The Key Player Problem[1]

Stephen P. Borgatti[2]

Introduction

The key player problem (KPP) consists of two separate sub-problems, which can be described at a general level as follows:

1. (KPP-1) Given a social network, find a set of *k* nodes (called a *kp-set* of order *k*) which, if removed, would maximally disrupt communication among the remaining nodes.

2. (KPP-2) Given a social network, find a kp-set of order *k* that is maximally connected to all other nodes.

Of course, these introductory definitions leave out what is meant precisely by "maximally disrupt communication" 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 would seem clear that KPP-1 involves fragmenting a network into components, or, failing that, making distances between nodes so large as to be practically disconnected. In contrast, KPP-2 involves finding nodes that can reach as many remaining nodes as possible via direct links or perhaps short paths.

The first problem, KPP-1, arises in a number of contexts. A prime example in the public health context is the immunization/quarantine problem. Given an infectious disease that is transmitted from person to person, and given that it is not feasible to immunize and/or quarantine an entire population, which subset of members should be immunized/quarantined so as to maximally hinder the spread of the infection? An example in the military context is target selection. Given a network of terrorists who must coordinate in order to mount effective attaches, and given that only a small number can be intervened with (e.g., by arresting or discrediting), which ones should be chosen in order to maximally disrupt the network?

The second problem, KPP-2, arises in the public health context 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 the organizational management context, the problem occurs 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. In the military context, KPP-2 translates to locating an efficient set of enemies to surveil, turn (into double-agents), or feed misinformation to.

At first glance, both KPP-1 and KPP-2 would appear to be easily solved by either employing some graph-theoretic concepts such as cutpoints and cutsets, or via the methods of social network analysis, such as measuring node centrality. It turns out, however, that none of the existing methods are adequate. This paper explains why and presents a new approach specifically designed for the key player problem.

Centrality Approach

[1] 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.

[2] Dept. of Organization Studies, Carroll School of Management, Boston College, Chestnut Hill, MA 02467

The centrality approach consists of measuring the centrality of each node in the network, then selecting the *k* most central nodes to comprise the kp-set. Since many measures of centrality exist, one question that arises is which measure to use. For KPP-1, we can expect the best measures to be those based on betweenness. For example, 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 responsible for connecting many pairs of nodes via the best path, and deleting that node should cause many pairs of nodes to be more distantly (if not completely disconnected).

For KPP-2, we can expect measures based on degree centrality and closeness centrality to be useful. Degree centrality is simply the number of nodes that a given node is adjacent to. Hence, depending on what the social relation represented by the graph is and assuming that adjacency implies potential for influence, a node with high degree can potentially directly influence many other nodes. Closeness centrality is defined as the sum of geodesic distances from a given node to all others, where geodesic distance refers to the length of the shortest path between two points. Thus, a node with a low closeness score (very central) should be able to influence, directly and indirectly, many others.

The centrality measures are plausible solutions for KPP. However, they are not optimal. There are two basic problems, which I refer to as the design issue and the group selection issue. Of the two, the group selection issue is the more serious.

The Design Issue

The design issue arises ultimately from the fact that centrality measures were not designed with KPP-1 and KPP-2 specifically in mind. Starting with KPP-1, consider the graph in Figure 1.

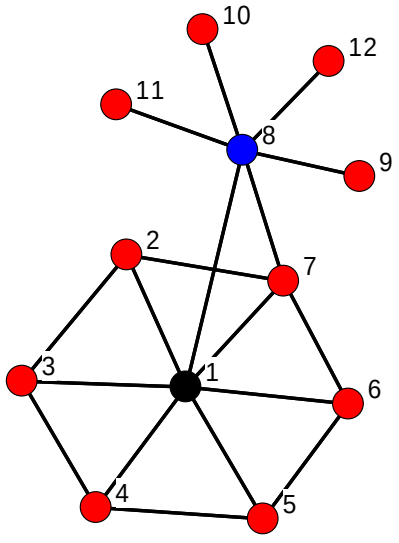

Figure 1.

