Chapter 2
Performance Evaluation of Social Network Using Data Mining Techniques

Mrutyunjaya Panda, Ajith Abraham, Sachidananda Dehuri, and Manas Ranjan Patra

Abstract Social network research relies on a variety of data sources, depending on the problem scenario and the questions, which the research is trying to answer or inform. Social networks are very popular nowadays and the understanding of their inner structure seems to be promising area. Cluster analysis has also been an increasingly interesting topic in the area of computational intelligence and found suitable in social network analysis in its social network structure. In this chapter, we use k-cluster analysis with various performance measures to analyse some of the data sources obtained for social network analysis. Our proposed approach is intended to address the users of social network, that will not only help an organization to understand their external and internal associations but also highly necessary for the enhancement of collaboration, innovation and dissemination of knowledge.

M. Panda (✉)
Department of ECE, Gandhi Institute for Technological Advancement (GITA), Bhubaneswar, Odisha, India
e-mail: mrutyunjaya74@gmail.com

A. Abraham
Machine Intelligence Research Labs (MIR labs), Scientific Network for Innovation and Research Excellence, Auburn, WA, USA
e-mail: ajith.abraham@ieee.org

S. Dehuri
Department of Computer and Information Technology, F.M. University, Balasore, India
e-mail: sachilapa@gmail.com

M.R. Patra
Department of Computer Science, Berhampur University, Ganjam, India
e-mail: mrpatra12@gmail.com

A. Abraham (ed.), *Computational Social Networks: Mining and Visualization*,
DOI 10.1007/978-1-4471-4054-2_2,

Introduction

A social network is an umbrella with nodes of individuals, groups, organizations and related systems that tie in one or more types of interdependencies. Social network data analysis is intended to understand the social structure, which subsists amongst entities in an organization [1–3]. Social network analysis is focused on uncovering the patterning of people's interaction. Network analysis is based on the intuition that these patterns are important features of the lives of the individuals who display them. Further, the social network approach has primarily involved in two important aspects such as: (1) it is guided by formal theory organized in mathematical terms, and; (2) it is grounded in the systematic analysis of empirical data.

Fining et al. [4] used a similar wave that is imminent under the generic banner of data mining tools that may stem from reality mining [5] and its link with social networking relationships.

With the proliferation of social media and online communities in networked world, large gamut of data have been collected and stored in databases. The rate at which such data is stored is growing at a phenomenal rate. As a result the classical method of data analysis is of rare interest. This chapter presents a study about the effectiveness of the data mining methods in analysing the social network data collected from various social network sites and UCI machine learning database. Our integrated framework leads to a term swam based data mining to address the users of social network, that will not only help an organization to understand their external and internal associations but also highly necessary for the enhancement of collaboration, innovation and dissemination of knowledge.

The rest of the chapter is organized as follows. A brief overview on the related research in this field is outlined in section "Related Research", followed by overview on social network analysis in section "Preliminaries". Section "Social Network Data" introduces the dataset obtained for our analysis, with the proposed methodology in section "Methodology and Experimental Setup". Section "Results and Discussion" discusses some results with discussion. Finally, we conclude the chapter in section "Conclusion" with a light on future direction of research.

Related Research

Boyd and Ellison [6] used social network sites (SNSs) that allow users to register, create their own profile page containing information about themselves which may either be real or virtual, in order to establish the public connections with the other members and to communicate with them. SNS like Facebook and MySpace are amongst the ten most popular websites in the world including Orkut, Cyworld and Mixi. Tufekci [7] emphasized the motivation behind the use of SNS as its sociability with a suggestion to consider that some types of people may never use Social network sites extensively [8]. Data mining emotion in social network

communication with MySpace is proposed in [9], where the authors confirm that MySpace is an emotion-rich environment and therefore suitable for the development of specialist sentiment analysis techniques. Apart from this, it is observed that using both age and gender in the emotion strategies for classification points to the difficulty in making accurate classifications and poses several challenges to the automatic classification with the existing methodologies. Zhou et al. [10] build up an social network mining solution to discover the social network, users' relationship, key figures and impaction to the organization on BBS website in order to understand the internal and external association of an organization so as to enhance the collaboration and disseminating knowledge. Kaufman et al. [11] introduce a new social network dataset site Facebook.com with findings to exemplify the scientific and pedagogical potential of this new network resource with a future prospect in this area of research. Since social network research embodies a range of expertise from anthropology to Computer Sciences, it is quite difficult to find the benchmarks for social networks [12]. Further, social networks have been measured on various dataset including online social networks [13] to sexual transmission networks [14]. The authors propose a method by using formal concept analysis in understanding the social networks with ease in analysis and visualizing the networks and propose to use bigger networks in future [15]. Social data mining is used to improve bioinspired intelligent systems with swarm optimization, ant colony and cultural algorithms are discussed in [16]. Research on social network analysis in the data mining community includes following areas: clustering analysis in [17] and [18], classification [19], link prediction [20] and [21]. Other achievements include PageRank [22, 23] and Hub-Authority [24] in web search engine. An experimental study on important set of "small world" problem is discussed by Milgram in [25] and [26], which gives an insight about the network structure rather than reconstructing the actual networks. In this, the authors tries to probe the distribution of path lengths in an acquaintance network by passing a letter to one of their first name acquaintance to an assigned target individual in an attempt. while doing this, many letters got missed while only six people could successfully targeted and passed on average through their hands, which subsequently led to the of the "six degrees of separation", coined by Guare [27].A review of the issues regarding controlling the possible sources of inconsistency in social network data gathered directly by using questionnaires and interviews, has been discussed in [28], citing the reason for which the researchers tried to adopt other possible methods for probing social network. One source of copious and relatively reliable data is collaboration networks. These are typically affiliation networks in which participants collaborate in groups of one kind or another, and links between pairs of individuals are established by common group membership. Examples of such a network include: collaboration network of film actors, which has been well documented online in Internet Movie Database[29], where actors collaborate in films and two actors are considered connected if they have worked in a film together, the statistical properties of this type of networks can be understood from the research done in [30, 31]; networks of co-authorships amongst academicians, where individuals are linked if they have coauthored in one or more papers as explained in [32, 33] and networks of board of directors

in which two directors are linked if they belong to the same board of directors at least in any one company, as discussed in [34, 35] to name a few. Aiello et al. [36, 37] have analysed a network of telephone calls made over the AT&T long-distance network on a single day. The vertices of this network represent telephone numbers and the directed edges calls from one number to another. Ebel et al. [38] have reconstructed the pattern of email communications between 5,000 students at Kiel University from logs maintained by email servers. In this network the vertices represent email addresses and directed edges represent a message passing from one address to another. Abraham et al. [39] addressed the computational complexity of social networks analysis and clarity of their visualization that uses combination of Formal Concept Analysis and well-known matrix factorization methods. The goal is to reduce the dimension of social network data and to measure the amount of information which is lost during the reduction. Singular value decomposition has already been used in the field of social network data [40] to determine the position of nodes in the network graph. The research made in [41] by Bulkley et al. examines hypotheses about the efficient and strategic uses of social networks by a specific group of white collar workers, in which they examined two existing theories relating network structure and tie strength to performance and put forward a new hypothesis. They used a unique data set containing email patterns and accounting records for several dozen executive recruiters and found statistically significant differences related to network (1) structure (2) flow and (3) age. Abraham et al. [42] try to consider a Web page as information with social aspects. Each Web page is the result of invisible social interaction. For the description of the social aspects of Web pages, they used the term MicroGenre with fundamental concepts of MicroGenre with an illustration to the experiments for the detection and usage of MicroGenres. Social Network Analysis (SNA) that is based on the data are collected from the students at the Faculty of Organization and Informatics, University of Zagreb is presented in [43], where they conclude firstly, that the position in a social network cannot be forecast only by academic success and, secondly, that part-time students tend to form separate groups that are poorly connected with full-time students.

