Submitted to *Management Science*
manuscript

Structural Search and Optimization in Social Networks

Milind Dawande, Vijay Mookerjee, Chelliah Sriskandarajah, Yunxia Zhu
School of Management, University of Texas at Dallas, Richardson, TX 75083,
milind@utdallas.edu, {vijaym, 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 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, etc), 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 while PP is strongly NP-hard. Motivated by practical considerations, we also consider a size-constrained variant of EGP and show that it is strongly NP-hard. Finally, we argue for the need to solve a resource allocation problem (to distribute limited resources among the chosen individuals) subsequent to the identification of an optimal (or near-optimal) solution to EGP or PP. We illustrate this problem for several social networks reported in the literature.

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). Often in social network analysis the interest is to explain 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. While some sites support the maintenance of pre-existing social networks, others help strangers connect based on their shared interests and/or activities. Some sites cater to diverse audiences while 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 each other for exchanging ideas, solving problems, companionship, and so on.

1.1. The Significance of Search

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 upon 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 like the web or the telephone network. In the web, the typical nature of search is to provide the user with a set of web sites that match based upon a list of search terms. There is usually no requirement that the web sites 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. Table 1 provides a snapshot of real-world social networks that have been constructed to conduct a variety of searches of interest.

From a technical perspective, when the results of a search need to meet (or exceed) a specified set measure (specifically, a non-additive 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.

1.2. Using Search for Operational Decision-Making

The ability to construct social networks of interest provides businesses with an opportunity to exploit them to improve their operational decisions. Consider, for example, a firm interested in the targeted marketing of its products to consumers. Probably for the first time ever, such a firm now has the ability to collect and analyze data on existing and potential customers and construct a network that incorporates features such as buying habits, geographical location, mutual influences, etc. Using this network to partition the consumer market into appropriate segments is an important

Domain	Purpose of the social network	Reference
Marketing	Search for opinion leaders among physicians to promote new drugs	Orgnet.com (2008)
	Find a group of users for targeted advertising	Sharma and Steel (2008)
Criminology	Search for key players in a criminal communication network	Morselli and Giguere (2006)
	Identify principal vulnerabilities in criminal networks	Sparrow (1991)
Politics	Search for a set of influential legislators to co-sponsor legislation	Fowler (2006)
Organizational Behavior	Search for influential staff in an educational institution	Hawe and Ghali (2008)
Epidemiology	Search the critical persons/places in a TB-outbreak network to limit the spread of the disease	Klovdahl et al. (2001)
Software Development	Examine the effects of network embeddedness on the success of open-source projects	Grewal et al. (2006)
	Research on how the demographic diversity of a team affects its performance	Reagans and Zuckerman (2001)
Bibliography	Find out the most influential papers on a subject	Kim and McMillan (2008)
Sociology	Examine the role of social networks in shaping individuals' ability to generate a creative outcome	Cattani and Ferriani (2008)
	Search the interaction network of animals to check the impact of age and gender	Everett and Borgatti (1999)

Table 1 A Snapshot of Applications of Search in Social Networks in Various Domains.

search problem of interest to the firm (Sharma and Steel 2008). The results of this search can be stored and used for operational decisions such as the scheduling of advertisement campaigns and deciding the acceptable risk level in approving credit applicants. As another example, the shipping department of a firm can use the network of its customers to search (and store) routes and preferred schedules and combine them with real-time traffic reports to obtain an overview of current deliveries and potential problems, and identify resources to resolve bottlenecks (IBM 2007).

In general, given the complex nature of the structural relationships that a set of individuals may need to satisfy in a search problem of interest, it is reasonable to assume that obtaining an

optimal (or even a feasible) solution may be challenging and time consuming. However, for most social networks, the need for solving such a search problem may arise only sporadically. Thus, the results of the search can be stored and used profitably for tactical decisions. We will revisit this issue again in Section 4.

1.3. Summary of Our Results

We formulate two specific search problems – the Elite Group Problem (EGP) and the Portal Problem (PP) – that represent two fundamental notions on which search in social networks is likely to be prevalent. We summarize our results below.

(a) 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, bipartite graphs, cycles, paths, etc), we show that an optimal solution to both EGP and PP is easy to obtain.

(b) We show that EGP is polynomially solvable on a general graph while PP is strongly NP-hard. Motivated by practical considerations, we also consider a size-constrained variant of EGP and show that it is strongly NP-hard.

(c) We argue for the need to solve a resource allocation problem (to distribute limited resources among the chosen players) subsequent to the identification of an optimal (or near-optimal) solution to EGP or PP. We justify this problem for several social networks reported in the literature.

The remainder of the paper is organized as follows. In Section 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 introduce two optimization problems – EGP and 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 describes instances of a resource-allocation problem that could arise subsequent to search. Section 5 concludes the 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 arises naturally: 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, since 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, Stephenson and Zelen 1989, Scott 2000). For set-based search too, 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 recent studies. For example, Carrington et al. (2005) and Everett and Borgatti (1999) discuss the notions of group (or set) betweeness and group degree centralities.

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 (Orgnet.com 2008). 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 purpose 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, a set of expert oncologists may be important to identify to arrive at 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, 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). It is important to point out that while a single centrally located member may be useful to identify for a variety of purposes, the value of identifying a centrally located set of members may be even higher. To cripple a terrorist group, it is often sub-optimal to spend resources by individually striking at isolated targets; rather a concerted effort at eliminating a set of centrally located targets may do the most damage to the effective functioning of the organization. Identifying a central group (rather than a central individual) also reminds one of possible externalities within the group: a router among a centrally located set of routers may acquire viruses from other members of the group that also support heavy traffic.

We now introduce two specific problems that correspond to influential sets and central sets. Following the definition of each problem, we discuss its origin and provide several examples of social networks where the problem is relevant.

2.1. The Elite Group Problem (EGP) and The Size-Constrained Elite Group Problem (SCEGP)

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 SCEGP, a positive integer $k \leq n$.

SOLUTION OF EGP: A set $W \subseteq V$ such that there does not exist a directed arc $(i, j) \in A$ with $i \in W,\ j \notin W$.

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 $(i, j) \in A$; 0 otherwise.

Note that in every graph $G(V, A)$, there is at least one feasible elite group V, which has score $\gamma_V = 0$.

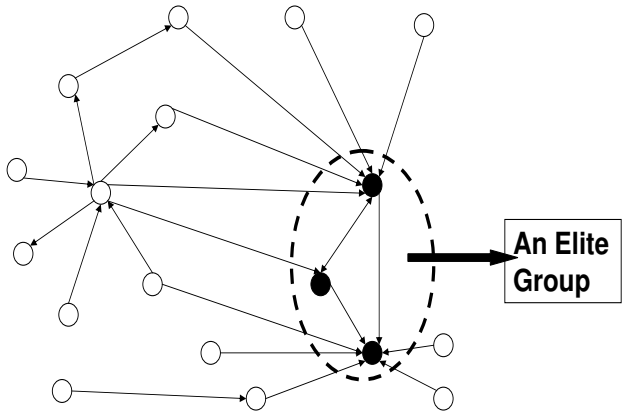

Figure 1 An "Influence" Network and an Elite Group.

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 the properties 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 opinion leaders. For instance, when analyzing the opinion-seeking network among physicians (Orgnet.com 2008), we find out that a physician who is not sure about a new medical treatment typically consults other physicians for advice. If a physician is consulted by a lot of peers, she can be regarded as a key opinion leader (an "elite" member) in this network. The notion of an elite group also appears in sociometric networks. For example, in Hoffman and Wilcox (1992), the members of a group are asked to nominate one of them as the project leader. In the resulting "trust network", all the members who get nominations consist of an elite group. This information is useful in the search for a champion of the project. Fowler (2006) analyzes the co-sponsorship network in the United States Senate. In this network, the prominent senators typically receive a significant amount of co-sponsorship. Thus, the set of these prominent senators constitute an (approximate) elite group.

2.2. The Portal Problem (PP) and The Exact-Size Portal Problem (ESPP)

Technical Definition

INSTANCE: n players; an 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 ESPP, a set $Q \subseteq V$ such that $|Q| = k$.

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

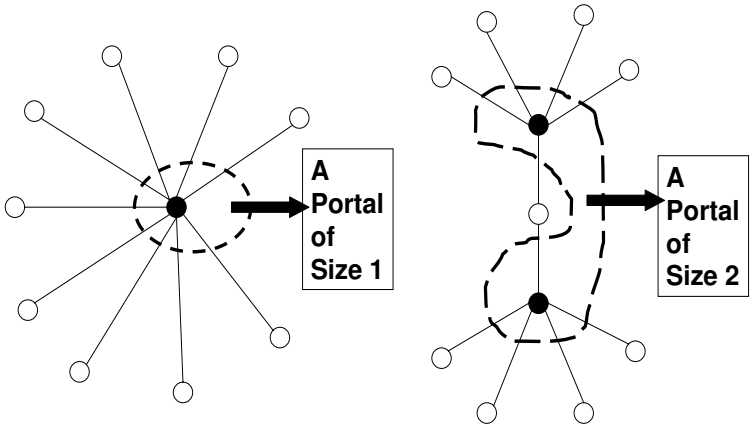

Figure 2 Optimal Portals in Two Simple Networks.