Node 1 has the highest centrality on all considered measures, including betweenness centrality. Yet deleting node 1 has relatively little effect on the network. Distances between certain pairs of nodes do increase, but it is clear that communication among all points remains possible as there is no fragmentation. In contrast, deleting node 8, which does not have the highest betweenness, is more effective. Removing 8 splits the graph into five unconnected fragments (components).

For KPP-2, the picture is a little brighter. If we formulate KPP-2 in terms of reaching the most nodes directly, degree centrality is optimal. If we formulate it in terms of reaching the most nodes in up to *m* steps, then we can readily define a new measure of centrality ("*m*-reach centrality") that counts the number of nodes within distance *m* of a given node.

The Group Selection Issue

The group selection issue, discussed as the group centrality problem in Everett and Borgatti (1999), refers to the fact that selecting a set of nodes which, as an ensemble, solves KPP-1 or KPP-2, is quite different from selecting an equal number of nodes that individually are optimal solutions for KPP. To start with, consider KPP-1. Figure 2 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 than deleting *i* alone. In contrast, deleting *m* with *h* does produce increased fragmentation, even though individually *m* is not nearly as effective as *i*. The reason *i* and *h* are not as good together as *i* and *m* is that *i* and *h* are redundant with respect to their liaising role – they connect the same third parties to each other. In a sense, the centrality of one is due to the centrality of the other, with the result being that the centrality of the ensemble "control" is quite a bit less than the sum of the centralities of each.

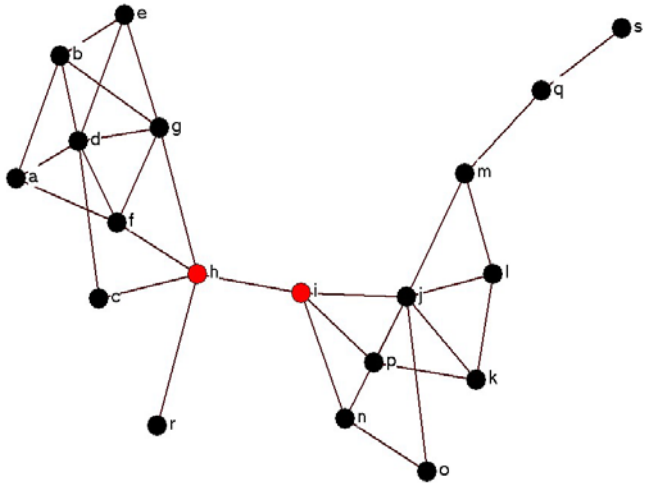

Figure 2

The redundancy principle also applies to KPP-2. Consider the graph in Figure 3. Nodes *a* and *b* are individually the best connected. Each reaches five other nodes, more than any other by far. But together they reach no more than either does alone. In contrast, if *b* is paired with *c* (which individually reaches only three nodes), the set reaches every node in the network.

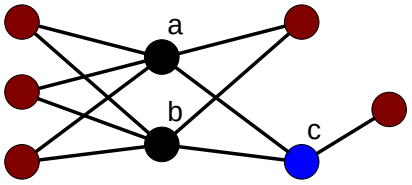

Figure 3.

Graph Theoretic Approaches

In addition to basic concepts such as components and distance, a number of graph-theoretic concepts are relevant to KPP. For KPP-1, the most obvious are the notions of cutpoint and bridge, which are nodes and lines respectively whose deletion would increase the number of components in the graph. These concepts do not address the group-selection issue. However, both have set-level counterparts in the form of cutsets. A cutset is a set of nodes (or lines) whose removal would increase the number of components in the graph. Most work has focused on minimum weight cutsets, which are smallest sets that have the cutset property. There are three difficulties with cutsets in the KPP context. 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 we are able to find a smallest set that achieves that level of success). In this sense, the graph theoretic approaches solve the inverse of the problem we seek to solve. Second, no account is taken of distances among nodes. Third, all solutions that increase the number of components are equal, even if one solution creates just two components while another creates ten.