Preliminaries

Social Network Analysis (SNA)

Social network analysis is used to understand the social structure, which exists amongst entities in an organization. The defining feature of social network analysis (SNA) is its focus on the structure of relationships, ranging from causal acquaintance to close bonds. This is in contrast with other areas of the social sciences where the focus is often on the attributes of agents rather than on the relations between them. SNA maps and measures the formal and informal relationships to understand what facilitates or impedes the knowledge flows that bind the interacting units i.e. who knows whom and who shares what information and how. Social

network analysis is focused on uncovering the patterning of people's interaction. SNA is based on the intuition that these patterns are important features of the lives of the individuals who display them. The network analysts believe that how an individual lives depends in large part on how that individual is tied into larger web of social connections. Moreover, many believe that the success or failure of societies and organizations often depends on the patterning of their internal structure, which is guided by formal concept analysis, which is grounded in systematic analysis of the empirical data. With the availability of powerful computers and discrete combinatorics (especially graph theory) after 1970, the study of SNA take off as an interdisciplinary speciality, the applications are found many folds that include: organizational behaviour, inter-organizational relations, the spread of contagious diseases, mental health, social support, and the diffusion of information and animal social organization [17].

SNA Basics

The two basic elements of SNA are links and nodes. Links are connections, or ties, between individuals or groups and nodes are the individuals or groups involved in the network. A nodes importance in a social network refers to its centrality. Central nodes have the potential to exert influence over less central nodes. A network that possesses just a few or perhaps even one node with high centrality is a centralized network. In this type of network all nodes are directly connected to each other. Subordinate nodes direct information to the central node and the central node distributes it to all other nodes. Centralized networks are susceptible to disruption because they have few central nodes and damage to a central node could be devastating to the entire network. A simple social network is shown in Fig. 2.1.

Decentralized networks are those that do not possess one central hub; but rather possess several important hubs. Each node is indirectly tied to all others and therefore the network has more elasticity. Consequently, these networks are more difficult to disrupt due to their loose connections and ability to replace damaged nodes. Consequently, terror networks choose this type of structure whenever possible.

Social network analysts use the term degrees in reference to the number of direct connections that a node enjoys. The node that possesses the largest number of connections is the hub of the network. The term betweenness refers to the number of groups that a node is indirectly tied to through the direct links that it possesses. Therefore, nodes with high a degree of betweenness act as liaisons or bridges to other nodes in the structure. These nodes are known as "brokers" because of the power that they wield. However, these "brokers" represent a single point of failure because if their communication flows is disrupted than they will be cut off to the nodes that it connects. Closeness measures the trail that a node would take in order to reach all other nodes in a network. A node with high closeness does not necessarily have the most direct connections; but because they are "close" to many members they maintain rapid access to most other nodes through both direct and indirect ties.

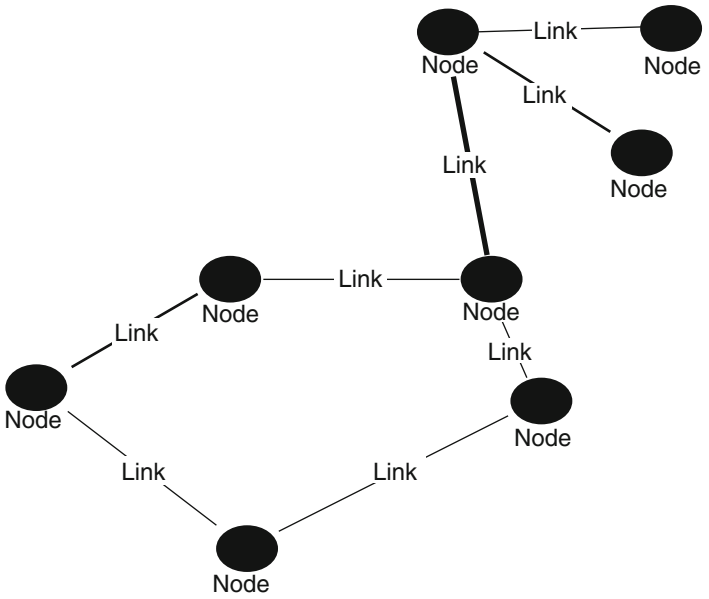

Fig. 2.1 Social network diagram

Strengths

- Provides visual representations. Social network analysis allows the researcher to visualize the structure of an organization through the use of link charts. It, also, allows researchers to identify previously unknown links. The knowledge gained through this type of analysis permits analysts to forecast activities of an actor/ organization and outline a possible attack strategy on criminal or terrorist organizations by focusing resources on important actors in the hopes of weakening the entire network.
- Provides data that is useful for analysis. SNA software, also, provides the researcher with data that can be analysed to determine the centrality, betweenness, degree, and closeness of each node.

Weaknesses

- Analysis is dependent upon data. Like any other software, SNA software will only analyse the data it is given. Therefore, if the data is incomplete or incorrect than the final product will be inaccurate. Furthermore, it will only provide a complete overview of a network unless ALL factors have been researched and entered, which is almost impossible. Consequently, it should be used as a tool only and not relied upon to provide an absolute depiction of a network.

- Time consuming. It takes a great deal of time to research a topic in order to find the appropriate information. A first time user of social network analysis will not only have to research the target vigorously; but, also become familiar with SNA which is a daunting task.
- Various resources are needed. A computer with an Internet connection is essential in conducting social network analysis. Specialized software is, also, needed to perform SNA.

Applications

With regard to purposes of knowledge management, social network analysis may help to evaluate availability and distribution of critical knowledge and thus facilitates:

- Strategic development of organizational knowledge,
- Transfer and sustainable conservation of implicit knowledge,
- Development of core competencies (like leadership development),
- Creation of opportunities to improve communication processes,
- Identification and support of communities of practice,
- Harmonization of knowledge networks (after mergers and acquisitions),
- Sustainable management of external relationships.

Step-by-Step Procedure

We provided a basic step-by-step procedure on how to design a social network model as below:

Step 1: The researcher must first identify a suitable target.

