Locating Key Actors in Social Networks Using Bayes' Posterior Probability Framework

D.M. Akbar Hussain and Daniel Ortiz-Arroyo

Department of Software Engineering & Media Technology
Esbjerg Institute of Technology
Niels Bohrs Vej 8, Esbjerg 6700, Denmark
akbar@aaue.dk, do@aaue.dk

Abstract. Typical analytical measures in graph theory like degree centrality, betweenness and closeness centralities are very common and have long history of their successful use. However, modeling of covert, terrorist or criminal networks through social graph dose not really provide the hierarchical structure of such networks because these networks are composed of leaders and followers. It is possible mathematically, for some graphs to estimate the probability that the removal of a certain number of nodes would split the networks into may be non functional network. In this research we investigate and analyze a social network using Bayes probability theory model to calculate entropy of each node present in the network to high light the important actors in the network. This is accomplished by observing the amount of entropy change computed by successively removing each node in the network.

Keywords: Social Networks Analysis, Bayes' Theorem, Entropy, Key Actors.

1 Introduction

A typical social network (social graph) shows the connections amongst various nodes representing actors (people) revealing many characteristics of these nodes for example active, semi active, passive and dormant nodes. The human social networks are similar to a large picture with fuzzy borders which may some time overlap with other social networks. Social interactions represent an important activity describing understanding, mutual common interests, including joint work/projects, hobbies, or simply common destinations in a physical environment. Similarly, drug dealers, terrorist and covert networks are also represented through social graphs. Since 9-11 terrorist attacks a great deal of research is taking place firstly to understand the dynamics of these terrorist networks (analysis) and secondly, developing methods to either destabilize or disintegrate these networks. Insight visualization of any social network typically focuses on the characteristics of the network structure. Social Network Analysis is a mathematical method for 'connecting the dots', SNA allows us to map and measure complex relationships/connections between human groups, animals, computers

D. Ortiz-Arroyo et al. (Eds.): EuroISI 2008, LNCS 5376, pp. 27–38, 2008.

or other information/knowledge processing entities and organizations [1]. These relationships can reveal unknown information about these dots and the network itself. Jacob Moreno invented "Sociometry" which is the basis of SNA, utilized "sociograms" to discover leaders and map indirect connections in 1934 [2]. The two basic elements of SNA are connections and nodes. Connections are ties between individuals or groups and nodes are the individuals or groups involved in the network. Typically, importance of a node 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 which case all subordinate nodes send information to the central node and the central node disseminate the information to all other nodes in the network [3,4,5]. Centralized networks are susceptible to disruption because damage to a central node is normally catastrophic to the entire network, similar in principle to a client server architecture. There are different dynamics of social networking for example Kin-based (father, husband), Role-based (office), Interactions (chatting) and Affiliations (clubs etc). Analysts have applied SNA in many fields to reveal hidden informal links between nodes [6]. For example in businesses SNA have been used to analyze email patterns to determine which employees are overloaded, similarly, law enforcement and national security organizations are using various method of SNA to identify important nodes and connections of terrorist organizations [7].

2 Literature Review

SNA has widely been used to study the networks for example in qualitative studies the facilitators of link establishment and in quantitative studies the use of statistical methods to measure existing network. Most studies in link establishment have been carried out in sociology and criminology [8]. Statistical analysis mostly dealt with exploring the key actors using standard centrality measures. In contrast to this, the dynamic social network analysis methods have been dealing with network recovery, network measurement and statistical analysis. In network recovery multiple instantaneous network representation are recovered from longitudinal data to model the evolving network. In dynamic network measurement three types of techniques are used, deterministic measure, probabilistic measures and Temporal measures. In deterministic measures network size, degree, betweenness and closeness measures are computed whereas in probabilistic measures degree distribution and clustering coefficient are measured. As the network development is a continuous process so temporal measure deals with this continuous process by considering a time variable. Statistical analysis typically studies and explains the topologies of networks. Paramjit and Swartz [9] have used random-effects models to incorporate dependence between the dyads, originally this idea was proposed by Wong [10] in which the likelihood of ties in terms of the nodal attributes rather than in terms of network structural properties for example transitivity and cyclicity are expressed. Bayesian approach has been used in network modeling. Markov chain Monte Carlo (MCMC)

