity at different points in time followed by similarity analysis of subgroup evolution over successive time periods. The DISSECT method can be supplemented with behavioral measures of sense of community where administration of a questionnaire is feasible. Finally, we conclude the chapter with a discussion on possible applications and use of the DISSECT method.

A. Chin (✉)
Nokia Research Center, Building 2, No. 5 Donghuan Zhonglu, Economic and Technological Development Area, Beijing 100176, China
e-mail: alvin.chin@nokia.com

M. Chignell
Department of Mechanical and Industrial Engineering, University of Toronto, 5 King's College Road, Toronto, Ontario, M5S 3G8, Canada
e-mail: chignell@mie.utoronto.ca

A. Abraham et al. (eds.), *Computational Social Network Analysis*, Computer Communications and Networks, DOI 10.1007/978-1-84882-229-0_4,

4.1 Introduction

For many people and organizations, the Internet has become a place where communication interactions occur, supplanting part of the functionality of face-to-face meetings and group interactions. 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. Vast social hypertexts [36] are forming on the World Wide Web as networks of documents are increasingly being connected explicitly and implicitly to networks of people, and on a grand scale. Rich online social networks may be inferred relatively easily from blogs, Web forums, e-mail, and the emergence of Web 2.0 [86]. Web sites incorporating around tags, social bookmarks, podcasts, and mashups are examples of social hypertexts, where Web pages are nodes in a social network and hyperlinks between pages form links (relationships) between the nodes [21]. Feedback on Web pages foster online conversation, creating explicit links between authors and readers. It is now possible to infer social behavior from studying the relationships between nodes and patterns of communication that are formed from these communication media, using social network analysis [46]. As people communicate with each other through networks of interconnected Web pages, common ties may be established and social interactions may develop [8, 31, 38, 61, 93, 115], leading in some cases to a sense of virtual community [8] that may sometimes be analogous to the sense of community that people develop in physical environments [17].

When analyzing the social networks formed from online interaction, the focus can be on using graph structure (nodes and links) or on semantics (content analysis and text analysis). Previous research has shown that cohesive subgroups form communities of interest [29], have weak ties [46], and have cohesive bonds that bring people together [90]. Research on finding cohesive subgroups of online interactions within social networks remains an interesting and open issue.

Finding cohesive subgroups can be used as a first step to identify communities of interest that people belong to. Tags can then be assigned to the subgroups by using semantic analysis to form meta-tags that describe the interests or the expertise of a group of people. The assigned tags can then be used for activities such as marketing research, advertising, and expertise location. For instance, someone who seems to be actively involved in a number of communities involving Linux and its applications is likely to have expertise that is relevant to Linux problems, or to know someone else who has that expertise.

Online social networks evolve over time, and research has looked into the temporal aspects of social networks changing over time (e.g. [72, 100]). It has been found that groups discovered in social networks differ in their cohesiveness or bonding [90], which can be in time or space. Since online social networks have time inherent in their structure, cohesive subgroups can be defined as those that are similar over time based on Social Identity Theory [103] where group members feel closer if they are similar to each other. Cohesive subgroups can also be considered as optimum subgroups by calculating the optimum number of clusters [62], modularity [85], or optimizing graphs [104]. Similarity measures are used in our research to assess the

cohesiveness of subgroups. Since they explicitly consider changes in subgroupings over time, similar measures take into account network membership dynamics in the social network. Different types of similar measures can be constructed depending on the particular network dynamics observed.

This chapter addresses the problem of tracking community in social networks inferred from online interactions, by expanding on the problem of finding subgroups initially explored through the SCAN method [22] and addressing the limitations of the SCAN method. The first section of the chapter provides the literature review of research that can be used for finding subgroups for tracking community. This forms the basis for the creation and explanation of our SCAN method in the second section, which addresses the drawbacks of previous research. In addition, we describe the applications and limitations of the SCAN method. This results in the creation of a new framework called Data-Intensive Socially Similar Evolving Community Tracker or DISSECT as described in the third section, where multiple known subgroups within a social network are tracked in terms of similarity-based cohesiveness over time. We also discuss the implications of DISSECT for community evolution, and the evaluation of DISSECT using behavioral measures. Practical applications of the DISSECT approach in expertise location, marketing, and information search are discussed in the conclusions section of the chapter.

4.2 Finding Subgroups for Tracking Community

In this section, we review previous literature on finding subgroups in social networks. This work can be differentiated according to the different types of measure or structure that are targeted: centrality, cohesive subgroups and clustering, and similarity.

4.2.1 Centrality

Network centrality (or centrality) [44] is used to identify the most important/active people at the center of a network or those that are well connected. Numerous centrality measures such as degree [40,45,79,116], closeness [19,70,75], betweenness [30,47,49,57,76,85,106,108], information [25,26,42], eigenvector [37,84,95], and dependence centrality [78,79] have been used for characterizing the social behavior and connectedness of nodes within networks. The logic of using centrality measures is that people who are actively involved in one or more subgroups will generally score higher with respect to centrality scores for the corresponding network.

Researchers have compared and contrasted centrality measures in various social networks (e.g. [19, 34, 66, 80]), however three centrality measures which have been referred to the most with respect to subgroup membership are degree, closeness, and betweenness centrality.

Degree centrality [44] measures the number of direct connections that an individual node has to other nodes within a network. Nodes with high degree centrality have been shown to be more active [45] and influential [79]. Degree distribution has been used to visualize the role of nodes within subgroups [116]. *Closeness centrality* [44] measures how many steps on average it takes for an individual node to reach every other node in the network. In principle, nodes with high closeness centrality should be able to connect more efficiently or easily with other nodes, making them more likely to participate in subgroups. Closeness centrality has been used to identify important nodes within social networks [26, 70, 75], and identify members with a strong sense of community [19,20]. *Betweenness centrality* measures the extent to which a node can act as an intermediary or broker to other nodes [44]. 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 [30, 47]. Betweenness centrality has been used to reveal the hierarchical structure of organizations [49, 106, 108], and to identify opinion leaders [76] and influential members with a strong sense of community in blogs [19, 20].

Betweenness centrality is mostly used to find and measure subgroup and community membership [47–49, 76, 85, 106, 108], whereas degree and closeness centrality are used for characterizing influential members. Although network centrality measures are easy to calculate using computer programs such as Pajek [28] and UCINET [9], there has been no consensus among researchers as to the most meaningful centrality measure to use for finding subgroup members [25]. In extremely large social networks, computational efficiency may become an issue in selecting which centrality measure to use. With respect to three commonly used centrality measures, degree centrality is the easiest to calculate, closeness centrality is more complex, and betweenness centrality has the highest calculation complexity [18].

4.2.2 Cohesive Subgroups and Clustering

Finding cohesive subgroups within social networks is a problem that has attracted considerable interest (e.g. [42, 92, 102, 114]), because cohesive subgroups can indicate the most active members within a community. There are two types of approaches to finding cohesive subgroups. In the first approach, 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). In the second approach, clustering and partitioning methods are used which 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.

4.2.2.1 Clique and *k*-Plex Analysis

Cliques and k-plexes have been used to characterize groupings in social networks [2, 5, 21, 32, 92, 102, 113]. Cliques are fully connected subgroups [113] where each member has a direct connection to every other member in the subgroup, thus forming a completely connected graph within the subgroup. Pure cliques tend to be rare in social networks [102] because the criterion of full connectedness tends to be overly strict, thus pure clique analysis will miss many meaningful subgroupings [2, 5, 113].

