Comput Math Organiz Theor (2006) 12: 21–34
DOI 10.1007/s10588-006-7084-x

Identifying sets of key players in a social network

Stephen P. Borgatti

Abstract A procedure is described for finding sets of key players in a social network. A key assumption is that the optimal selection of key players depends on what they are needed for. Accordingly, two generic goals are articulated, called KPP-POS and KPP-NEG. KPP-POS is defined as the identification of key players for the purpose of optimally diffusing something through the network by using the key players as seeds. KPP-NEG is defined as the identification of key players for the purpose of disrupting or fragmenting the network by removing the key nodes. It is found that off-the-shelf centrality measures are not optimal for solving either generic problem, and therefore new measures are presented.

Keywords Social networks · Centrality · Cohesion

1. Introduction

The problem of identifying key players in a social network is, at first glance at least, an old one. One stream of relevant research is node centrality (e.g., Bonacich, 1972; Freeman, 1979), which attempts to quantify the structural importance of actors in a network. In addition, work on identifying cores and peripheries (e.g., Seidman, 1983; Borgatti and Everett, 1999; Everett and Borgatti, 1999a) is relevant, as is work on group-level centrality (Everett and Borgatti, 1999b). Structural measures of social capital (e.g., Coleman, 1990; Burt, 1992; Borgatti, Jones and Everett, 1998) also tend to identify key players, although the perspective is reversed in that with social capital research one asks what features of the network contribute to the individual, whereas with key player research we ask which individuals are important for the network.

However, in this paper I attempt to show that existing measures and algorithms do not optimally solve the key player problem as I define it, and that new approaches are needed. The approach I explore is based on measuring explicitly the contribution of a set of actors to

S. P. Borgatti (✉)
Department of Organization Studies, Boston College
E-mail: borgatts@bc.edu

the cohesion of a network. In addition, I identify two separate conceptions or functions of key players which reflect different analytical goals, and develop separate measures of suitability for each type of goal. In this sense, I follow the problem-specific approach to centrality advocated by Friedkin (1991).

2. Defining the problem

I argue that there are two fundamentally different aspects of the key player issue, reflecting different kinds of purposes to which key player measurements and identifications are put. Effectively, there are two separate key player problems.

The first key player problem is defined in terms of the extent to which the network depends on its key players to maintain its cohesiveness. I refer to this as the "Key Player Problem/Negative" (KPP-Neg) because we measure importance in the breach—the amount of reduction in cohesiveness of the network that would occur if the nodes were not present. KPP-Neg arises in the public health context whenever we need to select a subset of population members to immunize or quarantine in order to optimally contain an epidemic. In the military or criminal justice context the problem arises when we need to select a small number of players in a criminal network to neutralize (e.g., by arresting, exposing or discrediting) in order to maximally disrupt the network's ability to mount coordinated action.

The second key player problem is defined in terms of the extent to which key players are connected to and embedded in the network around them. I refer to this as "Key Player Problem/Positive" (KPP-Pos). A practical application in which KPP-Pos arises in a public health context is when a health agency needs to select a small set of population members to use as seeds for the diffusion of practices or attitudes that promote health, such as using bleach to clean needles in a population of drug addicts. Another application arises in an organizational context when management wants to implement a change initiative and needs to get a small set of informal leaders on board first, perhaps by running a weekend intervention with them. Finally, the problem arises in a military or criminal justice context when one needs to select an efficient set of actors to surveil, to turn (as into double-agents), or to feed misinformation to. In all these cases, we are looking for a set of network nodes that are optimally positioned to quickly diffuse information, attitudes, behaviors or goods and/or to quickly receive the same.

A formal definition, then, of the key player problems is as follows. 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.

Of course, these introductory definitions leave out what is meant precisely by "least possible cohesion" and "maximally connected". Part of the process of solving these problems is providing definitions of these concepts that lead to feasible solutions and useful outcomes. However, it can be said at the outset that KPP-Neg involves fragmenting a network into components, or, failing that, making path lengths between nodes so large as to be practically disconnected. In contrast, KPP-Pos involves finding nodes that can reach as many remaining nodes as possible via direct links or perhaps short paths. However, a goal of my approach is to construct generic solutions to each problem such that they can be used no matter how cohesion or connectedness is defined.

