# Chapter 2
# Discovering Sets of Key Players in Social Networks

**Daniel Ortiz-Arroyo**

**Abstract** The discovery of single key players in social networks is commonly done using some of the centrality measures employed in social network analysis. However, few methods, aimed at discovering *sets* of key players, have been proposed in the literature. This chapter presents a brief survey of such methods. The methods described include a variety of techniques ranging from those based on traditional centrality measures using optimizing criteria to those based on measuring the efficiency of a network. Additionally, we describe and evaluate a new approach to discover sets of key players based on entropy measures. Finally, this chapter presents a brief description of some applications of information theory within social network analysis.

## 2.1 Introduction

*Social Network Analysis* (SNA) comprises the study of relations, ties, patterns of communication, and behavioral performance within social groups. In SNA, a social network is commonly modeled by a graph composed of nodes and edges. The nodes in the graph represent social actors and the links the relationship or ties between them. A graph consisting of $n$ nodes and $m$ edges is defined as $G = \{V, E\}$, where $V = \{v_1, v_2, \ldots, v_n\}$ is the set of nodes or vertex and $E = \{e_1, e_2, \ldots, e_m\}$ is a set of links or edges. In general, graphs where the edges do not have an associated direction are called *undirected graphs*. Graphs that contain no cycles are called *acyclic graphs*. For convenience, in the rest of this chapter, we will use the terms undirected acyclic graph, graph, and network as synonyms. Additionally, we will use indistinctly the term node, player, and actor.

One important issue in SNA is the determination of groups in complex social networks. Groups are disjoint collections of individuals who are linked to each other

D. Ortiz-Arroyo (✉)
Department of Electronic Systems, Esbjerg Institute of Technology, Aalborg University, Denmark
e-mail: do@aaue.dk

A. Abraham et al. (eds.), *Computational Social Network Analysis*, Computer Communications and Networks, DOI 10.1007/978-1-84882-229-0_2,

by some sort of relation or interaction. Within a group, members have different positions. Some of them occupy central positions, others remain in the periphery, and the rest lies somewhere in between. A group may have one or more key players. While this definition of a group is intuitive, a more mathematical description of a group is required to enable us analyzing systematically social networks. One possible definition of a social group is based on the concept of a *clique*. A clique of a graph $G$ is defined as a subgraph $H$ of $G$ in which every vertex is connected to every other vertex in $H$. A clique $H$ is called *maximal* if it is not contained in another subgraph of $G$. While this definition of a clique may be useful to study small social networks,[1] other more complex organizations have been analyzed using *semi-lattices* and a more recent extension of these mathematical structures called *Galois lattices* [2, 3].

Numerous studies in SNA have proposed a diversity of measures to study the communication patterns and the structure of a social network. One of the most studied measures is *centrality*. Centrality describes an actor's relative position within the context of his or her social network [4]. Centrality measures have been applied in a diversity of research works, for instance, to investigate influence patters in interorganizational networks, to study the power or competence in organizations, analyzing the structure of terrorist and criminal networks, analyzing employment opportunities, and many other fields [5].

The ability that centrality measures have to determine the relative position of a node within a network has been used in previous research work to discover *key players* [6–8] in social networks. Key players are these nodes in the network that are considered "important" with regard to some criteria. In general, the importance of a node is measured in a variety of ways depending on the application. In this chapter, we will define important nodes as those nodes that have a major impact on the cohesion and communication patterns that occur in the network.

One possibility for measuring the importance of a node given the previous criteria is to calculate how many links a node has with the rest of the network's nodes, this is called *degree centrality*. Nodes with high degree centrality have higher probability of receiving and transmitting whatever information flows in the network. For this reason, high degree centrality nodes are considered to have influence over a larger number of nodes and/or are capable of communicating quickly with the nodes in their neighborhood. Degree centrality is a *local* measure [9], as only the connections of a node with its neighbors are taken into account to evaluate node's importance.

Other centrality measures evaluate the degree with which a player controls the flow of information in the network. Messages sent through the network frequently pass through these players; they function as "brokers". A measure that models this property is called *betweenness*.

Another closely related method that has been used to evaluate the importance of a node within a network is based on measuring how close a node is located with respect to every other node in the network. The measure is called *closeness*. Nodes

[1] The use of cliques to model social groups has been criticized by some authors (e.g. [1, 2]) due to the strict mathematical definition of cliques.

with low closeness are able to reach (or be reached by) most or all other nodes in the network through geodesic paths.

Some other proposed centrality measures try to evaluate a player's degree of "popularity" within the network, i.e., they represent centers of large *cliques* in the graph. A node with more connections to higher scoring nodes is considered as being more important. The measure that captures this intuition is called *eigenvector centrality*.

Contrarily to a local measure such as degree centrality, metrics like betweenness, closeness, or eigenvector centrality are considered *global* measures [9] since they evaluate the impact that a node has on the global structure or transmission of information within the network.

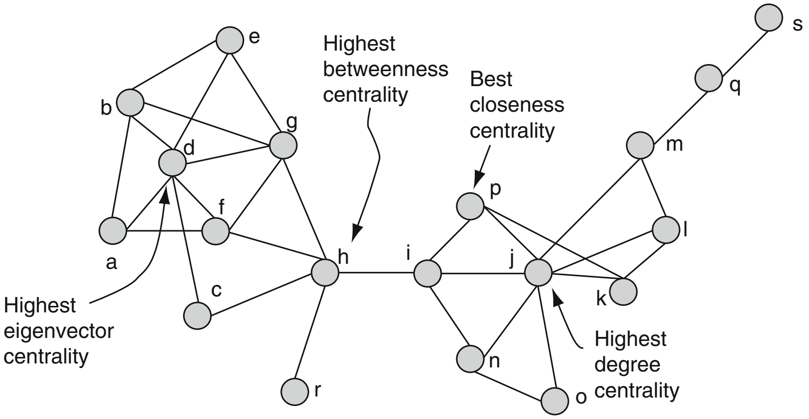
Degree centrality, betweenness, closeness, and eigenvector centrality are among the most popular measures used in SNA. However, over the years other measures have been proposed in the literature to overcome some of their limitations. Among these measures we can mention *information centrality*, *flow betweenness*, the *rush index*, and the *influence* [10], among others.

In spite of the relative simplicity of the centrality measures we have described, recent research has found that such metrics are robust in the presence of noise. Noise in this case refers to the possibility of including or excluding some nodes and links from a network during its construction due to the use of imprecise or incorrect information. In [11] Borgatti and Carley studied the performance of centrality measures under the conditions of imperfect data. Firstly, they generated random graphs with different densities. Afterward, it was measured the effect that the addition or removal of nodes and edges had on the accuracy of each of the centrality measures employed in the experiments. Borgatti et al. found out that, as expected, the accuracy of centrality measures decreases with an increasing error rate, but surprisingly, it does it in a predictable and monotonic way. This result means in principle that if one were able to estimate the percentage of errors made when a network is built, it could also be possible to estimate bounds on the accuracy of the results obtained by applying centrality measures. The other interesting finding reported in [11] was that all centrality measures performed with a similar degree of robustness. However, it must be remarked that the results of this study apply only to random graphs.