In a subgroup of n members, the full connectedness requirement of cliques (where every person in the clique is connected to $n - 1$ other members in the clique), may be relaxed by requiring fewer connections to other group members. In the *k-plex* approach, connectedness is expressed as the minimum number $n - k$ of connections that each person in the group must have to the other group members. In a k-plex, as the parameter k increases, the connectedness requirement is relaxed [54]. k-Plex analysis has been used for finding subgroup members [5, 21, 82, 102] in a network. However, as with clique analysis, finding k-plexes in large networks is a computationally expensive and exhaustive process because it scales exponentially with the number of nodes in the network and is an NP-complete problem [5]. An additional issue with k-plex analysis is that the most appropriate value of k for subgroup analysis in a particular social network may not be obvious.

4.2.2.2 Clustering and Partitioning

Clustering and other techniques such as link analysis [10, 65] and co-citation analysis [1, 41, 64, 67–69] can be used to detect subgroups within social hypertext networks. Hierarchical clustering automates the process of finding subgroups by grouping nodes into a cluster if the nodes are similar and then successively merging clusters until all nodes have been merged into a single remaining cluster. Techniques based on hierarchical clustering have been used to quantify the structure of community in Web pages [24, 30, 47, 73], blogs [88, 89], and discussion groups [50]. Hierarchical clustering (using such algorithms as in [55, 118]) results in a hierarchy (tree) being formed where the leaves of the tree are the nodes that are clustered and can be visualized as dendrograms. Nested clusters within the derived hierarchy may then be inferred to be subgroups.

In contrast to hierarchical cluster analysis, the groups formed in partitioning methods are not nested. 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. 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.

Orford [87] described a range of criteria for determining where to partition a dendrogram; however, the best criterion to use will generally vary with the problem context. Recent research has tended to assess specific measures for obtaining an optimal partition (e.g. [62]), using modularity (designated as Q) proposed by Newman and Girvan [85] for finding community structure in [7, 27, 33, 91, 94], vector partitioning [111] or normalized cut metrics [71, 97, 98, 117]. However, as noted by Radicchi et al. [91], it is not clear whether the "optimal" partitions are representative of real collaborations in the corresponding online communities. Van Duijn and Vermunt [110] noted that it is difficult to determine which measure is the most appropriate to use across a range of applications.

4.2.2.3 Summary

Hierarchical clustering has been shown to produce similar subgroupings as k-plex analysis and is less computationally intensive [22]. Modularity has been proposed as an optimizing method for partitioning dendrograms obtained from hierarchical clustering into subgroupings. Researchers (such as Lin et al. [74] and Traud et al. [105]) have combined 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.

4.2.3 Similarity

Cohesive subgroups should have a core group of people that remain the same over different time periods, because we hypothesize that subgroups will be cohesive to the extent that their members remain together based on "common fate" [109] where objects are more likely to be part of a grouping if they move together over time. However, 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. In general, cohesive families of subgroups at one time period should be similar to corresponding subgroups at a different time period.

Mathematically, similarity may be viewed as a geometric property involving the scaling or transformation necessary to make objects equivalent to each other. Several mathematical similarity measures have been defined such as Euclidean distance [35] and the cosine distance or dot product [112]. Other models of similarity are based on comparison of matching and mismatching features using a set-theoretic approach such as Tversky's feature contrast model [107], Gregson's content similarity model [51], and the Jaccard coefficient [59], which is defined as the size of the intersection divided by the size of the union of the objects being compared.

For assessing cohesion within subgroups, Johnson [60] proposed the ultrametric distance as a way of measuring distance within a hierarchy. For comparing

two different clustering hierarchies, one heuristic method for estimating similarity consists of converting each hierarchy to a matrix of ones and zeros where the ones represent the parent–child links in each hierarchy. The similarity between two hierarchies is then estimated as the correlation between the two corresponding matrices of ones and zeroes. A more formal approach is to use quadratic assignment [58] to assess the similarity between two partitions. Other related work by Falkowski et al. [39] focused on finding community instances using similarity.

From the preceding review, it can be seen that there are a number of similarity assessment approaches that may be applied to the problem of measuring the cohesiveness of subgroups over time. Unlike previous similarity methods, our method addresses the strength of cohesiveness of subgroups over time using a custom-built measure of similarity to meet the demands of identifying subgroups in a dynamic social network based on the content model of similarity [51].

4.3 Tracking Online Community Interactions Using the SCAN Method

A number of methods already exist for finding subgroups and clusters, but there is as of this writing no method that can potentially scale up to handle large social networks, or can identify subgroups that remain cohesive over time using an unsupervised learning approach. In this section, we explain the SCAN method and use an example to illustrate how to track online community interactions.

4.3.1 Social Cohesion Analysis of Networks (SCAN) Method

The Social Cohesion Analysis of Networks (SCAN) method was developed for automatically identifying subgroups of people in social networks that are cohesive over time [22]. The SCAN method is to be applied based on the premise that a social graph can be obtained from the online community interactions where the links are untyped (i.e., there are no associated semantics). In the social graph, each link represents an interaction between two individuals where one individual has responded to the other's post in the online community. The SCAN method has been designed to identify cohesive subgroups on the basis of social networks inferred from online interactions around common topics of interest. The SCAN method consists of the following three steps:

1. Select: Selecting potential members of cohesive subgroups from the social network.
2. Collect: Grouping these potential members into subgroups.
3. Choose: Choosing cohesive subgroups that have a similar membership over time.

4.3.1.1 Select

In the first step, the possible members of cohesive subgroups are identified. We set a cutoff value on a measure that is assumed to be correlated with likelihood of being a subgroup member, and then filter out people who fail to reach the cutoff value on that measure. We use betweenness centrality as this cutoff measure, since prior research has found that it does a fairly good job of identifying subgroup members although other centrality measures such as degree and closeness centrality could also be used. By selecting a cutoff centrality measure, we obtain a subgraph of the original social graph where all members that have a centrality below the cutoff centrality measure are removed, resulting in a list of potential active members of subgroups.

4.3.1.2 Collect

In the second step of the SCAN method, the objective is to recognize active subgroups from the subset of network members identified in the Select step. This is accomplished by forming subgroups of the selected members using cluster analysis, specifically weighted average hierarchical clustering. In general, hierarchical cluster analysis is more computationally efficient than k-plex analysis [22], and weighted average hierarchical clustering is a relatively efficient approach that has been used frequently by researchers. The output of hierarchical clustering is a set of nested, non-overlapping clusters, i.e., a tree, or dendrogram (a visual representation of a tree that is frequently used to represent a hierarchy of nested clusters visually). The extraction of the hierarchy shows potential cohesive subgroups, but it does not actually partition the people into a particular set of non-nested subgroups.

4.3.1.3 Choose

In the third step of the SCAN method, we identify the most cohesive subgroups over periods of time by computing the similarity of the possible cohesive subgroups between pairwise consecutive periods of time, and then selecting the cohesive subgroupings that result in the highest similarity. In comparing subgroups between pairwise consecutive periods of time, we need to consider networks in each period where membership is either fixed, where new members enter but existing members do not leave the network, or where new members enter and existing members leave the network at different time periods. The present version of the SCAN method only considers constant membership and new members that enter the network, through the application of two similarity measures.

