2011-9-15

Structural Search and Optimization in Social Networks

Milind Dawande, *et al.*

INFORMS Journal on Computing
Vol. 24, No. 4, Fall 2012, pp. 611–623
ISSN 1091-9856 (print) | ISSN 1526-5528 (online)

http://dx.doi.org/10.1287/ijoc.1110.0473

Structural Search and Optimization in Social Networks

Milind Dawande, Vijay Mookerjee, Chelliah Sriskandarajah, Yunxia Zhu
School of Management, University of Texas at Dallas, Richardson, Texas 75083
{milind@utdallas.edu, vijaym@utdallas.edu, chelliah@utdallas.edu, yunxia.zhu@student.utdallas.edu}

The explosive growth in the variety and size of social networks has focused attention on searching these networks for useful structures. Like the Internet or the telephone network, the ability to efficiently search large social networks will play an important role in the extent of their use by individuals and organizations alike. However, unlike these domains, search on social networks is likely to involve measures that require a *set* of individuals to collectively satisfy some skill requirement or be tightly related to each other via some underlying social property of interest.

The aim of this paper is to highlight—and demonstrate via specific examples—the need for algorithmic results for some fundamental set-based notions on which search in social networks is expected to be prevalent. To this end, we argue that the concepts of *an influential set* and *a central set* that highlight, respectively, the specific role and the specific location of a set are likely to be useful in practice. We formulate two specific search problems: the elite group problem (EGP) and the portal problem (PP), that represent these two concepts and provide a variety of algorithmic results. We first demonstrate the relevance of EGP and PP across a variety of social networks reported in the literature. For simple networks (e.g., structured trees and bipartite graphs, cycles, paths), we show that an optimal solution to both EGP and PP is easy to obtain. Next, we show that EGP is polynomially solvable on a general graph, whereas PP is strongly NP-hard. Motivated by practical considerations, we also discuss (i) a size-constrained variant of EGP together with its penalty-based relaxation and (ii) the solution of PP on balanced and full d-trees and general trees.

Key words: social networks; structural search; analysis of algorithms; computational complexity
History: Accepted by Karen Aardal, Area Editor for Design and Analysis of Algorithms; received November 2009; revised September 2010, April 2011; accepted May 2011. Published online in *Articles in Advance* September 15, 2011.

1. Introduction

A social network represents a social structure as a set of definite relationships between the members—entities or groups—of a social system. In its most commonly used representation, a social network can be viewed as a network of nodes (individuals, organizations, Web pages, etc.) related to one another using edges (friendship, commercial transactions, URL links, etc.). Over the years, social networks have been used to analyze social phenomena in a wide variety of domains, including sociology, epidemiology, social psychology, economics, anthropology, history, and human geography (Scott 2000, Wasserman and Faust 1994, Brandes and Erlebach 2005). Often in social network analysis, the researcher's interest is in explaining individual or group behavior in the context of the larger social structure in which the individual or group is situated.

More recently, social networking sites such as Facebook (http://www.facebook.com) and MySpace (http://www.myspace.com) have proliferated on the Internet and help users connect based on a wide range of interests and practices. Although some sites support the maintenance of preexisting social networks, others help strangers connect based on their shared interests and/or activities. Some sites cater to diverse audiences, whereas others attract people based on some shared identity (Boyd and Ellison 2007). Typically, the participants (players) of the network derive some utility from the network, for example, to find others for idea exchange, problem solving, companionship, and so on.

It should be clear that, like any other network-based phenomenon such as the telephone or the Internet, the ability of the individual or group to derive value depends on the ability to search the network for contacts. For example, searching the telephone network is facilitated by a phone directory, browsing the Internet requires a browser and a search engine, and so on. Many researchers believe that the advent of the Web browser and search engine was most influential to the explosive growth of the Internet. By analogy, it can be proposed that the utility of social networks to individuals and organizations will also depend on the ability to search the networks of interest for useful structures. For example, a participant in Facebook may want to

discuss a topic of interest and may need to call on a selected subset of friends to join the discussion. In the open source community, individual developers form a social network by virtue of having worked on common projects. In such a community, a developer or a firm may want to create a project team of members with certain specialized skills and access to resources.

Searching a social network often creates search problems that are different from those encountered in other network phenomena such as the Web or the telephone network. In the Web, the typical nature of search is to provide the user with a set of websites that match based on a list of search terms. There is usually no requirement that the websites returned by the search engine satisfy some complex relationship to one another other than, of course, the trivial relationship that they must all match (to varying degrees) with the list of search terms. On the other hand, search problems in a social network can be more complex. In particular, the search results may often need to satisfy a *set* measure. For example, in extracting a project team from a larger network, it may be important that the set of developers that are returned collectively satisfy some skill requirements but, in addition, are tightly related to one another by virtue of having worked on common projects. With the improvement in computing technology, the data and the tools needed to identify the network of interest are readily available. From a technical perspective, when the results of a search need to meet (or exceed) a specified set measure (specifically, a nonadditive measure), the search often becomes combinatorial in nature. Search problems in social networks therefore provide a challenging ground for researchers interested in applying graph-theoretic, algorithmic methods to the area. Our interest in this study stems from the new problems and opportunities that are likely to arise for the use of graph-theoretic methods to solve interesting search problems in social networks.

The remainder of this paper is organized as follows. In §2, we argue that two set-based notions—influential sets and central sets—are likely to provide a fundamental structural basis for important search problems arising in a variety of practical social networks, and we introduce two optimization problems—the elite group problem (EGP) and the portal problem (PP)—corresponding to these two notions. Section 3 investigates the complexities of these problems on several special graphs as well as on general graphs. Section 4 concludes our paper and provides directions for future research.

2. The Notions of Influential Sets and Central Sets

Given the significance of search in social networks and, consequently, the need for efficient algorithms, an important question naturally arises: What are some fundamental set-based notions on which search in social networks is expected to be prevalent? Traditionally, in social network analysis, two fundamental properties of individual members—their *location* and their *role* in the network—have proven to be fundamental. This is natural because these two properties provide insights into the groupings and interactions in the network. Accordingly, for individual members of a social network, network centrality measures, including *degree centrality*, *closeness centrality*, and *betweenness centrality*, have been heavily investigated and used (see, e.g., Freeman 1979; Brandes and Erlebach 2005, Chapters 3–5). For set-based search as well, structures and measures that highlight the specific role or specific location of a set are likely to be the most useful in practice. The need and use of such set-based measures has already been documented in other studies. For example, the notions of group (or other set) betweeness and group degree centralities are discussed in Chapter 4 of Carrington et al. (2005) and in Everett and Borgatti (1999).

The motivation to study the role played by members in a network has to do with understanding the influence a member can potentially cast over other members in the network. Such notions of influence exerted by a single member can intuitively be extended to the influence a *set* of members can potentially exert over the rest of the group. A set of influential members may be useful to identify for a variety of reasons, often having to do with wanting to promote an idea, product, or message to other members of the network. For example, a firm may wish to advertise a new product or service and use an influential group of members to help in this cause (Hill et al. 2006). Similarly, a welfare organization may want to disseminate ideas of social importance within a community of interacting members and use an influential set of members for spreading the message in an effective and timely manner. Another reason to study influential groups is often to identify a set of members who possess specialized knowledge or information pertaining to a specific domain—namely, the key *experts* in the group. For example, it may be important to identify a set of expert oncologists for devising an informed yet balanced plan of action to treat a difficult case. Here, a *set* of experts may be especially relevant to consult to eliminate or reduce bias as well as to surface fresh perspectives that can aid in problem solving.

The motivation to study the location of a member (or a set of members) is subtly different from that of examining member roles. Location is essentially a topological characteristic that has to do with a member or a set of members acting to facilitate contact between other interacting members of

the network. A centrally located member is *well connected*, or, in other words, has better access to other members by virtue of acting as a conduit that allows exchanges and flows of information or ideas in the network. A central location does not necessarily imply influence, and neither does an influential member necessarily need to be centrally located. Indeed, recent research in reality mining (Pentland 2004, Greene 2008, Hesseldahl 2008) and interaction within social networks reveals significant distinctions between these two concepts. For example, managers who may be influential within a business organization usually do not play a central role in the routing of communications between teams (Gloor et al. 2007, Thompson 2008). The players central for communication could, instead, be less influential employees. The question arises: What property does location convey that is useful to a problem solver? One benefit of identifying centrally located members is that it provides one with an understanding of the paths that are heavily used in the network so that sufficient resources can be made available at these locations to avoid communication bottlenecks from occurring. An interesting variant is one where the problem solver may *want* to thwart communication: the activities of a terrorist group may be significantly impaired by striking at locations or members that are central to the flow of communication within the network (Erickson 1981).

