Mining Structural Hole Spanners Through Information Diffusion in Social Networks

Tiancheng Lou[♯†*] and Jie Tang[†]
[♯]Google, Inc., Mountain View, CA 94043, US
[†]Tsinghua University, Beijing 100084, China
acrush@google.com, jietang@tsinghua.edu.cn

ABSTRACT

The theory of *structural holes* [4] suggests that individuals would benefit from filling the "holes" (called as structural hole spanners) between people or groups that are otherwise disconnected. A few empirical studies have verified that structural hole spanners play a key role in the information diffusion. However, there is still lack of a principled methodology to detect structural hole spanners from a given social network.

In this work, we precisely define the problem of mining top-k structural hole spanners in large-scale social networks and provide an objective (quality) function to formalize the problem. Two instantiation models have been developed to implement the objective function. For the first model, we present an exact algorithm to solve it and prove its convergence. As for the second model, the optimization is proved to be NP-hard, and we design an efficient algorithm with provable approximation guarantees.

We test the proposed models on three different networks: Coauthor, Twitter, and Inventor. Our study provides evidence for the theory of structural holes, e.g., 1% of Twitter users who span structural holes control 25% of the information diffusion on Twitter. We compare the proposed models with several alternative methods and the results show that our models clearly outperform the comparison methods. Our experiments also demonstrate that the detected structural hole spanners can help other social network applications, such as community kernel detection and link prediction. To the best of our knowledge, this is the first attempt to address the problem of mining structural hole spanners in large social networks.

Categories and Subject Descriptors

J.4 [**Social and Behavioral Sciences**]: Miscellaneous

General Terms

Algorithms, Experimentation

Keywords

Structural hole, Social network, Information diffusion, Minimal cut

1. INTRODUCTION

In sociology, there are a few well-established ideas on how positions in social networks benefit those people who occupy them [7]. One idea is that positions which act as an intermediary or a bridge between individuals of different groups tend to have access to a richer supply of information and have more control over their network relations. The notion forms the basis for the theory of *structural holes* [4], which suggests that advantages accrue to people who occupy such bridging positions. For example, if a researcher spans a structural hole, he/she could apply ideas and techniques from one group to problems faced by the other, or innovate by synthesizing ideas from different groups.

A series of empirical studies have demonstrated how structural holes positively relate to a wide range of indicators of social success [1, 5, 6, 28]. A few other papers use game theory to model the formation of structural holes in social networks. Goyal and Vega-Redondo [12] propose a model in which a node A potentially benefits from serving as an intermediary between nodes B and C even when it resides on an arbitrarily long B-C path. According to the presented model, the strategic link formation leads to a star network. However, in real world, many networks are not necessarily of the star topology. [7] explicitly models the notion of structural holes using a network formation game. It is based on the idea that A can only benefit from being an intermediary when A is on a length-two path between B and C. Kleinberg et al. [18] study the strategic and dynamic aspect to the theory of structural holes. They extend Burt's work [5] by modeling how social networks change over time if everyone is vying for those bridging positions. However, while much research has focused on studying the correlation between structural hole spanners and their success within an organization (as indicated by salary, reviews, promotion, and other measures), few work systematically investigates the problem in large online social networks.

We address the problem of identifying structural holes spanners: given a social network, who are the top-k users spanning structural holes in the social network, and what are the underlying patterns of these structural hole spanners? The problem could be considered as the reverse process of the study on strategic network formation with structural holes [7, 12, 18]. The latter problem is to design a game-theoretic model to study the evolution of network structure, while our problem is to detect who are likely to span structural holes in social networks based on the network structure.

The Model. We assume a setting in which a set V of n distinct users form l groups $\mathbf{C} = \{C_1, \cdots, C_l\}$ (called communities). A utility function $Q(v, \mathbf{C})$ is defined for each node to measure its degree to span structural holes. Formally, we have the following definition,

Definition 1. **Top-k Structural Hole Spanners.** Let $G = (V, E)$ denote a social network, where $V = \{v_1, v_2, \cdots, v_n\}$ is a set of n users, and $E \subseteq V \times V$ is a set of undirected social

*This work was done when the first author was studying in Tsinghua University.

WWW 2013, May 13–17, 2013, Rio de Janeiro, Brazil.
ACM 978-1-4503-2035-1/13/05.