Centrality measures make certain assumptions about the way the information flows in the network. Hence, as described in [10], the type of information flow assumed in the network determines which measure may be more appropriate to be applied in a specific problem. Figure 2.1[2] illustrates some nodes within a network that have different centrality values. This picture clearly illustrates that the type of flow that occurs within a network for an specific application domain must be determined before a centrality measure could be used correctly.

The literature on centrality measures is rather extensive; see for example [4,6,7], and [10]. However, very few methods have been proposed to find *sets of key players* capable of optimizing some performance criterion such as maximally disrupting the network or diffusing efficiently a message on the network.

[2] A similar figure is used in [12].



e
b
g
d
f
a
h
c
r
i
j
p
k
l
m
q
s
n
o
Highest betweenness centrality
Best closeness centrality
Highest eigenvector centrality
Highest degree centrality

**Fig. 2.1** Diverse centrality measures applied on an example network

Methods for discovering a set of key players in a social network have numerous applications. For instance, these methods may help intelligence agencies to disrupt criminal organizations or allocate human resources in a more effective way within formal organizations.

The problem of finding an individual key player is fundamentally different from that of finding a set of $k$-players. More specifically, the problem of getting an optimal set of $k$-players is different from the problem of selecting $k$ individuals that are each, individually optimal [12]. For this reason, applying naively centrality measures to find a set of key players will likely fail. A simple example that illustrates why this may happen is the case of a network with a few central nodes that are redundant. Eliminating the redundant nodes will have no effect on the network even if they have high centrality degree. Additionally, it is also possible to find nodes that in spite of not having a high centrality degree have in fact a greater impact in disrupting the network structure when removed. For instance, Fig. 2.1 illustrates that nodes $h$ and $i$ are redundant as the removal of any of them will fragment the network into two or three components. However, as is explained in Sect. 2.3, in this specific example node $h$ is more important than node $i$.

To simplify analysis, social networks are commonly considered *static* structures. However, most social interactions in reality do not remain static but rather evolve through time. *Dynamic network analysis* is an active area of research [13] that studies models of the evolution of social relations through time. Some of the methods employed to analyze dynamic networks comprise statistical process control and Markov chains among other techniques. Due to lack of space, in this chapter we will only focus on static networks.

This chapter presents a brief survey of methods that have been proposed in the literature recently to discover sets of key players in social networks. Additionally, a new method, based on Shannon's definition of entropy is introduced. To asses the performance of this method we have designed a simulation environment specially built for the purpose. The simulation environment allowed us to perform a comparative evaluation of the results obtained by entropy-based methods with those reported in the literature using other methods. Our preliminary results indicate that the entropy-based methods can be used effectively to identify sets of key players for certain type of networks.

The rest of this chapter is organized as follows. Section 2.2 presents a summary of related work on the use of information theory in SNA. Section 2.3 briefly describes some of the methods that can be used to discover sets of key players. Section 2.4 describes the proposed method based on entropy measures together with an evaluation of its preliminary performance results. Finally, Sect. 2.5 describes some possible research directions and provides some conclusions.

## 2.2 Information Theory in SNA

Information theory deals with the transmission, storage, and quantification of information. Concepts originally introduced in information theory have been successfully applied in a wide range of fields, ranging from digital communication systems, cryptography and machine learning to natural language processing, neurobiology and knowledge discovery in unstructured data.

One of the fundamental concepts employed in information theory is *entropy*. Entropy was originally proposed by Claude Shannon [14] as a measure to quantify the amount of information that can be transmitted through a noisy communication channel. In a complementary way, entropy is used to quantify the degree of uncertainty in the content of a message or in general the uncertainty within a system. Shannon's definition of entropy of a random variable $X$ that can take $n$ values is presented in Eq. 2.1.

$$H(X) = -\sum_{i=1}^{n} p(x_i) \times \log_2 p(x_i) \tag{2.1}$$

Given its wide applicability, concepts borrowed from information theory have been recently applied in SNA. For instance, in [15] a method capable of measuring centrality on networks that are characterized by *path-transfer flow* is described. In social networks characterized by path-transfer flow, information is passed from one node to other following a path. However, contrary to other patterns of communication, information is contained within a single node at a time, i.e., there is no parallel transfer of information. An example of this type of information flow appears in chain letters where each recipient add its name to the end of the letter and then sends it to other person within the network. Other examples include trading and smuggling networks.

The method introduced in [15] to determine the centrality of nodes in networks characterized by path-transfer flow basically consists in calculating the probability that the flow originated in a node stops at every other node in the network. The basic idea is to model the fact that highly central nodes may be identified by measuring how similar probabilities are that the flow originating in a node will stop at every other node within the network. In a highly central node, such as the one located in the center of a star graph, the probability that the flow starting in the central node ends in any other node in the network is exactly the same. Contrarily, the flow that starts in a node of a graph that is less central will have a more uneven distribution of probabilities. The definition of Shannon's entropy perfectly captures these two intuitions. Entropy is defined in terms of the downstream degree of a vertex, which is the number of eligible vertices to which the transfer can be next made. Then the transfer probability is defined as the inverse of the downstream degree of a node. Using the definition of transfer and stop probabilities in the calculation of Shannon's entropy and then normalizing it, finally provides the centrality measure for a vertex, as is described in [15].

In [16], Shetty and Adibi combine the use of cross-entropy and text-mining techniques to discover important nodes on the Enron corpora of e-mails. The corpora of e-mails is analyzed to create a social network representing the communication patterns among individuals in the company. The email messages in the Enron corpora were analyzed to determine their similarity regarding its contents. The degree of similarity in message content was used as an indication that the people sending these messages were talking about similar topic. Sequences of similar topic e-mails up of length two involving three actors $A, B, C$ sent for instance in the order $A_{sent}B_{sent}C$ were counted. Afterward, a method based on the calculation of cross-entropy for such sequences of messages was used to rank the importance of a node. Nodes that produced the highest impact in reducing the total cross-entropy when removed from the network were selected as the most important ones. The method proposed by Shetty and Adibi was designed specifically to discover the set of key players within the Enron scandal case. Their results show that the method was capable of finding some key players in the Enron company. However, these players were not necessarily participating in the Enron scandal.

The next section discusses other methods that can be used to discover sets of key players in other social networks.

## 2.3 Methods for Discovering Sets of Key Players

One naive approach that can be used to discover sets of key players is to measure the centrality of every single node in the network. Afterward, nodes are ranked according to their importance as measured by value of the specific centrality measure used. Finally, a subset of size $k$ of these nodes could be selected as the key players.