The twin notions of influence and centrality admit a variety of interpretations, depending on the context of the social network under consideration. Accordingly, there can be several meaningful measures to evaluate "good" influential and central sets. For example, to measure the centrality of a set of vertices, the classical measure of *betweenness centrality* of a single vertex has been extended to *group betweenness centrality* (Everett and Borgatti 1999) and *co-betweenness centrality* (Kolaczyk et al. 2009). Along this theme, we propose two specific measures: one for an influential set and the other for a central set. We now describe two optimization problems that correspond to these two measures, discuss their origins, and provide examples of social networks where these problems are relevant.

2.1. The Elite Group Problem and the Size-Constrained Elite Group Problem

2.1.1. Technical Definition

Instance: n players; an "influence" social network represented by a directed graph $G(V, A)$, $|V| = n$, in which the nodes represent the players and the set of arcs represent pairwise influences pertaining to a social property: a directed arc (i, j) indicates that i is influenced by j. For the size-constrained elite group problem (SCEGP), a positive integer $k \leq n$ is also given.

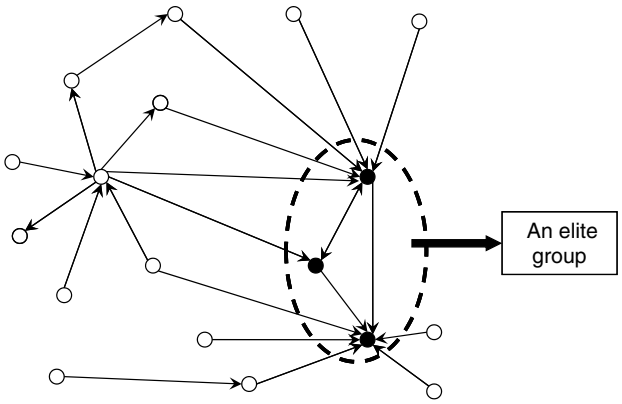

Figure 1 An "Influence" Network and an Elite Group

Solution of EGP: A set $W \subseteq V$ such that there does not exist a directed arc $(i, j) \in A$ with $i \in W$, $j \notin W$; see Figure 1.

Solution of SCEGP: Same as EGP, with the additional requirement that $|W| \leq k$.

Objective Function: Maximize the total number of directed arcs, γ_W, incident on any node in W from nodes in $V \backslash W$. More precisely, the *score* γ_W is defined as follows: $\gamma_W = \sum_{i \notin W,\, j \in W} a_{ij}$, where $a_{ij} = 1$ if $(i, j) \in A$, and 0 otherwise.

Note that in a graph $G(V, A)$, there is at least one feasible solution for EGP—namely, the complete set of nodes V, with score $\gamma_V = 0$. Also, observe that adding more nodes to an elite group does not necessarily increase the number of directed arcs into the group. If a node, say, j, is added to an elite group W, then to obtain $\gamma_{W \cup \{j\}}$, we (i) add to γ_W the number of arcs from $V \backslash \{W \cup \{j\}\}$ to j and (ii) subtract the number of arcs from node j to the nodes in W. Thus, depending on this trade-off, $\gamma_{W \cup \{j\}}$ can be larger or smaller or the same as γ_W.

2.1.2. Origin and Applications. The notion of an "elite" group originated from efforts to examine and understand social behavior within a close-knit community. In the 1980s, sociologist Li Fan analyzed the giving (and receiving) of gifts between the residents of a Mongolian town (Wellman et al. 2001) and found that one (elite) block of residents received gifts from the others but only exchanged gifts among each other. Thus, as a set, this group of residents only received gifts from the other members of the town. Another example of the notion of an elite group occurs in the analysis of the advice-seeking behavior of the members of a school, reported in Hawe and Ghali (2008). Here, the social network revealed that, together, the principal, the vice principal, and some key technical staff form a group with properties such that (i) most of the other staff members seek advice from one or more members of this group, and (ii) the members of the group typically seek advice only from (one or

more) members within the group. Thus, to influence opinion within the community in general, it may be beneficial to first convince this group of individuals.

In the context of social network analysis, the members of an elite group can be regarded as key players or opinion leaders. For instance, if a specific member of a community is often consulted by other members on (say) health issues, then she can be regarded as a key player (an elite member) in the opinion-seeking network of that community (Borgatti 2006). Another example is the cosponsorship network in the United States Senate (Fowler 2006). In this network, the prominent senators typically receive a significant amount of cosponsorship. Thus, the set of these prominent senators constitute an (approximate) elite group.

2.2. The Portal Problem and The Exact-Size Portal Problem

2.2.1. Technical Definition

Instance: n players; a connected, undirected graph $G = (V, E)$, $|V| = n$, in which the nodes represent the players and edges represent the pairwise connections between the players; a positive integer $k \leq n$.

Solution: For PP, a set $Q \subseteq V$ such that $|Q| \leq k$. For the exact-size portal problem (ESPP), a set $Q \subseteq V$ such that $|Q| = k$; see Figure 2.

Objective Function: Maximize $r(Q)$, defined as follows:

$$r(Q) = \frac{BC(Q)}{\binom{n-|Q|}{2}} \quad \text{and} \quad BC(Q) = \sum_{s \notin Q,\, t \notin Q,\, s \neq t} \frac{\sigma_{st}(Q)}{\sigma_{st}},$$

where σ_{st} is the total number of shortest paths from node s to node t, where $s, t \in V \setminus Q$, $s \neq t$, and $\sigma_{st}(Q)$ is the number of shortest paths from node s to node t that have at least one node in set Q as an internal node.

2.2.2. Previous Work and Applications. PP is a natural extension of the popular betweenness centrality (BC) measure (Freeman 1979, Scott 2000) for individual nodes (members) of a social network; for $k = 1$, an optimal solution to PP is a node with the highest BC. Everett and Borgatti (1999) extend the notion of BC to groups and illustrate the measure on a few examples. For a *given* set of nodes Q, Puzis et al. (2007) provide a polynomial-time algorithm to compute the nonnormalized measure $BC(Q)$ (referred to as "group betweenness centrality"). They prove that the problem of obtaining a set with the highest group betweenness centrality (i.e., $\max_{Q \subseteq V} BC(Q)$) is NP-hard and also propose a simple heuristic. Note that the behavior of our normalized measure $r(Q)$ can be fundamentally different from that of $BC(Q)$. In general, the set that maximizes the group betweenness centrality may not necessarily be an optimal solution of our model, and vice versa. Puzis et al. (2007) discuss an interesting application of a network of computers in which a limited number of virus-cleaning devices need to be placed at a subset of nodes (computers) to prevent the spread of viruses. To maximize the utility of the devices, it is beneficial to place them at the nodes of a portal of an appropriate size. Another interesting application where a portal may need to be identified is in a disease-outbreak network. For example, Klovdahl et al. (2001) describe a tuberculosis outbreak network and motivate the need to identify the critical members in this network to control the spread of the disease. Everett and Borgatti (1999) discuss the interaction network of animals (monkeys) and use the notion of a portal to determine a socially central set of animals.

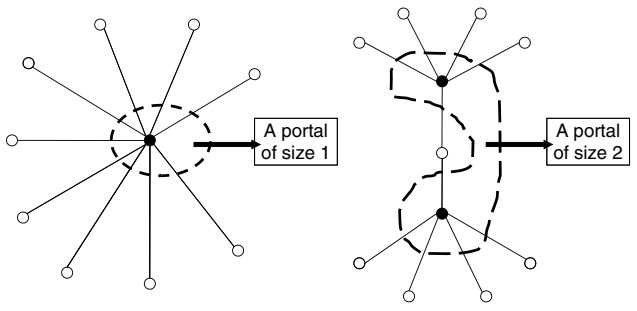

Figure 2 **Optimal Portals in Two Simple Networks**

3. Algorithmic Analysis

We now analyze EGP and PP. For a search problem, a basic question is that of its computational complexity. For simple networks, an optimal solution to both problems is easy to obtain. For EGP, we first illustrate this and then identify a structural property of an elite group that can help in reducing the size of the underlying graph. Then we show that EGP is polynomially solvable for a general network. Next, motivated by practical considerations, we introduce a size-constrained version of EGP together with its penalty-based relaxation and show that both are strongly NP-hard. For PP, we first show that PP is strongly NP-hard on a general graph. We then consider several special graphs on which PP is polynomially solvable. Finally, we discuss a heuristic for general trees.