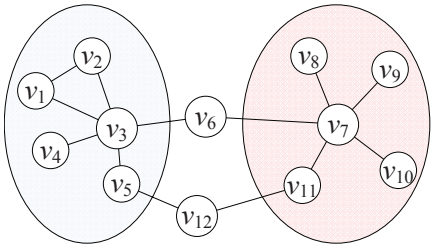


Figure 1: Illustration of structural holes. Nodes in the two ellipses form two communities. v_6 and v_{12} can be considered to span the structural hole between the two communities.

relationships between users. Further assume that the nodes of the social network can be grouped into l (overlapping) communities $\mathbf{C} = \{C_1, \cdots, C_l\}$, with $V = C_1 \cup \cdots \cup C_l$. Then, the top-$k$ structural hole spanners are defined as a subset of k nodes, denoted as V_{SH} in the network, which maximizes the following utility (quality) function:

$$\max_{V_{SH}} Q(V_{SH}, \mathbf{C}), \text{with} \quad |V_{SH}| = k \tag{1}$$

In the formulation, links can be directed or undirected. The utility function $Q(V_{SH}, \mathbf{C})$ is a general definition, which can be instantiated in different ways. Note that in the definition, we only consider the network information but not the content information. Our goal is to give a theoretical analysis for this problem. Combining the network information and the content information in practical mining algorithms is left as one of our future works.

We develop two instantiation models based on the above objective function. The general idea behind is to measure how a node bridges different communities. In the first model, we consider the importance of those connected nodes for mining the structural hole spanners. That is, if a node connects multiple important nodes (authoritative users), then the node is more likely to be a structural hole spanner. In the second model, we directly measure each node according to the theory of minimal cut on network. The problem is cast as finding k nodes such that after removing these nodes, the *decrease* of minimal cut for communities $\mathbf{C}$ in network G can be maximized. For both models, we provide theoretical analysis and develop efficient algorithms to solve with provable approximation guarantees. As far as we know, it is the first attempt to prove the NP-hardness of maximizing the *decrease* of minimal cut in an unweighted graph.

The problem poses a set of challenges. Figure 1 shows an example of structural holes with two communities. It is easy to see that v_6 and v_{12} can be viewed as structural hole spanners between the two communities. However, there are still several challenging questions: (1) Which node (v_6 or v_{12}) has a higher degree to span structural holes? How to quantify the degree of each node to span structural holes? (2) How to efficiently select top- k nodes to maximize the utility function (Eq. 1)? (3) How the detected structural hole nodes can help other social networking applications?

Results. In this work, we focus on studying the problem of mining top-k structural hole spanners in large-scale networks from both theoretical and empirical aspects. We test the proposed models on three different networks: Coauthor, Twitter, and Inventor. In Coauthor, we try to understand who act as bridges between different research communities; in Twitter, we attempt to detect who act as intermediaries for information diffusion; in Inventor, we study how technologies diffuse across different companies via inventors who span structural holes. Our study presents the following results:

- 1% of Twitter users who span structural holes control 25% of the information diffusion (retweeting). This provides a strong evidence for the theory of structural hole [4, 5].
- We compare the proposed models with several alternative methods for detecting structural hole spanners and the results show that our models clearly outperform (+20-40% for maximizing the information diffusion) the comparison methods.
- We apply the detected structural hole spanners to help communities detection [32] and link prediction [29], two important applications in social networks. Results demonstrate that the structural hole information can significantly improve the quality (+10% in terms of F1-score) of communities detection and improve the performance (+3-4% by F1-score) of predicting the type of relationships in two different networks.

Table 1: Statistics of the three networks. #Articles respectively indicates the number of publications, tweets, and patents in the three networks.

Dataset	#Users	#Relationships	#Articles
Coauthor [31]	815,946	2,792,833	1,572,277
Twitter [15]	112,044	468,238	2,409,768
Inventor [30]	2,445,351	5,841,940	3,880,211

Organization. Section 2 introduces the data sets used in our study and our observations over different networks. Section 3 presents the proposed model and describes the algorithm for solving the model; Section 4 describes potential applications of mining structural hole spanners. Section 5 and Section 6 present the results. Finally, Section 7 discusses related work and Section 8 concludes.

2. DATA AND OBSERVATIONS

Before proceeding, we first engage in some high-level investigation of structural holes in several different social networks.