Another more interesting approach to find key players is described in [17]. This approach is based on measuring the communication efficiency of a network. The efficiency $E$ of a network $G$ was defined in Eq. 2.2:

$$E(G) = \frac{\sum_{i \neq j \in G} \varepsilon_{ij}}{N(N-1)} = \frac{1}{N(N-1)} \sum_{i \neq j \in G} \frac{1}{d_{ij}} \tag{2.2}$$

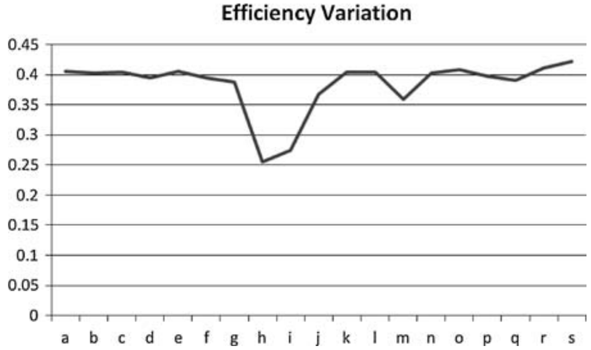


**Fig. 2.2** Efficiency variation of a graph taken from Borgatti's examples in [12]

where $N$ is the number of nodes in graph $G$ and $\varepsilon_{ij}$ is the communication efficiency, which is proportional to the inverse of $d_{ij}$ (the shortest path length between two nodes $i, j$). The equation calculates all shortest paths between all pairs of nodes normalized by the number of all possible paths that will be contained in a fully connected graph consisting of $N$ nodes. The method essentially consists in removing nodes one by one, recalculating the drop in network efficiency every time. These nodes that produce the largest impact in reducing the overall efficiency of a network are selected as the key players. The advantage of this method is that it can be easily implemented. Figure 2.2 shows the result of calculating graph efficiency using Eq. 2.2 for the example graph shown in Fig. 2.1.

Figure 2.2 shows that the method based on calculating graph efficiency will detect nodes $h$, $i$, and $m$ as being the key players, using an appropriate threshold value. However, the method fails at detecting that nodes $h$ and $i$ are in fact redundant.

The problem of previous two approaches is that they measure the effect that each single node has on the network independently. Hence, as previous example shows they will likely fail at identifying redundant nodes.

Another heuristic approach briefly sketched in [12] consists in selecting the top individual player using whatever centrality measure is appropriated for the task. Then, the nodes that are least redundant are added to the set of key players. The challenge of this approach will be to find an efficient procedure to determine which nodes are the least redundant.

The concept of centrality has been applied not only to single individuals within a network but also to groups of individuals. In [18], measures for degree centrality, closeness, and betweenness are defined for a group. Using these measures, groups having high centrality will be the key players. It must be remarked that group centrality can be used not only to measure how "central" or important a group is, but also in constructing groups with maximum centrality within an organization. For instance, a team of experts can be distributed within an organization in such a way that it has high group centrality. The idea is that this group will provide readily access to the expertise needed by other members of an organization.

In [19] a recent approach to discover a group of key players is presented. The method is based on the concept of *optimal inter-centrality*. Inter-centrality measure takes into account a player's own centrality and its contribution to the centrality of others. The individual optimal inter-centrality measure is then generalized to groups of players. The group with the highest inter-centrality measure is the key group.

Another approach to discover sets of key players, proposed by Borgatti in [12], consists in selecting simultaneously $k$ players via combinatorial optimization. In that work, Borgatti defines two problems related to discovering sets of key players as follows.

The *Key Player Problem Positive (KPP-Pos)* consists of identifying these $k$-players that could be used as seeds in diffusing optimally some information on the network.

The *Key Player Problem Negative (KPP-Neg)* goal consists of identifying those $k$-players that, if removed, will disrupt or fragment the network. A more formal definition of the two problems taken from [12] is

*"Given a social network(represented as an undirected graph), find a set of k nodes (called a kp-set of order k) such that,*

1. *(KPP-Neg) Removing the kp-set would result in a residual network with the least possible cohesion.*
2. *(KPP-Pos) The kp-set is maximally connected to all other nodes."*

Borgatti found that off-the-shelf centrality measures are not appropriate for the task of discovering sets of key players as defined by KPP-Pos and KPP-Neg problems. Hence, he proposes a new method based on combinatorial optimization and greedy heuristics. Additionally, to evaluate the solution to both KPP-Neg and KPP-Pos problems, Borgatti proposes new metrics to measure how successfully both problems are solved. One metric is called the *degree of reachability* described by Eq. 2.3:

$$D_F = 1 - 2\frac{\sum_{i>j} \frac{1}{d_{ij}}}{N(N-1)} \tag{2.3}$$

where $d_{ij}$ is the distance between nodes $i, j$, and $N$ the total number of nodes in the graph. The metric $D_F$ captures the fragmentation and relative cohesion of the components in the network.

The other metric proposed by Borgatti is the *weighted proportion* of nodes reached by the set of key players defined in Eq. 2.4:

$$D_R = \frac{\sum_j \frac{1}{d_{Kj}}}{N} \tag{2.4}$$

where $d_{Kj}$ is the distance from any member of the key player set to a node $j$ not in the set. This metric evaluates the degree with which the set of key players is isolated from the rest of the nodes.

The greedy heuristic presented in [12] seeks to select those nodes in the graph that maximize $D_F$ and $D_R$ metrics. The algorithm taken from [12] is presented as Algorithm 2.1.

**Algorithm 2.1** (taken from [12])

```
1: Select $k$ nodes at random to populate set $S$
2: Set F = fit using appropriate key player metric
3: for all nodes $u$ in $S$ and each node $v$ not in $S$ do
4:     DELTAF = improvement in fit if $u$ and $v$ were swapped
5: end for
6: Select pair with largest DELTAF
7: a. If DELTAF $\leq$ then terminate
8: b. Else, swap pair with greatest improvement in fit and set F = F + DELTAF
9: Go to step 3
```

Borgatti applied the proposed approach to two data sets, one terrorist network and a network of members of a global consulting company with advice-seeking ties. The results obtained by Borgatti show that the combinatorial optimization together with the use of the success metrics perform well on the two problems considered.

## 2.4 Discovering Sets of Key Players Using Entropy Measures

A new method aimed at finding sets of key players based on entropy measures that provide a simple solution to both the KPP-Pos and KPP-Neg problems will be introduced in this section.

The method based on entropy measures has some similarities with the method described in [16, 17]. However, contrarily to the approach described in [16], this method relies only on the structural properties of the network, uses Shannon's definition of entropy instead of cross-entropy. Additionally, the method described in [16] was specifically designed to detect important nodes on the Enron corpus, whereas the entropy-based method can be applied in many other problems.

The entropy-based method shares also shares some similarity with the one described in [17]. The main difference lies in the type of measure used which is Shannon's entropy instead of efficiency as defined in Eq. 2.2. Additionally, the entropy-based method is aimed at providing simple alternative solutions to both KPP-Pos and KPP-Neg problems. However, it must be remarked the entropy-based method does not aim at solving both problems optimally as was done in [12], but to provide an alternative simple solution that could be used to tackle both problems.