For the first similarity approach (as described in detail in [22]), the cohesion across all subgroups between two consecutive time periods T_1 and T_2 can be calculated according to Eq. 4.1:

$$Sim_{T_1,T_2} = \frac{2 * N(T_1 \cap T_2)}{N(T_1 \cup T_2)}, \tag{4.1}$$

where $N(T_1 \cap T_2)$ is the number of pairs where both members are in the same cluster in T_1 and in T_2. $N(T_1 \cup T_2)$ is the total number of pairs who are in the same cluster in either (or both) T_1 and T_2. The parameter 2 is added as a multiplier to the numerator of this expression so that the resulting similarity measure is scaled between 0 and 1.

The second similarity approach (as detailed in [18]) measures the cohesion of the largest individual subgroup. This examines all the possible pairwise relationships between members of the subgroup, and determines how many of the pairs still exist (i.e., are inside a subgroup) in the second time period. The similarity can then be calculated using the following formula according to Eq. 4.2:

$$Sim_{T_1,T_2} = \frac{N(S_1 \cap T_2)}{N(S_1)} , \tag{4.2}$$

where S_1 is the largest subgroup in T_1, $N(S_1 \cap T_2)$ is the number of common pairs from the largest subgroup S_1 that still exist in T_2, and $N(S_1)$ is the number of pairs in the largest subgroup S_1. As an example, if there was 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 ten pairs from the first time period and after that there would be three choose two or three pairs remaining together in one of the spin-off groups plus one pair together in the second spin-off for a total of four pairs remaining together in the second time period, so the similarity between the two time periods would then be $4/10 = 0.4$. It can be seen that this approach results in an intuitive measure that is easy to calculate.

These measures of similarity can then be used to assess cohesiveness over time with a betweenness centrality cutoff being chosen that maximizes this measure of similarity in the sample. The clusters for each time period T_1 and T_2 that are obtained from the highest similarity using the selected betweenness centrality cutoff, then form the cohesive subgroups, with the level of measured similarity between adjacent time periods providing an indicator of the amount of cohesiveness.

4.3.2 Application and Limitations of the SCAN Method

The SCAN method was tested on the TorCamp Google group (as explained in [22]) and a set of YouTube vaccination videos and its comments (as explained in [23]). The SCAN method was able to find a cohesive subgroup in the TorCamp Google group case study even though semantics of the links between members in the social network were not utilized in this task. In general, the cohesiveness criterion using similarity and the SCAN method worked well with the TorCamp Google group because the dataset reflected online interactions of topics of common interest, from which the discussions had clearly identifiable active members based on the number of responses.

However, the method was unable to distinguish between two types of activist groups (anti-vaccination and pro-vaccination) in a set of YouTube vaccination

Table 4.1 Content analysis of the comments made on anti-vaccination and pro-vaccination videos (using the anti- and pro-labeling of videos provided by Keelan et al. [63])

Network	Number of videos	Number of anti-comments	Number of pro-comments
Anti-vaccination	34	76	6
Pro-vaccination	66	67	13

videos because the conversational threads contained mixtures of people from those two types of activist groups (anti- and pro-vaccination) as shown in Table 4.1 [23].

In datasets where online interactions revolve around debate and activism, the SCAN method will not be able to find particular classified subgroups, because these types of people are intertwined within the discussions around social media (such as the YouTube vaccination videos) and many of the comments may not be related to the discussion.

One of the remaining challenges in applying the SCAN method concerns the selection of time periods across which subgroup cohesion is compared. It seems likely that the appropriate time period will depend on how quickly a subgroup evolves or grows, and also on the size of the subgroup and the rate of interaction that occurs within the subgroup. Within each time period selected for analysis, there needs to be social interaction to allow application of the SCAN method, while it should also be possible to construct multiple periods where there is a reasonable expectation that sufficient cohesion will occur across time periods to make similarity assessment viable. While past research has examined cohesion between pairs of time windows, it is possible that in long-lasting subgroups comparison of cohesion across multiple time windows may also provide more accurate estimates of cohesion within subgroups.

4.4 Data-Intensive Socially Similar Evolving Community Tracker (DISSECT)

In this section, we outline a framework for a new method for tracking community evolution based on the SCAN method. This method is referred to as the Data-Intensive Socially Similar Evolving Community Tracker (DISSECT). The steps in this method are listed and possible techniques for implementing each step are discussed. Detailed implementation of the steps in the DISSECT method are left as topics for future research.

In the SCAN method, the problem of finding cohesive subgroups across different time periods using similarity was addressed. The DISSECT methodology provides a framework for more extensive analysis of community formation in online

interactions. The DISSECT method addresses the following shortcomings of the SCAN method:

1. The SCAN method only focused on betweenness centrality; other centrality measures may be useful.
2. The SCAN method only looked into two types of similarity measures (constant membership and members entering the network); there is a need to examine for other types.
3. The time periods used in the SCAN method were defined ad hoc as a matter of convenience, without any systematic evaluation.
4. The SCAN method fails if semantic properties determine subgroup membership.

4.4.1 Framework for the DISSECT Method

The DISSECT method addresses the shortcomings identified in the SCAN method, and uses the following steps:

1. Find the initial time periods for analysis.
2. Label subgroups of people from the network dataset using content analysis and semantic properties. If possible, individuals are also labeled so as to facilitate later similarity analysis between subgroups at different time periods.
3. Select the possible members of known subgroups to be tracked using the Select step from the SCAN method.
4. Carry out cluster analysis of interaction data taken at snapshots in time and involving known subgroups of people (using the Collect step from the SCAN method).
5. Repeat steps 3 and 4 for different values of centrality (note that the DISSECT approach is agnostic in terms of which of the many available measures of centrality should be used).
6. Calculate similarity of subgroups for the designated time periods from step 1 using the clustering results of the previous step. In this case, the similarity measure can be augmented to take into account semantic labels assigned to different people. The advantage of using similarity measures that include semantics as well as link structure is that they can identify non-cohesive groups of people with heterogeneous viewpoints and affiliations who may yet have bursts of interaction during specific periods of debate.
7. Repeat steps 2 through 6 for different time period intervals and combinations.
8. Construct a chronological view of each subgroup showing how it changes over time (as the assigned semantic labels change), and also showing how subgroups merge and split in response to the changing interests of their members.

The ultimate goal in a DISSECT analysis is to trace the evolution of subgroups into communities. The DISSECT method does not stand as a theory of how communities form. However, logically it would seem that if communities do emerge out of online interactions then they are likely to evolve, initially, from smaller subgroups.

Some of the steps in the DISSECT method are now described further below.

4.4.1.1 Find the Initial Time Periods for Analysis

Here, the objective is to 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 so that there is enough data to distinguish potential subgroups, and there should be a sufficient number of them to estimate cohesion over time. Further research is needed to determine guidelines concerning what lengths and numbers of time periods should be used for different types of online interaction.

4.4.1.2 Label Subgroups of People from the Network Dataset Using Content Analysis and Semantic Properties

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 [52] and other natural language-processing techniques that analyze content of the posts [53] can be used for classifying the links and the nodes. For example, in the YouTube vaccination video study as mentioned in the previous section, people were labeled based on the comments they made, and videos were labeled based on whether they were anti-vaccination or pro-vaccination in viewpoint.