Data collections. We consider three different types of networks for studying the structural hole problem: Coauthor, Twitter and Inventor. Table 1 gives basic statistics of the three networks.

Coauthor is a network of authors, collected by ArnetMiner[1]. The data set is obtained from [31]. The network consists of 815,946 authors and 2,792,833 coauthorships. For the evaluation purpose, we create a sub-network, which contains coauthorships extracted from papers published at 28 major computer science conferences. These conferences cover six research areas: Artificial Intelligence (AI), Databases (DB), Data Mining (DM), Distributed Parallel Computing (DP), Graphics, Vision and HCI (GV), as well as Networks, Communications and Performance (NC)[2]. Each conference has a group of program committee (PC) members. We extracted the PC member information from respective conference websites from 2008 to 2010. Computer scientists who served as PC members at conferences of different areas are considered as spanning structural

[1]http://arnetminer.org, an academic search system.

[2]AI: IJCAI, AAAI, ICML, UAI, UMAP, NIPS, and AAMAS; DB: VLDB, SIGMOD, PODS, ICDE, ICDT, and EDBT; DM: SIGKDD and ICDM; DP: PPoPP, PACT, IPDPS, ICPP, and Euro-Par; GV: SIGGRAPH, CVPR, and ICCV; NC: SIGCOMM, PERFORMANCE, SIGMETRICS, INFOCOM, and MOBICOM.

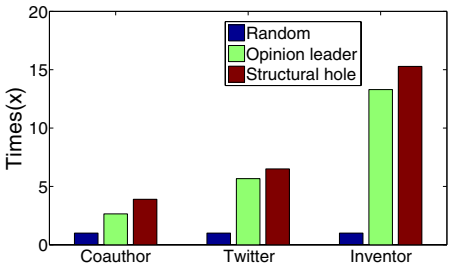


Figure 2: Structural hole spanners are more likely to connect important nodes than opinion leaders. Random is the average number of publications/tweets/patents authored by neighborhood nodes of a random user in the respective network. The average number is taken as a unit of measurement and the y-axis indicates the average score of different categories of users under this measurement.

holes across those areas. In total, we extracted 1,718 PC members, among whom 107 PC members span structural holes. Our goal is to identify those PC members who span structural holes from the coauthor network.

Twitter is crawled from Twitter.com, a widely used microblogging system. The data set is obtained from [15, 22]. The subnetwork is comprised of 112,044 users, 468,238 following links among them, and all tweets (2,409,768 tweets) posted by these users. Here, we examine the role of structural hole spanners in the information diffusion process on Twitter.

Inventor is a network of inventors, extracted from a large patent data set from USPTO[3]. The data set is obtain from [30]. The inventor network contains 2,445,351 inventors and 5,841,940 co-inventing relationships. Each company is considered as a community. We study how technologies spread across different companies via inventors who span structural holes.

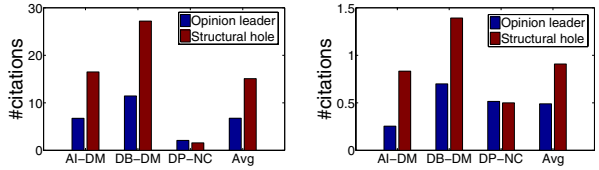
Observable analysis. We study the different behavior patterns between structural hold spanners and opinion leaders, and the interplay between structural hold spanners and information diffusion. Intuitively, we have the following questions:

- How likely would structural hole spanners connect with "opinion leaders"?
- How likely would structural hole spanners influence the information diffusion?

Structural hole spanner and Opinion leaders. We study the connectivity between opinion leaders and different categories of users (Opinion leaders, Structural holes, and Random). In Coauthor, we take all PC members as opinion leaders and PC members who serve at conferences of different areas as structural holes. In Inventor, we perform PageRank [27] and select the top 1% users with the highest PageRank scores as opinion leaders, and those top users who worked in different companies as structural hole spanners. In Twitter, we also perform PageRank on the following network and select the top 1% users as opinion leaders. We use the $(\alpha - \beta)$ community detection algorithm [14, 24] to find overlapping communities. Then those top users who exist in different communities are treated as structural hole spanners.