We first define the connectivity of a node $v_i \in V$ in a graph as:

$$\chi(v) = \frac{\deg(v_i)}{2N}, \quad N > 0 \tag{2.5}$$

where $\deg(v_i)$ is the number of incident edges to node $v_i$ and $N$ the total number of edges in the graph. We can use $\chi$ as the stationary probability distribution of random walkers in the graph [20]. This is called the *connectivity probability distribution* of the graph.

Another probability distribution can be defined in terms of the number of shortest or geodesic paths that have $v_i$ as source and the rest of nodes in the graph as targets:

$$\gamma(v) = \frac{spaths(v_i)}{spaths(v_1, v_2, \ldots, v_M)}, \quad spaths(v_1, v_2, \ldots, v_M) > 0 \tag{2.6}$$

where $spaths(v_i)$ is the number of shortest paths from node $v_i$ to all the other nodes in the graph and $spaths(v_1, v_2, \ldots, v_M)$ is the total number of shortest paths $M$ that exists across all the nodes in the graph. This is called the *centrality probability distribution* of the graph.

Using Eqs. 2.5 and 2.6 to define our probability distributions, we can obtain different entropy measures by applying the definition of entropy in Eq. 2.4. This procedure allows us to define *connectivity entropy* $H_{co}$ and *centrality entropy* measures $H_{ce}$ of a graph $G$ in the following way:

$$H_{co}(G) = -\sum_{i=1}^{n} \chi(v_i) \times \log_2 \chi(v_i) \tag{2.7}$$

$$H_{ce}(G) = -\sum_{i=1}^{n} \gamma(v_i) \times \log_2 \gamma(v_i) \tag{2.8}$$

It must be noticed that Eqs. 2.7 and 2.8 should be normalized to enable us to compare the centrality or connectivity entropies obtained from different types of networks. However, this is not done here since we will not be comparing different networks.

The connectivity entropy measure provides information about the connectivity degree of a node in the graph. In a fully connected graph, the removal of a node will decrease the total entropy of the graph in the same proportion as when any other node is removed. All nodes will have the same effect on the graph entropy leaving it still densely connected after a node is removed. However, in a graph with lower density, the removal of nodes with many incident edges will have a larger impact in decreasing the total connectivity entropy of the system, compared to the case when a node with a smaller connectivity degree is removed. This effect is illustrated in Figs. 2.3 and 2.4.

Centrality entropy provides information on the degree of reachability for a node in the graph. In a fully connected graph the removal of any node will have the same effect on centrality entropy as when any other node is removed. All nodes are equally important for the flow of information. This effect is illustrated in Fig. 2.4. Contrarily, in partially connected graphs, those nodes whose removal will split the graph in two or more parts or that will reduce substantially the number of geodesic paths available to reach other nodes when removed, will have a higher impact in decreasing the total centrality entropy. This effect is illustrated in Figs. 2.5 and 2.6 where the removal of node $v_5$ causes the disconnection of node $v_6$, and this event produces the largest change in centrality entropy for the graph.

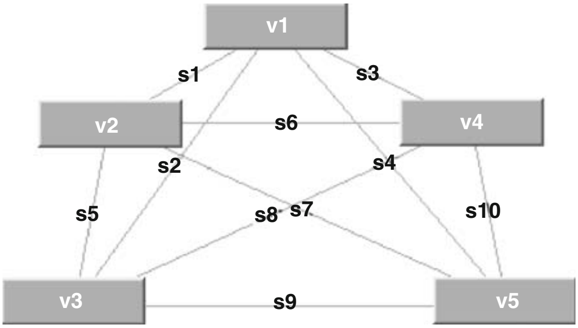


**Fig. 2.3** Fully connected graph

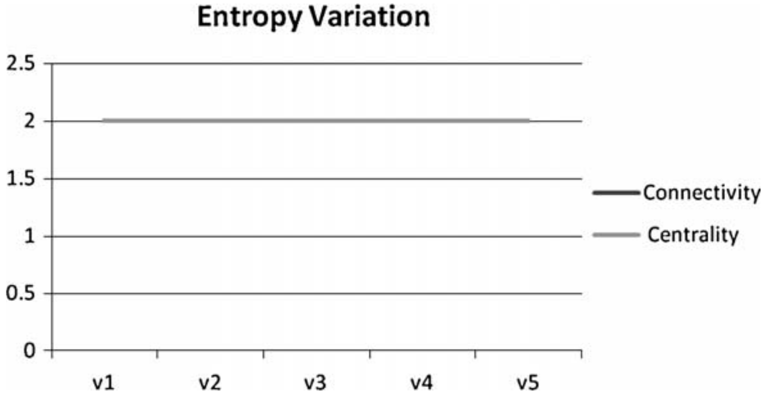


**Fig. 2.4** Entropy variation of a fully connected graph

Note that Figs. 2.4 and 2.6 also show that there is either perfect or very high correlation between the connectivity and centrality entropy measures when applied to the fully connected and partially-connected graph examples, respectively. This happens due to the fact that these graphs are very symmetric. Homogeneity is the strongest form of symmetry that a graph can posses. Therefore, the correlation among these two measures will decrease as the network becomes more and more heterogeneous. This fact will be illustrated in the following example graphs.

In general, centrality and connectivity entropies provide an average measure of network *heterogeneity* since they measure either the diversity of paths to reach the nodes within the graph or the diversity of link distribution in the graph, respectively. Heterogeneity in complex networks is identified by looking at the degree distribution $P_k$, which is the probability of a node having $k$ links [21]. The method introduced in this section additionally to degree distribution adds path distribution, which is the probability $P_l$ that a node is being reached by other nodes through $l$ different geodesic paths.

The entropy-based method introduced in this chapter is presented in Algorithm 2.2. In summary, the algorithm attempts to solve KPP-Pos and KPP-Neg