Step 2: Thoroughly research the subject matter. The researcher must explore who knows whom and how well do they know each other? They must, also, discover how does the information or resources flow within a network? Researchers should also analyse how members of a network know each other? Although these are not the only factors to consider they provide a good starting point. The actual factors will vary from project to project.

Step 3: The information gathered must be placed into a database. The most commonly used matrix is a node-by-node matrix. This type of matrix uses as many rows and columns as there are nodes in a network. The simplest method of data entry is done by placing O's or 1's into a spreadsheet. The O's represent no connection and 1's signify a link. A sample is shown in Fig. 2.2 below.

This is a binary matrix. A signed matrix assigns a nodes relationship with another with $+1$ (positive relationship), 0 (no relationship) and -1 (negative relationship). However, an ordinal matrix allows the researcher to assign a strength rating to each connection. In other words if a strong connections exists than instead of placing a

Fig. 2.2 Relation matrix

	A	B	C
A	0	1	1
B	0	0	0
C	1	0	0

1 in the spreadsheet, the researcher would place a number that signifies strength such as a 5 or a 10. Once the spreadsheet is complete the SNA software is able to generate, using various mathematical equations, interpretable data

Step 4: A researcher must investigate the basics of SNA before he/her can begin to analyse the compiled data. There are numerous articles, books, and websites available to the novice social network analyst that will provide everything from basic knowledge to expert advice. When the researcher purchases the appropriate SNA software many will come with tutorials that will explain basic social network theories. The tutorial should, also, provide the researcher with directions as to how to use the software itself; however, an equally effective way to learn this type of software is trial and error. In other words the researcher can simply "plays around" with the software functions and use the tutorial to interpret the results.

Step 5: The last step is to perform the actual analysis. The basic knowledge of SNA and of the specialized software that the researcher gained should make the actual analysis much less difficult. The analysed data may then be used to identify possible tactics to disrupt or improve a networks communications or resources.

Data Mining

There has been extensive research work on clustering in data mining. Traditional clustering algorithms [44] divide objects into classes based on their similarity. Objects in a class are similar to each other and are very dissimilar from objects in different classes. Social network clustering analysis, which is different from traditional clustering problem, divides objects into classes based on their links as well as their attributes. The biggest challenge of social network clustering analysis is how to divide objects into classes based on objects' links, thus we need find algorithms that can meet this challenge. A k-clustering scheme is used with $k = 5$ with Tabu search to build the SNA. Tabu search is a numerical method for finding the best division of actors into a given number of partitions on the basis of approximate automorphic equivalence. In using this method, it is important to explore a range of possible numbers of partitions unless one has a prior theory about this, to determine how many partitions are useful. Having selected a number of partitions, it is useful to re-run the algorithm a number of times to insure that a global, rather than

local minimum has been found. The method begins by randomly allocating nodes to partitions. A measure of badness of fit is constructed by calculating the sums of squares for each row and each column within each block, and calculating the variance of these sums of squares. These variances are then summed across the blocks to construct a measure of badness of fit. Search continues to find an allocation of actors to partitions that minimizes this badness of fit statistic. The Tabu search algorithm is provided below:

The Tabu Search Algorithm has traditionally been used on combinatorial optimization problems related to feature selection and has been frequently applied to many integer programming, routing and scheduling, traveling salesman and related problems. The basic concept of Tabu Search is presented by Glover [45] who described it as a meta-heuristic superimposed on another heuristic. It explores the solution space by moving from a solution to the solution with the best objective function value in its neighbourhood at each iteration even in the case that this might cause the deterioration of the objective. (In this sense, "moves" are defined as the sequences that lead from one trial solution to another.) To avoid cycling, solutions that were recently examined are declared forbidden or "Tabu" for a certain number of iterations and associated attributes with the Tabu solutions are also stored. The Tabu status of a solution might be overridden if it corresponds to a new best solution, which is called "aspiration". The Tabu lists are historical in nature and form the Tabu search memory. The role of the memory can change as the algorithm proceeds. For initializations at each iteration, the objective is to make a coarse examination of the solution space, known as "diversification", but as locations of the candidate solutions are identified, the search is more focused to produce local optimal solutions in a process of "intensification". Intensification and diversification are fundamental cornerstones of longer term memory in Tabu search. In many cases, various implementation models of the Tabu Search method can be achieved by changing the size, variability, and adaptability of the Tabu memory to a particular problem domain. In all, Tabu Search Algorithm is an intelligent search technique that hierarchically explores one or more local search procedures in order to search quickly for the global optimum. As one of the advanced heuristic methods, Tabu Search is generally regarded as a method that can provide a near-optimal or at least local optimal solution within a reasonable time domain.

What is being minimized is a function of the dissimilarity of the variance of scores within partitions. That is, the algorithm seeks to group together actors who have similar amounts of variability in their row and column scores within blocks. Actors who have similar variability probably have similar profiles of ties sent and received within, and across blocks – though they do not necessarily have the same ties to the same other actors. Unlike the other methods mentioned here, the Tabu search produces a partitioned matrix, rather than a matrix of dissimilarities. It also provides an overall badness of fit statistic. Both of these would seem to recommend the approach, perhaps combined with other methods.

Social Network Data

Dataset Used

We use Terrorist data available in UCINET [46] and Netlog Data [47] for building our social network analysis.

Terrorist Dataset

In this, we use terrorist dataset with their name and their relationships to each other. In this, 64 terrorist names from different terrorist organization are taken into consideration for the study of their social network with 0 for no relationship and 1 for having some kind of relationship. Their names include Hani Hanjour, Abu Walid, Madjid Sahoune, and Faisal Al Salmi etc. A sample of the terrorist network dataset is shown in Table 2.1 in terms of relation matrix.

Netlog Dataset

Netlog is an online platform where users can keep in touch with and extend their social network. It is an online social portal, specifically targeted at the European Youth. It is developed by Netlog NV, based in Ghent, Belgium. Netlog is currently available in 37 languages and has more than 72 million members throughout the Europe, and this number is increasing day by day.

On Netlog, One can create its own webpage with a blog, picture, videos, events and much more to share with their own needs. It is thus the ultimate tool for the young people to connect and communicate with their social network. Netlog NV has developed a unique localization technology ensuring that all content is geotargeted and personalized to each member's profile.

Table 2.1 Relation matrix for a sample of Terrorist Network dataset in 2-mode

	HH	MM	NA	SA*	KAM	MA	WA	WA	SS
Hani Hanjour	0	1	1	1	1	1	0	0	0
Majed Moqed	1	0	1	1	1	0	0	0	0
Nawaf Alhazmi	1	1	0	1	1	1	0	0	0
Salem Alhazmi*	1	1	1	0	1	0	0	0	0
Khalid Al-Mihdhar	1	1	1	1	0	0	0	0	0
Mohamed Atta	1	0	1	0	0	0	0	1	1
Waleed Alshehri	0	0	0	0	0	0	0	1	1
Wail Alshehri	0	0	0	0	0	1	1	0	1
Satam Suqami	0	0	0	0	0	1	1	1	0

Netlog collect the following types of personal data to publish the information intended to be made public by the people, under the conditions specified in their privacy settings.