4.4.1.3 Select the Possible Members of Known Subgroups that You Want to Track (from the Previous Step)

While betweenness centrality appears to be a useful filter for screening potential subgroup members, further research is required to assess when and how other centrality measurement methods might be used. In addition, more rigorous criteria (other than simple inspection of the frequency distribution) are needed for choosing the appropriate cutoff centrality in the Select step of the SCAN method. As a starting point, degree centrality may be a better criterion 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 lower computational complexity than the betweenness, and closeness centrality measures.

4.4.1.4 Cluster Analysis to Infer Possible Subgroups

Once the network is screened for potential subgroup members, hierarchical clustering (e.g., weighted average hierarchical clustering) is used to group the potential members into subgroups. The output of the cluster analysis is a hierarchy (dendrogram) that contains nested subgroups of people. The nested nature of these

subgroups means that hierarchical cluster analysis does not yield unambiguous subgroups (i.e., a partition) directly. In order to create a subgroup partition, the dendrogram has to be cut at a particular similarity value. Further research is needed on how best to cut dendrograms in order to obtain partitioned subgroups.

The issue of how to create subgroup partitions is likely to be particularly important in situations where multiple subgroups are expected. Orford [87] made the point that the best method for partitioning a dendrogram will depend on the type of data. It is suggested that while modularity has received recent attention in partitioning community data into subgroups, further research is needed that compares its effectiveness in this context to other potentially useful measures that have been proposed in the literature. For example, random walks have been suggested as an alternative approach for finding community structure [101].

4.4.1.5 Repeat the Steps in Sects. 4.4.1.3 and 4.4.1.4 for Different Values of Centrality

There are a number of ways in which centrality measurement and clustering or partitioning can be carried out. If we treat maximization of the subgroup similarities across time periods as the objective (criterion), then the problem of choosing the particular methodology to be used can be envisioned as a parameter search (maximization) problem where the goal is to obtain subgroupings that are maximally cohesive (self-similar) over time.

4.4.1.6 Calculating Similarity

As noted in the preceding subsection, similarity (as an indicator of cohesiveness) can be treated as an objective function to be maximized. Further research is needed to determine which similarity measures are best for finding cohesive subgroups, since Eqs. 4.1 and 4.2 were formulated from first principles and their use in this context is not, as yet, backed by research evidence. As noted elsewhere, the similarity measures that were developed in SCAN have to be modified to take into account incoming and outgoing actors from different time periods. When pairwise similarities are calculated, for each pairing of people, the pair should only be included in the analysis if both people in the pair were part of the social network for both time periods being assessed.

Other issues concern the problem of making the SCAN or DISSECT methodology scalable to very large online social networks. Subgroupings can be represented as matrices of ones and zeros (where an edge with one between two people indicates some kind of interaction with the challenge being to implement the set-theoretic similarity measure as a sequence of relatively simple operations on the two matrices of binary data). Block modeling may be one way of accomplishing this by identifying blocks of structurally similar actors within an adjacency matrix derived from social network data [3, 11, 99]. While blocks may indicate the presence of cohesive subgroups, Frank [43] has argued that blocks differ from cohesive subgroups since

blocks of actors engage in common patterns of interaction throughout the network, but do not necessarily engage in the direct interactions that occur between members of cohesive subgroups. Nonetheless, block modeling could be applied to the adjacency matrix to construct possible subgroups in an alternative fashion.

4.4.1.7 Repeat Steps in Sects. 4.4.1.2 Through 4.4.1.6 for All Possible Time Period Intervals and Combinations

How do we know that the time periods that are selected initially in step 1 provide the most optimum cohesive subgroups? We need to study the effect of varying the number of time periods and the size of each time period on subgroup cohesion and similarity. There are many other methods that can be used for analyzing networks over time such as probabilistic stochastic models, time windows, and time graphs and burst analysis. For example, SIENA [100] is a network package that uses stochastic, actor-oriented models for the evolution of social networks which could be used to take all the possible configurations of the network ties and then model the observed network dynamics (the actual ties that happened in different time periods). The results could then be compared with that found from the similarity analysis in order to evaluate its performance. In addition, time windows can be defined such as in Social Networks Image Animator (SoNIA) in order to visualize the subgroups [81] and sliding time windows can be used for varying the time window to determine their effect on the formation of subgroups and their cohesiveness [39]. By representing the social interactions as time graphs that have time added to the link indicating when the social interaction happened (forming typed links), burst analysis can be used in order to discover subgraphs of bursts that indicate cohesive subgroups around a specific event [67] and to study the formation of groups [4].

No matter which method is used to define the time periods, we need to repeat the steps in Sects. 4.4.1.2 through 4.4.1.6 for each time period selected in order to determine which one provides the optimal subgroup cohesion.

4.4.2 Using DISSECT to Identify Community Evolution

As motivation for identifying community evolution, we previously used the results of our SCAN method [22] to track the subgroups within the TorCamp Google group for 2 years. Figure 4.1 visualizes the cohesive subgroups from the TorCamp Google group and shows how members in the TorCamp Google group move in and out of subgroups at 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 indicate (as shown in the legend) in which time period the member first appeared as a member of the subgrouping.

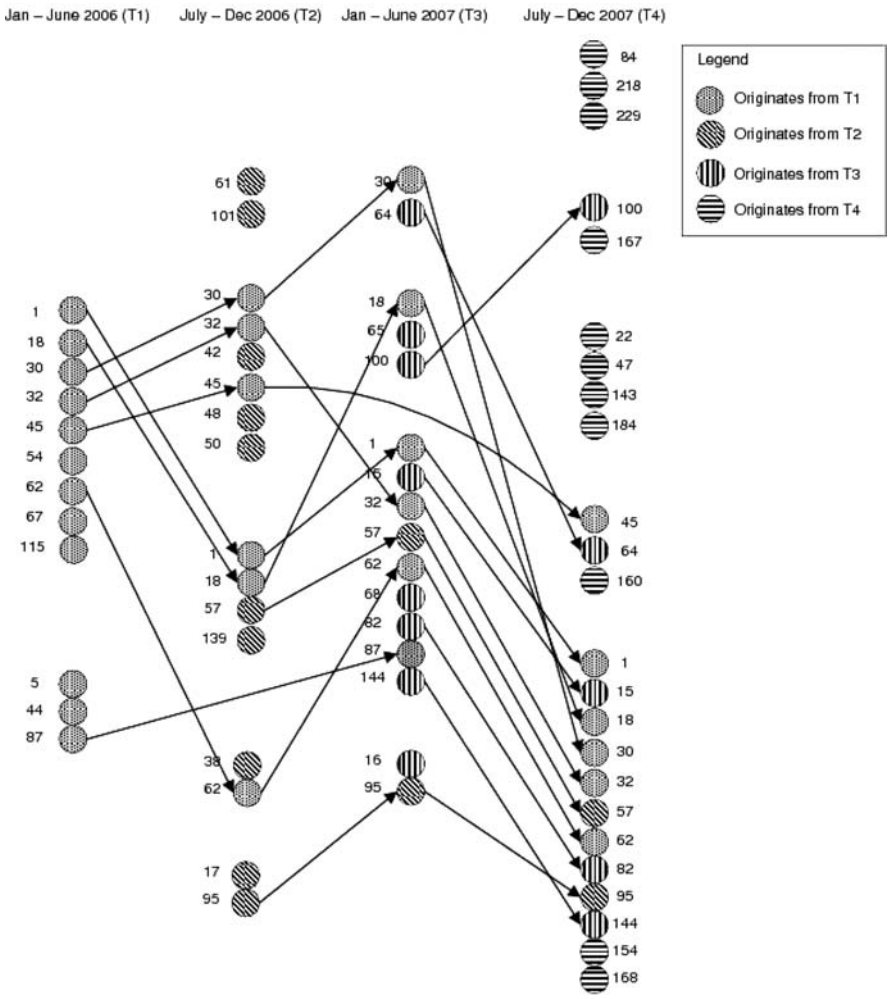

Fig. 4.1 Visualizing cohesive subgroupings in the TorCamp Google group from 2006 to 2007

Similarly, from the results of DISSECT, Communities might evolve in a number of ways. For instance:

1. A core subgroup growing until it becomes a highly centralized community

   By computing the degree and closeness centralities of each of the people in the network in each time period, we can see if the centralities and the number of members increase. This will determine which people from the core subgroup become more tightly connected to the group. If the number of members increases within the core subgroup is growing and the members are becoming more tightly connected, therefore forming into a highly centralized community.