3.1. The Elite Group Problem

Given a directed graph $G(V, A)$, recall that an elite group is a set $W \subseteq V$ such that there does not exist any directed arc $(i, j) \in A$ with $i \in W$, $j \notin W$. The objective of EGP is to maximize the total number (or score) γ_W of directed arcs incident on the nodes in W. For some simple networks, it may be

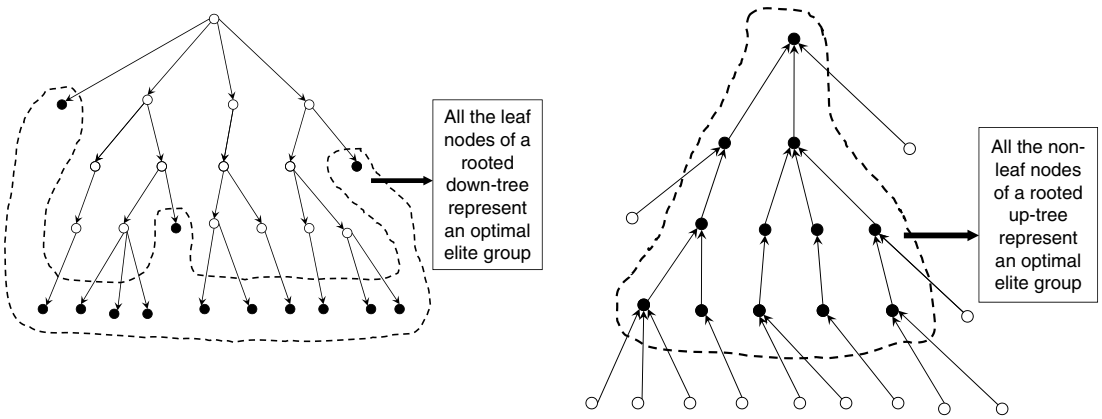

Figure 3 Optimal Elite Group for a Rooted Down-Tree and a Rooted Up-Tree

straightforward to prove the optimality of a specific elite group. Rooted up- and down-trees are especially useful networks to study because they represent hierarchically organized structures, e.g., reporting relationships in a department, natural taxonomies (Cross and Parker 2004). The proof of the following observation is included in the Online Supplement (available at http://dx.doi.org/10.1287/ijoc.1110.0473).

Observation 1. If the graph G is a rooted down-tree (i.e., each node in G, except the root, has a unique predecessor, and all arcs in G are directed downwards from the root to the leaf nodes; see Figure 3), then the elite group W^* consisting of all the leaf nodes of G is an optimal elite group. If the graph G is a rooted up-tree (i.e., each node in G, except the root, has a unique successor, and all arcs in G are directed upward from the leaf nodes toward the root; see Figure 3), then the elite group W^* consisting of all nonleaf nodes of G is an optimal elite group.

Our next result helps us "shrink" the strongly connected components (e.g., directed cycles) in G to single nodes in our search for an elite group. We will use this result later in the proof of Theorem 2. The proof of the following observation is included in the Online Supplement.

Observation 2. If G contains a strongly connected component, and at least one node on this component belongs to an elite group W (respectively, the complement $\bar{W} = V \backslash W$), then all the other nodes on the component must belong to W (respectively, $\bar{W}$).

Note that there are many polynomial algorithms to find a strongly connected component (if one exists) in a graph. If G contains a strongly connected component C, then, by using Observation 2, we can shrink C into a single node. An arc in the original graph between a node $v \in V \backslash C$ and a node of C is represented in the *shrunk graph* as between v and the (shrunk) node representing the component. Thus, in the shrunk graph, we use a separate new arc to represent each arc between a node in $V \backslash C$ and a node of C in the original graph. Therefore, in general, the shrunk graph becomes a multigraph because there may be parallel arcs between two nodes. We can continue this type of shrinking until there is no nontrivial strongly connected component in the modified shrunk graph. Consequently, we can assume, without loss of generality, that the network is a directed acyclic graph (DAG). Therefore, EGP translates into finding a sink set of maximum indegree in a DAG. The following result follows immediately from Observation 2.

Lemma 1. *There is a one-to-one correspondence between elite groups in G and elite groups in the shrunk graph: for every elite group W in G, we can get one in the shrunk graph with the same score by taking the nodes of the shrunk graph corresponding to all the strongly connected components in W. Conversely, for every elite group W' in the shrunk graph, we can get one in G with the same score by taking the nodes in all the strongly connected components corresponding to the nodes in W'.*

Next, we show that EGP is polynomially solvable.

Theorem 1. *The EGP is polynomially solvable.*

Proof. For $j \in V$, define $\pi_j \in \{0, 1\}$ as follows:

$$\pi_j = \begin{cases} 1, & \text{node } j \text{ belongs to the elite group } W; \\ 0, & \text{otherwise.} \end{cases}$$

Then, an integer programming (IP) formulation for EGP is as follows:

$$\begin{aligned} \max \quad & \sum_{(i,j) \in A} (\pi_j - \pi_i) \\ \text{s.t.} \quad & \pi_i - \pi_j \le 0 \quad \forall (i,j) \in A, \\ & \pi_i \in \{0,1\} \quad \forall i \in V. \end{aligned}$$

For a directed arc (i, j), if node i is in W (i.e., $\pi_i = 1$), then node j must also be in W (i.e., $\pi_j = 1$). Otherwise, if $\pi_i = 0$, then $\pi_j \in \{0, 1\}$. This is enforced by the

first constraint. In the objective function, $(\pi_j - \pi_i)$ represents the contribution of arc (i, j) to γ_W: if nodes i and j are both in W or both in $V \setminus W$ (i.e., $\pi_i - \pi_j = 0$), then the contribution is 0. If node i is in $V \setminus W$ and node j is in W (i.e., $\pi_i = 0$, $\pi_j = 1$), then the contribution is 1. The constraints of the IP can be written as $A\pi \leq 0$, where A is the node–arc incidence matrix of G and $\pi \in \{0, 1\}^{|V|}$. It is well known that the node–arc incidence matrix of a *directed* graph is totally unimodular (see, e.g., Hoffman and Kruskal 1956, Nemhauser and Wolsey 1988). Thus, the linear programming relaxation of the above IP results in an integer optimum. The result follows. □

Note that the shrinking of strongly connected components (Lemma 1) maintains the total unimodularity of the constraint matrix of the IP above. Thus, the size of a network containing strongly connected components can be reduced before formulating the EGP. The objective function of the IP above is to maximize the number of arcs directed into the elite group; instead, if we change the objective function to maximize a weighted linear combination of arcs directed into the elite group, the modified problem remains polynomially solvable.

Remark 1 (Penalizing the Size of the Elite Group). Note that EGP does not impose any constraint on the cardinality (i.e., the number of nodes) of the elite group. In §§3.1.1 and 3.1.2, we consider a hard constraint on the cardinality. An alternative is to impose a "soft" constraint by imposing a penalty $p > 0$ on the cardinality. In this case, the objective function in the IP above changes to $\max \sum_{(i,j)\in A} (\pi_j - \pi_i) - p \sum_{i \in V} \pi_i$. Because this modified objective is linear and the constraint matrix is totally unimodular, the modified problem remains polynomially solvable.

3.1.1. The Size-Constrained Elite Group Problem. Typically, the purpose of identifying an elite group is to use the members of this group to effectively influence the other members of the social network. Thus, for practicability in managing this subsequent task, the size of an elite group may need to be restricted. Motivated by this requirement, Theorem 2 discusses the complexity of the *size-constrained elite group problem* (*SCEGP*), defined as follows: Given a positive integer $k \leq n$, find an optimal elite group $W \subseteq V$ with $|W| \leq k$.

Theorem 2. *The decision problem corresponding to SCEGP is strongly NP-complete.*

Proof. The strongly NP-complete problem that we use in our reduction is the balanced biclique problem (Garey and Johnson 1979), defined as follows.

Balanced Biclique Problem (BBP)

Instance: An undirected bipartite graph $G = (U \cup V, E)$, with $|U| = |V| = n$; a positive integer $k \leq n$.