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. In fact, our normalized measure for a portal has been used earlier in the literature. For example, in Everett and Borgatti (1999) and Puzis et al. (2007), the term "Group Betweenness Centrality" is used to describe this measure.

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

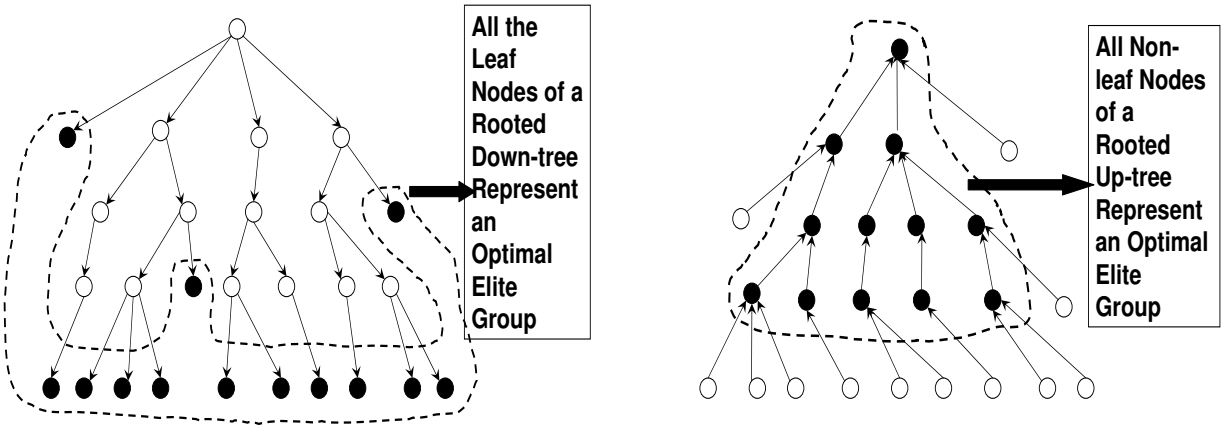

Figure 3 Optimal Elite Group for a Rooted Down-Tree and a Rooted Up-Tree.

TB-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 influential set of animals.

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 and show that it is strongly NP-hard. We consider several special graphs on which PP is polynomially solvable then show that PP is strongly NP-hard on a general graph.

3.1. The Elite Group Problem (EGP)

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 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, etc (Cross and Parker 2004).

LEMMA 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.*

Proof: First, note that the root is not included in an optimal elite group; for otherwise, each node of G is in the elite group and the score is 0, which is clearly a non-optimal solution for any non-trivial rooted down-tree G. Consider an optimal elite group W which contains a non-leaf node t such that the unique predecessor of t is not in W. Note that all descendants of t are also in W. Let $n_t \geq 1$ be the number of direct descendants of t in G. Then, removing t from W results in a feasible elite group $W^{'} = W - \{t\}$ with score $\gamma_{W'} = \gamma_W + (n_t - 1) \geq \gamma_W$. Continuing, we can similarly remove all non-leaf nodes from W without decreasing the score to obtain an elite group consisting only of leaf nodes. Thus, there exists an optimal elite group W^* consisting only of leaf nodes. Finally, note that W^* must contain *all* leaf nodes. This follows since including a leaf node strictly increases the score of an elite group. ∎

The proof of the following result is similar.

LEMMA 2. *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 upwards from the leaf nodes towards the root. See Figure 3) then the elite group W^* consisting of all non-leaf nodes of G is an optimal elite group.*

Our next result helps us "shrink" the 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 3.

LEMMA 3. *If G contains a directed cycle, and at least one node on this cycle belongs to an elite group W (respectively, the complement $\overline{W} = V \backslash W$), then all the other nodes on the cycle must belong to W (respectively, $\overline{W}$).*

Proof: Consider a directed cycle (v_1-v_2-…-v_n-v_1). Suppose, without loss of generality, $v_1 \in W$. Since there is a directed arc from v_1 to v_2, node v_2 must belong to W as well. Continuing this argument, nodes $v_3, v_4, \cdots, v_n$ must belong to W. Similarly, if, say, $v_1 \in \overline{W}$, then $\{v_2, v_3, \ldots, v_n\} \subseteq \overline{W}$. For otherwise, if $v_j \in W$ for some $j \in \{2, 3, \ldots n\}$, then $v_1 \in W$. The result follows. ∎

Note that there are many polynomial algorithms to find a directed cycle (if one exists) in a graph. If G contains a directed cycle C, then, by using Lemma 3, we can shrink C into a single node. All arcs from nodes in $V\backslash C$ to nodes in C are now incident to the *shrunk node* representing the cycle. We can continue this type of shrinking (possibly recursively) until there is no directed cycle in the modified *shrunk graph*. Thus, we can assume without loss of generality that the network contains no directed cycle. The following result follows immediately from Lemma 3.

THEOREM 1. If $\bar{G}$ is a shrunk graph and a shrunk node belongs to an elite group (respectively, complement of an elite group) in $\bar{G}$, then all the nodes on the directed cycle(s) corresponding to the shrunk node in the original graph G must belong to the elite group (respectively, complement of the elite group).

Next, we show that EGP is polynomially solvable.

THEOREM 2. 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:

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

The constraint matrix of the above IP is the node-arc incidence matrix of G. 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 directed cycles (Theorem 1) maintains the total modularity of the constraint matrix of the IP above. Thus, the size of a network containing directed cycles can be reduced before formulating the EGP.

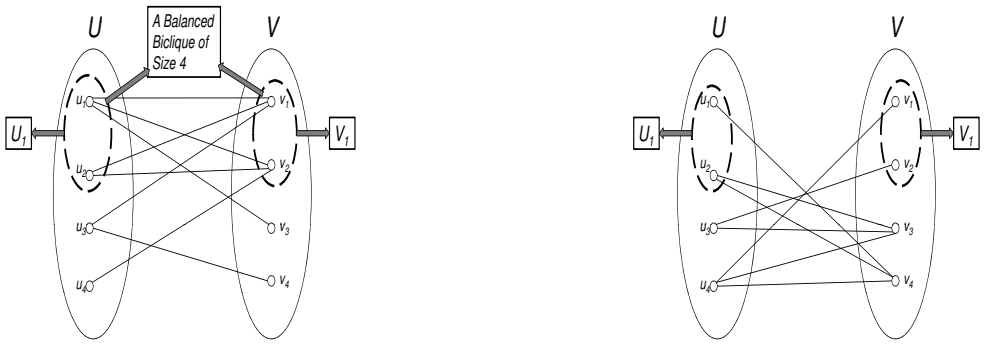

Figure 4 A Bipartite Graph G with a Balanced Biclique, and Its Bipartite Complement Graph G^c, Which is Used in the Proof of Theorem 3.

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 (see Section 4 for some illustrative examples). Thus, for practicability in managing this subsequent task, the size of an elite group may need to be restricted. Motivated by this requirement, Theorem 3 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 3. *The decision problem corresponding to SCEGP is strongly NP-Complete.*

The strongly NP-Complete problem which we use in our reduction is the Balanced Biclique Problem (Garey and Johnson 1979).

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:

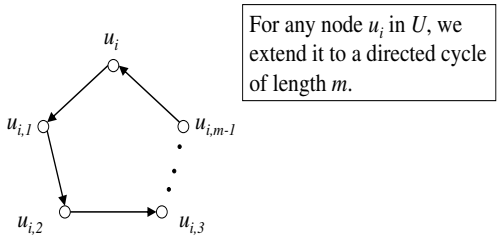

Figure 5 The Widget, a Directed Cycle with Length m, Used in the Proof of Theorem 3.

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 (resp., 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 $m-1$ additional nodes $(u_{i,1}, u_{i,2}, \cdots, u_{i,m-1})$ (see Figure 5). Let $U' = \{u_i, u_{i,1}, u_{i,2}, \cdots, u_{i,m-1} | u_i \subseteq U, i = 1, 2, \cdots, n\}$. 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 $SCEGP$:

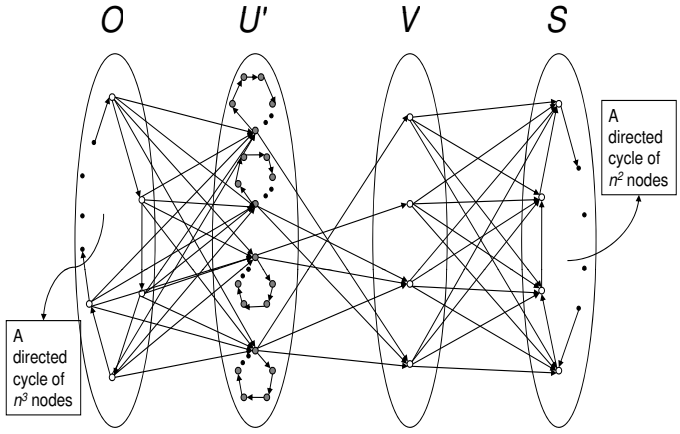

Figure 6 The Constructed Graph G' for $SCEGP$.