simulation technique has also been used to determine the characteristic marginal posterior distribution which allows for complicated modeling and inference independent of sample size. This is in contrast with analyses which focus only on the estimation of primary parameters and their asymptotic standard errors. MCMC has been used by Gill and Swartz for Bayesian analysis of round robin interaction data where the response variable was continuous [11,12]. Nowicki and Snijders [13] used MCMC Bayesian analysis for block model structures where the relationship between two nodes depends only on block membership. How the basic Bayesian model can be modified to cater to special settings are presented by Holland and Leinhardt [14]. Paramjit [9] demonstrated to introduced covariates and the stochastic block models to the basic Bayesian model [10] and how MCMC simulation output can be used in model selection for Bayesian analysis of directed graphs data. Our method of using Bayes posterior probability for statistical analysis is very straight forward as we compute the posterior probability for each node and then this probability is used in the evaluation of over all entropy of the network (explained later in section 4 and 5).

3 Network Structure and Analysis

Given any network where the nodes/agents are individuals, groups, organizations etc., a number of network measures such as centrality or cut-points are used to locate critical/important nodes/agents. Typically, social network analysis try to identify the following characteristics:

- Important individual, event, place or group.
- Dependency of individual nodes.
- Leader-Follower identification.
- Bonding between nodes.
- Vulnerabilities identification.
- Key players in the network.
- Potential threat from the network.
- Efficiency of overall network

Networks visualization is semantically presented in the form of a graph in which the nodes represent entities and the arcs represent relationship among nodes. Classification of nodes and its distinctiveness is a challenging task and one needs to discover the following characteristics [15].

- An individual or group that if given new information can propagate it rapidly.
- An individual or group that has relatively more power and can be a possible source of trouble, potential dissidents, or potential innovators.
- An individual or group where movement to a competing group or organization would ensure that the competing unit would learn all the core or critical information in the original group or organization (inevitable disclosure).
- An individual, group, or resource that provides redundancy in the network.

Many traditional social network measures and the information processing network measures can help in revealing importance and vulnerabilities of the nodes/ agents in the network [16,17,18,19]. Application of existing tools on these complex socio-technical networks/systems is very demanding to winkle out the required information. Most of the measures and tools work best when the data is complete; i.e., when the information is inclusive about the interaction among the nodes. However, the difficulty is that covert and terrorist networks are typically distributed across many boundaries for example from cities or countries and data about them is never complete-correct at a certain instant of time. Normally, a sampled snapshot data is available some of the links may be intentionally hidden. Also data is collected from multiple sources for example news (print/tv), open source internet data, security agencies, etc., and at different time instants. In addition inclusive and correct information may be prohibitive because of secrecy. Obviously, there could be other difficulties but even these provide little guidance for what to expect when analyzing these complex socio-technical systems with the developed tools. Following paragraph provides the strength and limitations of SNA.

– Strengths

The most fundamental strength of SNA is that it provides a visual representation of the network structure. It allows the analysts to compare and identify previously unknown links. The knowledge gained through this process even can be used to forecast not only the individual activities of the actors but also of network/organization.

– Limitations

SNA is data dependent like most analytical software, therefore, correct and up to date data is essential for true analysis of a network/organization, therefore, if the data is incomplete or incorrect final product will be inaccurate. Generally it is believed that SNA is used as a tool only and one should not be relied upon to provide an absolute depiction of a network. Another important point of its limitation is that it is time consuming it takes a great deal of time to research a topic in order to find the appropriate information.

In this paper we are analyzing social networks systematically using Bayes posterior probability to calculate the entropy of individual nodes [20]. Once the total entropy of the whole network is evaluated then successively one node from the network is removed each time and the effect in entropy is measured. The maximum entropy change is expected to occur for the most important key player node showing the level of uncertainty if that node is not present in the network. Section 4 provides a mathematical formulation of Bayes theorem in relation to our methodology, entropy calculations and experimental results are discussed in section 5 and finally conclusion is summarized in section 6.