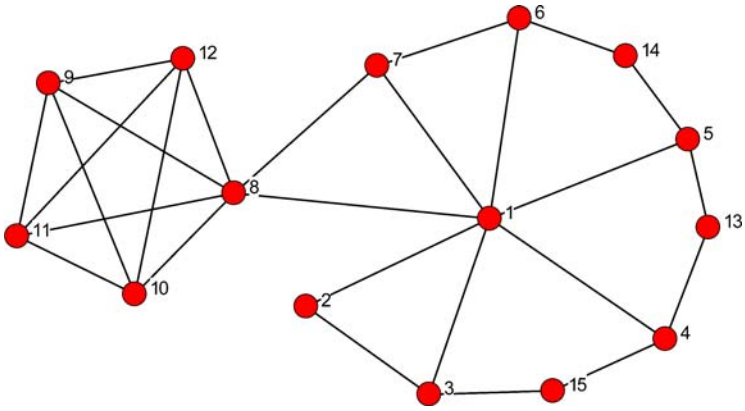

Fig. 1 Hypothetical network in which removing the most central node ("1") does not fragment the network

At first glance, both KPP-Neg and KPP-Pos would appear to be easily solved by selecting an appropriate measure of node centrality and selecting the *k* most central nodes to populate the kp-set. Alternatively, there are concepts in graph-theory that seem tailor-made for both KPP-Neg and KPP-Pos. However, it turns out that, for both KPP-Neg and KPP-Pos, these approaches fail for two separate reasons, which I label the goal issue and the ensemble issue. The goal issue refers to the fact that centrality measures were not designed with KPP-Neg or KPP-Pos specifically in mind, and hence are not necessarily optimal solutions. The ensemble issue refers to the fact that KPP-Neg and KPP-Pos are explicitly about selecting sets of nodes rather than individuals, and the optimal set for any task is not necessarily composed of the *k* most-optimal individuals when considered alone.

In the next two sections, I describe and address each issue.

3. The goal issue

For KPP-Neg, the most appropriate centrality measure would obviously be betweenness (Freeman, 1979). Freeman's betweenness measure sums the proportion of shortest paths from one node to another that pass through a given node. Thus, a node with high betweenness is along the shortest path between many pairs of nodes, and deleting that node should cause many pairs of nodes to become fully disconnected or at least more distantly connected.

However, the optimality of betweenness in identifying the node whose removal from the network would most reduce cohesion is not guaranteed under typical definitions of cohesion. Consider the network in Fig. 1. Node 1 has the highest centrality on all standard measures, including betweenness centrality. Yet deleting node 1 has no effect on disconnecting the network. Distances between nodes do increase somewhat, but if fragmentation is the goal, it is clear that removing node 1 is ineffective. In contrast, deleting node 8, which has lower centrality on all measures, does disconnect the graph. Removing node 8 splits the graph into two large fragments (i.e., components).

Whereas measures of centrality were not developed with problems like KPP-Pos and KPP-Neg in mind, a number of graph-theoretic concepts were. For example, much work has been done on the vulnerability of graphs to disconnection, which relates directly to KPP-Neg. In particular, there is the notion of a cutpoint, which is a node whose deletion would increase

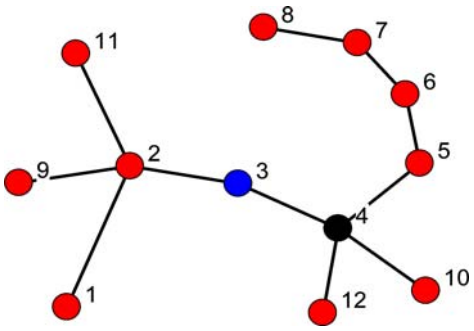

Fig. 2 Hypothetical network in which the most central node ("4") does not reach the most number of nodes in two steps or less

the number of components in the graph. However, there are three difficulties with using cutpoints in the KPP context. First, no account is taken of the size of components created by the removal of a cutpoint. If removing the cutpoint merely isolates a single node, leaving all other nodes connected, this is not seen as better than removing a cutpoint which divides the network into two equal sized components. Second, if the graph contains no cutpoints, the concept provides no way of choosing a node whose deletion would nearly disconnect the graph or which would make distances so long as to be practically disconnected. Third, cutpoints are a kind of discrete nominal classification rather than a measurement of the extent to which removing a node fragments a network.

Turning now to KPP-Pos, if we formulate the problem in terms of identifying a node that reaches the most nodes directly (i.e., paths of length 1), simple degree centrality is in fact optimal. However, if we formulate it in terms of reaching the most nodes in up to m steps, then the most appropriate standard centrality measure is closeness centrality, and this falls short. Closeness centrality is defined as the sum of graph-theoretic distances from a given node to all others in the network. For example, in the graph shown in Fig. 2, node 4 has the best closeness centrality (it is a total of 24 links away from all others). However, if we are interested in reaching the most nodes along paths of length 2 or less, node 3 would be a better choice since it can reach 8 nodes in addition to itself while node 4 can only reach 6 nodes.

4. The ensemble issue

The ensemble issue, which is discussed as the group centrality problem in Everett and Borgatti (1999b), refers to the fact that selecting a set of k nodes that, as an ensemble, optimally solves KPP-Pos or KPP-Neg, is quite different from selecting the k nodes that individually are optimal.

To start with, consider KPP-Neg. Figure 3 shows a graph in which nodes h and i are, individually, the best nodes to delete in order to fragment the network. Yet, deleting i in addition to h yields no more fragmentation, by any measure, than deleting i alone. In contrast, node m is not individually as effective as node i in separating pairs of nodes, but deleting m with h does produce more fragmentation than m or h alone. The reason that i and h are not as good together as h and m is that i and h are redundant with respect to their liaising role—they are equivalent in that they connect the same third parties to each other. Another way to look at it is that the centrality of one is due in part to the centrality of the other (i.e., their centralities are not independent), with the result being that the centrality of the ensemble combination is quite a bit less than the sum of the centralities of each.