DECISION QUESTION: Let $t = km + (n-k) + n^2$ and $D = kn^3 + kn^2$. Does there exist an elite group W in G' such that $|W| \leq t$, and $\gamma_W \geq D$?

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 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 $U_1^{'} = \{C_i | u_i \in U_1\}$, $U_2^{'} = \{C_i | u_i \in U_2\}$, $W = U_1^{'} \cup V_2 \cup S$, $\overline{W} = O \cup U_2^{'} \cup V_1$ (see Figure 7). We now show that the set W is an elite group that provides an affirmative answer to the decision question.

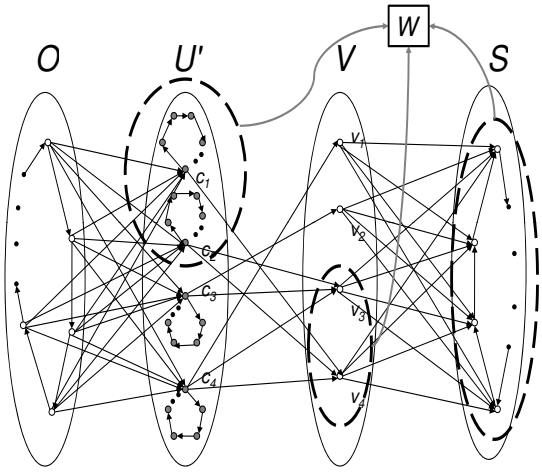

Figure 7 **Graph G' with Elite Group Set W.**

First we need to prove the set W is a valid elite group in $G^{'}$, i.e., there is no arc from W to $\overline{W}$. 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 $\overline{W}$. Similarly, there is no arc from V_2 to $\overline{W}$ and from S to $\overline{W}$. 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 $\overline{W}$ 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, $\gamma_W \geq kn^3 + kn^2 = D$. The result follows.

$\Longleftarrow$ Suppose W is an elite group in G' with $|W| \leq t$ and $\gamma_W \geq D$. Let $\overline{W} = N \backslash W$. 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 $\overline{W}$. Similarly, the nodes in S (respectively, O) either all belong to W or all belong to $\overline{W}$.*

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

Claim 2. *Each node in O must belong to $\overline{W}$. 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 $|W| \geq |O| + |U'| = n^3 + nm$. Since $n \geq 2$ and $n \geq k$, we have $n^3 > n^2 + n > n^2 + n - k$ and $nm \geq km$. So $n^3 + nm > (n^2 + n - k) + km$, which implies $|W| > t$. This contradicts the assumption that $|W| \leq t$. Thus, each node in O must belong to $\overline{W}$.

Suppose a node in S belongs to $\overline{W}$. Then, from Claim 1, each node in S belongs to $\overline{W}$. Also, each nodes in V must belong to $\overline{W}$. As shown above, each node in O is in $\overline{W}$. 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 $|W| \leq t = n^2 + km + n - k = km + m - k = (k+1)m - k$, we have $|W| = m|Q| \leq (k+1)m - k$, so $|Q| \leq k$. Thus $\gamma_W = n^3|Q| \leq n^3 k < kn^3 + kn^2 = D$, which contradicts the assumption that $\gamma_W \geq D$. 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 $\overline{W} = O \cup U_2^{'} \cup V_1$. Let $U_1 = \{u_i | C_i \in U_1^{'}\}$.

Claim 3. $|U_1| = k$.

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

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

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

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

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

Note that $U_1^{'} \subseteq W$, $V_1 \subseteq \overline{W}$. 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 $|U_1| = k, |V_1| \geq k$, there exists at least one balanced biclique of size $2k$ in G. This concludes the proof of Theorem 3. ■

3.2. The Portal Problem (PP)

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

As mentioned earlier, a portal is a natural extension to a set-based measure of the notion of Betweenness Centrality (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 *non-additive*. In other words, BCs of two distinct nodes in Q cannot, in general, be simply added when computing $r(Q)$. This is obvious, since 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 $r(Q)$.

In Section 3.2.3, we show that PP is strongly NP-hard. An efficient, polynomial-time algorithm for obtaining an optimal solution on general graphs is, therefore, unlikely. Even for highly structured graphs, e.g., paths and balanced binary trees, an optimal solution is not obvious. We now discuss these two graphs.

3.2.1. Special Trees: Paths and Balanced Binary 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

$V \backslash Q$; thus, $\sigma_{st} = 1$ (see Section 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 i^{th} connected component in $G'(Q)$, $i = 1, 2, \cdots, l$.

a_i: the *size* (i.e., the number of nodes) of component A_i, $i = 1, 2, \cdots, l$.

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,

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

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 illustrate the solution of this problem for paths and balanced binary trees.

• **Paths:**

If we remove k nodes from a path, then we obtain at most $l = k + 1$ connected components. If the deleted nodes contain two or more adjacent nodes, then the number of connected components is strictly less than $k + 1$; however, in this case, we can assume empty components (i.e., components with $a_i = 0$ nodes). Thus, without loss of generality, we can assume that exactly $k + 1$ components result from the deletion of k nodes (i.e., $l = k + 1$). We will solve PP by first getting an optimal solution for ESPP.

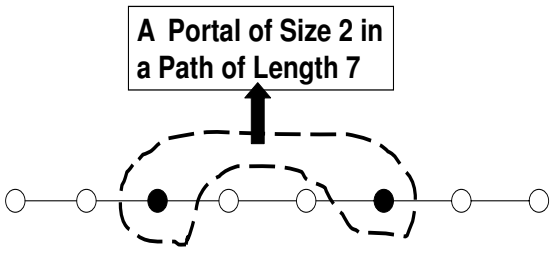

Figure 8 Optimal Portal in a Path.

LEMMA 4. *Let G be a path v_1-v_2-...-v_n. If Q^* is an optimal solution of ESPP on G, then any pair of distinct connected components A_p and A_q in $G'(Q^*)$ satisfies $|a_p - a_q| \leq 1$.*

Proof: Let Q be an optimal solution, and let there be connected components A_p and A_q in $G'(Q)$ such that $a_p - a_q \geq 2$. It is easy to construct Q' such that, in $G'(Q')$, we have $a_p' = a_p - 1, a_q' = a_q + 1$, $a_i' = a_i, i \in \{1, 2, \ldots, l\} \backslash \{p, q\}$. In fact, for *any* set of desired (component) sizes $\tilde{a}_i, i = 1, 2, \cdots, l = k+1$, satisfying $\sum_{i=1}^{l} \tilde{a}_i = n - k, \tilde{a}_i \in Z^+$, setting $Q = \{v_{1+\tilde{a}_1}, v_{2+\tilde{a}_1+\tilde{a}_2}, \ldots, v_{l-1+\tilde{a}_1+\tilde{a}_2+\ldots+\tilde{a}_{l-1}}\}$ generates connected components of the required sizes in $G'(Q)$.

Using (2), we have

$$\begin{aligned} BC(Q') - BC(Q) &= \frac{(n-k)^2 - \sum_{i=1}^{l} a_i'^2}{2} - \frac{(n-k)^2 - \sum_{i=1}^{l} a_i^2}{2} \\ &= \frac{\sum_{i=1}^{l} a_i^2 - \sum_{i=1}^{l} a_i'^2}{2} \\ &= \frac{a_p^2 + a_q^2 - a_p'^2 - a_q'^2}{2} \\ &= a_p - a_q - 1 \end{aligned}$$

Since $a_p - a_q \geq 2$, we have $BC(Q') - BC(Q) > 0$. This contradicts the optimality of Q. The result follows. ∎

THEOREM 4. *Let $\mu = \frac{n-k}{k+1}, c = n - k - (k+1)\lfloor \mu \rfloor$. Then, Q^* is an optimal solution of ESPP if and only if $G'(Q^*)$ has exactly c connected components of size $\lfloor \mu \rfloor + 1$ and exactly $(k+1-c)$ connected components of size $\lfloor \mu \rfloor$.*

Proof:

$\Longrightarrow$ Let Q^* be an optimal solution of ESPP. Then, from Lemma 4, the number of nodes in each connected component in $G'(Q^*)$ is either $a_i = \lfloor \mu \rfloor$ or $a_i = \lfloor \mu \rfloor + 1$. Since $\sum_{i=1}^{k+1} a_i = n - k = c(\lfloor \mu \rfloor + 1) + (k+1-c)\lfloor \mu \rfloor$, exactly c (resp., $k+1-c$) connected components have size $\lfloor \mu \rfloor + 1$ (resp., $\lfloor \mu \rfloor$).