4 Bayes Theorem

Bayes' Theorem is a simple mathematical formula used for calculating conditional probabilities. Bayes' Theorem originally stated by Thomas Bayes and it

has been used in a wide variety of contexts, ranging from marine biology to the development of "Bayesian" Spam blockers for email systems. Through the use of Bayes' Theorem precise measures can be obtained by showing how the probability that a theory is correct is affected by new evidence [21,22]. In a Bayesian framework the conditional and marginal probabilities for stochastic events for example A and B are computed through this relationship:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} \tag{1}$$

$$P(A|B) \propto P(B|A)\ P(A) \tag{2}$$

Where P(A) is the prior probability or marginal probability of A, $P(A|B)$ is the conditional probability given B also called posterior probability. $P(B|A)$ is conditional probability given A, P(B) is prior probability and considered as normalizing constant. $L(A|B)$ is the likelihood of A given fixed B, here $P(B|A)$ is equal to $L(A|B)$ however, at times likelihood L can be multiplied by a factor so that it is proportional to, but not equal probability P. It should be noted that probability of an event A conditional on another event B is generally different from the probability of B conditional on event A, however, there is a unique relationship between the two which is provided by Bayes theorem. We can formulate the above relationship as:

$$posterior = \frac{likelihood\ \times\ prior}{normalizing\ constant} \tag{3}$$

We can re-write equation 1 as the ratio $P(B|A)/P(B)$ which is typically called as standardized likelihood or normalized likelihood so it can be written as:

$$posterior = normalized\ likelihood\ \times\ prior \tag{4}$$

Suppose we have a network of nodes (graph) and we are interested in calculating the posterior probability $P(A|B)$ of a node to see if it is the most important node of the network. Bayes probability theory provides such possibility through its conditional probability theorem, for this reason we have expanded the above expression for convenience to interpret various terms according to our implementation model. Therefore,

$$P(A|B) = \frac{P(B|A)\ P(A)}{P(B|A)\ P(A)\ +\ P(B|N)\ P(N)} \tag{5}$$

P(A) is the prior probability or marginal probability of node A regardless of any information, which is computed by considering the total number of nodes present in the network. For example if there are ten nodes in the network then each node has 10 % chance of being the key actor node, however, if we assume that the node under consideration is the key actor node then it must have probability value of 90 % or above for being the key actor node (prior probability) as the other 9 nodes are not important actors in the network. P(N) is the probability

that the node is not a key-player given by (1 - P(A)). $P(B|A)$ is conditional probability given A, meaning that node is a key-player, which is computed based on the number of links incident on that particular node, so if there are n nodes in the network then to be the central node of the network it has to be linked with other (n - 1) nodes. $P(B|N)$ is conditional probability given N meaning that node is not a key-player, which is obtained by computing $(1-P(B|A))$. Bayesian approach have been used in dynamic SNA issues, statistical analysis and Network measurement [22,23,24,25], our approach here is different, it is straight forward and much simpler. Basically, here we are interested in evaluating the theory or hypothesis (equation 1) for A based on B which is the new information (evidence) that can verify the hypothesis and P(A) is our best estimate of the probability (known as the prior probability of A) prior to considering the new information. What we are interested is to discover the probability that A is correct (true) with the assumption that the new information (evidence) is correct. We are using the Bayes probability values obtained through the relationship given by equation 5 in our mathematical derivation of uncertainty level entropy formula given in the next section.

5 Shannon's Entropy as Uncertainty

Uncertainty is observed in most situations where probability theory is applied or used, for example tossing a fair coin or rolling a fair dice, one cannot guarantee what will be the outcome [26]. However, one can describe the scenario with a probability distribution for example in the case of fair coin: Pr(coin=head)=0.5; Pr(coin=tail)=0.5 and in the case of fair dice: Pr(dice=1)=1/6; Pr(dice=2)=1/6; Pr(dice=3)=1/6; Pr(dice=4)=1/6; Pr(dice=5)=1/6; Pr(dice=6)=1/6 but what if the coin and dice are biased then it will have different distribution for example; Pr(coin=head)=0.4; Pr(coin=tail)=0.6