2. A number of subgroups coalescing into a multipolar community

   At a specific time period, we can have various numbers of subgroups within an online community. These subgroups can combine to form a multipolar community in which various subgroups can be connected to each other. During the merging and community formation process, leaders or activity members within the contributing (merging) subgroups will act as bridge connectors, and the betweenness centrality of these members will likely be the highest in the overall network. One can then follow how the subgroups change with time, and track which members are more central at different times as the network evolves.
3. Coalesced subgroups forming clusters which then become affiliated in a multilayered community which may have both centralized and multipolar aspects

   The community formed can have a core subgroup that is highly centralized but with subgroups that coalesce into a multipolar community. To find these types of subgroups, we can use techniques based on the previous steps as starting points.

Presumably many different methods of community formation occur online. The unsolved problem is how to come up with reliable and flexible algorithms for tracking a broad range of community evolution processes. To determine how well the algorithms find subgroups and the central members, we can use behavioral measures such as Sense of Community, Social Network Questionnaire, and Frequency of Ties [18] for evaluating the subgroups found in the DISSECT method.

4.4.2.1 Behavioral Measures for Evaluating Subgroups

The Sense of Community inventory [77] assesses the internal perceptions that a person has about his or her role in a community. The inventory consists of four components: membership, need, influence, and shared emotional connection. Chavis [16] created the Sense of Community Index to provide a quantitative measure for sense of community and its components. Members of cohesive subgroups have been shown to have high scores of sense of community (calculated in [16]) from our case study of the TorCamp Google group [21].

The Social Network Questionnaire (developed by Chin and Chignell [18]) is based on the Social Network List (SNL) developed by Hirsch [56] and the Social Support Questionnaire (SSQ) developed by Sarason et al. [96]. These tools are based on name generator techniques [6, 12–14, 113] in which participants are asked to name all the people that they know or communicate with (considered as ties). The Social Network Questionnaire is designed to obtain a social network of acquaintances that participants have by asking them which members that they know from the subgroups identified in DISSECT and how close their relationship is with those members.

The Frequency of Ties questionnaire (also developed by Chin and Chignell [18]) is designed to provide a more quantitative interpretation of the frequency of communication for each network tie. Participants are asked the frequency of communication using approximate number of exchanges that they have with a list of members

identified as part of subgroups from the DISSECT method. This helps to determine which members of the subgroups are the most influential based on communication with other members in the network.

All three measures described above can be used to characterize the behavior of cohesive subgroup members by making correlations with centrality and the edge weights in the network (number of interactions between two members). For example, we performed correlations between sense of community, number of ties known with other members and network centrality. Table 4.2 summarizes these results.

Figure 4.2 shows that degree centrality in the TorCamp Google group tends to increase as the number of ties known increases.

According to the TorCamp Google group case study that we conducted [18], cohesive subgroup members have higher centrality scores, send more messages, have weak ties, have greater number of known ties, and have a higher sense of community

Table 4.2 Pearson correlations between Sense of Community subscales, number of ties known, and network centrality ($N = 9$) in the TorCamp Google group

Sense of community subscale	Number of ties known	Degree centrality	Betweenness centrality	Closeness centrality
Membership	0.737	0.589	−0.408	0.542
Emotional connection	0.634	0.292	−0.552	0.453
Influence	0.211	0.096	−0.398	0.182
Needs	0.089	0.173	−0.563	0.196

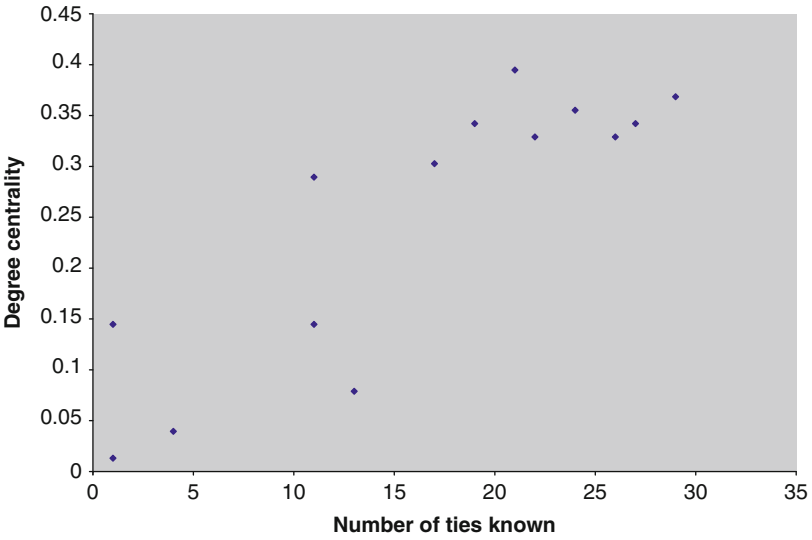

Fig. 4.2 Association of number of ties known on degree centrality

than other members that are not part of cohesive subgroups. Further research, however, is still needed to show how the results generalize to other data sets.

4.5 Conclusion

In this chapter, we propose a framework for tracking community evolution in an online community called DISSECT or Data-Intensive Socially Similar Evolving Community Tracker. This framework is an expanded and enhanced version of the SCAN method [22], for finding cohesive subgroups in online interactions. The framework is designed to be a step-by-step process to track the evolution of community members. This chapter has discussed the steps in the framework, and has raised research issues that need to be considered.

4.5.1 Applications

Given that a method now exists for tracking cohesive subgroups from large networks (of at least 200 nodes) automatically, there are many applications to which it can be used of which three examples will be briefly mentioned here

1. Marketing
2. Expertise location
3. Information search

4.5.1.1 Marketing

Knowledge of where people live has long been used to develop targeted marketing campaigns on the assumption that "birds of a feather flock together." For instance, Claritas Corporation of San Diego developed the PRIZM system which assigns one or several of 66 clusters to each US zipcode. The clusters range from #1 Upper Crust, composed mainly of multimillionaires, to the relatively impoverished #66 Low-Rise Living, with many lifestyles in between. Clearly, it would be beneficial to have similar online maps of where people "live" on the Internet that can predict their lifestyle and interests. Instead of the zipcodes where people "hang out" physically, online clusters could be based on who people interact with online and the subgroups that they belong to. Products and advertisements could then be marketed to relevant members and interested subgroups using the DISSECT method. For example, based on a subgroup of friends who are avid watchers of the television series *Lost* (identified through the DISSECT method), and given that some of those friends also watch another television series called *Heroes*, then the system would recommend the user to watch *Heroes*.