- *Public information uploaded by the person*
  - Information in individual profile, blog, shouts, pictures, videos, events, music, links
  - Messages sent to other users, as well as ratings, shouts and contributions to another user's guest book
  - Links to your friends and groups
- *Private information uploaded by the person*
  - Settings and administrative data such as user name and password, skin, credits and shortcuts
- *History and Logs*
  - Time, date and URL of all Netlog pages visited by you
  - The URL of the referring websites
  - The searches one performs on the website.

A sample data with eight users having 16 features to analyse the social network amongst the users is shown in Table 2.2.

Methodology and Experimental Setup

We conducted our research on social network analysis to reveal the relationships amongst the terrorist using the terrorist network dataset. We also use Netlog dataset to understand the social networking of a common man. All experiments are conducted in a Pentium-4 Machine with 2.86 GHz CPU, 40 GB HDD, 512 MB RAM. We use UCINET 6.0 tool [46] for our performance evaluations in social network analysis. The Netdraw software produces a visualization of the 2-mode data with all the actors and event nodes. The actor nodes are circles and the event nodes are the squares. The Netdraw visualization works directly from the 2-mode data set in UCINET tool used. More details about the network analysis in 2-mode network can be obtained from Borgatti and Everett [48].

It is evident from Table 2.1 with the concept of relation metric as discussed earlier in section "Step-by-Step Procedure" that Hari Hanjour has some relationship with Majed Moged Nawaf Alhazmi, Salem Alhazmi*, Khalid Al-Mihdhar and Mohamed Atta, where as poses no relationship with Waleed Alshehri, Wail Alshehri and Satam Suqami and so on for all others.

From Table 2.2, we can understand the data as: the person with user_id 1 with a current time value is online now, is a female, aged about 44 years citizen of "Tn" staying in city of Greeneville. He is a member of NETLOG since November 1, 2007

Table 2.2 A sample of Netlog dataset in 1-mode network

UI	CT	ON	MS	F	G	O_S	A	City	C	Ph	B	GB	O	SO	SOF
1	########	TRUE	664 visitors since 1 November 2007	23	Female	Online	44 years	Greeneville	Tn	18	1	17	1	0	42,617
2	########	TRUE	232 visitors since 16 May 2007	8	Male	Online	67 years	Stockton	United States	6	1	1	1	0	38,510
6	########	TRUE	663 visitors since 6 November 2006	23	Male	Away	33 years	Catterick	United Kingdom	5	1	36	1	0	42,024

Where, *UI* user_id, *CT* current time, *ON* online_now, *MS* member since, *F* friends, *G* gender, *A* age, *C* country, *Ph* photos, *O_S* online_now_state, *O* objects, *SO* swfobject, *SOF* size_of_profile, *GB* guestbook, *B* blog

with 664 numbers of visitors visited his profile with 23 friends, having 18 photos, 1 blog, 17 guest books with 42,617 as size of her profile.

Metrics in Social Network Analysis

The following measures (Metrics) are used in social network analysis.

- *Centrality:* This measure gives a rough indication of the social power of a node based on how well they "connect" to the network. "Betweenness", "Closeness", and "Degree" are considered to fall under the measures of centrality.
- *Betweenness:* It is defined as the extent to which a node lies between other nodes in the network. Here, the connectivity of the node's neighbours is taken into account in order to provide a higher value for nodes which bridge clusters. This metrics reflects the number of people who are connecting indirectly through direct links.
- *Closeness:* This refers to the degree with which an individual is nearer to all others in a network either directly or indirectly. Further, it reflects the ability to access information through the "grapevine" of network members. In this way, the closeness is considered to be the inverse of the sum of the shortest distance (sometimes called as geodesic distance) between each individual and all other available in the network.
- *Degree:* It is the count of the number of ties to other actors in the network.
- *Clustering coefficient:* This provides the likelihood that two associates of a node are associates with themselves. A higher clustering coefficient indicates a greater "cliquishness".
- *Centralization:* It is calculated as the ratio between the numbers of links for each node divided by maximum possible sum of differences. While a centralized network will have many of its links dispersed around one or a few nodes, the decentralized network is one in which there is little variation between the number of links each node possesses.
- *Density:* It is the degree that measures the respondent's ties to know one another. The density may be sparse or dense network depends upon the proportion of ties in a network relative to the total number of possibilities.

Results and Discussion

In order to understand Closeness one must understand geodesic distance; which is the number of relations in the shortest possible "walk" from one actor to another. Therefore, geodesic distance is the most commonly used measure of Closeness. For instance whose inCloseness score of 100 has the lowest total of geodesic distances

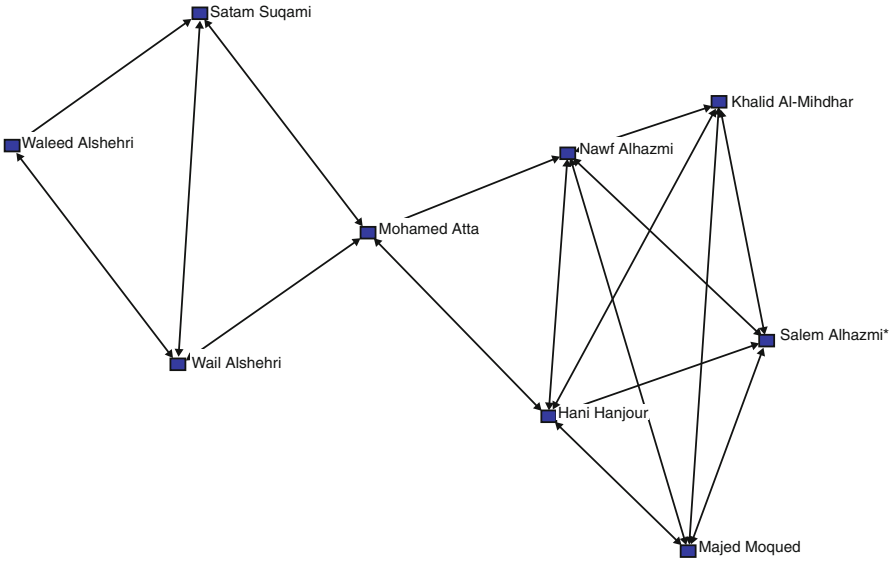

Fig. 2.3 Social network for Terrorist Network dataset (1-mode network)

from other actors; Nearness can be re-expressed as farness. In other words, actors because of their inFarness score of 210 from Netlog Data have the largest total of geodesic distances from other actors. High closeness centrality indicates the greater autonomy of an individual person, since he or she is able to reach the other members easily (and vice versa). Low closeness centrality indicates higher individual member dependency on the other members, i.e. the willingness of other members to give access to the network's resources.