Therefore, probability distributions are not created equal which implies that each of these distributions have different uncertainty and interestingly, fair dice or fair coin has the highest uncertainty as we are in more doubt about the outcome. Shannon converted this uncertainty into a quantitative measure (real number) H[X] for a random variable X, which takes the probability distribution as [27]; X = Pr(1); Pr(2); Pr(3); Pr(4); Pr(5); .Pr(n) this states that X can assume value from n possible choices. The quantitative measure H should be uniformly distributed for complete uncertainty meaning each outcome has equal likelihood to occur. H also has to be continuous function of probabilities so a small change in probability should always bring a small change in H, finally probabilities can be grouped in different ways so H is a function of the distribution and not a function based on our grouping within the distribution. Based on these assumption entropy of a random variable is given as;

$$H\left[X\right] = k\sum_{r=1}^{n} P_r(x) log P_r(x) \tag{6}$$

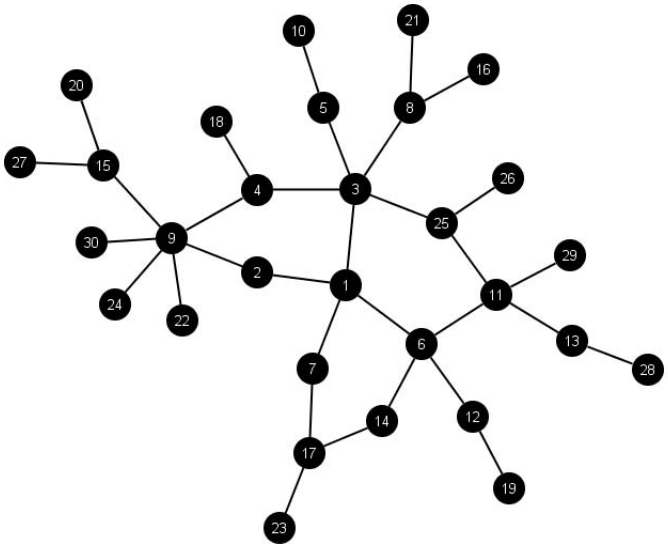

Fig. 1. Example 30 Node Network

Where k is an arbitrary constant, which is taken as -1, we can rewrite the above formula as [26];

$$H[X] = -1 \sum_{r=1}^{n} P_r(x) log P_r(x) \tag{7}$$

It should be noted that H[X] is not a function of random variable X rather it is a function of probability distribution of X [22].

Example Network 1

Our first example model is shown in figure 1, which shows a network of relatively less complex interactions, this network is a hierarchical structure similar to a typical small organization. This network has 30 nodes, first of all we need to determine the Bayes priori probability for each node using the computation explained earlier in section 4 then Bayes posterior probability for each node is computed after substitutions of corresponding terms. These probability values are then substituted in the entropy formula equation (7) and networks overall entropy is computed. Now successively a single node is physically removed from the network and the system computes its adjacency matrix based on the new structure (removal of a node) and then same cycle of computation starts for Bayes probabilities and entropy calculations and results are stored in a vector. Once this process for each of the node in the network is completed the entropy vectors corresponding to each node are plotted as a mesh matrix shown in figure 2, the color coded bar (value from 0.38 to 0.48) is used to indicate the amount of change in entropy from the over all network entropy. It can be seen in figure 2 that key player nodes are shown with a shade having small value 0.38 where as less important players have shade with