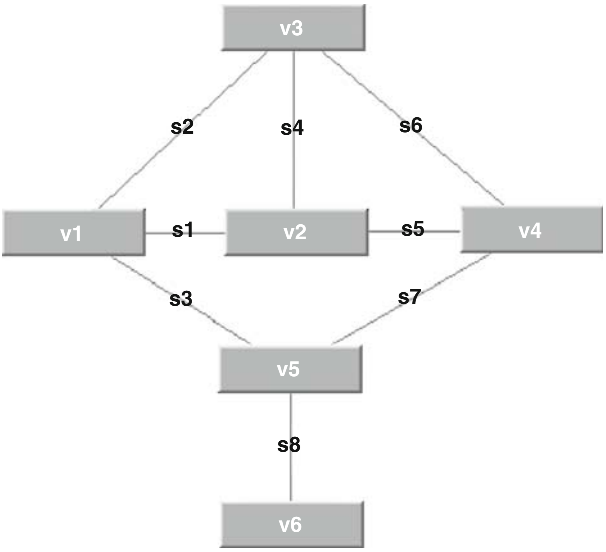


**Fig. 2.5** Partially connected graph

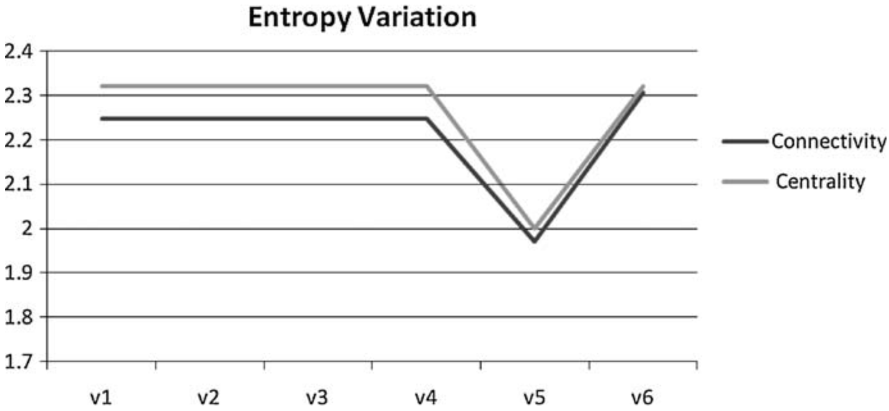


**Fig. 2.6** Entropy variation of a partially connected graph

problems using connectivity entropy and centrality entropy. The basic idea is to find those nodes that produce the largest change in connectivity or centrality entropy when removed from the graph. These nodes should be included in the set of key players as they have the largest impact in the structure (information content) of the network. The value of $\delta_i$, allows us to control how many players should be included in the set.

Since centrality entropy is based on the calculation of all the unweighted shortest paths in the network, it has the highest effect in the complexity of Algorithm 2.2. The complexity of Dijkstra's shortest path algorithm (from a single source node to all others) is $O(n^2)$.[3] However, given that Algorithm 2.2 needs to calculate all the shortest paths from every single node in the graph its overall complexity is $O(n^3)$.

**Algorithm 2.2** Entropy-based method

```
Calculate initial total entropy $H_{co_0}(G)$ and $H_{ce_0}(G)$
for all $nodes \in$ graph $G$ do
    Remove node $v_i$, creating a modified graph $G'$
    Recalculate $H_{co_i}(G')$ and $H_{ce_i}(G')$, store these results
    Restore original graph $G$
end for
To solve the KPP-Pos problem select those nodes that produce the largest change in graph
    entropy $H_{co_0}$-$H_{co_i} \geq \delta_1$
To solve the KPP-Neg problem select those nodes that produce the largest change in graph
    entropy $H_{ce_0}$-$H_{ce_i} \geq \delta_2$
```

Figure 2.8 shows the results of applying Algorithm 2.2 to the graph in Fig. 2.7. The graph is provided as an example by Borgatti in [12]. Our results show that centrality entropy is capable of detecting redundant nodes such as $h$ and $i$. Node $i$ is redundant as its removal will not have any impact on the number of partitions

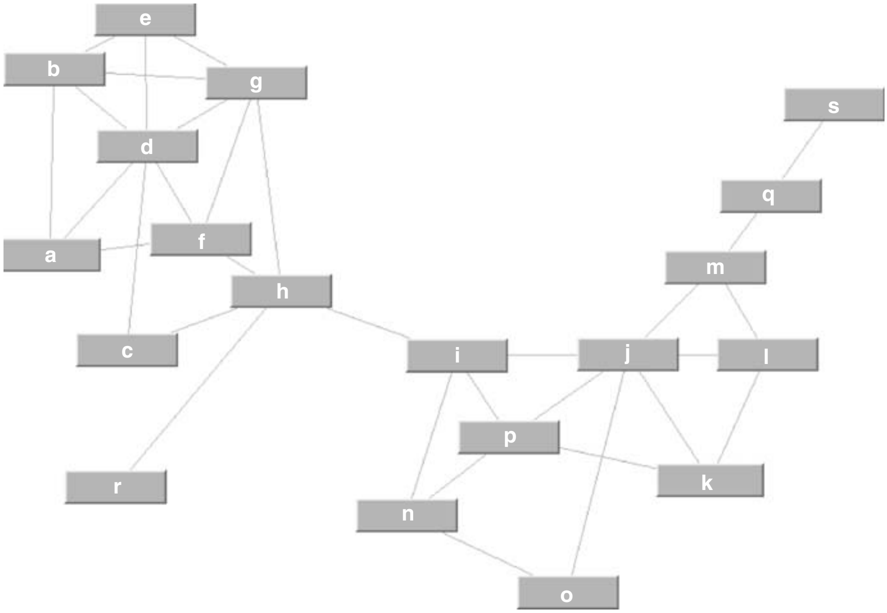


**Fig. 2.7** Graph taken from Borgatti's examples in [12]

[3] The complexity is calculated assuming that an adjacency matrix is used to represent the graph, other implementations using other more efficient data structure representations perform better.

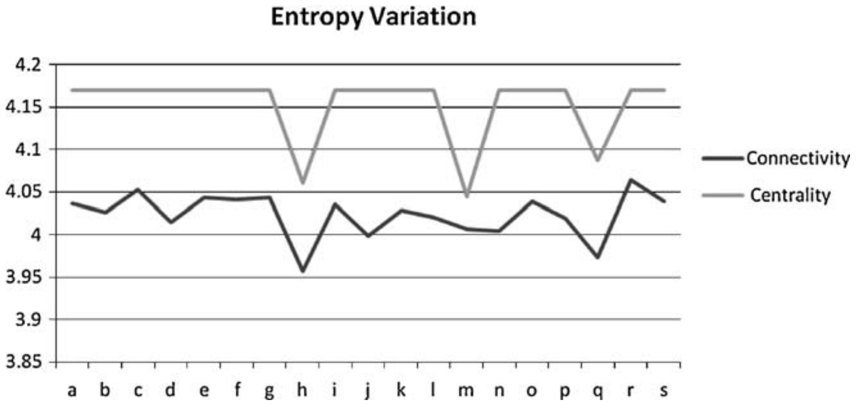


**Fig. 2.8** Entropy variation of a graph taken from Borgatti's examples in [12]

created, once $h$ has been removed. This happens in spite of $i$ having a high centrality value. The reason this occurs is that when node $h$ is disconnected it leaves node $r$ isolated from the rest of the graph, fragmenting the network into three components. The paths that go from $r$ to the rest of the nodes contribute significantly to the overall centrality entropy of the graph. Contrarily, when node $i$ is removed, the graph will be fragmented into two components. However, as node $r$ will remain connected it will still be able to communicate with the subnetwork to which it is attached, contributing with these paths to the total entropy calculation. In this simple example, the algorithm will find the set of key players consisting of $\{h, m, q\}$. By adjusting the value of $\delta_i$ we can control how many nodes we will include in the final set of key players.