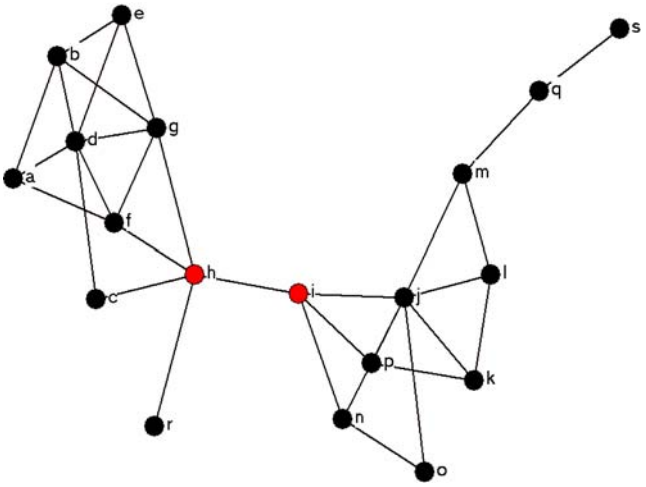

Fig. 3 Network in which removing the two most central nodes ("*i*" and "*j*") is not as disruptive as removing a different pair of nodes ("*i*" and "*j*")

Similarly, the graph-theoretic concept of a vertex cut-set generalizes the cutpoint to address the ensemble issue directly. A vertex cutset is a set of nodes whose removal would increase the number of components in the graph. Most graph-theoretic work has focused on minimum weight cutsets, which are smallest sets that have the cutset property. However, cutsets retain many of the same difficulties as cutpoints when applied to the KPP-Neg problem. First, we cannot specify the number of nodes in the set and then seek the set of that size that does the best job (rather, the measure of success is fixed and cutset methods are able to find a smallest set that achieves that level of success). In this sense, the graph-theoretic approach solves the inverse of the problem we seek to solve. Second, no account is taken of the size of components created by the cut. A cut that isolates a single node is no better than one that divides the graph into equal size components. Third, no account is taken of distances among nodes. Hence a set whose removal would not only cut the network in half but also lengthen distances within each half would not be considered better than one that merely cut the network in half.

A redundancy principle also applies to KPP-Pos. Consider the graph in Fig. 4. Nodes *a* and *b* are individually the best connected. Each is adjacent to five other nodes, more than any other by far. But together they reach no more than either does alone. In contrast, if *a* is paired with *c* (which individually reaches only three nodes), the ensemble reaches every node in the network. The reason that {a,c} is more effective than {a,b} is that *a* and *c* are less structurally equivalent (Lorraine and White, 1971; Burt, 1976) than are *a* and *b*. Structural equivalence refers to the extent to which two nodes have overlapping neighborhoods—i.e., are connected to the same third parties. Structurally equivalent nodes are, by definition, redundant with

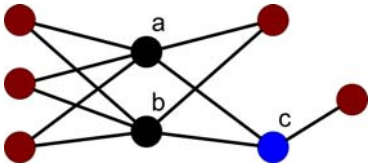

Fig. 4 Network in which the two most central nodes taken together ("*a*" and "*b*") are adjacent to fewer nodes than a different set of nodes ("*a*" and "*c*")

respect to adjacency and distance. Thus, the redundancy relevant to KPP-Pos is with respect to adjacency and distance, whereas the redundancy relevant to KPP-Neg is with respect to bridging (i.e., linking the same third parties).

For KPP-Pos, applicable graph-theoretic concepts include vertex covers and dominating sets. A vertex cover is a set of nodes whose members are incident upon every edge in the graph. A dominating set is a (typically minimal) set of nodes whose members are adjacent to all other nodes in the graph. For our purposes these are equivalent and fail for exactly the same reasons that cutsets fail for KPP-Neg. The focus of graph-theoretic research has been on finding the smallest cover or dominating set that achieves a fixed goal (reaching all nodes) perfectly. Our problem is the reverse: finding a set of fixed size that achieves the goal as well as possible. In addition, we would prefer to measure the extent to which a set reaches all nodes, so that we can evaluate our success.

5. Proposed solution

5.1. KPP-Neg: Fragmentation

The fundamental concept implicit in KPP-Neg is graph fragmentation. What is needed to solve the problem is a direct measure of graph fragmentation. With that we can then evaluate any candidate set of nodes in terms of how successful it is in solving KPP-Neg.