Solution: An induced subgraph $G_1 \subseteq G$ such that $G_1 = (U_1 \cup V_1, E_1)$, $U_1 \subseteq U$, $V_1 \subseteq V$, $|U_1| = |V_1| = k$, $E_1 \subseteq E$, and $u_1 \in U_1$, $v_1 \in V_1$ implies that $\{u_1, v_1\} \in E_1$. The *size* of the biclique is $2k$.

Given an arbitrary instance of BBP specified by G, we construct an instance of SCEGP on a related graph G'. The construction of G' is done in two steps. First, we obtain G^c, the bipartite complement graph of G. Then, we add two additional node sets O and S, extend each node in U into a directed cycle, and give directions to all edges to get G'. We now explain our construction and illustrate with an example of G in Figure 4:

Step 1. Get G^c, the bipartite complement graph of G (see Figure 4).

Step 2. We add two node sets O and S consisting, respectively, of n^3 and n^2 nodes. The nodes of O (respectively, S) form a directed cycle. There is a directed arc from each node $o_i \in O$ to each node in U. There is a directed arc from each node in V to each node $s_i \in S$. Let $m = n + n^2$. Next, we extend each node $u_i \in U$ into a length m directed cycle C_i by adding

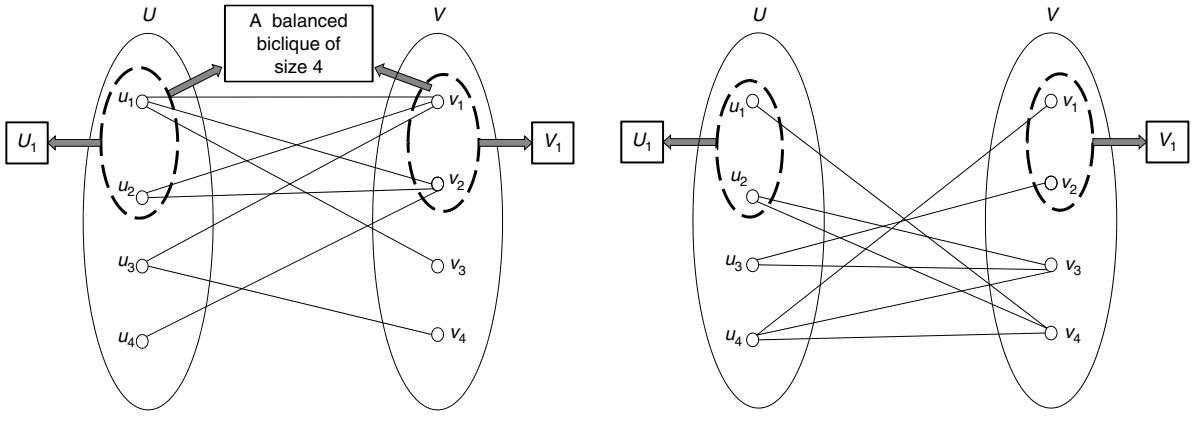

Figure 4 A Bipartite Graph G with a Balanced Biclique, and Its Bipartite Complement Graph G^c, Which Is Used in the Proof of Theorem 2

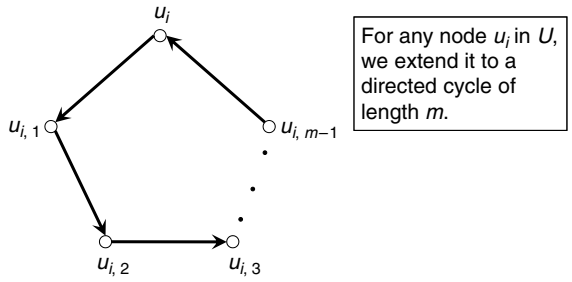

Figure 5 The Widget, a Directed Cycle with Length m, Used in the Proof of Theorem 2

$m-1$ additional nodes $(u_{i,1}, u_{i,2}, \ldots, u_{i,m-1})$ (see Figure 5). Let $U' = \{u_i, u_{i,1}, u_{i,2}, \ldots, u_{i,m-1} \mid u_i \in U,\ i = 1, 2, \ldots, n\}$. The edges between O and U' are directed from O to U', those between U' and V are directed from U' to V, and those between V and S are directed from V to S. The construction of G' is now complete (see Figure 6). Let $N = O \cup U' \cup V \cup S$. On G', consider the following decision question for *SCEGP*:

Decision Question: Letting $t = km + (n-k) + n^2$ and $D = kn^3 + kn^2$, does there exist an elite group W in G' such that $|W| \le t$, and $\gamma_W \ge D$?

Note that the construction of the decision problem from the given instance of the BBP is polynomially bounded. That is, the total number of nodes in G' is bounded by a polynomial in n, as is the time necessary to construct a description of the input of the decision problem. The decision problem is clearly in class NP. We now show that the decision question has an affirmative answer if and only if the original graph G contains a balanced biclique of size $2k$ (i.e., $|U_1| = |V_1| = k$).

$\Longrightarrow$ Suppose $U_1 \cup V_1$ is a balanced biclique of size $2k$ in G. Let $U_2 = U \backslash U_1$, $V_2 = V \backslash V_1$. In G', let $U_1' = \{C_i \mid u_i \in U_1\}$, $U_2' = \{C_i \mid u_i \in U_2\}$, $W = U_1' \cup V_2 \cup S$, and

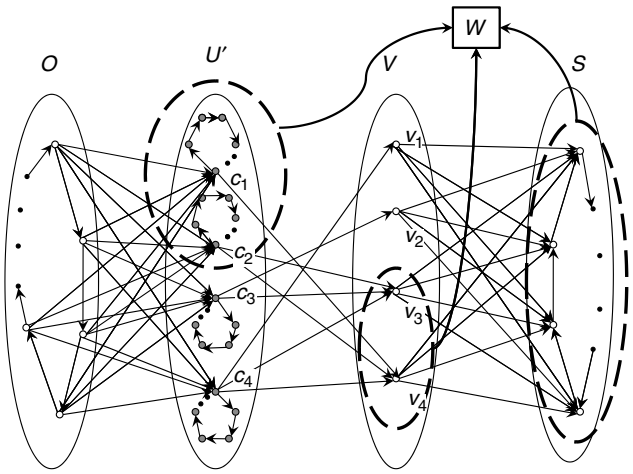

Figure 7 Graph G' with Elite Group Set W

$\bar{W} = O \cup U_2' \cup V_1$ (see Figure 7). We now show that the set W is an elite group that provides an affirmative answer to the decision question.

First, we need to prove the set W is a valid elite group in G'; i.e., there is no arc from W to $\bar{W}$. Since $U_1 \cup V_1$ is a biclique of G, then there is no arc from U_1' to V_1 in G'. Since G is bipartite, there is no arc between U_1' and U_2'. Also, by construction, there is no arc from U_1' to O. Thus, there is no arc from U_1' to $\bar{W}$. Similarly, there is no arc from V_2 to $\bar{W}$ and from S to $\bar{W}$. Thus, W is a valid elite group.

Next, observe that $|W| = |U_1'| + |V_2| + |S| = km + (n-k) + n^2 = t$. Finally, note that γ_W is the number of arcs from $\bar{W}$ to W. The number of arcs from O to U_1' (respectively, V_1 to S) is kn^3 (respectively, kn^2). Also, the number of arcs from U_2' to V_2 is nonnegative. Thus, $\gamma_W \ge kn^3 + kn^2 = D$. The result follows.

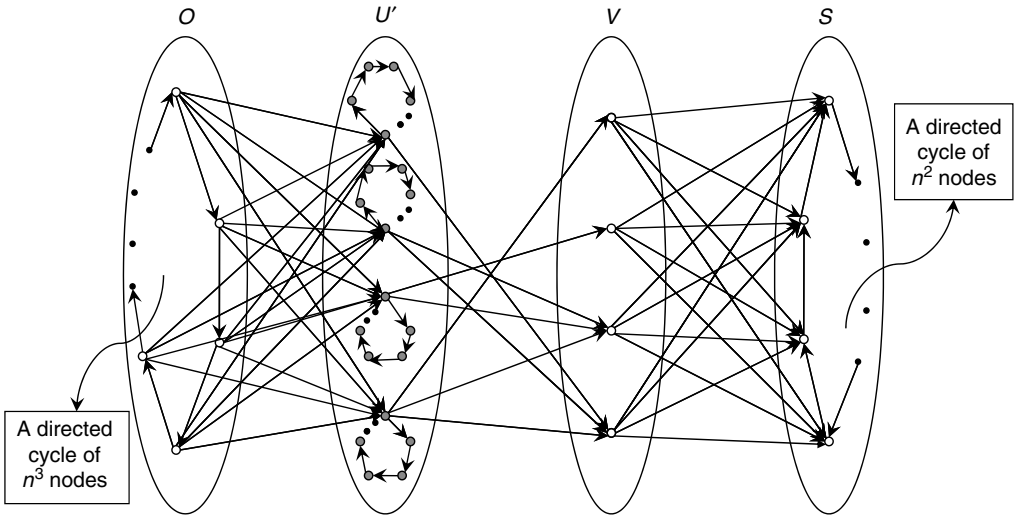

Figure 6 The Constructed Graph G' for *SCEGP*

$\Longleftarrow$ Suppose W is an elite group in G' with $|W| \le t$ and $\gamma_W \ge D$. Let $\bar{W} = N \backslash W$. The following claims characterize the set W.