For KPP-2, applicable 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 set of nodes whose members are adjacent to all other nodes in the graph.[3] For our purposes these are equivalent and fail for exactly the same reasons as cutsets.

Measuring Success

In order to optimally solve KPP, we must have a clear definition of success. I propose that for KPP-1 there are two network properties we want to maximize by removing the kp-set: fragmentation and distance. For KPP-2, we want to measure the distance-based reach of the kp-set into the network around it. Therefore, we need measures for each of these concepts.

Fragmentation

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 by dividing by the number of nodes.

$$C = \frac{K}{n}$$

Eq. 1

The problem with this measure is that it doesn't take into account the sizes of the components. For example, in Figure 2, deleting node *m* would break the network into two components, but the vast majority of nodes remain connected. In contrast, deleting node *i* would also result in just two components, but more pairs of nodes would be disconnected from each other.

This suggests another measure 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 follows:

[3] Graph theorists differ on whether these sets are understood to be minimal or not.

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

Eq. 2.

Since, by definition, nodes within a component are mutually reachable, the *F* measure can be rewritten more economically in terms of the sizes (s_k) of each component (indexed by *k*):

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

Eq. 3.

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 it is defined as follows:

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

Eq. 4.

One difference between *F* and *H* is that while both achieve minimum values of 0 when the network consists of a single component, when the network is maximally fragment (all isolates) the *H* measure can only achieve 1-1/*n*. If we normalize *H* by dividing by 1-1/*n*, we obtain the *F* measure (and seeing *F* as a normalization of *H* points us to a slightly faster computing formula).

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

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

Eq. 5.

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:

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

Eq. 6.

Distance

While the fragmentation measure *F* is very satisfactory for what it does, it does 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 (Figure 4a) is seen as equally fragmented as a network divided into two

components of size 5 in which each component is a line (Figure 4b). Yet distances and therefore transmission times are much higher in the latter network. Further, if we can delete only so many nodes, it may be that there is no possible selection of nodes whose removal would disconnect the graph. In such cases, we would still like some way of evaluating which sets are better than others.

An obvious solution would be to measure the total distance between all pairs of nodes in the network. 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 sum 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 Equation 2, that weights by reciprocal distance.

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

Eq. 7.

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 Figure 4a has two components of size 5 and the ${}^{D}F$ measure is 0.556. The graph in Figure 4b, 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.

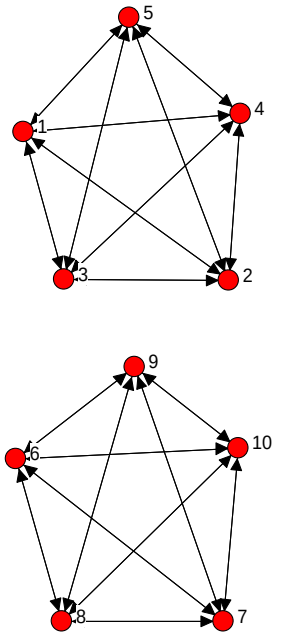

Figure 4a. ${}^{D}F = 0.556$

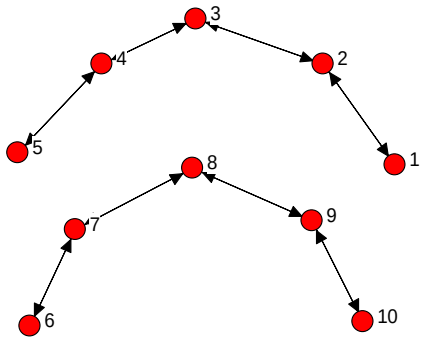

Figure 4b. ${}^{D}F = 0.715$

Distance-Based Reach

The measures discussed for KPP-1 are graph-level measures that we apply to a kp-set by removing the set and measuring the fragmentation in the remaining graph. For KPP-2, we seek a set of nodes that, as a group, is maximally connected to all other nodes. Hence, we need a direct measure of the centrality of the kp-set as a group. The concept of group centrality has already been elaborated by Everett and Borgatti (1999), but only degree, closeness, betweenness and eigenvector group centrality measures have been discussed. As noted earlier, these measures are not optimal for the KPP-2 problem. Hence, we must develop new ones based on the concept of reach.