Perhaps the most obvious measure of network fragmentation is a count of the number of components. If the count is 1, there is no fragmentation. The maximum fragmentation occurs when every node is an isolate, creating as many components as nodes. For convenience, we normalize the count (labeled C in Eq. (2)) by dividing by the number of nodes (labeled n).

$$^{COMP}F = \frac{C}{n} \tag{2}$$

The problem with this measure is that it doesn't take into account the sizes of the components. For example, in Fig. 3, deleting node m would break the network into two components, but the vast majority of nodes remain happily connected. In contrast, deleting node i (or h) would also result in just two components, but more pairs of nodes would be separated from each other.

This suggests another measure of fragmentation that simply counts the number of pairs of nodes that are disconnected from each other. Given a matrix R in which $r_{ij} = 1$ if i can reach j and $r_{ij} = 0$ otherwise, we can define the new measure as shown in Eq. (3). This measure is the same as subtracting Krackhardt's (1994) measure of connectivity from unity.

$$F = 1 - \frac{2\sum_i \sum_{j<i} r_{ij}}{n(n-1)} \tag{3}$$

One problem with Eq. (3) is that it is relatively expensive to compute (at least in the optimization context that is introduced in a later section of this paper). However, since nodes within a component are mutually reachable, and since components of graph can be enumerated extremely efficiently, the F measure can be computed more economically by rewriting it in terms of the sizes (s_k) of each component (indexed by k):

$$F = 1 - \frac{\sum_k s_k(s_k-1)}{n(n-1)} \tag{4}$$

The F measure is remarkably similar to a diversity measure known variously as heterogeneity, the concentration ratio, the Hirschman-Herfindahl index, or the Herfindahl index. Applied to the current context, that measure is defined as follows:

$$H = 1 - \sum_k \left(\frac{s_k}{n}\right)^2 \tag{5}$$

One difference between F and H is that while both achieve minimum values of 0 when the network consists of a single component, the H measure can only achieve a maximum value of $1 - 1/n$ when the network is maximally fragmented (all isolates). Interestingly, if we try to normalize H by dividing by $1 - 1/n$, we obtain the F measure, as shown in Eq. (6).

$$H^* = \frac{1 - \sum_k \left(\frac{s_k}{n}\right)^2}{1 - n^{-1}} = 1 - \frac{\sum_k s_k(s_k - 1)}{n(n-1)} = F \tag{6}$$

An alternative approach is information entropy. Applied to this context, the measure is defined as

$$E = -\sum_k \frac{s_k}{n} \ln\left(\frac{s_k}{n}\right) \tag{7}$$

The measure is bounded from below at zero, but is unbounded from above. We can bound it by dividing it by its value when all nodes are isolates:

$$E = \frac{\sum_k \frac{s_k}{n} \ln\left(\frac{s_k}{n}\right)}{\sum_k \ln\left(\frac{s_k}{n}\right)} \tag{8}$$

While the fragmentation measure F and the entropy measure E are very satisfactory for what they do, they do not take into account the shape—the internal structure—of components. A network that is divided into two components of size 5 in which each component is a clique (Fig. 5(a)) is seen as equally fragmented as a network divided into two components of size 5 in which each component is a line (Fig. 5(b)). Yet distances and therefore transmission/transportation times are much higher in the latter network. As Granovetter (1973) noted, nodes don't have to be truly disconnected in order to be practically disconnected—if distances are long enough, the nodes are effectively separated.

In addition, there is another problem which is that in some cases the required size of the kp set is small enough that no set of that size disconnects the graph. Yet we would still like some way of evaluating which sets are better than others in terms of nearly disconnecting many pairs.

An obvious solution would be to measure the total distance between all pairs of nodes in the network, and take this as a measure of virtual disconnection. However, this only works in the case where the graph remains connected. Otherwise, we must sum infinite distances. A practical alternative is to base the measure on the sum of the reciprocals of distances, observing the convention that the reciprocal of infinity is zero. In that case we can create a version of F, based on Eq. (2), that weights by reciprocal distance. Effectively, we replace the simple r_{ij} in the equation (which has values of 0 or 1 indicating whether a pair is mutually