It must be noted that in a graph similar to the one in Fig. 2.7, but where node $r$ is eliminated, our algorithm will still be able to determine that node $h$ is more important than node $i$. This is due to the fact that there are more nodes in that part of the graph where node $i$ is the "gatekeeper" and therefore more paths leading to that subnetwork. Figure 2.9 shows the result of applying the entropy-based algorithm to a graph similar to the one in Fig. 2.7 but not containing node $r$. The set of key players in this case will still be $\{h, m, q\}$ as these are the nodes that produce the largest change in centrality entropy.

Figure 2.9 also shows that node $h$ has the largest impact on connectivity entropy when removed from the graph. Interestingly, the same graph also shows that node $q$ has more effect on connectivity entropy, when compared to node $m$. The reason is that removing $m$ leaves still a connected graph composed of nodes $q$ and $s$, which contributes to the total entropy. Contrarily, removing $q$ leaves the single node $s$ isolated.

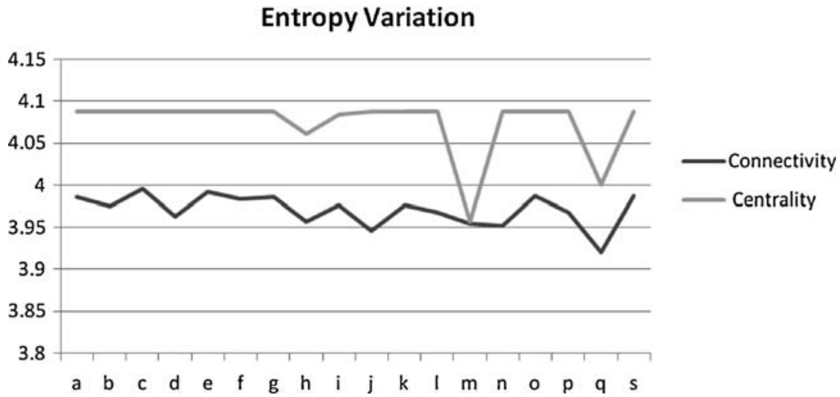


**Fig. 2.9** Entropy variation of modified graph taken from Borgatti's examples in [12]

## 2.4.1 Applying Entropy Measures to More Complex Social Networks

Figure 2.11 shows the results of applying Algorithm 2.2 using centrality and connectivity entropy to the terrorist graph in Fig. 2.10. The graph is a simplification of the graph provided by Krebs in [7]. Figure 2.11 shows that centrality entropy identifies a set of key players consisting of $\{atta, nalhazmi, darkazalni\}$, since these are the nodes that produce the biggest changes in entropy when removed, with *atta* producing the largest change. It must be noticed that nodes *nalhazmi* and *darkazanli* have the same effect on centrality entropy. This is because if we look at Fig. 2.10 we can see that both nodes will disconnect a single node if removed. However, removing *nalhazmi* will also cause a major impact in connectivity entropy, contrarily to the case when *darkazanli* is removed. This indicates that *nalhazmi* may be indeed more important than node *darkazanli*, even if both produce a similar effect on centrality entropy. This factor can also be used to grade the importance of a node in the graph.

Removing the set consisting of $\{atta, nalhazmi, darkazalni\}$ causes the network to be fragmented into five components. The optimization algorithm proposed by Borgatti produces a fragmentation of seven components.

Our Algorithm 2.2 finds that the set of nodes in Fig. 2.10 that solves KPP-Pos problem consists of $\{nalhazmi, halghamdi, salghamdi, atta\}$, as these are the nodes that will have the biggest impact on connectivity entropy when removed from the graph. The optimization algorithm proposed by Borgatti found that only three nodes are needed to reach 100% of the graph.

Previous results show that when entropy measures are applied to the terrorist network we can find similar results as those obtained by Borgatti. However, it must be remarked that the graph used by Borgatti in his experiments (derived from the one made available by Krebs in [7]) contains 63 nodes, whereas the network employed in our experiments (also derived from Krebs graph) contains only 34 nodes.

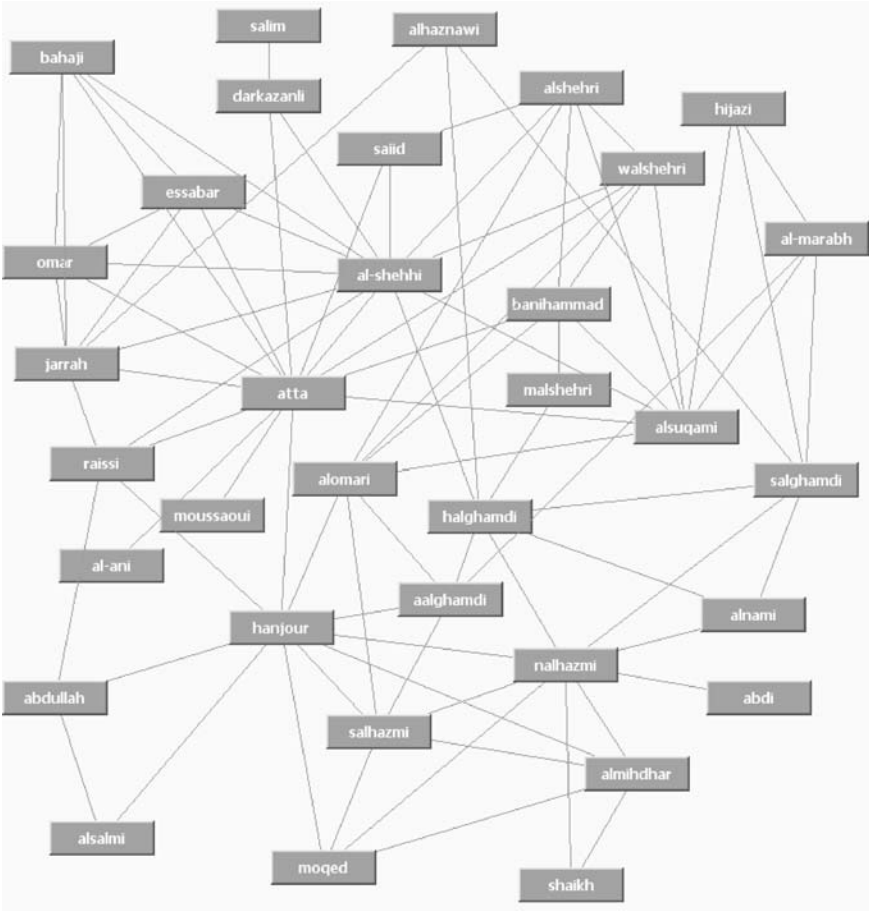


**Fig. 2.10** Terrorist network

Figure 2.12 shows the result of calculating the efficiency of the terrorist network in Fig. 2.10. The figure illustrates that the key players detected by the graph efficiency calculation are $\{atta, nalhazmi, darkazalni, hanjour\}$. The graph efficiency calculation finds *hanjour* as key player contrarily to centrality entropy measure. However this node does not cause a fragmentation in the network. Interestingly, it is connectivity entropy which also finds *hanjour* as key player since this node will cause a major disruption in the connectivity of the key players with the rest of the network.