Fig. 2. Entropy Mesh Matrix

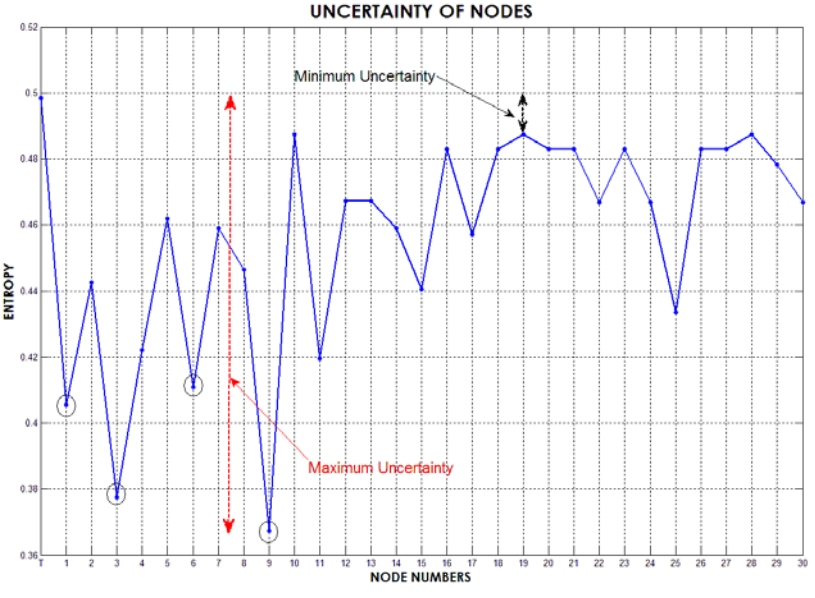

Fig. 3. Nodes Uncertainty

high value 0.48. The uncertainty is computed after removal of each node which is then compared against the networks overall entropy in figure 3, here it can be seen clearly that the important key actors in the network are node 1, 3, 6 and 9, more precisely 3 and 9 are the most important nodes.

Example Network 2

Next we took an example of a larger and more complex random network with 60 nodes as shown in figure 4. This network is more complex then the previous