4.5.2 Expertise Location

Well-cataloged and organized maps of online communities and subgroups would also facilitate tasks such as expertise location, either by inferring expertise directly based on subgroup membership, or else by asking relevant subgroups to nominate a suitable expert who could respond knowledgeably to an inquiry once irrelevant members such as spammers have been removed. For many applications, finding the right neighborhood may be the main problem (after which "locals" can act as guides to detailed information) and labeled subgroups can serve as entry points into neighborhoods. An example of such an application could be an expert-finding application for finding experts and leaders in online environments using the SCAN method (such as blogs, video, and social networking platforms such as Facebook and MySpace). This would then allow users to contact and connect to those experts to grow their social network and engage in constructive debate and conversation.

4.5.3 Information Search

The DISSECT method may also be useful for enhancing information search practice, including search within a social network and community-based search. The first approach deals with searching for content or searching for people, using a social network as a form of filtered or targeted network compared to the broad search that is carried out today on search engines such as Google. This approach is needed because according to Cervini, "without the ability to execute directed searches, through a social network, the transition cost of finding other users within the system is simply too high to warrant using the system" [15]. Applications such as Social Network and Relationship Finder (SNARF) [83] have been created for visualizing and ranking the most relevant contacts to the user based on contact interactions in e-mail, for example. However, these types of applications are very general and do not take into account how to identify the experts within the social network and how to recommend specific people to the user. The DISSECT method can be used to find the "expert" people within the social network. A contact search application could be created as an interface to search for information from relevant people and subgroups in a user's social network, obtained from e-mail or some other online collaborative medium or even people's search histories.

The second approach is community-based search and involves the development of community-based search engines that use knowledge of labeled subgroups to improve search performance. This can be done both by labeling information (e.g., blogs or Web pages) in terms of the subgroups that authors and readers belong to, and also in terms of modifying queries based on the subgroups that the people composing those queries belong to. Pages that reflect matching communities or interests could then be ranked higher in search results, and could be labeled according to the communities that they are associated with.

Acknowledgements We would like to thank the TorCamp group for allowing us to use their Google Groups site for data analysis and the participants for completing the behavioral surveys. The authors would also like to thank Jennifer Keelan and Kumanan Wilson for providing us with the content analysis information from the YouTube vaccination videos shown in Table 4.1.

References

1. Adar E, Li Z, Adamic LA, Lukose RM (May 2004) Implicit structure and the dynamics of blogspace. In: Workshop on the weblogging ecosystem, 13th international World Wide Web conference
2. Alba RD (2003) A graph-theoretic definition of a sociometric clique. J Math Sociol 3:113–126
3. Anderson CJ, Wasserman S, Faust K (1997) Building stochastic blockmodels. Social Networks 14:137–161
4. Backstrom L (2006) Group formation in large social networks: membership, growth, and evolution. In: KDD 06: Proceedings of the 12th ACM SIGKDD international conference on knowledge discovery and data mining, ACM Press, pp 44–54
5. Balasundaram B, Butenko S, Hicks I, Sachdeva S (2007) Clique relaxations in social network analysis: the maximum k-plex problem. Technical report, Texas A and M Engineering
6. Bass LA, Stein CH (1997) Comparing the structure and stability of network ties using the social support questionnaire and the social network list. J Soc Pers Relat 14:123–132
7. Bird C (2006) Community structure in oss projects. Technical report, University of California, Davis
8. Blanchard AL, Markus ML (2004) The experienced "sense" of a virtual community: characteristics and processes. SIGMIS Database 35(1):64–79
9. Borgatti SP, Everett GM, Freeman CL (2002) Ucinet for windows: software for social network analysis. Analytic Technologies, Harvard, USA
10. Brin S, Page L (1998) The anatomy of a large-scale hypertextual web search engine. In: WWW7: Proceedings of the 7th international conference on World Wide Web 7. Elsevier Science BV, Amsterdam, the Netherlands, pp 107–117
11. Burt R (1982) Toward a structural theory of action: network models of social structure, perception, and action. Academic, New York
12. Burt R (1984) Network items and the general social survey. Social Networks 6:293–339
13. Campbell KE, Barret AL (1991) Name generators in surveys of personal networks. Social Networks 13:203–221
14. Carrington PJ, Scott J, Wasserman S (2006) Models and methods in social network analysis. Cambridge University Press, New York, NY, USA
15. Cervini AL (2003) Network connections: An analysis of social software that turns online introductions into offline interactions. Master's thesis, New York University, New York, NY
16. Chavis DM (2008) Sense of community index. http://www.capablecommunity.com/pubs/Sense%20of%20Community%20Index.pdf. Accessed 30 September 2008
17. Chavis DM, Wandersman A (1990) Sense of community in the urban environment: a catalyst for participation and community development. Am J Commun Psychol 18(1):55–81
18. Chin A (January 2009) Social cohesion analysis of networks: a method for finding cohesive subgroups in social hypertext. PhD thesis, University of Toronto
19. Chin A, Chignell M (2006) A social hypertext model for finding community in blogs. In: Proceedings of the 17th international ACM conference on hypertext and hypermedia: tools for supporting social structures. ACM, Odense, Denmark, pp 11–22
20. Chin A, Chignell M (2007) Identifying communities in blogs: roles for social network analysis and survey instruments. Int J Web Based Commun 3(3):345–363
21. Chin A, Chignell M (2007) Identifying subcommunities using cohesive subgroups in social hypertext. In: HT '07: Proceedings of the 18th conference on hypertext and hypermedia. ACM, New York, NY, USA, pp 175–178