In the same way, betweenness refers to the number of groups that a node has indirect ties to through the direct links that it possesses. In other words, it represents the number of times that a node lies along the shortest path between two others. UCINET will calculate the betweenness for each node in a dataset automatically using the formula above. Interpreting the results is relatively easy; the larger the number the higher the betweenness the node possesses. UCINET will automatically place them in order of highest to lowest. All these results are presented in sections "1-Mode and 2-Mode Network" and "SNA with Netlog Data" with Netlog Data and Terrorist Network data respectively.

Further, the social network structure for the proposed analysis using Terrorist Network dataset are shown in Figs. 2.3 and 2.4 with 1-mode and 2-mode respectively. The brief description about the 1-mode and 2-mode network is provided below. The similar type of social network structure using Netlog dataset is shown in Fig. 2.5.

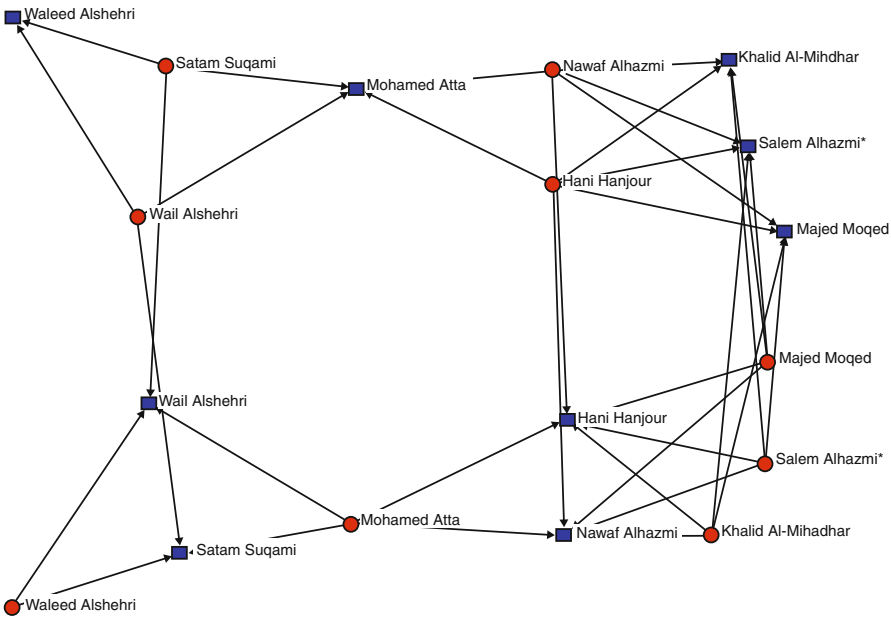

Fig. 2.4 Same for 2-mode network for Terrorist Network dataset

1-Mode and 2-Mode Network

A (2-dimensional) matrix is said to be 2-mode if the rows and columns index different sets of entities (e.g., the rows might correspond to persons while the columns correspond to organizations). In contrast, a matrix is 1-mode if the rows and columns refer to the same set of entities, such as a city-by-city matrix if distances. In social network analysis, 2-mode data refers to data recording ties between two sets of entities. In this context, the term "mode" refers to a class of entities – typically called actors, nodes or vertices – whose members have social ties with other members (in the 1-mode case) or with members of another class (in the 2-mode case). Most social network analysis is concerned with the 1-mode case, as in the analysis of friendship ties among a set of school children or advice-giving relations within an organization. The 2-mode case arises when researchers collect relations between classes of actors, such as persons and organizations, or persons and events. For example, a researcher might collect data on which students in a university belong to which campus organizations, or which employees in an organization participate in which electronic discussion forums. These kinds of data are often referred to as affiliations. Co-memberships in organizations or participation in events are typically thought of as providing opportunities for social relationships among individuals (and also as the consequences of pre-existing relationships). At the same time, ties between organizations through their members are thought to be conduits through which organizations influence each other.

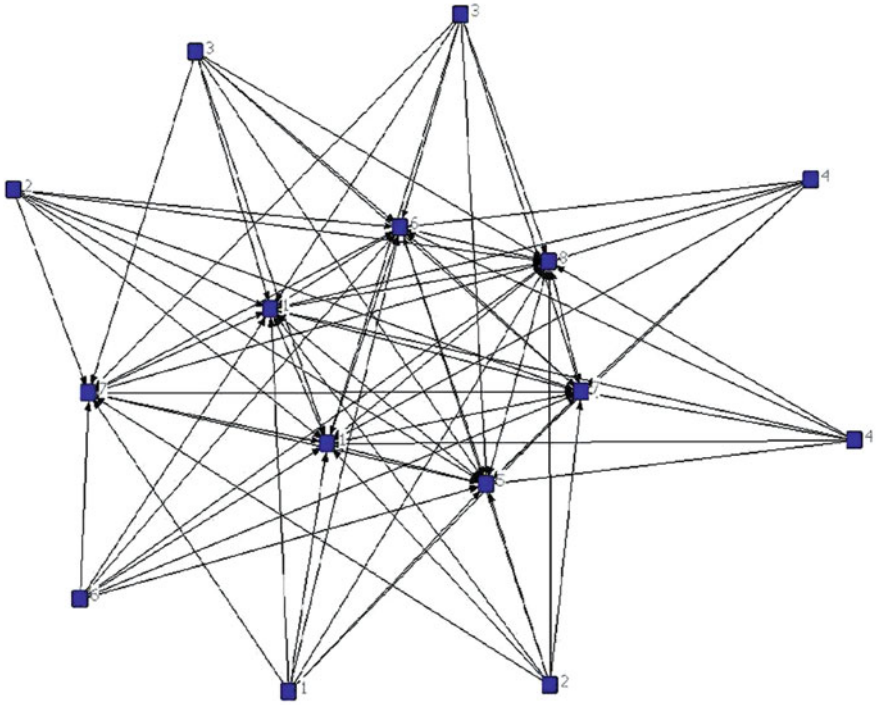

Fig. 2.5 Social network structure using Netlog1 data_16inst_8users

Table 2.3 Closeness centrality

		1	2	3	4
		inFarness	outFarness	inCloseness	outCloseness
1	1	14.000	126.000	100.000	11.111
12	5	14.000	126.000	100.000	11.111
10	7	14.000	127.000	100.000	11.024
4	1	14.000	126.000	100.000	11.111
7	1	210.000	112.000	6.667	12.500
9	4	210.000	113.000	6.667	12.389
3	2	210.000	112.000	6.667	12.500

SNA with Netlog Data

Closeness Centrality with Netlog Data

Closeness Centrality Measures can be understood from Table 2.3:

Closeness centrality approaches emphasize the distance of an actor to all others in the network by focusing on the geodesic distance from each actor to all others.

Table 2.4 Statistics of the closeness centrality measures

		1	2	3	4
		inFarness	outFarness	inCloseness	outCloseness
1	Mean	118.733	118.733	49.046	11.831
2	Standard deviation	97.571	6.942	45.501	0.687
3	Variance	9520.063	48.196	2070.352	0.471

One should use either directed or undirected geodesic distances among actors. The sum of these geodesic distances for each actor is the "farness" of the actor from all others. We can convert this into a measure of nearness or closeness centrality by taking the reciprocal of the farness and norming it relative to the most central actor.