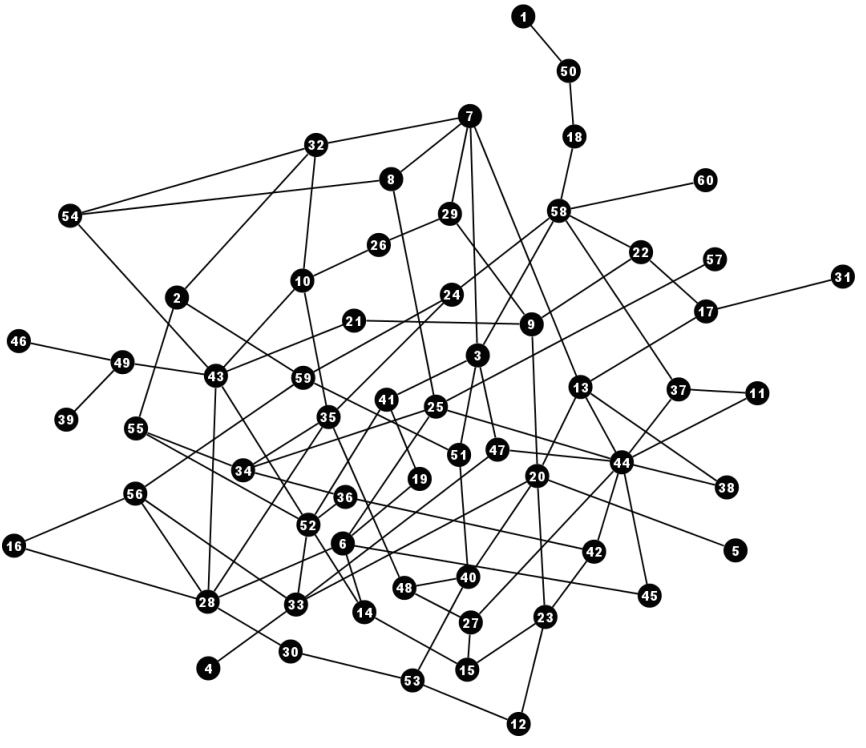

Fig. 4. Random 60 Node Network

network because the numbers of links and the nodes have been doubled. After, evaluating the posterior probability and subsequently the uncertainty, results are plotted in figure 5, it can be seen that node 6, 7, 13, 20, 28 and 44 seems to be the important key actors, 44 being the most important node (shade color value of 0.4). The uncertainty of nodes against the over all network entropy is plotted in figure 6 and it also indicates similar nodes to be the important players in the network, however, if you look at the network more closely very interesting results are actually present in figure 6. By visual inspection it can be seen in the network of figure 4, that there are 5 (20, 28, 43, 52 and 58) nodes having same number of links (degree centrality = 6). However, our system reveals that node 58 has less importance than the rest of four nodes, which is evident if we look at the placement of node 58 in the network. This shows that our method is more robust and efficient in predicting the key actor node as its entropy change is more compared with other nodes having same standing in the network. Also, there are 7 (3, 6, 7, 13, 25, 33 and 35) nodes having the same numbers of links (degree centrality = 5) but our system have shown that node 13 has the greater uncertainty among them making it the key actor node for this group.

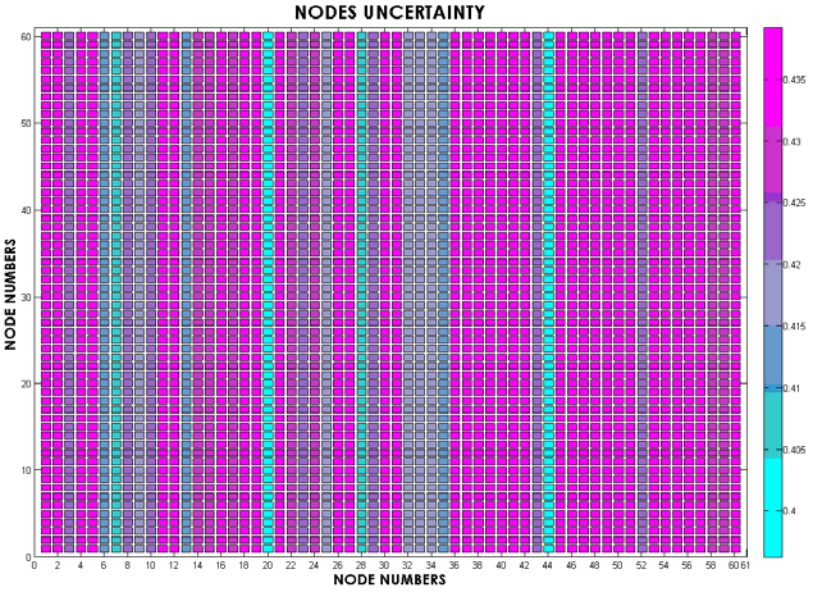

Fig. 5. Entropy Mesh Matrix

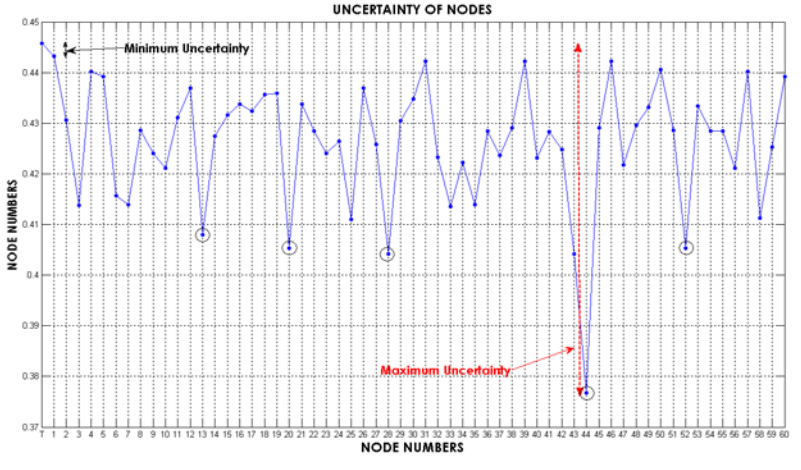

Fig. 6. Nodes Uncertainty

6 Conclusion

The standard statistical solution for SNA has been matured for long time now and used in the studying social behavior however, elucidating the pattern of connections in social structure is very challenging. The reason being that some of the existed links within the network are not visible or cannot be seen or may

be concealed by individuals so the conventional social network analysis cannot be applied. The real world social networks including small world networks have varying complexity. The purpose of this paper is to investigate and locate the important actors in such networks. The idea of using such model is based on the underlying assumption philosophy of Bayesian Posterior Probability that uncertainty and degree of belief can be measured as probability. We have shown through simulation that Bayes approach combined with information entropy model is very useful in revealing the key players/actors in a social network. We have computed results for many networks having varying degree of complexity but results for two such networks are presented although all of them shown consistency in revealing the important information. In our future work we would like to extend this framework by incorporating additional information for computing the prior probability from its simple total network number to information type and message contents.