The simplest group reach measure, termed *m-reach*, 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 face validity. The disadvantage is that 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 that all paths longer than *m* are wholly irrelevant.

A more sensitive measure, called distance-weighted reach, can be defined as the sum of the reciprocals of distances from the kp-set *S* to all nodes (see Equation 8). As described by Everett and Borgatti (1999), the distance from a set to a node outside the set can be usefully defined in a number of ways, such as taking the maximum distance from any member of the set to the outside node, taking the average distance, or taking the minimum distance. For KPP-2, the minimum distance is appropriate since the fact that the distance to an outside node might be large for a given member of the set will usually be irrelevant.

$$ {}^{D}R = \frac{\sum_{j} \frac{1}{d_{Sj}}}{n} $$

Eq. 8.

In the equation, the distance from kp-set *S* to node *j* is indicated by d_{Sj}. In addition, it should be noted that the summation is across all nodes and the distance from the set to a node within the set is defined to be 1. As before the reciprocal of an infinite distance is defined to be 0. Taking some interpretive license, we can view ${}^{D}R$ as the proportion of all nodes reached by the set, where nodes are weighted by their 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 every node is an isolate.

Selecting a KP-Set

For KP-sets of size 1, the measures presented above can used straightforwardly to select key players by simply choosing the one with the highest score on any given measure. Thus, they can be regarded as new measures of node centrality that are optimized for the key player problem.

For kp-sets of size k > 1, however, there is no simple procedure for choosing an optimal set. Some heuristic procedures may be of value. For example, for KPP-2, we start by selecting the node with the highest ${}^{D}R$ score. Then, for each of the remaining k-1 selections, we choose the node with the highest score that is not adjacent to any of the nodes already selected. This algorithm, a variant of the first fit decreasing algorithm for the bin-packing problem, is fast and easy, but often yields solutions that are considerably less than optimal.

Other heuristic approaches specific to the KPP can be constructed, but the fact that we have clear measures of success that are easily computed recommends a generic combinatorial optimization

algorithm, 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 5 outlines the method, which is normally repeated using dozens of random starting sets.

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 = improvement 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

Figure 5. Greedy optimization algorithm.

Empirical Trials

The operation of the algorithm is illustrated using two datasets drawn from the public health (AIDS) and military (terrorism) contexts. Both cases are approached from both KPP-1 and KPP-2 points of view.