Claim 1. *In G', the nodes in C_i either all belong to W or all belong to $\bar{W}$. Similarly, the nodes in S (respectively, O) either all belong to W or all belong to $\bar{W}$.*

Proof of Claim 1. The nodes in C_i (respectively, S, O) form a directed cycle. The result follows from Observation 2. □

Claim 2. *Each node in O must belong to $\bar{W}$. Similarly, each node in S must belong to W.*

Proof of Claim 2. Suppose a node in O belongs to W. Then, from Claim 1, each node in O belongs to W. Also, from the definition of elite group, each node in U' must belong to W. Consequently, $|W| \ge |O| + |U'| = n^3 + nm$. Since $n \ge 2$ and $n \ge k$, we have $n^3 > n^2 + n > n^2 + n - k$ and $nm \ge km$. So $n^3 + nm > (n^2 + n - k) + km$, which implies $|W| > t$. This contradicts the assumption that $|W| \le t$. Thus, each node in O must belong to $\bar{W}$.

Suppose a node in S belongs to $\bar{W}$. Then, from Claim 1, each node in S belongs to $\bar{W}$. Also, each node in V must belong to $\bar{W}$. As shown above, each node in O is in $\bar{W}$. Thus, only a subset $Q' \subseteq U'$ can belong to W. Let $Q = U \cap W$. Note that $|W| = |Q'| = m|Q|$. Since $m = n + n^2$ and $|W| \le t = n^2 + km + n - k = km + m - k = (k+1)m - k$, we have $|W| = m|Q| \le (k+1)m - k$, so $|Q| \le k$. Thus $\gamma_W = n^3|Q| \le n^3 k < kn^3 + kn^2 = D$, which contradicts the assumption that $\gamma_W \ge D$. Thus, each node in S belongs to W. □

As a consequence of Claim 2, we have $W = U_1' \cup V_2 \cup S$ and $\bar{W} = O \cup U_2' \cup V_1$. Let $U_1 = \{u_i \mid C_i \in U_1'\}$.

Claim 3. $|U_1| = k$.

Proof of Claim 3. We first show that $|U_1| \le k$. Suppose $|U_1| \ge k+1$; then $|W| \ge |U_1'| = |U_1| m \ge (k+1)m = km + m$. Since $m = n + n^2 > (n-k) + n^2$, we have $|W| \ge km + m > km + n - k + n^2 = t$, which contradicts the assumption that $|W| \le t$. Thus, $|U_1| \le k$.

Next, we show that $|U_1| \ge k$. Suppose $|U_1| \le k-1$. Let $|V_1| = h$. Then, $|V_2| = |V| - |V_1| = n - h$. Recall that γ_W is the number of arcs from $\bar{W}$ to W.

The number of arcs from O to U_1' (respectively, from V_1 to S and from U_2' to V_2) is $n^3|U_1| \le n^3(k-1)$ (respectively, hn^2 and $\le n|V_2| = n(n-h)$). Thus $\gamma_W \le n^3(k-1) + hn^2 + n(n-h) = kn^3 - n^3 + n^2 + h(n^2 - n)$. Since $n^2 - n > 0$ and $0 \le h \le n$, $(n^2 - n)h$ reaches its maximum when $h = n$. Thus $kn^3 - n^3 + n^2 + h(n^2 - n) \le kn^3 - n^3 + n^2 + n(n^2 - n) = kn^3 < kn^3 + kn^2 = D$. Thus, $\gamma_W < D$, contradicting the assumption that $\gamma_W \ge D$. Thus, $|U_1| \ge k$. The result follows. □

Claim 4. $|V_1| \ge k$.

Proof of Claim 4. Note that $|W| = |U_1'| + |V_2| + |S| = km + |V_2| + n^2 \le t = km + (n-k) + n^2$. Thus, $|V_2| \le n - k$. Since $|V_1| = n - |V_2|$, we have $|V_1| \ge k$. □

Note that $U_1' \subseteq W$, $V_1 \subseteq \bar{W}$. Then, from the definition of an elite group, there is no arc from U_1' to V_1 in G'. Since G' is the bipartite complement graph of G, there is an edge between each node in U_1 and each node in V_1 in G. Since $|U_1| = k$, $|V_1| \ge k$, there exists at least one balanced biclique of size $2k$ in G. This concludes the proof of Theorem 2. □

3.1.2. Relaxing the Structure of the Elite Group. Because of the combined requirements of cardinality and structure, the SCEGP of §3.1.1 is not guaranteed to always have a nontrivial feasible solution. However, in practice, we may sometimes prefer to identify a nonempty group of players who can influence a large number of players outside the group but are also influenced by a few outsiders. To enable such solutions, we relax the constraint that forbids directed arcs from nodes of an elite group W to $V \backslash W$. Instead, for each arc, we impose a penalty $p > 0$ that is specified as part of the input. The objective function is the number of incoming arcs into W minus p times the number of arcs coming out of W. We refer to this problem as the *size-constrained elite group problem with penalties* (*SCEGPP*).

Technical Definition

Instance: n players; an "influence" social network represented by a directed graph $G(V, A)$, $|V| = n$, in which the nodes represent the players and the set of arcs represent pairwise influences pertaining to a social property: a directed arc (i, j) indicates that i is influenced by j. A positive integer $k \le n$. A positive number p.

Solution: A set $W \subseteq V$ such that $|W| \le k$.

Objective Function: Maximize γ_W, the number of arcs from $V \backslash W$ to W minus p times the number of arcs from W to $V \backslash W$. That is, $\gamma_W = \sum_{i \notin W,\, j \in W} a_{ij} - p \sum_{i \in W,\, j \notin W} a_{ij}$, where $a_{ij} = 1$ if $(i, j) \in A$, and 0 otherwise.

IP Formulation. For $j \in V$, we define $\pi_j \in \{0, 1\}$ as follows:

$$\pi_j = \begin{cases} 1, & \text{node } j \text{ belongs to the elite group } W; \\ 0, & \text{otherwise.} \end{cases}$$

For $(i, j) \in A$, we define $x_{ij} \in \{0, 1\}$ and $y_{ij} \in \{0, 1\}$ as follows:

$$x_{ij} = \begin{cases} 1, & \text{arc } (i, j) \text{ is from } V \backslash W \text{ to } W; \\ 0, & \text{otherwise;} \end{cases}$$

$$y_{ij} = \begin{cases} 1, & \text{arc } (i, j) \text{ is from } W \text{ to } V \backslash W; \\ 0, & \text{otherwise.} \end{cases}$$

An IP formulation for SCEGPP is as follows:

$$\begin{aligned}
\max\ & \gamma_W = \sum_{(i,j)\in A} (x_{ij} - p y_{ij}) \\
\text{s.t.}\ & \pi_i - \pi_j = y_{ij} - x_{ij} \quad \forall (i,j) \in A, \\
& y_{ij} + x_{ij} \le 1 \quad \forall (i,j) \in A, \\
& \sum_{i\in V} \pi_i \le k, \\
& x_{ij}, y_{ij} \in \{0,1\} \quad \forall (i,j) \in A, \\
& \pi_i \in \{0,1\} \quad \forall i \in V.
\end{aligned}$$

If node $i \in W$ and node $j \in V\backslash W$ (i.e., $\pi_i = 1$, $\pi_j = 0$), then the first constraint enforces that $y_{ij} = 1$, $x_{ij} = 0$. Similarly, if node $i \in V\backslash W$ and node $j \in W$ (i.e., $\pi_i = 0$, $\pi_j = 1$), then $y_{ij} = 0$, $x_{ij} = 1$. For $\pi_i = \pi_j = 0$ or $\pi_i = \pi_j = 1$, we have $y_{ij} - x_{ij} = \pi_i - \pi_j = 0$, and hence, $y_{ij} = 0$, $x_{ij} = 0$, because of the second constraint. Thus, in the objective function, $(x_{ij} - p y_{ij})$ represents the contribution of arc (i,j) to γ_W: (i) if arc (i,j) is from $V\backslash W$ to W, then the contribution is 1; (ii) if arc (i,j) is from W to $V\backslash W$, then the contribution is $-p$; and (iii) if arc (i,j) is from W to W (or from $V\backslash W$ to $V\backslash W$), then the contribution is 0. The following theorem discusses the computational complexity of SCEGPP.