$\Longleftarrow$ This follows since *any* Q for which $G'(Q)$ has exactly c (resp., $k+1-c$) connected components of size $\lfloor \mu \rfloor + 1$ (resp., $\lfloor \mu \rfloor$) provides the same objective function value for ESPP:

$$BC(Q) = \frac{(n-k)^2 - \sum_{i=1}^{l} a_i^2}{2} = \frac{(n-k)^2 - c(\lfloor \mu \rfloor + 1)^2 - (k+1-c)(\lfloor \mu \rfloor)^2}{2}.$$

■

Thus, given an explicit description of the path v_1-v_2-...-v_n and a positive integer k, an optimal solution for ESPP is $Q^* = \{v_{i(\lfloor\mu\rfloor+2)}, i = 1, 2, \cdots, c\} \cup \{v_{c(\lfloor\mu\rfloor+2)+j(\lfloor\mu\rfloor+1)}, j = 1, 2, \cdots, k-c\}$. The optimal objective function value is $r(Q^*) = BC(Q^*)/\binom{n-k}{2}$, where

$$BC(Q^*) = \frac{(n-k)^2 - c(\lfloor\mu\rfloor+1)^2 - (k+1-c)(\lfloor\mu\rfloor)^2}{2}.$$

An optimal solution for PP is also easy to obtain: we simply solve ESPP for each $\bar{k} \leq k$. Since $k \leq n$, this requires time polynomial in the size of the input.

- **Balanced Binary Trees:**

On a rooted Balanced Binary Tree, each node (except the leaf nodes) has two distinct successors, each node (except root) has a unique predecessor. All leaf nodes have the same distance (height) to the root node. For a binary tree, if we remove any node other than the root node, we will add two more connected components into the remaining graph. So if we remove k nodes from a binary tree, we will have at most $l = 2k+1$ connected components left.

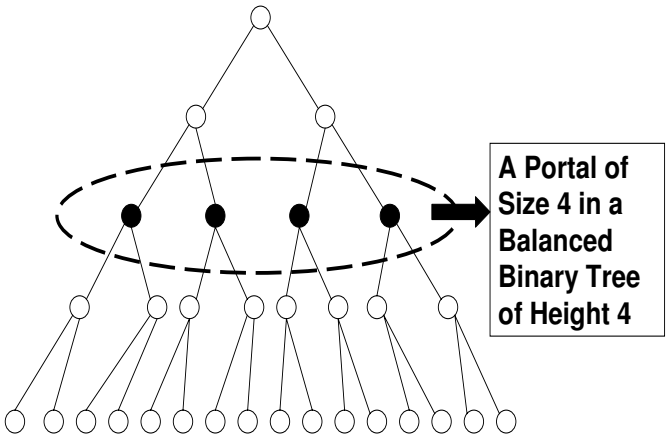

Figure 9 Optimal Portal in a Balanced Binary Tree.

THEOREM 5. *Let G be a balanced binary tree with height $h \geq 2$. For an instance of PP defined by G and a positive integer k, let $t = \min\{\lceil h/2\rceil, \lfloor \log_2 k\rfloor\}$ and let $\bar{Q}$ denote the set of nodes on the t^{th} level of G. Then, $\bar{Q}$ provides an asymptotically optimal solution to PP, with $r(\bar{Q}) \geq (1 - \frac{1}{2^{t+1}})$.*

Proof: First, since the t^{th} level of a balanced binary tree has 2^t nodes, $|\bar{Q}| = 2^t \leq k$. Note that $G'(\bar{Q})$ has exactly $l = 2|\bar{Q}| + 1$ connected components. Of these, we have (i) $2|\bar{Q}|$ identical components,

say $A_i, i = 1, 2, \ldots, 2|\bar{Q}|$, each with $2^{h-t} - 1$ nodes. Thus, $a_1 = a_2 = \cdots = a_{2|\bar{Q}|} = 2^{h-t} - 1$, and (ii) one component, say A_l, with $2^t - 1$ nodes. Thus, $a_l = 2^t - 1$.

From (1),

$$\begin{aligned} BC(\bar{Q}) &= \sum_{1 \le i < j \le l} a_i a_j \\ &= 2|\bar{Q}| a_l a_1 + a_1^2 \binom{2|\bar{Q}|}{2} \\ &= 2^{t+1}(2^t - 1)(2^{h-t} - 1) + (2^{h-t} - 1)^2 \frac{2^{t+1}(2^{t+1} - 1)}{2} \\ &= (2 - \frac{1}{2^t})2^{2h} - 2^{t+h+1} + 2^t. \end{aligned}$$

$$\begin{aligned} \text{Also, } \binom{n - |\bar{Q}|}{2} &= \binom{2^{h+1} - 1 - 2^t}{2} \\ &= 2^{2h+1} - (2^{t+1} + 3)2^h + 0.5(2^t + 1)(2^t + 2). \end{aligned}$$

$$\begin{aligned} \text{Thus, } BC(\bar{Q}) - (1 - \frac{1}{2^{t+1}}) \binom{n - |\bar{Q}|}{2} &= (2 - \frac{1}{2^t})2^{2h} - 2^{t+h+1} + 2^t - \\ &\quad (1 - \frac{1}{2^{t+1}})[2^{2h+1} - (2^{t+1} + 3)2^h + 0.5(2^t + 1)(2^t + 2)] \\ &= 2^h(2 - (3)2^{-(t+1)}) - 2^{2t-1} - 2^{t-2} + 2^{-(t+1)} - 0.25 \end{aligned}$$

Since $t = \min\{\lceil h/2 \rceil, \lfloor \log_2 k \rfloor\}$ and $\lceil h/2 \rceil \le (h+1)/2$, we have $t \le \lceil h/2 \rceil \le (h+1)/2$. Thus, $2^h \ge 2^{2t-1}$. Consequently, we have

$$\begin{aligned} BC(\bar{Q}) - (1 - \frac{1}{2^{t+1}}) \binom{n - |\bar{Q}|}{2} &\ge 2^{2t-1}(2 - (3)2^{-(t+1)}) - 2^{2t-1} - 2^{t-2} + 2^{-(t+1)} - 0.25 \\ &= 2^{2t-1} - 2^t + 2^{-(t+1)} - 0.25 \\ &= 2^{-(t+1)}(2^{t-1} - 1)(2^{2t+1} - 1) \\ &\ge 0 \end{aligned}$$

Thus,

$$r(\bar{Q}) = \frac{BC(\bar{Q})}{\binom{n - |\bar{Q}|}{2}} \ge (1 - \frac{1}{2^{t+1}})$$

.

Since $t = \min\{\lceil h/2 \rceil, \lfloor \log_2 k \rfloor\}$, the ratio $r(\bar{Q}) \to 1$ with an increase in the size of G and k. ∎

3.2.2. Other Graphs: Cycles, Cliques, Bicliques

- **Cycles:**

Given a cycle $G(V,E)$ and $Q \subseteq V$, let $G'(Q)$ denote the induced subgraph obtained by removing all the nodes in Q from G. If $k=1$, any node $v \in V$ is an optimal solution for PP. If $k \geq 2$, then in general $G'(Q)$ is a forest with disjoint paths as its connected components. Let G be a cycle v_1-v_2-...-v_n-v_1; the *length* of the cycle is n. We will use the same notation as in Section 3.2.1. The removal of $k \geq 2$ nodes from a cycle results in at most $l=k$ connected components. If the deleted nodes contain two or more adjacent nodes, then the number of connected components is strictly less than k; however, in this case, we can assume empty components (i.e., components with $a_i = 0$ nodes (see Section 3.2.1). Thus, without loss of generality, we can assume that exactly k components result from the deletion of k nodes (i.e., $l=k$). We will solve PP by first getting an optimal solution for ESPP. Without loss of generality, we let $a_1 \geq a_2 \geq \ldots \geq a_k$.

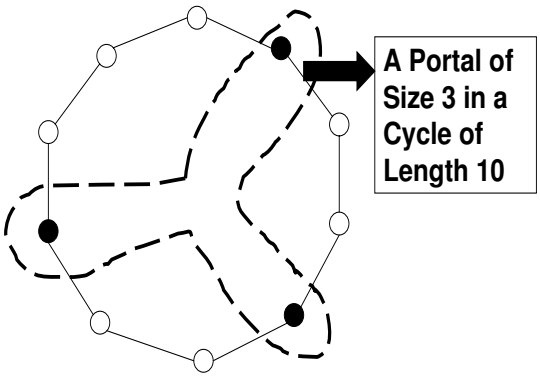

Figure 10 Optimal Portal in a Cycle.