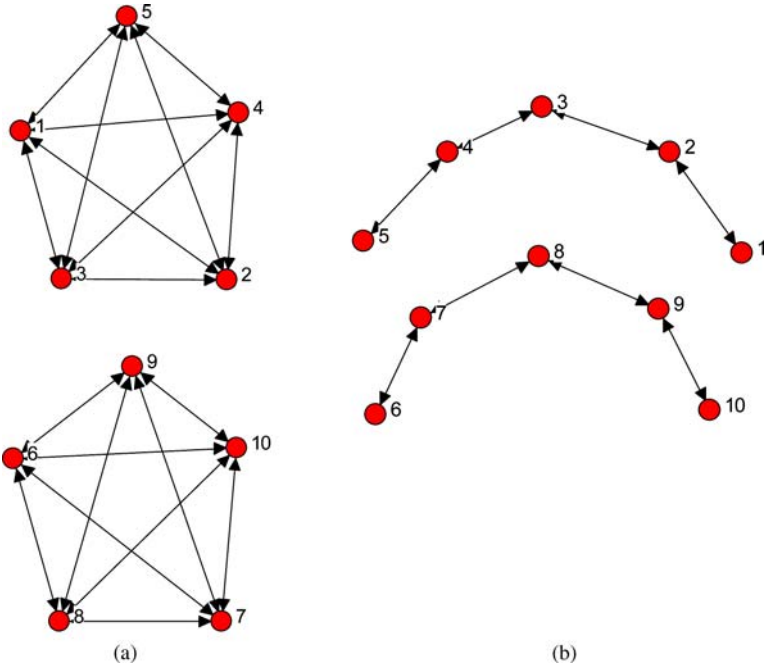

Fig. 5 (a) ${}^{D}F = 0.556$ and (b) ${}^{D}F = 0.715$

reachable or not) with $1/d_{\mathrm{ij}}$, which provides a degree of reachability that varies from 0 to 1.

$$ {}^{D}F = 1 - \frac{2\sum_{i>j}\frac{1}{d_{ij}}}{n(n-1)} \tag{9} $$

The ${}^{D}F$ measure is identical to F when all components are complete (i.e., each component is also a clique). However, when distances within components are greater than 1, the measure captures the relative cohesion of the components. For example, the graph in Fig. 5(a) has two components of size 5 and the ${}^{D}F$ measure is 0.556. The graph in Fig. 5(b), which is less cohesive, also has two components of size 5, but the ${}^{D}F$ measure is 0.715, indicating much less cohesion. Like the F measure, ${}^{D}F$ achieves its maximum value of 1.0 when the graph consists entirely of isolates.

5.2. KPP-Pos: Inter-set cohesion

The fundamental concept implicit in KPP-Pos is the connection or cohesion that members of one set of nodes (the kp set) have with members of another (the remainder of the network). To solve the problem, we need a direct measure of the amount of connection between a set and the rest of the graph.

Thus, we want to define a function, C_K, which gives the amount of cohesion between members of the set K and the remainder of the network (V-K). As a starting point, we might define the following measure (Borgatti, Everett and Shirey, 1992), in which $\mathrm{a}_{ij} = 1$ if node

i is adjacent to node j and $a_{ij} = 0$ otherwise:

$$C_K = \sum_{i \in K, j \in V-K} a_{ij} \tag{10}$$

However, as Everett and Borgatti (1999b) note, this simplistic approach ignores the structural equivalence of set members, essentially double-counting ties to the same individuals. Instead, we would like to define a slightly more sophisticated type of measure, as follows:

$$C_K = \sum_{j \in V-K} \bigcup_{i \in K} a_{ij} \tag{11}$$

In the equation, the operation $\cup$ is a non-specific aggregation function such as taking the minimum or the maximum. If $\cup$ is the maximum function, then C_K is defined as the number of distinct nodes outside of K that members of K are adjacent to. This is identical to Everett and Borgatti's (1999b) notion of group degree centrality.

In addition to this measure, we would like another measure that incorporates the notion of distances, as we did with fragmentation. The reasons are the same: Two groups of a given size may be adjacent to the same number of nodes, but with one group the unreached nodes may only be one link away, while with the other they may be very distant—so distant as to preclude diffusion in a reasonable amount of time.

A simple approach, termed the m-reach measure, is to replace adjacency with reachability, such that ${}^{\mathrm{m}}r_{ij} = 1$ if i can reach j via a path of length m or less, and ${}^{\mathrm{m}}r_{ij} = 0$ otherwise. If we take the $\cup$ operation to be the maximum function, then we can express the m-reach measure as follows:

$$C_K = \sum_{j \in V-K} \bigcup_{i \in K} {}^{m}r_{ij} \tag{12}$$

M-reach, then, is a count of the number of unique nodes reached by any member of the kp-set in m links or less. The advantage of this measure is its ease of interpretation. The disadvantages are that (a) it assumes that all paths of length m or less are equally important (when in fact a path of length 1 is likely to be more important than a path of length 2), and (b) that all paths longer than m are wholly irrelevant.

A more sensitive measure, to be called distance-weighted reach, can be defined as the sum of the reciprocals of distances from the kp-set S to all nodes, where distance from the set to a node is defined as the minimum distance. This measure is given in Eq. (13).

$$C_K = \sum_{j \in V-K} \bigcup_{i \in K} \frac{1}{d_{ij}} \tag{13}$$

For convenience of interpretation, it is useful to regard all distances within the KP set to be unity, and let the summation occur over all nodes. It is also convenient to normalize the measure to run between 0 and 1. At the same time, we can also simplify the notation by defining d_{Kj} to be the minimum distance from any member of K to node j, yielding the final measure shown in Eq. (14).

$${}^{D}R = \frac{\sum_j \frac{1}{d_{Kj}}}{n} \tag{14}$$

1. Select k nodes at random to populate set S
2. Set F = fit using appropriate key player metric
3. For each node u in S and each node v not in S
    a. DELTAF=improvment in fit if u and v were swapped
4. Select pair with largest DELTAF
    a. If DELTAF <= then terminate
    b. Else, swap pair with greatest improvement in fit and set F=F + DELTAF
5. Go to step 3

Fig. 6 Greedy optimization algorithm

Taking some interpretive license, we can view ${}^{D}R$ as the weighted proportion of all nodes reached by the set, where nodes are weighted inversely by their minimum distance from the set and only nodes at distance 1 are given full weight. Hence, ${}^{D}R$ achieves a maximum value of 1 when every outside node is adjacent to at least one member of the kp-set (i.e., the kp-set is a dominating set). The minimum value of 0 is achieved when no member of the kp-set belongs to the same component as any node outside the kp-set—i.e., the kp-set is completely isolated.