Theorem 3. *The decision problem corresponding to SCEGPP is strongly NP-complete.*

Proof. We again use BBP in our reduction. The notation is as defined in §3.1.1. The construction of the graph G' is exactly the same as in the proof of Theorem 2. The decision question, however, is different.

Decision Question: Letting $t = km + (n-k) + n^2$, $p = n^4 + 3n^3 + 3n^2$, and $D = kn^3 + kn^2$, does there exist an elite group W in G' such that $|W| \le t$, and $\gamma_W \ge D$?

The decision problem is clearly in class NP. It is easy to see that the decision question has an affirmative answer if and only if the original graph G contains a balanced biclique of size $2k$ (i.e., $|U_1| = |V_1| = k$).

$\Longrightarrow$ This part is exactly the same as in the proof of Theorem 2.

$\Longleftarrow$ Suppose W is an elite group in G' with $|W| \le t$ and $\gamma_W \ge D$.

Claim 5. *In G', there does not exist any arc from W to $\bar{W}$.*

Proof of Claim 5. If there exists at least one arc from W to $\bar{W}$, then $\gamma_W \le \sum_{(i,j)\in A} x_{ij} - p$. Since $\sum_{(i,j)\in A} x_{ij} \le |G'| \le p$, we have $\gamma_W \le 0$, which contradicts $\gamma_W \ge D$. □

With Claim 5, the remainder of the argument is the same as in the proof of Theorem 2. □

3.2. The Portal Problem

Given a connected, undirected graph $G(V,E)$ and a positive integer k, recall from §2 that an optimal portal is a set $Q \subseteq V$, $|Q| \le k$ such that $r(Q)$ is maximized.

As mentioned earlier, a portal is a natural extension to a set-based measure of the notion of BC for a single node. For $k = 1$, PP reduces to the well-known betweenness centrality problem, which is polynomially solvable (Everett and Borgatti 1999). Thus, PP is polynomially solvable when $k = 1$. However, for higher values of k, finding an optimal solution is often a challenging task. The primary difficulty is that the measure $r(Q)$ is *nonadditive*. In other words, BCs of two distinct nodes in Q cannot, in general, be simply added when computing $r(Q)$. This is obvious because a specific path between nodes i and j, $i, j \in V\backslash Q$ with two or more internal nodes in Q is counted only once in the computation of $r(Q)$.

We first show that PP and ESPP are strongly NP-hard in §3.2.1. An efficient polynomial-time algorithm for obtaining an optimal solution on general graphs is, therefore, unlikely. Then, we address special graphs (bicliques and balanced and full d-trees) in §3.2.2. Finally, in §3.2.3, we analyze a heuristic for general trees.

3.2.1. Proof of Hardness of PP and ESPP. Puzis et al. (2007) use the vertex cover problem (Garey and Johnson 1979) to show the hardness of the nonnormalized measure $BC(Q)$ (see §2.2), which is the numerator of our measure $r(Q)$. Furthermore, given G and k, the variant considered in Puzis et al. (2007) requires that the solution have exactly k nodes. The strongly NP-complete problem that we use in our reduction is the independent set problem (Garey and Johnson 1979).

Independent Set Problem (ISP)

Instance: A connected, undirected graph $G = (V,E)$; a positive integer $k \le |V|$.

Solution: A set of nodes, $I \subseteq V$, $|I| \ge k$, such that no two nodes in I are connected by an edge in E.

Theorem 4. *The decision problem corresponding to PP is strongly NP-complete.*

Proof. Given an arbitrary instance of ISP, specified by $G(V,E)$, we consider the following decision problem:

Decision Question: Does there exist a portal $Q \subseteq V$ in $G(V,E)$ such that $|Q| \le |V| - k$ and $r(Q) \ge 1$?

Note that the decision problem is clearly in class NP. We now show that ISP has an affirmative answer if and only if the above decision question has an affirmative answer.

Suppose I^* is an independent set in G with at least k^* nodes. Let $Q^* = V\backslash I^*$. Then, $|Q^*| \le |V| - k^*$. From the definition of an independent set, it follows that all

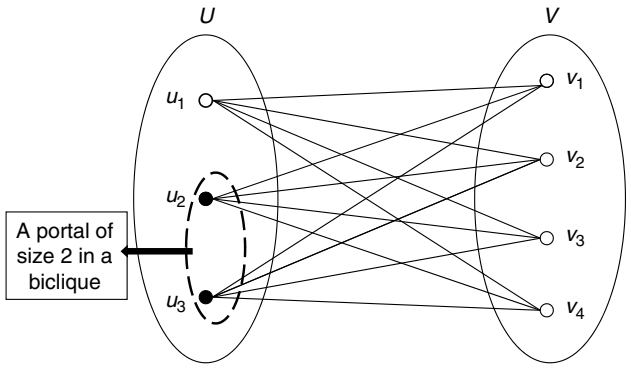

Figure 8 Optimal Portal in a Biclique

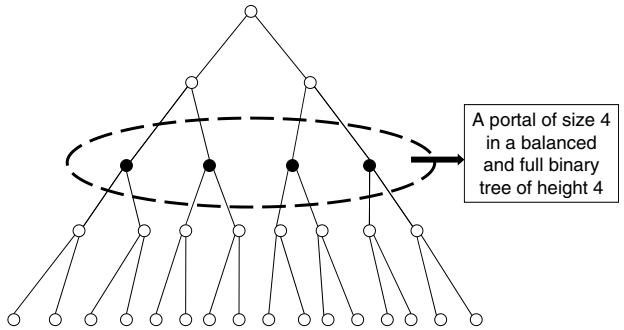

Figure 9 Optimal Portal in a Balanced and Full Binary ($d = 2$) Tree

paths in G between any two nodes in I^* have at least one node in Q^* as an internal node. Thus, $r(Q^*) = 1$, and the decision question has an affirmative answer. Conversely, if there exists $Q \subseteq V$ with $|Q| \le |V| - k$ and $r(Q) \ge 1$, then the set $V \setminus Q$ is an independent set of at least k nodes. □

Corollary 1. *The decision problem corresponding to ESPP is strongly NP-complete.*

3.2.2. Results on Specific Families of Graphs. We now discuss two specific families of graphs: bicliques and balanced and full d-trees.

Bicliques. Let $G = (U \cup V, E)$ be a biclique: $n_1 = |U| \le |V| = n_2$, and $u \in U$, $v \in V$ implies that $\{u, v\} \in E$. The *size* of the biclique is $n_1 + n_2$. Let $Q_1, Q_2 \subseteq U \cup V$. If $|Q_1 \cap U| = |Q_2 \cap U|$ and $|Q_1 \cap V| = |Q_2 \cap V|$, then $BC(Q_1) = BC(Q_2)$. Thus, for $Q \subseteq U \cup V$, the objective function $r(Q)$ depends only on two numbers: $k_1 = |Q \cap U|$ and $k_2 = |Q \cap V|$. Theorem 5 (respectively, Corollary 2) provides an optimal solution to PP (respectively, ESPP); see Figure 8. The proof of Theorem 5 is included in the Online Supplement.

Theorem 5. *Let $G = (U \cup V, E)$ be a biclique with $n_1 = |U| \le |V| = n_2$. Let $Q \subseteq U \cup V$. Let $k_1 = |Q \cap U|$, $k_2 = |Q \cap V|$. Then,*

(a) *For $1 \le k \le n_1 - 1$, any set Q that satisfies $k_1 = k$ and $k_2 = 0$ is an optimal solution of PP.*

(b) *For $k \ge n_1$, then $Q = U$ is an optimal solution of PP.*

We also summarize the solution of ESPP.