In a different example of social network, Fig. 2.14 shows the result of applying centrality and connectivity entropy to the graph in Fig. 2.13. The graph describes the advise-seeking ties between members of a company and was obtained from [12].

Applying Algorithm 2.2 to this network, we found that the set of most important players for solving KPP-Neg consists of $\{HB, BM, WD, NP, SR\}$. In this same

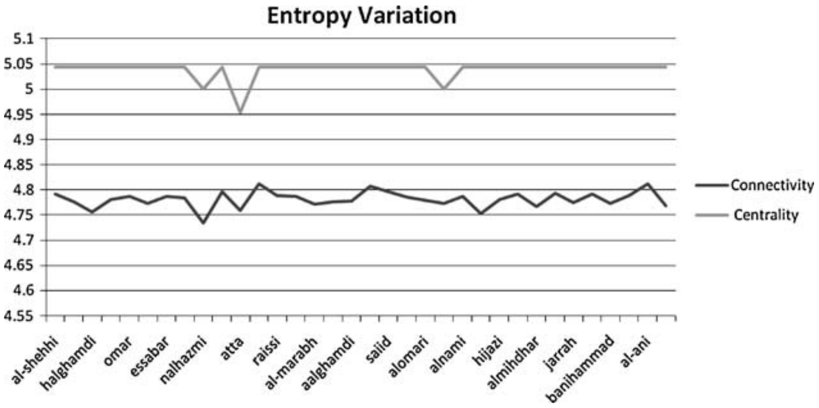


**Fig. 2.11** Entropy variation of terrorist network

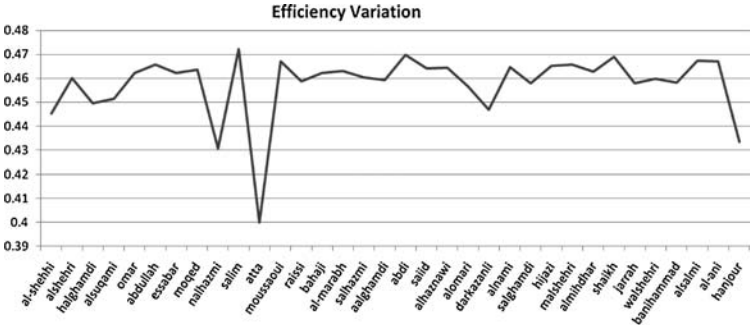


**Fig. 2.12** Efficiency variation of terrorist network

example, Borgatti obtained a set of key players consisting of $\{HB, BM, WD\}$ [12]. This is the set of players that if removed will divide the network into six components. Our algorithm finds the same elements additionally to *NP* and *SR*. However, it must be remarked that contrarily to [12], the centrality entropy-based algorithm does not try to optimize any specific metric.

In KPP-Pos problem, we are asked to find the smallest set of nodes that are well connected to the entire network. This set of players are the ones that if used as "seeds" will reach 100% of the network.

If we look only at the connectivity entropy chart in Fig. 2.14 we notice that Algorithm 2.2 will select nodes $\left\{BM, DI, HB, BW, CD, BS', NP, TO, BS\right\}$ as the key players when a set of size $k = 9$ is selected. These are the nodes that when removed will produce the largest changes in connectivity entropy. This list indicates that connectivity entropy allows us to get 89% of the key players found by Borgatti for a similar set size. However, if we add to the set, the 10th node that produces the next largest change in connectivity entropy, we will obtain a set consisting of $\left\{BM, DI, HB, BW, CD, BS', NP, TO, BS, PS\right\}$. This new set contains 100% of the nodes that Borgatti found as the key players in [12].

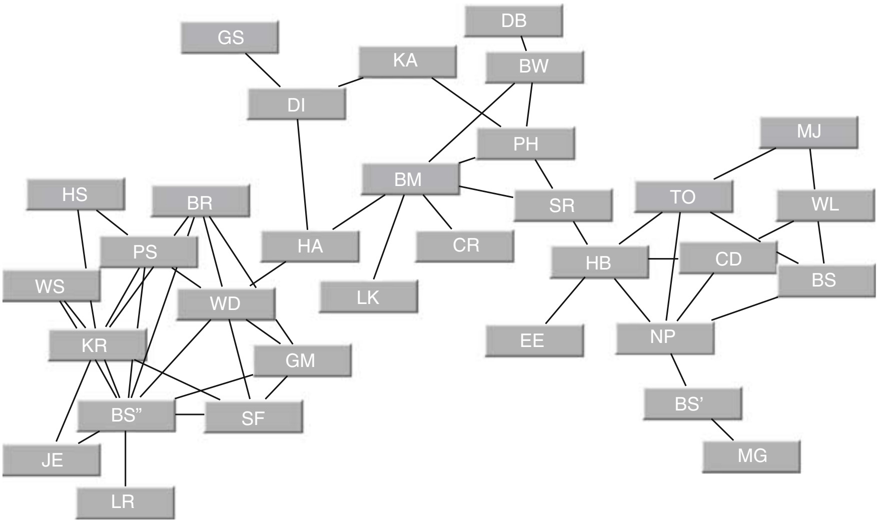


**Fig. 2.13** Company ties network

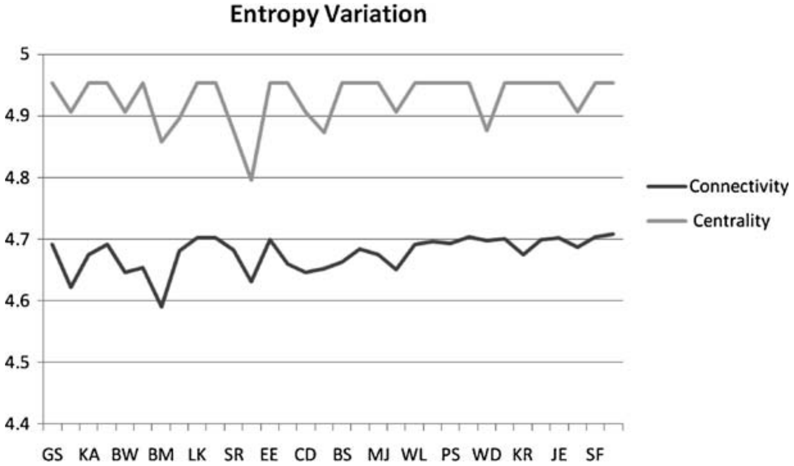


**Fig. 2.14** Entropy variation of company ties network

In this last example it must be noted that the graph used in these experiments is exactly the same that represents the "company ties" problem described in [12].