5.3. Selecting a KP-set via combinatorial optimization

Since the straightforward heuristic of simply choosing the top k players fails, we must seek another way. One approach is to seek modifications of the top k player heuristic that address some of its weak points. For example, we could begin by choosing the top individual player, and then add the next best individual player that is least redundant with those already selected. Another approach is to simultaneously select the k members of the kp-set via combinatorial optimization. This is the approach taken in this paper.

If we represent a solution to either KPP-Pos or KPP-Neg as a string S of 1s and 0s where $s_i = 1$ if node i is a member of the proposed kp-set and $s_i = 0$ otherwise, then it is easy to apply a number of off-the-shelf optimization algorithms to find S, such as tabu-search (Glover, 1986), K-L (Kernighan and Lin, 1970), simulated annealing (Metropolis et al., 1953) or genetic algorithms (Holland, 1975). Initial experiments suggest that all of these do an excellent job on KPP, and so I present only a simple greedy algorithm. Figure 6 outlines the method, which is normally repeated using dozens of random starting sets.

6. Proof of concept

The operation of the algorithm is illustrated empirically using two datasets. The first is a network of acquaintances among known terrorists, the second is a network of advice-seeking within a consulting company.

6.1. Terrorist dataset

The terrorist dataset, compiled by Krebs (2002), consists of a presumed acquaintance network among 74 suspected terrorists. For the purposes of this analysis, only the main component is used, consisting of 63 individuals.

The first question we ask (KPP-Neg) is which persons should be isolated from the network in order to maximally disrupt the network. Let us assume that we can only isolate three people.

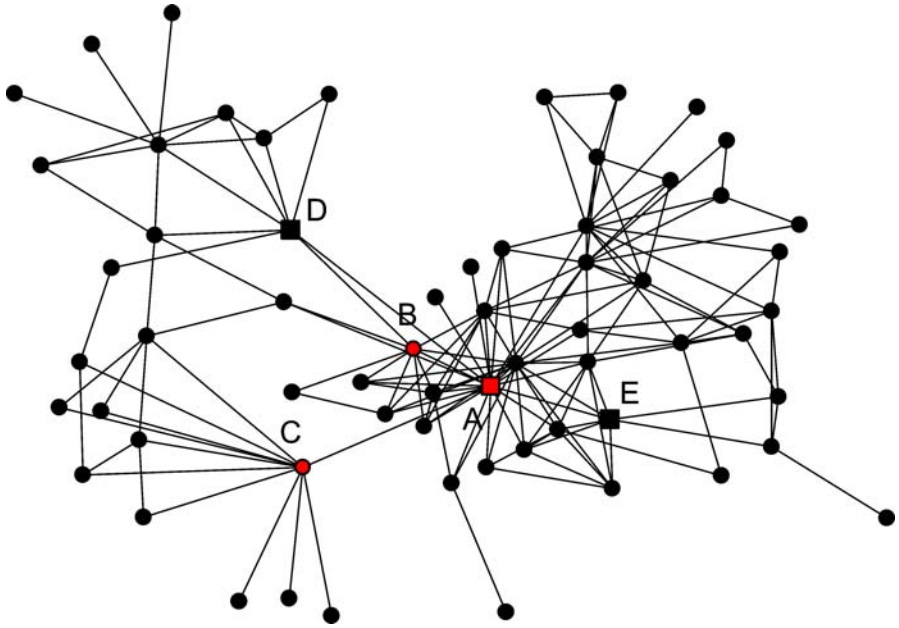

Fig. 7 Terrorist network compiled by Krebs (2002)

A run of the algorithm using F measure selects the three red nodes identified in red in Fig. 7 (nodes A, B and C). Removing these nodes yields a fragmentation measure of 0.59, and breaks the graph into 7 components (including two large ones comprising the left and right halves of the graph).

The second question we ask (KPP-Pos) is, given that we would like to diffuse certain information, which actors would we want to be exposed to the information so as to potentially reach all other actors quickly and surely? Let us assume that information that travels more than two links tends to degrade or be viewed with suspicion. Hence we want the smallest set of nodes that can reach all others within two links or less (i.e., we use the m-reach criterion with $m = 2$). The algorithm finds that a set of just three nodes (the square nodes in Fig. 7, labeled A, C and D) reaches 100% of the network.