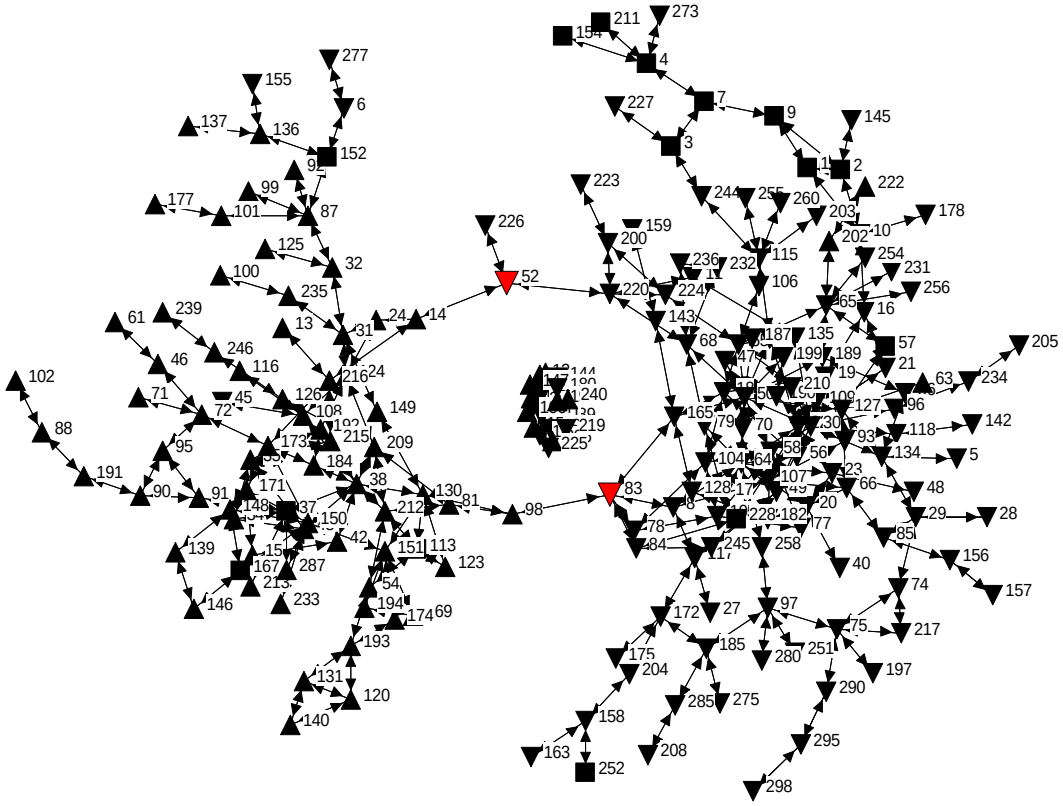

Figure 6. Acquaintance network. Upward triangles indicate African-Americans, downward triangles indicate Puerto-Ricans, and squares identify all others.

AIDS Example

The AIDS dataset consists of an acquaintance network among 293 drug injectors on the streets of Hartford, CT. The data are described in Weeks et al (2002). The network consists of one large main component (193 nodes), and many very small components. As shown in Figure 6, the main component of the network has a very clear structure. It consists of two groups, one african-american (with higher HIV+ proportion), and the other largely puerto-rican (with lower HIV+ proportion). Connection between the two groups is limited by just a few acquaintances and this bottleneck helps maintain the lower HIV+ rate in the Puerto-Rican part of the network. Whether through immunization (should that become possible) or quarantining, it is clear that one thing we would want to do early on is to isolate the two groups from one another, both because we want to maintain the low HIV+ rates in Puerto-Rican cluster, and because risk is a function of the size of the component a node is part of. Thus, we have a case of KPP-1.

The network provides a useful first test of the key player optimization algorithm for two reasons. First, the structure of the network makes it quite vulnerable to disconnection, and easy to check the results visually. If the algorithm fails this test, it is clearly not adequate. Second, the network is already fragmented, providing noise that could confuse some algorithms.

Based on visual inspection, it is clear that immunizing or quarantining just two nodes would separate the main component into two nearly equal halves. So for the first run of the algorithm we seek a kp-set of size 2. The starting fragmentation index for the graph is 0.567. The algorithm selected two nodes, identified in black in Figure 6, which, if deleted, would increase fragmentation to 0.658 (a proportional increase in fragmentation of 16%). The selected nodes match our intuition and divide the main component in two big chunks.

Turning to KPP-2, we are also interested in selecting a small group of nodes to be subjects of an intervention – specifically, to be trained as peer educators (known as Peer Health Advocates or PHAs) to disseminate and demonstrate HIV prevention practices. Weeks et al (2002) did this by hand, laboriously selecting the smallest group that could reach more than 50% of the main component of the network. The winning set contained 14 nodes. Running Key Player on the same data, yields the same result, as shown in Table 1.

Group Size	Number Reached	Percent Reached
1	16	8.3
2	27	14.0
3	36	18.7
4	43	22.3
5	49	25.4
6	55	28.5
7	61	31.6
8	67	34.7
9	72	37.3
10	77	39.9
11	82	42.5
12	87	45.1
13	92	47.7
14	97	50.3

Table 1.

As might be expected, the number of people reached increases as a fractional power of the group size, but fuller consideration of this phenomenon is outside the scope of this paper.