Finally, Fig. 2.15 shows the result of calculating the efficiency of the company ties network in Fig. 2.14. The figure illustrates that the key players discovered by the graph efficiency calculation are $\{HA, SR, HB, WD\}$. In this case the efficiency calculation finds two of the three key players that were also found by Borgatti's optimization method and our centrality entropy-based calculations.

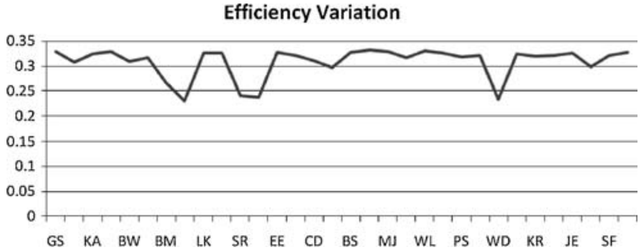


**Fig. 2.15** Efficiency variation of company ties network

It must be remarked that being connectivity and centrality entropies average measures of the heterogeneity of a graph, these measures will not be useful when applied to more homogeneous graphs. This fact is partially shown in Fig. 2.4 for the fully connected graph shown in Fig. 2.3. When a network obtained from Enron's e-mail corpora was constructed, it was found that the network was very homogeneous. Because of this, results showed that the entropy-based centrality measure had very little variations when nodes were removed from the graph.

## 2.5 Conclusions and Future Work

In this chapter we have described methods aimed at discovering sets of key players in social networks. A new method that finds the set of key players within a network using entropy measures was introduced. The method provides a simple solution to the KPP-Pos problem, selecting the set of nodes that produce the largest change in connectivity entropy when removed from a graph. Similarly, to solve KPP-Neg centrality entropy is used, measuring how the overall entropy changes when a node is removed from the graph. The main advantage of this method is its simplicity. We have shown the application of an entropy-based method in discovering sets of key players to two examples of social networks: a terrorist organization and a company. Our experimental results show that these methods are capable of obtaining comparable results with those described in [12], where combinatorial optimization algorithm and special performance metrics are used. However, one of the disadvantages of entropy-based methods is that these methods only work on non-dense heterogeneous networks.

We created a special simulation environment to asses the performance of the some of the methods presented. The simulation environment accepts as input the description of a graph in the standard XML-based file format for graphs called GraphML. The development process of the simulation environment was substantially reduced by using open source libraries. To create the mathematical models and representation of a graph we use the jGraphT library. JGraphT is an extension to jGraph, a popular graphic visualization library that has been optimized to handle several data models and algorithms. The algorithms provided by jGraphT allow us to traverse and analyze the properties of a graph. To show the simulation results we used jChart and jFreeChart. Finally, as jGraph does not provide a free graph layout algorithm we have implemented a variation of the well-known spring algorithm [22]. The whole simulation environment was designed using design patterns and was written in the Java language.

A possible extension to the study of entropy-based measures of centrality is to investigate their robustness, using a method similar to the one described in [11] on both random and real graphs. The entropy-based approach may also be extended with heuristics targeted at optimizing some specific metrics, similarly as it was done in [12]. Other measures borrowed from information theory such as mutual information may be used to provide insights into the dependencies between the nodes in the graph.

Finally, we plan to investigate techniques aimed at reducing the current overall complexity ($O(n^3)$) of the algorithms employed to find all the shortest paths within the network more efficiently. This is one of the weaknesses not only of the entropy-based measures described in this chapter but also of other similar methods that require to find all possible shortest paths between pairs of nodes within a network. In this regard we are exploring a simple approach that finds simultaneously all the shortest paths within the nodes in the graph on the multicore shared memory personal computers that are widely available today. The entropy-based algorithms will be implemented in the programming language Erlang, a functional language that provides parallel-processing capabilities based on the message passing model.

## References

1. Wasserman S, Faust K (1994) Social network analysis. Cambridge University Press, Cambridge.
2. Freeman LC (1996) Cliques, galois lattices, and the human structure of social groups. Social Networks 18(3):173–187
3. Falzon L (2000) Determining groups from the clique structure in large social networks. Social Networks 22(2):159–172
4. Friedkin NE (1991) Theoretical foundations for centrality measures. Am J Sociol 96(6):1478–1504
5. Borgatti SP, Everett MG (2006) A graph-theoretic framework for classifying centrality measures. Social Networks 28(4):466–484
6. Freeman LC (1977) A set of measures of centrality based on betweenness. Sociometry 40(1):35–41
7. Krebs V (2002) Uncloaking terrorist networks.First Monday 7(4). http://firstmonday.org/htbin/cgiwrap/bin/ojs/index.php/fm/article/view/941/863
8. Borgatti SP (2003) The key player problem. In: Breiger R, Carley K, Pattison P (eds) In dynamic social network modeling and analysis: workshop summary and papers. National Academy of Sciences Press, Washington, DC, pp 241–252
9. Scott J (2000) Social network analysis: a handbook. Sage, London
10. Borgatti SP (2004) Centrality and network flow. Social Networks 27(1):55–71
11. Borgatti SP, Carley K, Krackhardt D (2006) Robustness of centrality measures under conditions of imperfect data. Social Networks 28:124–1364
12. Borgatti SP (2006) Identifying sets of key players in a network. Comput Math Organ Theory 12(1):21–34

13. McCulloh IA, Carley KM (2008) Social network change detection. Technical report, Carnegie Mellon University
14. Shannon C (1948) A mathematical theory of communication. Bell Syst Tech J 17:379–423, 623–656
15. Tutzauer F (2006) Entropy as a measure of centrality in networks characterized by path-transfer flow. Social Networks 29(2):249–265
16. Shetty J, Adibi J (2005) Discovering important nodes through graph entropy the case of enron email database. In: LinkKDD '05: Proceedings of the 3rd international workshop on Link discovery. ACM, New York
17. Latora V, Marchiorib M (2003) How the science of complex networks can help developing strategies against terrorism. Chaos Soliton Fract 20(1):69–75
18. Everett MG, Borgatti SP (2005) Extending centrality. In: Carrington P, Scott J, Wasserman S (eds) Models and methods in social network analysis. Cambridge University Press 28:57–76
19. Ballester C, Calvo-Armengol A, Zenou Y (2005) Who's Who in Networks Wanted – The Key Player. CEPR Discussion Paper No. 5329. Centre for Economic Policy Research, London. Available at http://ssrn.com/abstract=560641
20. Doyle PG, Snell LT (1984) Random walks and electric networks. Mathematical Association of America, Washington, DC
21. Solé RV, Valverde S (2004) Information theory of complex networks: on evolution and architectural constraints. In: Lecture notes in physics, vol 650, pp 189–207. Springer, Berlin/Heidelberg
22. Kamada T, Kawai S (1989) An algorithm for drawing general undirected graphs. Inform Process Lett 31:7–15