LEMMA 5. *Let G be a cycle v_1-v_2-...-v_n-v_1. If Q^* is an optimal solution of ESPP on G, then Equation (1) (see Section 3.2.1) holds for Q^*.*

Proof: We consider two cases: (a) n is odd and (b) n is even.

Case a: n is odd. There is a unique shortest path in G connecting any two distinct nodes s and t in $V \backslash Q$. If $a_1 \leq \frac{n-1}{2}$, it is easy to verify that (1) holds. If $a_1 \geq \frac{n+1}{2}$, $BC(Q)$ has the following form:

$$BC(Q) = \sum_{s \notin Q, t \notin Q, s \neq t} \frac{\sigma_{st}(Q)}{\sigma_{st}} = \sum_{1 \leq i < j \leq l} a_i a_j + 0.5(a_1 - \frac{n-1}{2})(a_1 - \frac{n+1}{2}) \quad (3)$$

Next, we prove the following claim:

CLAIM 5. *Let G be a cycle v_1-v_2-…-v_n-v_1 and let n be odd. If Q^* is an optimal solution of ESPP on G which satisfies $a_1 \geq \frac{n+1}{2}$, then $a_1 = \frac{n+1}{2}$ in $G'(Q^*)$.*

Proof of Claim 5: Let Q be an optimal solution with $a_1 > \frac{n+1}{2}$. It is easy to construct $Q^{'}$ such that, in $G'(Q')$, we have $a_1^{'} = a_1 - 1, a_k^{'} = a_k + 1,\ a_i^{'} = a_i, i \in \{2, \ldots, k-1\}$: For *any* set of desired (component) sizes $\tilde{a}_i, i = 1, 2, \cdots, k$, satisfying $\sum_{i=1}^{k} \tilde{a}_i = n - k, \tilde{a}_i \in Z^+$, setting $Q = \{v_{1+\tilde{a}_1}, v_{2+\tilde{a}_1+\tilde{a}_2}, \ldots, v_{k-1+\tilde{a}_1+\tilde{a}_2+\ldots+\tilde{a}_{k-1}}, v_n\}$ generates connected components of the required sizes in $G'(Q)$. Using (3), we have

$$\begin{aligned} BC(Q') - BC(Q) &= \frac{(n-k)^2 - \sum_{i=1}^{l} a_i^{'2}}{2} + 0.5(a_1^{'} - \frac{n-1}{2})(a_1^{'} - \frac{n+1}{2}) \\ &\quad - \frac{(n-k)^2 - \sum_{i=1}^{l} a_i^2}{2} - 0.5(a_1 - \frac{n-1}{2})(a_1 - \frac{n+1}{2}) \\ &= \frac{n - 1 - 2a_k}{2} \end{aligned}$$

Since $a_k < \frac{(n-1)}{2}$, we have $BC(Q') - BC(Q) > 0$. This contradicts the optimality of Q. The result follows. □

Finally, note that when $a_1 = \frac{n+1}{2}$, (3) is equivalent to (1). Thus (1) holds for an optimal solution of ESPP on G.

Case b: n is even. For any pair $(v_i, v_{i+\frac{n}{2}}) \in V \backslash Q, i = 1, 2, \cdots, \frac{n}{2}$, there are two shortest paths in G connecting the two nodes of that pair. If $a_1 \leq \frac{n}{2}$, (1) holds. If $a_1 \geq \frac{n}{2}$, $BC(Q)$ has the following form:

$$BC(Q) = \sum_{s \notin Q, t \notin Q, s \neq t} \frac{\sigma_{st}(Q)}{\sigma_{st}} = \sum_{1 \leq i < j \leq l} a_i a_j + 0.5(a_1 - \frac{n}{2})^2 \tag{4}$$

Next, we prove the following claim:

CLAIM 6. *Let G be a cycle v_1-v_2-…-v_n-v_1 and let n be even. If Q^* is an optimal solution of ESPP on G which satisfies $a_1 \geq \frac{n}{2}$, then $a_1 = \frac{n}{2}$ in $G'(Q^*)$.*

Proof of Claim 6: Let Q be an optimal solution of ESPP on G with $a_1 > \frac{n}{2}$. Construct $Q^{'}$ such that, in $G'(Q')$, we have $a_1^{'} = a_1 - 1, a_k^{'} = a_k + 1,\ a_i^{'} = a_i, i \in \{2, \ldots, k-1\}$. Using (4), we have

$$BC(Q') - BC(Q) = \frac{(n-k)^2 - \sum_{i=1}^{l} a_i^{'2}}{2} + 0.5(a_1^{'} - \frac{n}{2})^2 - (\frac{(n-k)^2 - \sum_{i=1}^{l} a_i^2}{2} + 0.5(a_1 - \frac{n}{2})^2)$$

$$= \frac{n-1-2a_k}{2}$$

Since $a_k < \frac{n-1}{2}$, we have $BC(Q') - BC(Q) > 0$. This contradicts the optimality of Q. The result follows. □

When $a_1 = \frac{n}{2}$, (4) is equivalent to (1). Thus, (1) holds for an optimal solution of ESPP on G. Combining the two cases above, we conclude that if G is a cycle and Q^* is an optimal solution of ESPP on G, then (1) holds for Q^*. ■

Lemma 5 implies the following result (Lemma 6), which, in turn, implies Theorem 6. We avoid providing the detailed proofs since they are similar to those of Lemma 4 & Theorem 4, respectively.

LEMMA 6. *Let G be a cycle v_1-v_2-…-v_n-v_1. If Q^* is an optimal solution of ESPP on G, then any pair of distinct connected components A_p and A_q in $G'(Q^*)$ satisfies $|a_p - a_q| \leq 1$.*

THEOREM 6. *Let $\mu = \frac{n-k}{k}, c = n - k - k\lfloor\mu\rfloor$. Then, Q^* is an optimal solution of ESPP if and only if $G'(Q^*)$ has exactly c connected components of size $\lfloor\mu\rfloor + 1$ and exactly $(k-c)$ connected components of size $\lfloor\mu\rfloor$.*

Thus, given an explicit description of the cycle v_1-v_2-…-v_n-v_1 and a positive integer $k \geq 2$, an optimal solution for ESPP is $Q^* = \{v_{i(\lfloor\mu\rfloor+2)}, i = 1, 2, \cdots, c\} \cup \{v_{c(\lfloor\mu\rfloor+2)+j(\lfloor\mu\rfloor+1)}, j = 1, 2, \cdots, k-1-c\} \cup \{v_n\}$. The optimal objective function value is $r(Q^*) = BC(Q^*)/\binom{n-k}{2}$, where

$$BC(Q^*) = \frac{(n-k)^2 - c(\lfloor\mu\rfloor+1)^2 - (k-c)(\lfloor\mu\rfloor)^2}{2}.$$

An optimal solution for PP is also easy to obtain: we simply solve ESPP for each $\bar{k} \leq k$. Since $k \leq n$, this requires time polynomial in the size of the input.

- **Cliques:**

Let $G = (V, E)$ be a clique and let $Q \subseteq V$. For any two nodes $s, t \in V \backslash Q$, the unique shortest path in G between s and t is of length 1 and exists in $G'(Q)$. Thus, no shortest path between any two nodes in $V \backslash Q$ has an internal node in Q. Thus, $\sigma_{st}(Q) = 0$. It follows that $r(Q) = BC(Q) \equiv 0$ for any $Q \subseteq V$. In other words, any subset of nodes is an optimal solution for PP and ESPP.

• **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 7 (resp., Corollary 1) provides an optimal solution to PP (resp., ESPP). However, we first need to prove several intermediate results.

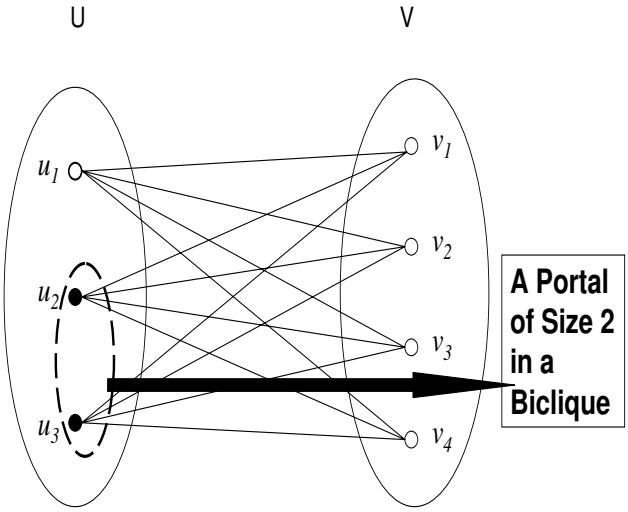

Figure 11 Optimal Portal in a Biclique.