References

1. Krebs, V.: Connecting the dots, tracking two identified terrorists (2002)
2. Moreno, J.L.: Sociometry, experimental method and the science of society, an approach to a new political orientation. Beacon house (1951)
3. Freeman Linton, C.: A set of measures of centrality based on betweenness. Sociometry 40, 35–41 (1971)
4. Freeman Linton, C.: Centrality in social networks: Conceptual clarification. Social networks 1, 215–239 (1979)
5. Anthonisse, J.M.: The rush in a graph, university of amsterdam mathematical centre, amsterdam (1971)
6. Kutcher, C.: Social network analysis - linking foreign terrorist organizations (2008)
7. Akbar Hussain, D.M.: Destabilization of terrorist networks through argument driven hypothesis model. Journal of software 2(6), 22–29 (2007)
8. Kaza, S., Hu, D., Chen, H.: Dynamic social network analysis of a dark network: Identifying significant facilitators. In: ISI, pp. 40–46 (2007)
9. Gill, P.S., Swartz, T.B.: Bayesian analysis of directed graphs data with applications to social networks. Appl. statist. 53, part 2, 249–260 (2004)
10. Wong, G.Y.: Bayesian models for directed graphs. J. am. statist. ass. 82, 140–148
11. Gill, P.S., Swartz, T.B.: Statistical analyses for round robin interaction data. Can. j. statist. 29, 321–331
12. Gill, P.S., Swartz, T.B.: Bayesian analysis for dyadic designs in psychology
13. Nowicki, K., Snijders, T.A.B.: Estimation and prediction for stochastic blockstructures. J. am. statist. ass. 96, 1077–1087
14. Holland, P.W., Leinhardt, S.: An exponential family of probability distributions for directed graphs. J. am. statist. ass. 76, 33 – 65
15. Carley, K.M., Lee, J.-S., Krackhardt, D.: Destabilizing networks, dept. of social and decision sciences, carnegie mellon university, pittsburgh, pa 15143 (November 2001)
16. Bavelas, A.: A mathematical model for group structures. Human organization 7, 16–30 (1948)
17. Shaw, M.E.: Group structure and the behaviour of individuals in small groups. Journal of psychology 38, 139–149 (1954)

18. Scott, J.: Social networks analysis, 2nd edn. Sage publications, London (2003)
19. Newman, M.E.J.: A measure of betweenness centrality based on random walks, cond-mat/0309045 (2003)
20. Hayter, A.J.: Probability and statistics for engineers and scientists, 2nd edn. (2002) ISBN 0-534-38669-5
21. Ibe, O.C.: Fundamentals of applied probability and random processes. Elsevier/Academics press (2005) ISBN 0-12-088508-5
22. Montgomery, D.C., Runger, G.C.: Applied statistics abd probability for engineers, 4th edn. John Wiley and Sons, Chichester (2006)
23. Koskinen, J.H., Snijders, T.A.B.: Bayesian inference for dynamic social network data. Journal of statistical planning and inference 137, 3930–3938 (2007)
24. Siddarth, K., Daning, H., Chen, H.: Dynamic social network analysis of a dark network: Identifying significant facilitators. In: Proceedings of IEEE international conference on intelligence and security informatics, ISI 2007, New Brunswick, New Jersey, USA, May 23 - 24 (2007)
25. Rhodes, C.J., Keefe, E.M.J.: Social network topology: a bayesian approach. Journal of the operational research society 58, 1605–1611 (2007)
26. Feldman, D.: A brief introduction to information theory, excess entropy and computational mechanics
27. Shannon, C.E.: A mathematical theory of communication. Bell systems tech. J. 27, 379–423