6.2. Advice-seeking dataset

These data consist of advice-seeking ties among members of a global consulting company, reported by Cross, Borgatti and Parker (2002). The data were collected on a 1 to 5 strength-of-tie scale, but for this analysis we examine only the strongest ties (rated 5). A diagram is shown in Fig. 8.

We begin with a KPP-Neg analysis, and seek a small set of nodes to remove so as to disconnect the graph. As shown in the figure, when we request a set of two nodes using the distance-weighted fragmentation criterion, the algorithm selects the set {HB, WD}, which gives a ${}^{\mathrm{D}}F$ score of 0.817 and a division into four components (including one isolate). A search for a set of three nodes yields {HB, WD, BM} with a score of 0.843 and a division into six components (including three isolates).

Turning now to a KPP-Pos analysis, we seek a small set of nodes that are well connected to the entire network. To begin, we use the criterion of simple adjacency. Table 1 shows the

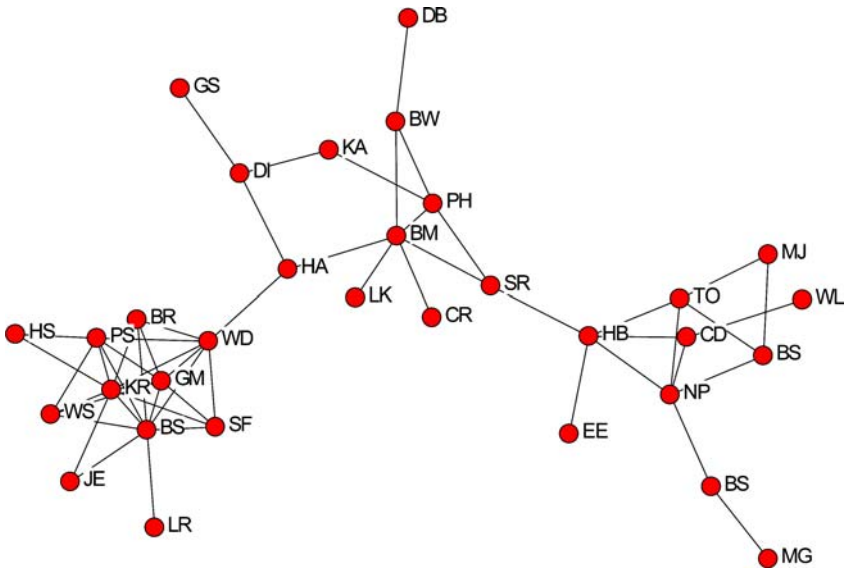

Fig. 8 Strong advice-seeking ties in global consulting company

Table 1 Proportion of nodes reached via paths of length 1 by kp-sets of size 1 to 9

K	Nodes Reached	% of Net Reached	KP-Set
1	10	31	{KR}
2	17	53	{BM,BS}
3	23	72	{BM,BS,NP}
4	26	81	{BM,BS,DI,NP}
5	27	84	{BM,BS,DI,KR,NP}
6	29	91	{BM,BS,DI,HB,KR,TO}
7	30	94	{BM,BS,BS2,DI,HB,PS,TO}
8	31	97	{BM,BS,BS2,CD,DI,HB,PS,TO}
9	32	100	{BM,BS,BW,BS2,CD,DI,HB,PS,TO}

proportion of the network reached by the optimal kp-set of sizes 1 through 9 via paths of length 1. An examination of the results for a 3-node kp-set shows that the algorithm picks a high degree node from each main cluster in the network, as one would expect.

Results are similar for the distance-weighted reach criterion. For example, for a 3-node set, the algorithm selects {BM, KR, NP}, which is different in one node from the simple adjacency criterion, but retains the pattern of selecting one well-connected node from each cluster. For a 2-node set, the algorithm returns {HB, KR} which differs strongly from the {BM, BS} pattern favored by the adjacency criterion. The {BM, BS} choice indicates a strategy of picking the centers of the most populous clusters. In contrast, the {HB, KR} selection indicates a strategy of being jointly close to as many nodes as possible, at the expense of making a selection from the core of a cluster.

7. Discussion

In this paper I have defined the key player problem and demonstrated why existing graph-theoretic methods along with the naïve centrality-based heuristic fail to solve the problem. Basically, two issues are observed with respect to the centrality approach: the goal issue and the ensemble issue. The goal issue refers to the fact that centrality measures were not designed with either key player problem in mind, and hence are not optimal. The graph-theoretic approaches solve the ensemble issue, and in many ways address the goal issue. However, the problem they solve is in some ways the opposite of the problem we seek to solve, in the sense that they fix the quality of the solution and search for the smallest solution, whereas we wish to fix the size of the solution set and search for the best quality.