LEMMA 7. *Let $G=(U\cup V,E)$ be a biclique with $n_1=|U|\le|V|=n_2$.*

1. *If $k=1$, then any node $u\in U$ is an optimal solution to PP;*

2. *If $k\ge n_1$, then $Q=U$ is an optimal solution to PP.*

Proof: If $k=1$, then let $v\in V$. It is easy to verify that $r(\{u\})\ge r(\{v\})>0=r(\emptyset)$. Thus, $\{u\}$ is an optimal solution of PP.

If $k\ge n_1$, then let $Q=U$. It follows that $G'(Q)=V$. Also, the shortest path (in G) between any two nodes in V has exactly one node in $Q=U$ as an internal node. Thus, $BC(Q)=\binom{|V|}{2}=\binom{n_2}{2}$. Also, $\binom{n-|Q|}{2}=\binom{n_2}{2}$. Thus, $r(Q)=1$, which is its maximum possible value. Thus $Q=U$ is an optimal solution of PP. ∎

To obtain an optimal solution of PP for $2\le k\le n_1-1$, we first obtain an optimal solution of the corresponding instance of ESPP in the following lemma.

LEMMA 8. *Let $G = (U \cup V, E)$ be a biclique with $n_1 = |U| \leq |V| = n_2$. Let $\bar{k} = \lfloor \frac{1+2n_1-\sqrt{8n_2-7}}{2} \rfloor$. Then,*

(i) If $\bar{k} < 2$, for $2 \leq k \leq n_1 - 1$, any set Q with $k_1 = |Q \cap U| = k$ and $k_2 = |Q \cap V| = 0$ is an optimal solution of ESPP.

(ii) If $\bar{k} \geq 2$,

(a) For $2 \leq k \leq \bar{k}$, any set Q which satisfies $k_1 = k-1$ and $k_2 = 1$ is an optimal solution of ESPP.

(b) For $\bar{k} < k \leq n_1 - 1$, any set Q which satisfies $k_1 = k$ and $k_2 = 0$ is an optimal solution of ESPP.

Proof: Since n_1, n_2 and k are given, $\binom{n_1+n_2-k}{2}$ is fixed. Thus, maximizing $r(Q)$ is equivalent to maximizing $BC(Q)$.

(i) Let Q_1 (resp., Q_2) be a set which satisfies $k_1 = k$ and $k_2 = 0$, (resp., $k_1 = 0$ and $k_2 = k$). We compare $BC(Q_1)$ and $BC(Q_2)$. We have $BC(Q_1) = \binom{n_2}{2}$, $BC(Q_2) = \binom{n_1}{2}$. Since $n_2 \geq n_1$, we have $BC(Q_1) \geq BC(Q_2)$.

(ii) For $k_1 \geq 1$ and $k_2 \geq 1$, $BC(Q) = \binom{n_1-k_1}{2} + \binom{n_2-k_2}{2}$. Using $k_2 = k - k_1$, we obtain $BC(Q) = 0.5[2k_1^2 - 2(k+n_1-n_2)k_1 + (n_1^2+n_2^2-n_1-n_2-2n_2k+k+k^2)]$. It is easy to verify that $BC(Q)$ reaches its maximum at $k_1 = k-1$.

(iii) Using (i) and (ii), an optimal Q satisfies $k_1 = k-1$ or $k_1 = k$. Let $\hat{Q}$ (resp., $\bar{Q}$) be a set which satisfies $k_1 = k-1$ and $k_2 = 1$, (resp., $k_1 = k$ and $k_2 = 0$). We compare $BC(\hat{Q})$ and $BC(\bar{Q})$. We have $BC(\hat{Q}) = \binom{n_1-k_1}{2} + \binom{n_2-1}{2}$, $BC(\bar{Q}) = \binom{n_2}{2}$.

$$\begin{aligned} BC(\hat{Q}) - BC(\bar{Q}) &= \binom{n_1-k_1}{2} + \binom{n_2-1}{2} - \binom{n_2}{2} \\ &= 0.5[(n_1-k+1)(n_1-k) + (n_2-1)(n_2-2) - n_2(n_2-1)] \\ &= 0.5[k^2 - (2n_1+1)k + (n_1^2+n_1-2n_2+2)] \end{aligned}$$

Let $g(k) = k^2 - (2n_1+1)k + (n_1^2+n_1-2n_2+2)$. Since the discriminant $\Delta = (2n_1+1)^2 - 4(n_1^2 + n_1 - 2n_2 + 2) = 8n_2 - 7 > 0$, $g(k) = 0$ has two roots,

$$k' = \frac{1+2n_1-\sqrt{8n_2-7}}{2} \text{ and } k'' = \frac{1+2n_1+\sqrt{8n_2-7}}{2}$$

Observe that $k'' > n_1$ and can, therefore, be ignored. For $n_2 \geq 2$, $k' \leq n_1 - 1$. Let $\bar{k} = \lfloor k' \rfloor$. Then

(a) If $\bar{k} < 2$, for $2 \leq k \leq n_1 - 1$, $BC(\hat{Q}) < BC(\bar{Q})$, any set Q with $k_1 = k$ and $k_2 = 0$ is an optimal solution of ESPP.

(b) If $\bar{k} \geq 2$, for $2 \leq k \leq \bar{k}$, $BC(\hat{Q}) > BC(\bar{Q})$, any set Q with $k_1 = k - 1$ and $k_2 = 1$ is an optimal solution of ESPP. For $\bar{k} < k \leq n_1 - 1$, $BC(\hat{Q}) < BC(\bar{Q})$, any set Q with $k_1 = k$ and $k_2 = 0$ is an optimal solution of ESPP. ∎

Lemma 8 provides an optimal solution of ESPP for $2 \leq k \leq n_1 - 1$. In our next result, we show that the optimal value, $r(Q)$, of ESPP increases with $|Q|$ for $2 \leq |Q| \leq n_1 - 1$. Thus, given k, $2 \leq k \leq n_1 - 1$, an optimal solution of PP can be obtained by solving the corresponding instance of ESPP.

LEMMA 9. *Let $G = (U \cup V, E)$ be a biclique with $n_1 = |U| \leq |V| = n_2$. Let Q^* denote an optimal solution of an instance of PP defined by G and a positive integer k, $2 \leq k \leq n_1 - 1$. Then, $|Q^*| = k$. Consequently, the optimal solution can be obtained by solving the corresponding instance of ESPP.*

Proof: Let Q be an optimal solution of ESPP with parameter $\hat{k}$. Using Lemma 8, we have the following two possibilities:

(i) If Q satisfies $k_1 = \hat{k}$ and $k_2 = 0$, then

$$r(Q) = \frac{BC(Q)}{\binom{n - |Q|}{2}} = \frac{\binom{n_2}{2}}{\binom{n_1 + n_2 - \hat{k}}{2}}.$$

For $2 \leq \hat{k} \leq k$, $r(Q)$ reaches its maximum at $\hat{k} = k$.

(ii) If Q satisfies $k_1 = \hat{k} - 1$ and $k_2 = 1$, then

$$\begin{aligned} r(Q) &= \frac{\binom{n_1 - (\hat{k} - 1)}{2} + \binom{n_2 - 1}{2}}{\binom{n_1 + n_2 - \hat{k}}{2}} \\ &= \frac{\hat{k}^2 - (2n_1 + 1)\hat{k} + n_1^2 + n_1 + n_2^2 - 3n_2 + 2}{\hat{k}^2 - (2n_1 + 2n_2 - 1)\hat{k} + (n_1 + n_2)(n_1 + n_2 - 1)} \\ &= 1 - \frac{2(n_2 - 1)}{n_1 + n_2 - \hat{k} - 1} + \frac{2(n_2 - 1)^2}{(n_1 + n_2 - \hat{k})(n_1 + n_2 - \hat{k} - 1)} \end{aligned}$$

CLAIM 7. *For $2 \le \hat{k} \le n-1$, $r(Q)$ is non-decreasing with an increase in $\hat{k}$.*

Proof of Claim 7: Let $f_1(\hat{k}) = \frac{2(n_2-1)}{n_1+n_2-\hat{k}-1}$, $f_2(\hat{k}) = \frac{2(n_2-1)^2}{(n_1+n_2-\hat{k})(n_1+n_2-\hat{k}-1)}$. Thus, $r(Q) = 1 - f_1(\hat{k}) + f_2(\hat{k})$. Note that $-f_1(\hat{k})$ decreases with $\hat{k}$, $f_2(\hat{k})$ increases with $\hat{k}$. The derivative of $r(Q)$ with respect to $\hat{k}$ is

$$\begin{aligned} r'(Q) &= -f_1'(\hat{k}) + f_2'(\hat{k}) \\ &= -\frac{2(n_2-1)}{(n_1+n_2-\hat{k}-1)^2} + \frac{2(n_2-1)^2(2n_1+2n_2-2\hat{k}-1)}{(n_1+n_2-\hat{k})^2(n_1+n_2-\hat{k}-1)^2} \\ &= \frac{-2(n_2-1)(\hat{k}^2-2(n_1+1)\hat{k}+(n_1^2-n_2^2+2n_1+3n_2-1))}{(n_1+n_2-\hat{k})^2(n_1+n_2-\hat{k}-1)^2} \end{aligned}$$

For $h(\hat{k}) = \hat{k}^2 - 2(n_1+1)\hat{k} + (n_1^2 - n_2^2 + 2n_1 + 3n_2 - 1)$, the discriminant $\Delta = 4(n_1+1)^2 - 4(n_1^2 - n_2^2 + 2n_1 + 3n_2 - 1) = 4(n_2^2 - 3n_2 + 2)$. For $n_2 > 2$, $\Delta = 4(n_2-2)(n_2-1) > 0$, so $h(\hat{k}) = 0$ has two roots: $k' = n_1 + 1 - \sqrt{n_2^2 - 3n_2 + 2}, k'' = n_1 + 1 + \sqrt{n_2^2 - 3n_2 + 2}$. Thus, $h(\hat{k}) > 0$ for $\hat{k} < k'$ or $\hat{k} > k''$; $h(\hat{k}) < 0$ for $k' < \hat{k} < k''$. Note that $k'' = n_1 + 1 + \sqrt{n_2^2 - 3n_2 + 2} > n_1$ and can, therefore, be ignored.