Corollary 2. *Let $G = (U \cup V, E)$ be a biclique with $n_1 = |U| \le |V| = n_2$. Let $Q \subseteq U \cup V$. Let $k_1 = |Q \cap U|$, $k_2 = |Q \cap V|$.*

1. For $1 \le k \le n_1 - 1$, any set Q that satisfies $k_1 = k$ and $k_2 = 0$ is also an optimal solution of ESPP.

2. For $n_1 \le k \le n_1 + n_2 - 1$, then any set Q that satisfies $k_1 = n_1$ and $k_2 = k - n_1$ is an optimal solution of ESPP.

Balanced and Full d-Trees. Given a tree $G(V, E)$ and $Q \subseteq V$, let $G'(Q)$ denote the induced subgraph obtained by removing all the nodes in Q from G. In general, $G'(Q)$ is a forest with disjoint trees as its connected components. Since G is a tree, there is a unique path in G connecting any two distinct nodes s and t in $V \setminus Q$; thus, $\sigma_{st} = 1$ (see §2.2). We first define some notation for a general tree $G(V, E)$:

n: the number of nodes in G (i.e., $n = |V|$).
k: the number of nodes in Q (i.e., $k = |Q|$).
l: the number of connected components in $G'(Q)$.
A_i: the ith connected component in $G'(Q)$, $i = 1, 2, \ldots, l$.
a_i: the *size* (i.e., the number of nodes) of component A_i, $i = 1, 2, \ldots, l$.

Consider a connected component, say, A_i, of $G'(Q)$. In G, there is a unique path from any node in A_i to each node in every other connected component in $G'(Q)$. Thus,

$$BC(Q) = \sum_{s \notin Q,\, t \notin Q,\, s \ne t} \frac{\sigma_{st}(Q)}{\sigma_{st}} = \sum_{1 \le i < j \le l} a_i a_j. \tag{1}$$

Since $\sum_{i=1}^{l} a_i = |V| - |Q| = n - k$, we have

$$BC(Q) = \frac{(n-k)^2 - \sum_{i=1}^{l} a_i^2}{2}. \tag{2}$$

Thus, for fixed n and k, maximizing $BC(Q)$ is equivalent to minimizing $\sum_{i=1}^{l} a_i^2$. We next illustrate the solution of this problem for balanced and full d-trees; see Figure 9.

On a rooted balanced and full d-tree, each node (except the leaf nodes) has d distinct successors, and each node (except root) has a unique predecessor. All leaf nodes have the same distance (height) to the root node. For a d-tree, if we remove any node other than the root node and leaf nodes, we will add d more connected components into the remaining graph. So if we remove k nodes from a d-tree, we will have at most $l = dk + 1$ connected components left. The proof of Theorem 6 is included in the Online Supplement.

Theorem 6. *Let G be a balanced and full d-tree with height $h \ge 2$. For an instance of PP defined by G and a*

positive integer k, let $t = \min\{\lceil h/2 \rceil, \lfloor \log_d k \rfloor\}$, and let $\bar{Q}$ denote the set of nodes on the tth level of G. Then, $\bar{Q}$ provides an asymptotically maximal solution to PP, with $r(\bar{Q}) \geq (1 - 1/d^{t+1})$.

Since $t = \min\{\lceil h/2 \rceil, \lfloor \log_d k \rfloor\}$, the ratio $r(\bar{Q}) \to 1$ with an increase in the size of G and k. Thus, the resulting family of solutions $\bar{Q}$ can be referred to as an asymptotically maximal family. Note that the solution $\bar{Q}$ is not necessarily optimal. Thus, in cases where $\bar{Q}$ is not optimal, there can exist a solution that is superior to $\bar{Q}$.

Remark 2. The solutions of PP and ESPP on paths, cycles, and cliques are straightforward to obtain. Therefore, these results are listed without proofs in the Online Supplement.

3.2.3. General Trees. An important question that arises naturally is that of the solution of PP on a general tree. The tree structure occurs frequently in real-world social networks. Wein (2009) describes the milk supply chain as a tree. The author argues that for a potential terrorist attack, it is enough to introduce a small amount of toxin (botulinum) at a few key nodes of the tree. We believe that such nodes can be identified by finding an optimal or near-optimal portal in the tree. In Perer and Wilson (2007), the authors discuss the underground distribution network of steroids among players of Major League Baseball. Investigators have used this network to determine the role of each individual in the distribution of steroids. Again, knowledge of a good portal in this network should help identify key members. Interfirm collaboration networks, studied in Schilling and Phelps (2007), are approximately trees. In their search for firms with a higher innovative output, the authors find a set of nodes that is a near-optimal portal. In Hanaki et al. (2007), the authors argue that locally, a large and sparse random network often resembles a pure branching tree.

The computational complexity of PP for a tree is an open problem. We now present a simple heuristic (Theorem 7) to find a portal in a tree and obtain a lower bound on its performance. To describe the heuristic, we need the following labeling procedure.

Labeling Procedure
Input: A tree $G(V, E)$.
Initialization: Let $i = 0$.
Step 1: Select all the leaf nodes of G, label them as being on level i, and include them in set $S(i)$. Let $G = G \backslash S(i)$.
Step 2: If $G = \varnothing$, terminate; otherwise, let $i = i + 1$ and go to Step 1.

We record the highest level we get in this labeling procedure as h, and refer to it as the *height* of the tree. Let $n[j] = |S(j)|$, the number of nodes in level j. It is easy to see that $n[0] \geq n[1] \geq n[2] \geq \cdots \geq n[h]$ and $n[h] \in \{1, 2\}$.

Theorem 7. *Let G be a tree with height $h \geq 2$. For an instance of PP defined by $G(V, E)$ and a positive integer $k \geq 2$,*

1. *If $n[1] \leq k$, then set $\bar{t} = 1$.*
2. *If $n[1] > k$, then set $\bar{t}$ such that $n[\bar{t} - 1] > k$, $n[\bar{t}] \leq k$.*

Let $t = \max\{\lceil h/2 \rceil, \bar{t}\}$. Let $\bar{Q}$ denote the set of nodes on the level t of G. Then, $\bar{Q}$ provides a solution to PP with $r(\bar{Q}) \geq (b(b-1)t^2 + 2bt(h-t))/((n-b)(n-b-1))$, where $b = n[t]$ and $n = |V|$. Moreover, this bound is tight.

Proof. Since $\bar{Q}$ includes all the nodes on level t of G, we have $|\bar{Q}| = n(t)$. Since $t = \max\{\lceil h/2 \rceil, \bar{t}\} \geq \bar{t}$, we have $n(t) \leq n[\bar{t}] \leq k$. Thus, $|\bar{Q}| \leq k$. Note that $G'(\bar{Q})$ has $l \geq b + 1$ connected components. Of these, we have (i) at least b components, say, A_i, $i = 1, 2, \ldots, b$, each with at least t nodes: from level 0 to level $t - 1$, A_i has at least one node from each level, and thus, $a_i = |A_i| \geq t$, $i = 1, 2, \ldots, b$; and (ii) one component, say, A_l, with at least $h - t$ nodes: from level $t + 1$ to level h, A_l has at least one node from each level, and thus, $a_l = |A_l| \geq h - t$. From (1) defined above, we have

$$\begin{aligned} BC(\bar{Q}) &= \sum_{1 \leq i < j \leq l} a_i a_j \\ &\geq \sum_{1 \leq i < j \leq b} a_i a_j + \sum_{1 \leq i \leq b} a_i a_l \\ &\geq \binom{b}{2} t^2 + bt(h-t) \\ &= \frac{b(b-1)}{2} t^2 + bt(h-t). \end{aligned}$$

Also, $\binom{n-|\bar{Q}|}{2} = ((n-b)(n-b-1))/2$. Thus, we have

$$r(\bar{Q}) = \frac{BC(\bar{Q})}{\binom{n-|Q|}{2}} \geq \frac{b(b-1)t^2 + 2bt(h-t)}{(n-b)(n-b-1)}.$$

To show the tightness of the bound, let G be a path of length 4. Thus, $n = 5$, $h = 2$, $n[0] = n[1] = 2$, and $n[2] = 1$. For $k = 2$, if we apply the heuristic, we have $\bar{Q}$ as all the nodes on level 1 of G. Then, $t = 1$, $b = n[1] = 2$, and $(b(b-1)t^2 + 2bt(h-t))/((n-b) \cdot (n-b-1)) = 1$, which implies that $\bar{Q}$ is an optimal solution to PP on G. □

When G is a balanced and full d-tree, the procedure in Theorem 7 is the same as the one in Theorem 6, which was shown to provide an asymptotically maximal solution to PP.