22. Chin A, Chignell M (2008) Automatic detection of cohesive subgroups within social hypertext: A heuristic approach. New Rev Hypermed Multimed 14(1):121–143
23. Chin A, Keelan J, Pavri-Garcia V, Tomlinson G, Wilson K, Chignell M (2009) Automated delineation of subgroups in web video: A medical activism case study. Journal of Computer-Mediated Communication. In Press
24. Clauset A (2005) Finding local community structure in networks. Phys Rev E 72:026132
25. Costenbader E, Thomas WV (October 2003) The stability of centrality measures when networks are sampled. Social Networks 25:283–307
26. Crucitti P, Latora V, Porta S (2006) Centrality measures in spatial networks of urban streets. Phys Rev E 73:036125
27. Danon L, Duch J, Diaz-Guilera A, Arenas A (2005) Comparing community structure identification. J Stat Mech Theor Exp: P09008
28. de Nooy W, Mrvar A, Batagelj V (2005) Exploratory social network analysis with Pajek. Cambridge University Press, New York, USA
29. Dixon J (1981) Towards an understanding of the implications of boundary changes – with emphasis on community of interest, draft report to the rural adjustment unit. Technical report, University of New England, Armidale
30. Donetti L, Munoz AM (2004) Detecting network communities: a new systematic and efficient algorithm. J Stat Mech Theor Exp 2004(10):P10012
31. Driskell BR, Lyon L (2002) Are virtual communities true communities? Examining the environments and elements of community. City and Community 1(4):373–390
32. Du N, Wu B, Pei X, Wang B, Xu L (2007) Community detection in large-scale social networks. In WebKDD/SNA-KDD '07: Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 workshop on Web mining and social network analysis. ACM, New York, NY, USA, pp 16–25
33. Duch J, Arenas A (2005) Community detection in complex networks using extremal optimization. Phys Rev E (Stat Nonlinear Soft Matter Phys) 72(2):027104
34. Dwyer T, Hong HS, Koschutzki D, Schreiber F, Xu K (2006) Visual analysis of network centralities. In: APVis '06: Proceedings of the 2006 Asia-Pacific symposium on information visualisation. Australian Computer Society, Darlinghurst, Australia, pp 189–197
35. Elmore LK, Richman BM (March 2001) Euclidean distance as a similarity metric for principal component analysis. Month Weather Rev 129(3):540–549
36. Erickson T (1996) The world-wide-web as social hypertext. Commun ACM 39(1):15–17
37. Estrada E, Rodriguez-Velazquez AJ (2005) Subgraph centrality in complex networks. Phys Rev E 71:056103
38. Etzioni A, Etzioni O (2001) Can virtual communities be real? In: Etzioni A (ed) The Monochrome Society, Princeton University Press, Princeton, pp 77–101
39. Falkowski T, Bartelheimer J, Spiliopoulou M (2006) Community dynamics mining. In: Proceedings of 14th European conference on information systems (ECIS 2006). Gteborg, Sweden
40. Fisher D (2005) Using egocentric networks to understand communication. IEEE Internet Comput 9(5):20–28
41. Flake WG, Lawrence S, Giles LC, Coetzee MF (2002) Self-organization and identification of web communities. IEEE Computer 35(3):66–71
42. Fortunato S, Latora V, Marchiori M (2004) Method to find community structures based on information centrality. Phys Rev E (Stat Nonlinear, Soft Matter Phys) 70(5):056104
43. Frank AK (1997) Identifying cohesive subgroups. Social Networks 17(1):27–56
44. Freeman CL (1978) Centrality in social networks: Conceptual clarification. Social Networks 1:215–239
45. Frivolt G, Bielikov M (2005) An approach for community cutting. In: Svatek V, Snasel V (eds) RAWS 2005: Proceedings of the 1st International workshop on representation and analysis of Web space, Prague-Tocna, Czech Republic, pp 49–54
46. Garton L, Haythornthwaite C, Wellman B (1997) Studying online social networks. J Comput Mediated Commun 3(1):1–30
47. Girvan M, Newman EJM (2002) Community structure in social and biological networks. Proc Natl Acad Sci USA 99:7821

48. Gloor AP (2005) Capturing team dynamics through temporal social surfaces. In: Proceedings of the 9th international conference on information visualisation (InfoVis 2005). IEEE, pp 939–944
49. Gloor AP, Laubacher R, Dynes BCS, Zhao Y (2003) Visualization of communication patterns in collaborative innovation networks – analysis of some w3c working groups. In: CIKM '03: Proceedings of the 12th international conference on information and knowledge management, ACM Press, New York, NY, USA, pp 56–60
50. Gómez V, Kaltenbrunner A, López V (2008) Statistical analysis of the social network and discussion threads in slashdot. In: WWW '08: Proceedings of the 17th international conference on World Wide Web. ACM, New York, NY, USA, pp 645–654
51. Gregson AMR (1975) Psychometrics of similarity. Academic, NY, USA
52. Gruzd A, Haythornthwaite C (2007) A noun phrase analysis tool for mining online community. In: Proceedings of the 3rd international conference oncommunities and technologies, East Lansing, Michigan, USA, pp 67–86
53. Gruzd A, Haythornthwaite C (2008) Automated discovery and analysis of social networks from threaded discussions. Paper presented at the International Network of Social Network Analysis, St. Pete Beach, FL, USA
54. Hanneman AR, Riddle M (2005) Introduction to social network methods (online textbook). University of California, Riverside, CA
55. Hartigan J (1975) Clustering algorithms. Wiley, New York, NY, USA
56. Hirsch JB (1979) Psychological dimensions of social networks: A multimethod analysis. Am J Commun Psychol 7(3):263–277
57. Hoskinson A (2005) Creating the ultimate research assistant. Computer 38(11):97–99
58. Hubert JL, Schultz J (1976) Quadratic assignment as a general data analysis strategy. Brit J Math Stat Psychol 29:190–241
59. Jaccard P (1901) Distribution de la flore alpine dans le bassin des dranses et dans quelques rgions voisines. Bulletin del la Socit Vaudoise des Sciences Naturellese, 37:241–272
60. Johnson CS (1967) Hierarchical clustering schemes.Psychometrika, 32
61. Jones Q (1997) Virtual-communities, virtual settlements and cyber-archaeology: A theoretical outline. J Comput Supported Coop Work 3(3)
62. Jung Y, Park H, Du DZ, Drake LB (2003) A decision criterion for the optimal number of clusters in hierarchical clustering. J Global Optim 25(1):91–111
63. Keelan J, Pavri-Garcia V, Tomlinson G, Wilson K (2007) Youtube as a source of information on immunization: a content analysis. JAMA: J Am Med Assoc 298(21):2482–2484
64. Kleinberg J (2002) Bursty and hierarchical structure in streams. In: KDD '02: Proceedings of the 8th ACM SIGKDD international conference on knowledge discovery and data mining. ACM, New York, NY, USA, pp 91–101
65. Kleinberg MJ (1999) Authoritative sources in a hyperlinked environment. J ACM 46(5): 604–632
66. Koschtzki D, Schreiber F (2004) Comparison of centralities for biological networks.In: Giegerich R, Stoye J (eds) Proceedings of the German conference on bioinformatics (GCB'04), Bielefield, Germany, pp 199–206
67. Kumar R, Novak J, Raghavan P, Tomkins A (2003) On the bursty evolution of blogspace. In: WWW '03: Proceedings of the 12th international conference on World Wide Web. ACM, New York, NY, USA, pp 568–576
68. Kumar R, Novak J, Raghavan P, Tomkins A (2004) Structure and evolution of blogspace. Commun ACM 47(12):35–39
69. Kumar R, Raghavan P, Rajagopalan S, Tomkins A (1999) Trawling the web for emerging cyber-communities. Computer Networks 31(11–16), pp 1481–1493
70. Kurdia A, Daescu O, Ammann L, Kakhniashvili D, Goodman RS (November 2007) Centrality measures for the human red blood cell interactome. Engineering in Medicine and Biology Workshop. IEEE, Dallas, pp 98–101
71. Leskovec J, Lang JK, Dasgupta A, Mahoney WM (2008) Statistical properties of community structure in large social and information networks. In: WWW '08: Proceedings of the 17th international conference on World Wide Web. ACM, New York, NY, USA, pp 695–704