From the above sample of the whole result obtained, Let us examine the in-farness, in-closeness, out-farness and out-closeness of the points as a measure of who is "central" or "influential" in this network. In undirected data, actors differ from one another only in how many connections they have. With directed data, however, it can be important to distinguish centrality based on in-degree from centrality based on out-degree. If an actor receives many ties or relationships with others in the network, they are often said to be prominent, or to have high prestige. That is, many other actors seek to have direct ties with them, which may indicate their importance. Actors who have unusually high out-degree are actors who are able to exchange with many others, or make many others aware of their views. Actors who display high out-degree centrality are often said to be influential actors.

From Table 2.3, we can observe that actors 7 & 1, 9 & 4 and 3 & 2 donot have much prominent relationship between each other with low inCloseness value of 6.667 and high inFarness value of 210.000. At the same time, higher outCloseness and lower outFarness between actors 7 & 1 and 3 & 2 are more influential among all relationships present. Relationship between 1 & 1, 12 & 5, 10 & 4 and 4 & 1 are closest or most central with inCloseness of 100.

From Table 2.4, the mean for inCloseness is 49.046 which says that the mean strength of ties across all possible ties (ignoring self-ties). Since the data are binary, this means that 49% of all possible ties are present (i.e. the density of the matrix). The standard deviation is a measure of how much variation there is among the elements. If all elements were one, or all were zero, the standard deviation would be zero, i.e. no variation. Here, the average variability from one element to the next is 45.501, almost same as the mean. So, we would say that there is, relatively, a great deal of variation in ties. With binary data, the maximum variability in ties – or the maximum uncertainty about whether any given tie is likely to be present or absent is realized at a density of .50. As density approaches either zero or unity, the standard deviation and variance in ties will decline accordingly.

Freeman Betweenness Centrality with Netlog Data

Supposing that somebody wants to influence you by sending you information, or make a deal to exchange some resources. But, in order to talk to you, he must go

Table 2.5 Betweenness centrality measures

		1	2
		Betweenness	nBetweenness
1	1	0.600	0.330
12	5	0.600	0.330
13	8	0.600	0.330
4	1	0.600	0.330
15	6	0.600	0.330
3	2	0.000	0.000

Table 2.6 Descriptive statistics for each measure of betweenness centrality

		1	2
		Betweenness	nBetweenness
1	Mean	0.200	0.110
2	Standard Deviation	0.283	0.155
3	Variance	0.080	0.024

through a mediator. For example, let's suppose that I wanted to try to convince the Director of my institute to provide a laptop. According to the rules of the institution, I must forward my request through department head and then administrative officer. Each one of these people could delay the request, or even prevent my request from getting through. This gives the people who lie "between" me and the Director power with respect to me. Having more than one channel makes me less dependent, and, in a sense, more powerful. Betweenness centrality views an actor as being in a favoured position to the extent that the actor falls on the geodesic paths between other pairs of actors in the network. That is, the more people depend on me to make connections with other people, the more power I have. If, however, two actors are connected by more than one geodesic path, and I am not on all of them, I lose some power. Using the computer, it is quite easy to locate the geodesic paths between all pairs of actors, and to count up how frequently each actor falls in each of these pathways. If we add up, for each actor, the proportion of times that they are "between" other actors for the sending of information in the dataset used by us in this chapter, we get the a measure of actor centrality. We can norm this measure by expressing it as a percentage of the maximum possible betweenness that an actor could have had. The results for this are shown in Tables 2.5 and 2.6.

Here, we get Network Centralization Index = 0.24%, a very low one with Unnormalized centralization of 6.000.

We can see from Table 2.5 that there is a variation in actor betweenness (from zero to 0.6), and that there is a variation (std. dev. = 0.283 relative to a mean betweenness of 0.2) and the overall network centralization is very low. This is important in the sense that, because we know that most of the connections can be made in this network with the aid of any intermediary – hence there can be a lot of "betweenness" with same value of 0.600. In the sense of structural constraint, the network contains lots of power that could be important for group formation and stratification.

Table 2.7 Degree centrality measure

		1	2	3
		Degree	NrmDegree	Share
15	6	581320.000	90.316	0.264
1	1	561374.938	87.217	0.255
12	5	86341.547	13.414	0.039
6	3	85508.977	13.285	0.039
5	3	85508.977	13.285	0.039

Table 2.8 Descriptive statistics for degree centrality

		1	2	3
		Degree	NrmDegree	Share
1	Mean	146766.125	22.802	0.067
2	Standard deviation	166598.547	25.883	0.076
3	Variance	27755075584.000	669.951	0.006

Freemans Degree Centrality Measures

Actors who have more ties to other actors are considered to be in advantageous positions for having many ties, many alternative ways to satisfy needs, and hence are less dependent on other individuals. Since they have many ties, they may have access to, and be able to call on more of the resources of the network as a whole. Also, they can act as third parties and deal makers in exchanges among others, and finally able to get benefit from this brokerage. So, degree centrality measure provides a very simple, but often very effective measure of an actor's centrality and power potential in a social network analysis. The results obtain for this is provided in Tables 2.7 and 2.8. Table 2.6 shows that actor 12 & 5 has highest degree with a value of 86341.547 and mean of 146766.125 and deviation of 166598.547. In this, Network Centralization = 77.90%, Heterogeneity = 15.26%, Normalized = 9.20%.

The degree of points is important because it tells us how many connections an actor has. Actors that receive information from many sources may be prestigious with high value (other actors want to be known by the actor, so they send information), and actors that receive information from many sources may also be more powerful. But, actors that receive a lot of information could also suffer from "information overload" or "noise and interference" due to contradictory messages from different sources. Hence, choosing a degree is of paramount importance in order to build a good social network model.

K-Clusters Using Tabu Search

As discussed earlier in section "Data Mining", Tabu search is a numerical method for finding the best division of actors into a given number of partitions on the basis of approximate automorphic equivalence. With this, it is important to explore a range

Table 2.9 Density table

	1	2	3	4	5
1	19410.510	19.778	4.750	−0.500	0.000
2	13561.036	15332.223	0.333	−0.500	0.000
3	13058.829	2.500	22987.500	−0.500	0.000
4	13066.239	4.667	0.500	22987.250	0.000
5	13310.836	6.000	−0.500	−0.500	22987.500

of possible numbers of partitions, one has to determine intelligently on how many partitions are useful. Having selected a number of partitions, it is useful to re-run the algorithm a number of times to insure that a global, rather than local minimum has been obtained. The detail parameter settings used in our experiments using Tabu Search algorithm is provided below.