Terrorist Example

The terrorist dataset, compiled by Krebs (2001), 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.

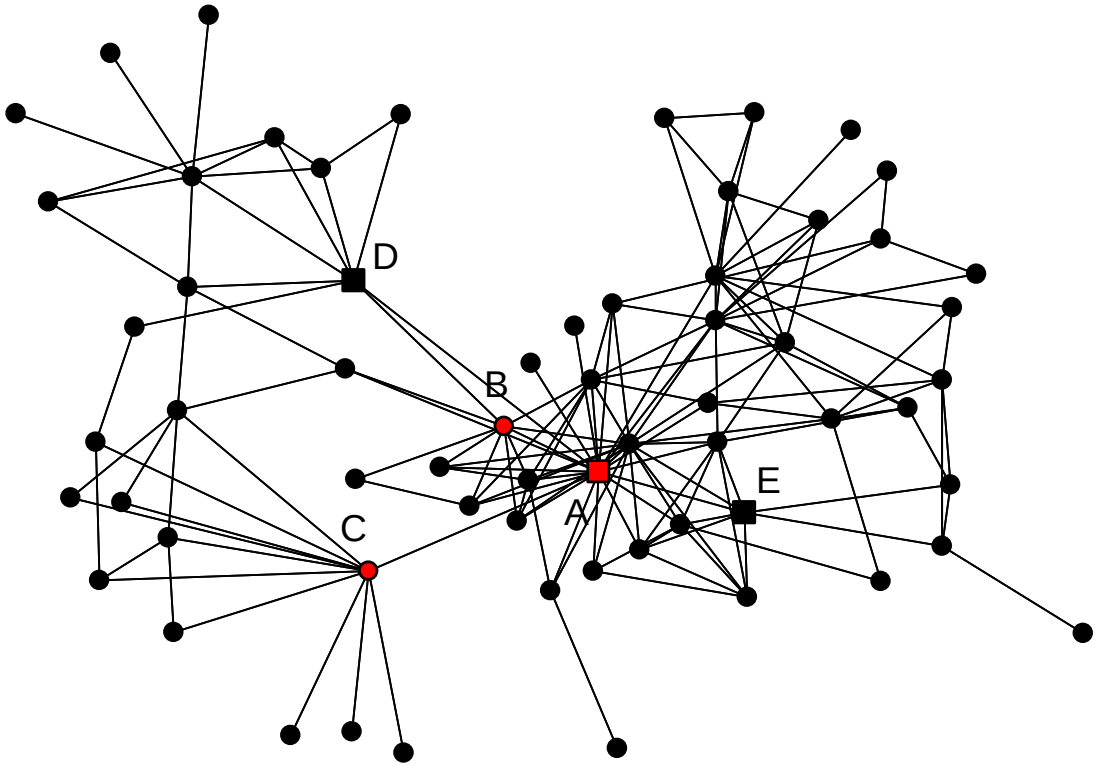

Figure 7. Terrorist network compiled by Krebs (2002).

The first question we ask is which persons should be isolated in order to maximally disrupt the network. Let us assume that we can only isolate three people. A run of the KeyPlayer program selects the three red nodes identified in red in Figure 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 is, given that we would like to diffuse certain information, which nodes would we want to be exposed to the information so as to potentially reach all other nodes quickly and with certainty? 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. The KeyPlayer algorithm finds that a set of three nodes (the square nodes in Figure 7, labeled A, C and D) reaches 100% of the network (including themselves).

Discussion

In this paper we have defined the KeyPlayer problem and demonstrated why the naïve centrality-based approach and more sophisticated graph-theoretic approaches fail to solve the problem. We have introduced new metrics for measuring success, and implemented a combinatorial optimization algorithm to maximize these quantities. Applications in both health and military areas were demonstrated.