72. Leydesdorff L, Schank T, Scharnhorst A, de Nooy W (2008) Animating the development of social networks over time using a dynamic extension of multidimensional scaling
73. Li X, Liu B, Yu SP (2006) Mining community structure of named entities from web pages and blogs. In: AAAI Spring Symposium Series. American Association for Artificial Intelligence
74. Lin RY, Chi Y, Zhu S, Sundaram H, Tseng LB (2008) Facetnet: a framework for analyzing communities and their evolutions in dynamic networks. In: WWW '08: Proceedings of the 17th international conference on World Wide Web. ACM, New York, NY, USA, pp 685–694
75. Ma W-H, Zeng PA (2003) The connectivity structure, giant strong component and centrality of metabolic networks. Bioinformatics 19(11):1423–1430
76. Marlow C (2004) Audience, structure and authority in the weblog community. In: International communication association conference, New Orleans, LA
77. McMillan WD, Chavis DM (1986) Sense of community: a definition and theory. J Commun Psychol 14(1):6–23
78. Memon N, Harkiolakis N, Hicks LD (2008) Detecting high-value individuals in covert networks: 7/7 London bombing case study. In Proceedings of the 2008 IEEE/ACS International Conference on computer systems and applications. IEEE Computer Society, Washington DC, USA, 4–31 April 2008, pp 206–215
79. Memon N, Larsen LH, Hicks LD, Harkiolakis N (2008) Detecting hidden hierarchy in terrorist networks: Some case studies. Lect Notes Comput Sci 5075:477–489
80. Mizruchi SM, Mariolis P, Schwartz M, Mintz B (1986) Techniques for disaggregating centrality scores in social networks. Sociol Methodol 16:26–48
81. Moody J, McFarland AD, Bender-deMoll S (2005) Visualizing network dynamics. Am J Sociol: Jan 2005
82. Mukherjee M, Holder LB (2004) Graph-based data mining on social networks. In: Proceedings of the 10th ACM SIG conference on knowledge discovery and data mining, ACM, Seattle, USA, pp 1–10
83. Neustaedter C, Brush AJ, Smith AM, Fisher D (2005) The social network and relationship finder: Social sorting for email triage. In: Proceedings of the 2nd conference on E-mail and anti-spam (CEAS 2005), California, USA
84. Newman EJM (2006) Modularity and community structure in networks. Proc Nat Acad Sci 103(23):8577–8582
85. Newman EJM, Girvan M (2004) Finding and evaluating community structure in networks. Phys Rev E 69:026113
86. O'Reilly T (2005) What is web 2.0? http://www.oreillynet.com/pub/a/oreilly/tim/news/2005/09/30/what-is-web-20. Accessed 30 September 2008
87. Orford DJ (1976) Implementation of criteria for partitioning a dendrogram. Math Geol 8(1):75–84
88. Paolillo CJ, Wright E (2004) The challenges of foaf characterization. http://stderr.org/~elw/foaf/. Accessed 30 September 2008
89. Paolillo CJ, Wright E (2005) Social network analysis on the semantic web: Techniques and challenges for visualizing foaf. http://www.blogninja.com/vsw-draft-paolillo-wright-foaf.pdf. Accessed 30 September 2008
90. Piper EW, Marrache M, Lacroix R, Richardsen MA, Jones BD (1983) Cohesion as a basic bond in groups. Hum Relat 36(2):93–108
91. Radicchi F, Castellano C, Cecconi F, Loreto V, Parisi D (2004) Defining and identifying communities in networks. Proc Natl Acad Sci USA 101(9):2658–2663
92. Reffay C, Chanier T (2003) How social network analysis can help to measure cohesion in collaborative distance learning. In: Proceedings of computer supported collaborative learning 2003. Kluwer, ACM, Dordrecht, NL, pp 343–352
93. Rheingold H (1993) The virtual community: homesteading on the electronic frontier. Addison-Wesley, Toronto, ON, Canada
94. Ruan J, Zhang W 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), Omaha, Nebraska, USA, 28–31 October 2007, pp 643–648

95. Ruhnau B (October 2000) Eigenvector-centrality – a node-centrality? Social Networks 22(4):357–365
96. Sarason GI, Levine HM, Basham BR, Sarason RB (1983) Assessing social support: the social support questionnaire. J Pers Social Psychol 44:127–139
97. Schaeffer ES (2007) Graph clustering. Comput Sci Rev 1(1):27–64
98. Shi J, Malik J (2000) Normalized cuts and image segmentation. IEEE Trans Pattern Anal Mach Intell 22(8):888–905
99. Snijders ABT, Nowicki K (1997) Estimation and prediction for stochastic block models for graphs with latent block structure. J Classif 14:75–100
100. Snijders AB Tom, Christian EG Steglich, Schweinberger M (2007) Modeling the co-evolution of networks and behavior. In: Kees van Montfort, Han Oud, Albert Satorra (eds) Longitudinal models in the behavioral and related sciences, Routledge Academic, England, pp 41–71
101. Steinhaeuser K, Chawla VN (2008) Is modularity the answer to evaluating community structure in networks. In: International workshop and conference on network science (NetSci'08), Norwich Research Park, UK
102. Sterling S (2004) Aggregation techniques to characterize social networks. Master's thesis, Air Force Institute of Technology. Ohio, USA
103. Tajfel H, Turner CJ (1986) The social identity theory of inter-group behavior. In: Worchel S, Austin LW (eds) Psychology of intergroup relations. Nelson-Hall, Chicago, USA
104. Tantipathananandh C, Berger-Wolf YT, Kempe D (2007) A framework for community identification in dynamic social networks. In: KDD '07: Proceedings of the 13th ACM SIGKDD international conference on knowledge discovery and data mining. ACM, New York, NY, USA, pp 717–726
105. Traud LA, Kelsic DE, Mucha JP, Porter AM (2009) Community structure in online collegiate social networks, American Physical Society, 2009 APS March Meeting, March 16–20, pp 1–38
106. Tremayne M, Zheng N, Lee KJ, Jeong J (2006) Issue publics on the web: Applying network theory to the war blogosphere. J Comput Mediated Commun 12(1), article 15. http://jcmc.indiana.edu/vol12/issue1/tremayne.html
107. Tversky A (1977) Features of similarity. Psychol Rev 84(4):327–352
108. Tyler RJ, Wilkinson MD, Huberman AB (2005) E-mail as spectroscopy: Automated discovery of community structure within organizations. Inform Soc 21(2):143–153
109. Uttal RW, Spillmann L, Sturzel F, Sekuler BA (2000) Motion and shape in common fate. Vision Res 40(3):301–310
110. van Duijn1 AJM, Vermunt KJ (2005) What is special about social network analysis? Methodology 2:2–6
111. Wang G, Shen Y, Ouyang M (2008) A vector partitioning approach to detecting community structure in complex networks. Comput Math Appl 55(12):2746–2752
112. Wang H, Wang W, Yang J, Yu SP (2002) Clustering by pattern similarity in large data sets. In: SIGMOD '02: Proceedings of the 2002 ACM SIGMOD international conference on management of data. ACM, New York, NY, USA, pp 394–405
113. Wasserman S, Faust K (1994) Social network analysis: methods and applications. Cambridge University Press, United Kingdom
114. Wellman B (2003) Structural analysis: from method and metaphor to theory and substance. In: Wellman B, Berkowitz SD (eds) Social structures: a network approach, Cambridge University Press, UK, pp 19–61
115. Wellman B, Guilia M (1999) Net surfers don't ride alone: virtual communities as communities. In: Wellman B (ed) Networks in the global village: life in contemporary communities, Westview Press, Colorado, US
116. Welser TH, Gleave E, Fisher D, Smith M (2007) Visualizing the signatures of social roles in online discussion groups. J Soc Struct 8, http://www.cmu.edu/joss/content/articles/volume8/Welser

117. Zahn TC (1971) Graph-theoretical methods for detecting and describing gestalt clusters. IEEE Trans Comput C-20(1):68–86
118. Zhao Y, Karypis G (2002) Evaluation of hierarchical clustering algorithms for document datasets. In: CIKM '02: Proceedings of the 11th international conference on information and knowledge management. ACM, New York, NY, USA, pp 515–524