(a) If $n_1 < n_2$, it is easy to verify that $k' = n_1 + 1 - \sqrt{n_2^2 - 3n_2 + 2} < 2$. Thus, for $2 \le \hat{k} \le n-1$, we have $h(\hat{k}) < 0$, which implies $r'(Q) = -f_1'(\hat{k}) + f_2'(\hat{k}) > 0$. It follows that $r(Q)$ increases with $\hat{k}$.

(b) If $n_1 = n_2$, it is easy to verify that $2 < k' = n_1 + 1 - \sqrt{n_1^2 - 3n_1 + 2} < 3$. Also, for $S, S' \subseteq U \cup V$ with $|S| = 2$ and $|S'| = 3$,

$$r(S) - r(S') = \frac{\binom{n_1-1}{2} + \binom{n_1-1}{2}}{\binom{2n_1-2}{2}} - \frac{\binom{n_1-1}{2} + \binom{n_1-2}{2}}{\binom{2n_1-3}{2}} = 0$$

Thus, $r(S) = r(S')$. To conclude, for $2 \le k \le n_1 - 1$, $r(\hat{k})$ reaches its maximum at $\hat{k} = k$.

□

Finally, note that when $\hat{k}$ changes from $\bar{k}$ to $\bar{k}+1$ (from Lemma 8, $\bar{k} = \lfloor \frac{1+2n_1-\sqrt{8n_2-7}}{2} \rfloor$), the optimal solution of ESPP changes form from $k_1 = \hat{k} - 1$ and $k_2 = 1$ to $k_1 = \hat{k}$ and $k_2 = 0$. Let Q' (resp., Q'') be an optimal solution of ESPP for $\hat{k} = \bar{k}$ (resp., $\hat{k} = \bar{k}+1$). Then Q' satisfies $k_1 = \bar{k} - 1$ and $k_2 = 1$ and Q'' satisfies $k_1 = \bar{k} + 1$ and $k_2 = 0$. From (i) and (ii) above, it is easy to verify that $r(Q'') > r(Q')$. The result follows. ■

To summarize, the results of Lemmas 7, 8, and 9, provide a complete solution of PP. We formally state the solution below.

THEOREM 7. *Let $G=(U\cup V,E)$ be a biclique with $n_1=|U|\le|V|=n_2$. Let $Q\subseteq U\cup V$. Let $k_1=|Q\cap U|$, $k_2=|Q\cap V|$, $\bar{k}=\lfloor\frac{1+2n_1-\sqrt{8n_2-7}}{2}\rfloor$.*

(a) If $k=1$, then any node $u\in U$ is an optimal solution to PP.

(b) If $\bar{k}\ge 2$, then for $2\le k\le\bar{k}$, any set Q which satisfies $k_1=k-1$ and $k_2=1$ is an optimal solution of PP.

(c) If $\bar{k}\ge 2$, then for $\bar{k}<k\le n_1-1$, any set Q which satisfies $k_1=k$ and $k_2=0$ is an optimal solution of PP.

(d) If $\bar{k}<2$, then for $2\le k\le n_1-1$, any set Q which satisfies $k_1=k$ and $k_2=0$ is an optimal solution of PP.

(e) If $k\ge n_1$, then $Q=U$ is an optimal solution to PP.

We also summarize the solution of ESPP.

COROLLARY 1. *Let $G=(U\cup V,E)$ be a biclique with $n_1=|U|\le|V|=n_2$. Let $Q\subseteq U\cup V$. Let $k_1=|Q\cap U|$, $k_2=|Q\cap V|$, $\bar{k}=\lfloor\frac{1+2n_1-\sqrt{8n_2-7}}{2}\rfloor$.*

1. For (a) *$k=1$,* (b) *$\bar{k}\ge 2$ and $2\le k\le\bar{k}$,* (c) *$\bar{k}\ge 2$ and $\bar{k}<k\le n_1-1$, and* (d) *$\bar{k}<2$ and $2\le k\le n_1-1$, the corresponding solutions described in Theorem 7 are also optimal for ESPP.*

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

3.2.3. Proof of Hardness of PP and ESPP

The strongly NP-Complete problem which we use in our reduction is Independent Set Problem.

Independent Set Problem (ISP)

Instance. An 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 8. *The decision problem corresponding to PP is strongly NP-Complete.*

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

DECISION QUESTION: Does there exist a portal $Q \subseteq V$ in $G(V, E)$ such that $|Q| \leq |V| - k$ and $r(Q) \geq 1$?

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, $|Q^*| \leq |V| - k^*$. From the definition of an independent set, it follows that all 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 $|Q| \leq |V| - k$ and $r(Q) \geq 1$, then the set $V \backslash Q$ is an independent set of at least k nodes. ∎

COROLLARY 2. *The decision problem corresponding to ESPP is strongly NP-Complete.*

4. Strategic Analysis and Operational Resource Allocation

Typically, structural search is a strategic issue. Unless there are frequent and significant changes in the topology of the network, an influential or central group of individuals is likely to maintain their collective role over a reasonable time period. For instance, law enforcement agencies, in attempting to combat the activities of sophisticated criminal organizations, often need to identify key groups of members or identify principal vulnerabilities in criminal networks (Sparrow 1991). The results of these searches typically continue to be of interest for several years. In some cases, structural search might be costly; e.g., for consumer-goods marketers, locating and identifying opinion leaders is a difficult and expensive undertaking (Weimann 1994, Robertson et al. 1984). Thus, structural search may be needed to be performed sporadically. On the other hand, a secondary problem that uses the result of the search and typically needs to be resolved more frequently is the efficient allocation of resources among the members of the chosen group. To illustrate the need for such a problem, we provide several examples of social networks from the literature.

• Orgnet.com (2008) discusses the importance of an elite group in a social network and provides an example of pharmaceutical firms that are interested in identifying a group of physicians who are key opinion leaders in their social network. The firms aim to focus their marketing efforts to the members of this group. For such firms, finding an appropriate elite group is the first step. Subsequently, a critical task is the allocation of valuable marketing resources targeted towards influencing these opinion leaders to promote their drugs.

• Klovdahl et al. (2001) illustrate the importance of a portal for a tuberculosis outbreak network. Here, the first goal is to identify a set of locations where medical relief needs to be made available immediately to help prevent the spread of the disease elsewhere. Thus, depending on the relative needs at these critical locations, an optimal allocation of the government's limited resources is an important problem.

• In lobbying networks, a lobbyist may have access only to a certain set of legislators (Carpenter et al. 1998). Thus, a customer (usually a company or an interest group) wanting to lobby influential decision makers to favor a bill may consider hiring several lobbyists. After identifying a set of influential decision makers, the task is to "cover" them using an appropriate set of lobbyists: each decision maker is approached by one or more lobbyists and each lobbyist is assigned the task of influencing one or more decision makers. Naturally, the customer is interested in hiring the most effective team of lobbyists from those that are feasible under its budget constraints and appropriate for its strategic objectives. A similar problem arises during the lobbying of super-delegates in the democratic presidential nomination in the United States (Nagourney and Hulse 2008).