Number of clusters: 5
Type of data: Similarities/Strengths/Cohesion
Method: correlation
Starting fit: 1.174; Starting fit: 0.609; Fit: 0.608; Fit: 0.605; Fit: 0.607; Fit: 0.605 (smaller values indicate better fit). r-square = 0.156

Clusters:
1: 1 3 3 5 8 6
2: 1 1 7
3: 4 7
4: 2 2
5: 4 6

Table 2.9 provides the density table for analysing the relationships amongst the actors present in the network, where Density is defined as the total number of ties divided by the total number of possible ties. It is especially relevant for knowledge community building within and between organizations, for a thorough understanding about the overall linkage between network members. The more the value of the density, the better will be the knowledge flow and denser will be the network. However, a negative value indicates less dense network.

SNA with Terrorist Network Data

Followed by the usefulness and description of all performance measures in Netlog dataset, we here use the terrorist dataset obtained from UCINET tool for our social network analysis. The closeness centrality is shown in Table 2.10.

Table 2.10 Closeness centrality measures using a sample of terrorist dataset

		1	2
		Farness	nCloseness
6	Mohamed Atta	107.000	57.944
11	Marwan Al-Shehhi	134.000	46.269
1	Hani Hanjour	141.000	43.972
3	Nawaf Alhazmi	141.000	43.972
21	Zacarias Moussaoui	144.000	43.056
51	Jean-Marc Grandvisir	250.000	24.800
52	Abu Zubeida	250.000	24.800
35	Nabil Almarabh	257.000	24.125

Table 2.11 Closeness centrality statistics

		1	2
		Farness	nCloseness
1	Mean	183.460	34.773
2	Standard deviation	30.430	6.081
3	Variance	925.963	36.978

Closeness Centrality

As discussed in section "SNA with Netlog Data" with Netlog dataset, more the closeness and less the farness value, better is the relationship between the terrorists in the network. From Table 2.10, it is evident that Nabil Almarabh is having more distant relationship in the network. The various statistics with mean farness of terrorist in the network is 183.460 with a deviation of 30.430, which is shown in Table 2.11 with a Network Centralization of 47.48%

Freeman Betweenness Centrality

In this, the Un-normalized centralization for the network using terrorist dataset is 65911.478. From Table 2.12, it can be observed that Essid Sami Ben Khamais has a high value of betweenness measure with 470.473, which indicates that among all others, he can work as a most vital mediator of knowledge flows with a high potential of control on the indirect relations of the other members in the network. The corresponding descriptive statistics are shown in Table 2.13. In the whole process of measuring betweenness centrality, our simulation gives a Network Centralization Index of 56.22%.

K-Clusters Using Tabu Search

Finally, we use Tabu search algorithm with 5-cluster to analyse our proposed social network analysis with following parameters, same as to that of taken for analysing Netlog dataset.

Table 2.12 Betweenness centrality measures for terrorist dataset

		1	2
		Betweenness	nBetweenness
6	Mohamed Atta	1106.944	58.538
37	Essid Sami Ben Khemais	470.473	24.880
21	Zacarias Moussaoui	434.533	22.979
3	Nawaf Alhazmi	287.580	15.208
1	Hani Hanjour	233.759	12.362
46	Djamal Beghal	195.683	10.348

Table 2.13 Descriptive statistics for each measure of betweenness centrality

		1	2
		Betweenness	nBetweenness
1	Mean	60.730	3.212
2	Standard deviation	162.565	8.597
3	Variance	26427.225	73.904

Table 2.14 Density table analysis

	1	2	3	4	5
1	0.433	0.029	0.024	0.005	0.059
2	0.029	0.260	0.050	0.009	0.020
3	0.024	0.050	0.260	0.018	0.013
4	0.005	0.009	0.018	0.355	0.000
5	0.059	0.020	0.013	0.000	0.333

Number of clusters: 5
Type of data: Similarities/Strengths/Cohesion
Method: correlation
Starting fit: 1.050; Starting fit: 0.539; Fit: 0.543; Fit: 0.550; Fit: 0.543; Fit: 0.539 (smaller values indicate better fit). r-square = 0.212
Cluster 1: Mohamed Atta Waleed Alshehri Wail Alshehri Satam Suqami Abdul Aziz Al-Omari* Marwan
Cluster 2: Zacarias Moussaoui Kamel Daoudi Mamduh Mahmud Salim Faisal Al Salmi Bandar Alhazmi
Cluster 3: Raed Hijazi Nabil Almarabh Nizar Trabelsi Djamal Beghal Abu Qatada Ahmed Khalil
Cluster 4: Essid Sami Ben Khemais Haydar Abu Doha Mohamed Bensakhria Tarek Maaroufi Lased Ben
Cluster 5: Hani Hanjour Majed Moqed Nawaf Alhazmi Salem Alhazmi* Khalid Al-Mihdhar Ahmed

The density table after Tabu search algorithm applied in the dataset is provided below in Table 2.14. The information flow will be better for high value of density.

Conclusions

We presented the cluster based data mining method to extract relationships in social network analysis using Netlog data and Terrorist network data. We used different performance measures in order to build a useful social relationship amongst the users. The social network analysis using cluster analysis made it very much useful for organizations to analyse the social network to understand the internal and external association of an organization, which further will be of immense use for collaborative work for innovation and dissemination of knowledge.

References

1. Wasserman, S., Faust, K.: Social Network Analysis: Methods and Applications (Structural Analysis in the Social Sciences). Cambridge University Press, Cambridge (1994)
2. Schtt, J.: Social Network Analysis: A Handbook. Sage Publications, Newbury Park (1991)
3. Hanneman Robert, A., Mark, R.: Introduction to Social Network Methods. University of California, Riverside, Riverside (2005)
4. Finin, T., Joshi, A., Kolari, P., Java, A., Kale, A., Karandikar, A.: The information ecology of social media and online communities. Artif. Intell. Mag. **28**, 1–12 (2008)
5. Eagle, N., Portland, A.: Reality mining: sensing complex social systems. Pers. Ubiquit. Comput. **10**, 255–268 (2005)
6. Boyd, D.M., Ellison, N.B.: Social network sites: definition, history and scholarships. J. Comput. Mediat. Commun. **13**(1), 210–230 (2008). Wiley
7. Tufekci, Z.: Grooming, gossip, Facebook and MySpace: what can we learn about these sites from those who won't assimilate? Inf. Commun. Soc. **11**(4), 544–564 (2008)
8. Sheldon, P.: The relationship between unwillingness to communicate and students Facebook use. J. Media Psychol: Theor. Method. Appl. **20**(2), 67–75 (2008)
9. Thelwall, M., Wilkinson, D., Uppal, S.: Data mining emotions in social network communications: gender differences in MySpace. J. Am. Soc. Sci. Technol **61**, 1–14 (2009). Wiley
10. Zhou, L., Ding, J., Wang, Y., Cheng, B., Cao, F.: The social network mining of BBS. J. Netw. **4**(4), 298–305 (2009)
11. Lewis, K., Kaufman, J., Gonzalez, M., Wimmer, A., Christakis, N.: Tastes, ties and time: a new social network dataset using FaceBook.com. Soc. Netw. **30**, 330–342 (2008). Elsevier
12. Vaidyanathan, A., Shore, M., Billinghurst, M.: Data in social network analysis. In: Computer-Mediated Social Networking. Lecture Notes in Computer Science, vol. 5322, pp. 134–149. Springer, Berlin/New York (2009)
13. Lusseau, D., Schneider, K., et al.: Behavioural ecology and sociology, vol. 54. Addison-Wesley, US (2003)
14. Academic, L.A., Glance, N.: The political blogosphase and the 2004 US election. In: Proceedings of the 2005 Workshop on the Weblogging Ecosystem, Chiba, Japan, 10–14 May 2005
15. Snasel, V., Horak, Z., Abraham, A.: Understanding social networks using formal concept analysis. In: IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology, WI-IAT '08, Vol. 3, pp. 390–393, 2008
16. Choa, A., Hernandez, A., Gonzalez, S., Castro, A., Gelbukh, A., Hernandez, A., Iztebegovic, H.: Social data mining to improved bio-inspired intelligent systems. In: Giannopoulou, E.G. (ed.) Data Mining in Medical and Biological Research, pp. 291–320. I-Tech, Vienna (2008)
17. Bhattacharya, I., Getoor, L.: Iterative record linkage for clearing and integration. In: Proceedings of the SIGMOD 2004 Workshop on Research Issues on Data Mining and Knowledge Discovery, Paris, France, pp. 1–18, 13 June 2004