The metrics for measuring success in the KPP-1 problem are essentially measures of graph cohesion that may be 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 fact that KPP-1 and KPP-2 have different solutions is interesting. For many, centrality is either a unitary concept with many highly correlated measures, or a fully multidimensional concept in which each measure indicates a different kind of centrality. I would suggest an intermediate view that divides notions of nodal importance into two basic types, corresponding to the KPP-1 and KPP-2 problems. KPP-1 measures quantify the extent to which a graph's cohesion is reduced by the removal of that node, while KPP-2 measures quantify the extent to which a node's ties reach into the network. It is the second type that is most directly about centrality: a node that is well-connected. The first type is not as much about the node's well-connectedness as about the connectedness of the rest of the graph in the absence of the node. From the graph's point of view, there is a loose analogy to the distinction in operant conditioning between positive reinforcement (KPP-2) and negative reinforcement (KPP-1): positive reinforcement provides benefit (connectivity) by directly providing a boon (the node's connections), while negative reinforcement provides benefit by relieving a stressor (the node holds together the otherwise fragmented graph). From the node's point of view, a node achieves its highest values on KPP-2 measures when the graph is highly cohesive. In contrast, high values on KPP-1 measures will normally occur only when the graph is not very cohesive. Actors high on KPP-2 measures lend themselves to maximizing utilization of resources flowing through the network, while actors high on KPP-1 measures lend themselves to maximizing the benefits of brokerage, gatekeeping and playing actors off each other.

Limitations and Next Steps

There are significant dimensions to the problem that I have ignored. Perhaps the most important is the issue of data quality. If this 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 will be explored. 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, suppose we knew, for each pair of nodes, the independent likelihood that a tie exists between them. Then, in principle, we could work out the expected distance (including infinity) between the nodes across all possible networks.[4] KPP measures based on distance and reachability could then be computed substituting expected distance for observed distance. The practical challenge here is to find shortcut formulas for expected distance and connectedness that enable fast computation.

[4] Note that the problem being addressed is certainty of observed data values, not the existence of ties. 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.

Another issue concerns the use of geodesic distance at all. As discussed by Borgatti (2002), different kinds of flows processes have different kinds of characteristic trajectories. For example, infections that immunize (or kill) the host don't return to nodes they have previously infected. Hence, they travel along graph theoretic paths. In contrast, gossip transmitted via email can easily reach the same node multiple times, but in general not from the same sources (i.e., a person doesn't usually tell the same confidential story to the same person more than once). Hence, stories travel along trails. Neither infections nor stories necessarily travel via the shortest paths to each node. Consequently, for those processes, the expected distance travelled by something flowing through the network is not equal to geodesic distance, and it would make sense to substitute this other expected distance in the equations.

For simplicity of exposition, this paper has assumed undirected simple graphs with non-valued edges. The extension to directed graphs is straightforward, but the extension to valued edges will require some development. An exception is the class of fragmentation measures, which generalize nicely along the lines of minimum weight cutsets, losing only the computational shortcuts available with non-valued data. The distance-based metrics will require different generalizations depending on the kinds of values, which could range from distances to capacities to probabilities of transmission. Multiple relations, recorded as separate graphs with a common node set, might be handled by summing the success criterion across all relations.

Finally, it may be of interest in the future to incorporate actor attributes. 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 criteria we would want to consider in fragmenting a network is maximizing separation of actors with certain attributes.

References

Borgatti, S.P. 2002. Stopping terrorist networks: Can social network analysis really contribute? *Sunbelt International Social Networks Conference*. 13-17 February. New Orleans.

Everett, M. G., & Borgatti, S. P. 1999. The centrality of groups and classes. *Journal of Mathematical Sociology*. 23(3): 181-201

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, B.W., and 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): April. http://www.firstmonday.dk/issues/issue7_4/krebs/index.html

Metropolis,N., A. Rosenbluth, M. Rosenbluth, A. Teller, E. Teller. 1953. Equation of state calculations by fast computing machines", *J. Chem. Phys.*,21, 6, 1087-1092.

Weeks, M.R., Clair, S., Borgatti, S.P., Radda, K., and Schensul, J.J. 2002. Social networks of drug users in high risk sites: Finding the connections. *AIDS and Behavior* 6(2): 193-206.