Due to the tactical nature of the resource allocation problems that might arise subsequent to structural search, it is both natural and convenient to first solve the search problem and then address the resource allocation on the result of the search. As the examples above indicate, the resource allocation subproblems are typically (but not necessarily) constrained *bipartite* assignment problems, with the chosen members and the resources as the two sides of the bipartition. A wide variety of bipartite assignment problems have been investigated in the literature (see, e.g., Ahuja et al. 1993, Garey and Johnson 1979, Nemhauser and Wolsey 1988). Thus, an efficient algorithm may

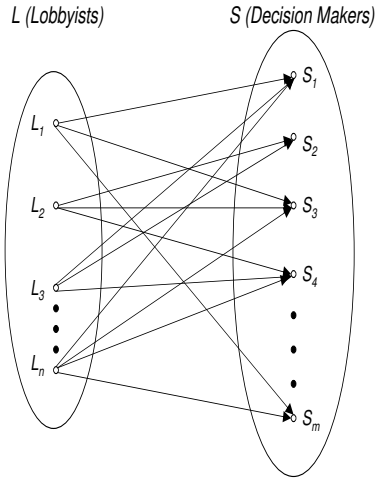

Figure 12 The Assignment Problem in a Lobbyist Network.

be available to solve a resource allocation problem. Consider, for example, the problem mentioned above in a lobbyist network. Here, the lobbyists (say, a total of n) and the decision makers (say, a total of m) represent nodes for the two sides of the bipartition (Figure 12).

The edge set A represents feasible assignments of lobbyists to decision makers; the weight w_{ij} of an edge (i,j) indicates a normalized numerical measure (e.g., the desirability and/or the cost) of assigning Lobbyist L_i to decision-maker S_j. Let $x_{ij} = 1$ if L_i is assigned to S_j; 0 otherwise. Then, given $G(L \cup S, A)$, the following problem of obtaining a cost-minimizing assignment of the lobbyists to the decision makers such that (i) each decision maker is contacted by at least $r \geq 1$ different lobbyists, and (ii) each hired lobbyist contacts at most $t \geq 1$ decision makers, is an instance of the well-known (and efficiently solvable) bipartite assignment problem.

$$\begin{aligned} & & \min & \sum_{(i,j)\in A} w_{ij}x_{ij} & \\ \text{s.t.} & \sum_i x_{ij} & \geq & r & \forall j \\ & \sum_j x_{ij} & \leq & t & \forall i \\ & x_{ij} & \in & \{0,1\} & \forall i,j. \end{aligned}$$

5. Conclusions and Future Research Directions

The ability to find useful structures in social networks will undoubtedly benefit their users and 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 – the Elite Group Problem (EGP) and the Portal Problem (PP) – derived, respectively, from the notions of influence and centrality. We demonstrate the relevance of these problems on a variety of social networks and show that (i) the basic EGP is polynomially solvable, (ii) the PP and a size-constrained version of EGP are both strongly NP-hard. We also analyze these problems on a few special networks. Finally, we highlight the need for solving a resource allocation problem – to distribute limited resources among the chosen players – subsequent to the identification of a solution to the search problem.

Popular social networks have experienced an explosive growth in recent years. For example, social networking sites such as Facebook and MySpace have typically added more than a million users each month in recent years; currently, both services attract about 115 million users to their sites each month (Arrington 2008). The ability to conduct efficient structural searches in such networks will undoubtedly play a key role in improving their utility for members and organizations. From the point of view of ordinary users, the availability of efficient structural search provides an opportunity to extend their social contacts, e.g., a user might want to check if she is "connected" to another user by a path of pairwise acquaintances. Organizations can profitably use search to identify teams of interest, e.g., a project manager in need for a limited number of members with appropriate, and typically complementary, skills. Similarly, the networking sites could benefit from making search available to special-interest groups. For example, as in Sharma and Steel (2008), an advertising agency may want to find groups of users who would likely be interested in its products and focus on targeted advertising to these groups.

In the industry, the focus, thus far, has been on developing "social search engines" to search social media and user-generated content, e.g., Twitter (http://search.twitter.com/), Social Mention (http://www.socialmention.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. Since 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).

References

Ahuja, R., T. Magnanti, J. Orlin. 1993. *Network Flows: Theory, Algorithms, and Applications*, Second Edition, Pearson Education.

Arrington, M. 2008. Facebook No Longer The Second Largest Social Network, TechCrunch, available at http://www.techcrunch.com/2008/06/12/facebook-no-longer-the-second-largest-social-network/

Boyd, D. M., N. B. Ellison. 2007. Social Network Sites: Definition, History, and Scholarship, *Journal of Computer-Mediated Communication*, **13**(1).

Carpenter, D. P., K. M. Esterling, D. M. J. Lazer. 1998. The Strength of Weak Ties in Lobbying Networks: Evidence from Health-Care Politics in the United States, *Journal of Theoretical Politics*, **10**(4), 417-444.

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

Cattani, G., S. Ferriani. 2008. A Core/Periphery Perspective on Individual Creative Performance: Social Networks and Cinematic Achievements in the Hollywood Film Industry, *Organization Science*, **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.

Erickson B. H. 1981. Secrect Societies and Social Structure, *Social Forces*, **60**(1), 188-210.

Everett M. G., S. P. Borgatti. 1999. The Centrality of Groups and Classes, *Journal of Mathematical Sociology*, **23**(3), 181-201.

Fowler J. H. 2006. Legislative Cosponsorship Networks in the US House and Senate, *Social Networks*, **28**, 454-465.

Freeman, L. C. 1979. Centrality in Social Networks: Conceptual Clarification, *Social Networks*, **1**(3), 215-239.

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

Gloor, P. A., D. Oster, J. Putzke, K. Fischback, D. Schoder, K. Ara, T. J. Kim, R. Laubacher, A. Mohan, D. O. Olguin, A. Pentland, B. N. Waber. 2007. Studying Microscopic Peer-to-Peer Communication Patterns, *Americas Conference on Information Systems*, 2007.

Greene, K. 2008. 10 Emerging Technologies 2008, *Technology Review*, 2008.

Grewal, R., G.L. Lilien, G. Mallapragada. 2006. Location, Location, Location: How Network Embeddedness Affects Project Success in Open Source Systems, *Management Science*, **52**(7), 1043-1056.

Hawe, P., L. Ghali. 2008. Use of Social Network Analysis to Map the Social Relationships of Staff and Teachers at School, *Health Education Research*, **23**, 62-69.

Hesseldahl, A. 2008. There's Gold in 'Reality Mining', *Business Week*, **23**, March 24, 2008.

Hoffman, A. J., J. B. Kruskal. 1956. Integral Boundary Points of Convex Polyhedra, *Linear Inequalities and Related Systems, Annals of Mathematics Studies*, **38**, 223-246.

Hoffman, C. C., L. Wilcox. 1992. Sociometric Applications in a Corporate Environment, *Journal of Group Psychotherapy*, **45**(1), 3-14.

IBM. 2007. Connect with Customers: Social Networking and Decision Making. Available at https://www.ibm-304.com/jct03004c/businesscenter/smb/us/en/contenttemplate/!!/gcl_xmlid=114836

Kim J., S. J. McMillan. 2008. Evaluation of Internet Advertising Research, *Journal of Advertising*, **37**(1), 99-112.

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, *Social Science and Medicine*, **52**, 681-694.

Morselli, C., C. Giguere. 2006. Legitimate Strengths in Criminal Networks, *Crime, Law and Social Change*, **45**, 185-200.

Nagourney, A., C. Hulse. 2008. Neck and Neck, Democrats Woo Superdelegates, *The New York Times*, Feb 10, 2008.

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

Orgnet.com. 2008. Finding Key Opinion Leaders and Influentials Using Social Network Analysis, avalilable at http://orgnet.com/KOL.html

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 Science*, **48**(1), 24-43.

Pentland, A. 2004. 'Reality Mining' the Organization, *Technology Review*, March, 2004.

Puzis R., Y. Elovici, S. Dolev. 2007. Fast Algorithm for Successive Computation of Group Betweenness Centrality, *Physical Review E*, **76**(5), 056709.

Reagans, R. E. Zuckerman. 2001. Networks, Diversity, and Productivity: The Social Capital of Corporate R & D Teams. *Organization Science*, **12**(4), 502-517.

Robertson, T. S., J. Zielinski, S. Ward. 1984. *Consumer Behavior*, Scott, Foresman, and Company.

Scott, J. 2000. *Social Network Analysis: A Handbook, Second Edition*, Sage Publications, Thousand Oaks, CA.

Sharma A., E. Steel. 2008. Targeted-Ad Initiative Is Crucial for MySpace, *The Wall Street Journal*, Aug 4, 2008.

Sparrow, M. K. 1991. The Application of Network Analysis to Criminal Intelligence: An Assessment of the Prospects, *Social Networks*, **13**, 251-274.

Stephenson, K. A. and Zelen, M. 1989. Rethinking Centrality: Methods and Examples, *Social Networks*, **11**, 1-37.

Thompson, C. 2008. Real-World Social Networks vs. Facebook 'Friends', *Wired Magazine*, **16**(8), 2008.

Weimann, G. 1994. *The Influentials: People Who Influence People*, State University of New York Press.

Wellman, B., W. Chen, W. Dong. 2001. *Networking Guanxi, Social Networks in China: Institutions, Culture, and the Changing Nature of Guanxi*, Cambridge University Press.