4. Conclusions and Future Research Directions

The ability to find useful structures in social networks will undoubtedly benefit their users as well

as other stakeholders—the businesses that use these networks and the sites that host them. Unlike the Internet, structural search on social networks is set-based and offers a rich variety of interesting combinatorial optimization problems. In this paper, our effort is to identify and analyze specific instances of such problems. We consider two problems—EGP (the elite group problem) and PP (the portal problem)—derived, respectively, from the notions of influence and centrality. We demonstrate the relevance of these problems on a variety of social networks and show the following results: (i) The basic EGP is polynomially solvable, whereas its size-constrained variant is strongly NP-hard. We also show the hardness of a penalty-based relaxation of the size-constrained version. (ii) PP is strongly NP-hard. We discuss the solution of PP on several special networks—bicliques, balanced and full d-trees, paths, cycles, and cliques—and propose a heuristic for general trees.

In the industry, the focus thus far has been on developing "social search engines" to search social media and user-generated content, e.g., Social Mention (http://www.socialmention.com), Twitter (http://search.twitter.com), and Delver (http://www.delver.com). Some networks do facilitate simple search; e.g., MySpace allows a user to find other users with similar interests. However, to our knowledge, there is little or no sophisticated structural search available to ordinary users of social networks. Because this type of search is typically combinatorial in nature, the resulting problems are expected to be challenging. One idea is to provide an easy-to-use modeling language to enable members to specify complex, constrained search and then use sophisticated solvers (e.g., CPLEX) or heuristics to solve the resulting problems. Another possibility is to develop a repository—that could evolve over time—of efficient algorithms for the typical combinatorial searches that users specify. The notions of an elite group and a portal studied in this paper are extensions to set-based measures of, respectively, indegree and betweenness centralities for individual members of a social network. Similarly, useful structures based on extensions of other popular centralities, e.g., the more general degree centrality or closeness centrality (Carrington et al. 2005), could also be investigated. Applications of such set-based measures have been discussed for several social networks (see, e.g., Cattani and Ferriani 2008, Owen-Smith et al. 2002, Morselli and Giguere 2006).

The ideas of search developed in this paper naturally flow into other operational problems of interest. One such problem is targeted online advertising. For example, in the Twitter network, it is possible for one member to "follow" another, suggesting a directed link in the network. The identification of an elite group within Twitter could, therefore, be used to target promotional material to members of this group. For instance, an advertisement could be targeted using keywords exchanged by two members during a conversation on Twitter.

Electronic Companion

An electronic companion to this paper is available as part of the online version at http://dx.doi.org/10.1287/ijoc.1110.0473.

References

Borgatti, S. P. 2006. Identifying sets of key players in a social network. *Comput. Math. Organ. Theory* **12**(1) 21–34.

Boyd, D. M., N. B. Ellison. 2007. Social network sites: Definition, history, and scholarship. *J. Comput.-Mediated Comm.* **13**(1) 210–230.

Brandes, U., T. Erlebach, eds. 2005. *Network Analysis: Methodological Foundations*. Lecture Notes in Computer Science, Vol. 3418. Springer-Verlag, Berlin.

Carrington, P. J., J. Scott, S. Wasserman. 2005. *Models and Methods in Social Network Analysis*. Cambridge University Press, New York.

Cattani, G., S. Ferriani. 2008. A core/periphery perspective on individual creative performance: Social networks and cinematic achievements in the Hollywood film industry. *Organ. Sci.* **19**(6) 824–844.

Cross, R., A. Parker. 2004. *The Hidden Power of Social Networks: Understanding How Work Really Gets Done in Organizations*. Harvard Business School Press, Boston.

Erickson, B. H. 1981. Secret societies and social structure. *Soc. Forces* **60**(1) 188–210.

Everett, M. G., S. P. Borgatti. 1999. The centrality of groups and classes. *J. Math. Sociol.* **23**(3) 181–201.

Fowler, J. H. 2006. Legislative cosponsorship networks in the US House and Senate. *Soc. Networks* **28**(4) 454–465.

Freeman, L. C. 1979. Centrality in social networks: Conceptual clarification. *Soc. Networks* **1**(3) 215–239.

Garey, M. R., D. S. Johnson. 1979. *Computers and Intractability: A Guide to the Theory of NP-Completeness*. W. H. Freeman, San Francisco.

Gloor, P. A., D. Oster, J. Putzke, K. Fischback, D. Schoder, K. Ara, T. J. Kim et al. 2007. Studying microscopic peer-to-peer communication patterns. *Americas Conf. Inform. Systems (AMCIS 2007)*, Association for Information Systems, Atlanta.

Greene, K. 2008. 10 emerging technologies 2008. *Tech. Rev.* (March/April) http://www.technologyreview.com/specialreports/specialreport.aspx?id=25.

Hanaki, N., A. Peterhansl, P. S. Dodds, D. J. Watts. 2007. Cooperation in evolving social networks. *Management Sci.* **53**(7) 1036–1050.

Hawe, P., L. Ghali. 2008. Use of social network analysis to map the social relationships of staff and teachers at school. *Health Ed. Res.* **23**(1) 62–69.

Hesseldahl, A. 2008. There's gold in "reality mining." *Business Week* (March 24) http://www.businessweek.com/technology/content/mar2008/tc20080323_387127.htm.

Hill, S., F. Provost, C. Volinsky. 2006. Network-based marketing: Identifying likely adopters via consumer networks. *Statist. Sci.* **21**(2) 256–276.

Hoffman, A. J., J. B. Kruskal. 1956. Integral boundary points of convex polyhedra. H. Kuhn, A. Tucker, eds. *Linear Inequalities and Related Systems*, Annals of Mathematical Studies, Vol. 38. Princeton University Press, Princeton, NJ, 223–246.

Klovdahl, A. S., E. A. Graviss, A. Yaganehdoost, M. W. Ross, A. Wanger, G. J. Adams, J. M. Musser. 2001. Networks and tuberculosis: An undetected community outbreak involving public places. *Soc. Sci. Med.* **52**(5) 681–694.

Kolaczyk, E. D., D. B. Chua, M. Barthelemy. 2009. Group betweenness and co-betweenness: Inter-related notions of coalition centrality. *Soc. Networks* **31**(3) 190–203.

Morselli, C., C. Giguere. 2006. Legitimate strengths in criminal networks. *Crime, Law Soc. Change* **45**(3) 185–200.

Nemhauser, G. L., L. A. Wolsey. 1988. *Integer and Combinatorial Optimization*. John Wiley & Sons, New York.

Owen-Smith, J., M. Riccaboni, F. Pammolli, W. W. Powell. 2002. A comparison of U.S. and European university-industry relations in the life sciences. *Management Sci.* **48**(1) 24–43.

Pentland, A. 2004. "Reality mining" the organization. *Tech. Rev.* (March 31) http://www.technologyreview.com/web/13517.

Perer, A., C. Wilson. 2007. The steroids social network: An interactive feature on the Mitchell report. *Slate* (December 21) http://www.slate.com/id/2180392/.

Puzis, R., Y. Elovici, S. Dolev. 2007. Fast algorithm for successive computation of group betweenness centrality. *Phys. Rev. E* **76**(5, Part 2) 056709.

Schilling, M. A., C. C. Phelps. 2007. Interfirm collaboration networks: The impact of large-scale network structure on firm innovation. *Management Sci.* **53**(7) 1113–1126.

Scott, J. 2000. *Social Network Analysis: A Handbook*, 2nd ed. Sage Publications, Thousand Oaks, CA.

Thompson, C. 2008. Real-world social networks vs. Facebook "friends." *Wired Magazine* **16**(8) http://www.wired.com/techbiz/people/magazine/16-08/st_thompson.

Wasserman, S., K. Faust. 1994. *Social Network Analysis: Methods and Applications*, Cambridge University Press, Cambridge, UK.

Wein, L. M. 2009. Homeland security: From mathematical models to policy implementation: The 2008 Philip Mccord Morse lecture. *Oper. Res.* **57**(4) 801–811.

Wellman, B., W. Chen, W. Dong. 2001. Networking Guanxi. T. Gold, D. Guthrie, D. Wank, eds. *Social Networks in China: Institutions, Culture, and the Changing Nature of Guanxi*. Cambridge University Press, Cambridge, UK, 221–242.