To address these issues, I have introduced a combinatorial optimization algorithm together with a set of success metrics specifically designed for KPP-Neg and KPP-Pos. The metrics for measuring success in the KPP-Neg problem are essentially measures of graph cohesion that are useful descriptively in a number of contexts besides the key player problem. Typical applications might be the comparison of similar organizations or using cohesion as a predictor of group performance. The KPP-Pos metrics can similarly be adapted for use in measuring both individual and group centrality. Actors occupying positions high on KPP-Pos measures are well-placed to maximize utilization of resources flowing through the network, while actors occupying positions high on KPP-Neg measures have the opportunity to maximize the benefits of brokerage, gatekeeping and playing actors off each other.

The research reported here opens a number of avenues for future exploration. One area of special interest concerns data quality. If the key player approach is to yield a practical tool, we cannot simply assume perfect data. Rather, the method should be robust in the face of errors in the data. Two approaches seem promising. First, there is the notion of not optimizing too closely to the observed dataset. If the data are known to vary from the truth by a given magnitude (e.g., 10% of observed ties don't actually exist and 10% of observed non-adjacent pairs are in fact adjacent), then we can randomly vary the data by this magnitude and optimize across a set of "adjacent" datasets obtained in this way. The result is a kp-set that is not necessarily optimal for the observed dataset, but will represent a high-quality solution for the neighborhood of the graph as a whole. An alternative approach is to treat knowledge of ties as probabilistic, modifying the KeyPlayer metrics accordingly. For example, if we knew the probability of a tie between any two nodes, we could, in principle, work out the expected distance (including infinity) between the nodes across all possible networks.[1] KPP measures based on distance and reachability could then be computed by substituting expected distance for observed distance. The practical challenge here is to find shortcut formulas for expected distance and connectedness that enable fast computation.

In addition, it is of interest to incorporate actor attributes into the key player metrics. In the military context, communication among actors with redundant skills may sometimes be less important than communication between actors with complementary skills. In the public health context, it is helpful in slowing epidemics to minimize mixing of different populations (such as when married women are linked to commercial sex workers via their husbands). Hence, an additional criterion we would want to consider in fragmenting a network is maximizing separation of actors with certain attributes.

[1] Note that the problem being addressed here is certainty of observed data values, not probability that a tie exists at a given moment. It is assumed in this approach that ties are fixed and not probabilistically emerging as a function of node attributes or other ties. The dynamic nature of ties is a different phenomenon that wants its own models.

Finally, an interesting line of research concerns the interaction of the network structure with key player metrics, and ultimately the ability of the algorithms to extract optimal sets. For example, it appears that nodes achieve the highest values on the KPP-Pos metrics when they are embedded in highly cohesive graphs. In such graphs, even small, easy-to-find kp-sets will have relatively large scores. In contrast, high values on KPP-Neg measures will normally occur only when the graph is not very cohesive. In such graphs, inexpensive heuristic methods can yield results as good as those obtained by costlier optimization methods.

Acknowledgements This research is supported by Office of Naval Research grant number N000140211032. Thanks to Scott Clair for leading me to this problem, Mark Newman for suggesting reciprocal distances, Kathleen Carley for useful discussions, and Valdis Krebs for providing illustrative data.

References

Borgatti SP (2002) Stopping terrorist networks: Can social network analysis really contribute? Sunbelt International Social Networks Conference 13–17. New Orleans

Borgatti SP, Everett MG (1999) Models of core/periphery structures. Social Networks 21:375–395

Borgatti MG, Jones C, Everett MG (1998) Network measures of social capital. Connections 21(2):27–36

Burt RS (1992) Structural Holes: The social structure of competition. Harvard University Press, Cambridge

Coleman J (1990) Foundations of social theory. Belknap Press, Cambridge, MA

Everett MG, Borgatti SP (1999a) Peripheries of cohesive subsets. Social Networks 21:397–407

Everett MG, Borgatti SP (1999b). The centrality of groups and classes. Journal of Mathematical Sociology. 23(3):181–201

Freeman LC (1979) Centrality in social networks: Conceptual clarification. Social Networks 1:215–239

Friedkin NE (1991) Theoretical foundations for centrality measures. American Journal of Sociology 96:1478–504

Glover F (1986) Future paths for integer programming and links to artificial intelligence. Computers and Operations Research 5:533–549

Holland J (1975) Adaptation in natural and artificial systems. University of Michigan Press

Kernighan BW, Lin S (1970) Efficient heuristic procedure for partitioning graphs. Bell Systems Technical Journal 49(2):291–297

Krebs V (2002) Uncloaking terrorist networks. First Monday 7(4) http://www.firstmonday.dk/issues/ issue7_4/krebs/index.html

Metropolis N, Rosenbluth A, Rosenbluth M, Teller A, Teller E (1953) Equation of state calculations by fast computing machines. J Chem Phys 21(6):1087–1092

Morris M, Kretzschmar M (1997) Concurrent partnerships and the spread of HIV. AIDS 11:641–648

Seidman S (1983) Network structure and minimum degree. Social Networks 5:269–287

Stephen P. Borgatti is Professor of Organization Studies at the Carroll School of Management, Boston College. His research is focused on social networks, social cognition and knowledge management. He is also interested in the application of social network analysis to the solution of managerial problems.