18. Kubica, J., Moore, A., Schneider, J.: Tractable group detection on large link datasets. In: Proceedings of the 3rd IEEE International Conference on Data Mining, Melbourne, FL, pp. 573–576, 19–22 Dec 2003
19. Lu, Q., Geetoor, G.: Link based classification. In: Proceedings of the 2003 International Conference on Machine Learning, Washington, DC, pp. 496–503, 21–24 Aug 2003
20. Liben-Nowell, D., Kleinberg, J.: The link prediction problem for social networks. In: Proceedings of the 2003 International Conference on Information and Knowledge Management, New Orleans, LA, pp. 556–559, 2–8 Nov 2003
21. Krebs, V.: Mapp. Netw. Terror. Cell Connect. **24**, 43–52 (2002)
22. Page, L., Brin, S., Motwani, R., Winograd, T.: The PageRank Citations Ranking: Bringing Order to the Web, Technical Report. Stanford University, Stanford (1998)
23. Kleinberg, J.: Autoritive success in a hyper linked environment. J. ACM **5**, 604–632 (1999)
24. Wasserman, S., Faust, K.: Social Network Analysis. Cambridge University Press, Cambridge (1994)
25. Milgram, S.: The small world problem. Psychol. Today **2**, 60–67 (1967)
26. Travers, J., Milgram, S.: An experimental study of the small world problem. Sociometry **32**, 425–443 (1969)
27. Guare, J.: Six Degrees of Separation: A Play. Vintage, New York (1990)
28. Marsden, P.V.: Network data and measurement. Ann. Rev. Sociol. **16**, 435–463 (1990)
29. Katzir, L., Liberty, E., Somekh, O.: Estimating sizes of social networks via biased clustering. In: Proceedings of the International Conference on World Wide Web (WWW-2011), Hyderabad, India, 28 Mar–1 Apr 2011, pp. 597–605. ACM Press, New York (2011)
30. Amaral, L.A.N., Scala, A., Barthélémy, M., Stan-ley, H.E.: Classes of small-world networks. Proc. Natl. Acad. Sci. USA **97**, 11149–11152 (2000)
31. Newman, M.E.J., Strogatz, S.H., Watts, D.J.: Random graphs with arbitrary degree distributions and their applications. Phys. Rev. E **64**, 026118 (2001)
32. Newman, M.E.J.: The structure of scientific collaboration networks. Proc. Natl. Acad. Sci. USA **98**, 404–409 (2001)
33. Barabási, A.-L., Jeong, H., Ravasz, E., Néda, Z., Schu-berts, A., Vicsek, T.: Evolution of the social net-work of scientific collaborations. Physica A **311**, 590–614 (2002)
34. Davis, G.F., Greve, H.R.: Corporate elite networks and governance changes in the 1980s. Am. J. Sociol. **103**, 1–37 (1997)
35. Mariolis, P.: Interlocking directorates and control of cor-porations: the theory of bank control. Soc. Sci. Quart. **56**, 425–439 (1975)
36. Aiello, W., Chung, F., Lu, L.: A random graphmodel for massive graphs. In: Proceedings of the 32nd Annual ACM Symposium on Theory of Computing, Portland, 21–23 Mar 2000, pp. 171–180. Association of Computing Machinery, New York (2000)
37. Aiello, W., Chung, F., Lu, L.: Random evolution of massive graphs. In: Abello, J., Pardalos, P.M., Resende, M.G.C. (eds.) Handbook of Massive DataSets, pp. 97–122. Kluwer, Dordrecht (2002)
38. Ebel, H., Mielsch, L.-I., Bornholdt, S.: Scale-freetopology of e-mail networks. Phys. Rev. E **66**, 035103 (2002)
39. Abraham, A., et al.: Reducing social network dimensions using matrix factorization methods. In: Proceedings of the 2009 Advances in Social Network Analysis and Mining, 19 Jan 2009, pp. 348–351. IEEE press, Piscataway (2009)
40. Freeman, L.C.: Graphical techniques for exploring social network data. In: Carrington, P.J., Scott, J., Wasserman, S. (eds.) Models and Methods in Social Network Analysis. Cambridge University Press, Cambridge (2005)
41. Bulkley, N., Alstyne, V., Marshall, W.: An Empirical Analysis of Strategies and Efficiencies in Social Networks, 1 Feb 2006. Boston University School of Management Research Paper No. 2010–29, MIT Sloan Research Paper No. 4682–08. Available at SSRN: http://ssrn.com/abstract=887406

42. Abraham, A., et al.: Social aspects of web page contents. In: Abraham, A., Snásel, V., Wegrzyn-Wolska, K. (eds.) Proceedings of the International Conference on Computational Aspects of Social Networks, CASoN 2009, Fontainebleau, France, 24–27 June 2009, pp. 80–87. IEEE Computer Society, Washington, DC (2009)
43. Divjak, B., Peharda, P.: Social network analysis of study environment. JIOS **34**(1), 67–80 (2010)
44. Han, J., Kamber, M.: Data Mining: Concepts and Techniques, 2nd edn. Morgan Kauffman, San Francisco (2006)
45. Glover, F.: Heuristics for integer programming using surrogate constraints. Decis. Sci. **8**(1), 156–166 (1977)
46. Halgin, D.: An introduction to UCINET and NetDraw. In: 2008 NIPS UCINET and NetDraw Workshop, Harvard University, Cambridge, 13–14 June 2008, pp. 1–47
47. Gyarmati, L., Tuan, A.T.: Measuring user behaviour in online social networks. IEEE Netw. **24**(5), 26–31 (2010)
48. Borgatti, S.P., Everett, M.G., Freeman, L.C.: Network analysis of 2-mode data. Soc. Netw. **19**, 243–269 (2002). Elsevier