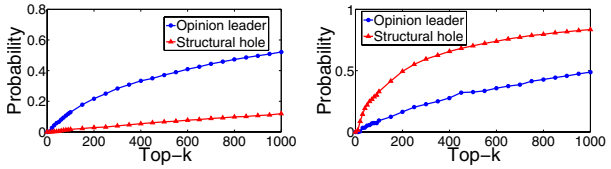
Figure 2 shows that structural hole spanners are more likely (+15-50%) to connect important nodes than opinion leaders. In

[3]http://uspto.gov/, the US patent and trademark office.



(a) Cross domain (b) Outer domain

Figure 3: Information diffusion on the Coauthor network. (a) Average number of cross-domain citations received by structural hole spanners and opinion leaders, e.g., in the AI-DM case, for opinion leaders in AI, only citations from DM are considered. (b) Average number of outer-domain (those domains other than AI and DM) citations received by structural hole spanners and opinion leaders.



(a) Inner domain (b) Cross domain

Figure 4: Information diffusion on the Twitter network. x-axis indicates top k opinion leaders/structural hole spanners; and y-axis indicates the probability of opinion leaders (or structural hole spanners) appearing on a tweet-forwarding path spread within a domain or across different domains.

Figure 2, Random stands for the average number of publications/tweets/patents authored by neighborhood nodes of a random user in the respective network. The average number is taken as a unit of measurement and the y-axis indicates the average score of different categories of users under this measurement. The analysis provides evidence in support of the first proposed model (Cf. §3).

Structural hole spanner and information diffusion. We perform another analysis of information diffusion in the Coauthor and the Twitter networks. In Coauthor, we consider the citation as a type of information diffusion. We count the average numbers of cross-domain citations received by opinion leaders and structural hole spanners. (Within the same domain, opinion leaders receive higher citations than others.) We find a striking phenomenon (Figure 3): contrast to the inner-domain citation, the average number of cross-domain citations received by the structural hole spanners almost doubles the number of opinion leaders. It seems that people from a research community C_1, if they want to follow the work related to community C_2, are more likely to refer (cite) the work of a researcher who spans the structural hole between C_1 and C_2, rather than an opinion leader in community C_2.

For the analysis of information diffusion in Twitter, we estimate the probability of opinion leaders (or structural hole spanners) appearing on a tweet-forwarding path. Figure 4 shows some interesting patterns on Twitter: opinion leaders play a key role in spreading information within a community (Figure 4(a)), while structural hole spanners are more important for spreading information between communities (Figure 4(b)). Another striking phenomenon is that the top 1% ($k = 1,000$) of structural hole spanners control almost 80% of the information diffusion between different communities, and 25% of all the information diffusion on Twitter.

The above observations constitute the intuition behind the following proposed models.

3. MODELS AND ALGORITHMS

We develop two instantiation models for the utility function (Eq. 1). The first model considers the connectivity between opinion leaders and structural hole spanners (Cf. Figure 2). The user's importance (authority) and the degree of spanning structural hole are defined in terms of one another in a mutual recursion. The intuition of the second model is from Figures 3 and 4. The model is defined based on the theory of information flow. We provide theoretical analysis for the two models and prove the NP-hardness for the second model, and develop efficient algorithms to achieve provable approximation guarantees.

3.1 Model One: HIS

The intuition of the first model can be also explained using the two-step information flow theory [16, 20], which suggests that ideas (innovations) usually flow first to opinion leaders, and then from the opinions leaders to a wider population. In this sense, if a user is connected with many "opinion leaders" in different communities, then the user is more likely to span structural holes. For example, in Figure 1, nodes v_3 and v_7 act as opinion leaders respectively in the two communities, thus they have more power to spread information than other nodes (such as v_5 and v_{11}). v_6 and v_{12} can be considered as the bridge to connect the two communities. By comparing v_6 with v_{12}, we see v_6 connects the two opinion leaders of the two communities, while v_{12} only connects two ordinary users. According to the information flow theory, v_6 would have a higher informational advantage (i.e., higher degree to span structural holes) than v_{12}. On the other hand, it would be natural to enhance the power of information spread through the connection to structural hole spanners. Thus, a node is more likely to act as an opinion leader, if it connects with structural hole spanners. Based on this intuition, we develop the first model, referred to as HIS. To begin with, we first give the following definition:

Definition 2. Given a network G, let $\mathbf{C} = \{C_1, \cdots, C_l\}$ denote l communities in the network G; let $I(v, C_i) \in [0, 1]$ be the importance of v in community C_i. Then for each subset of communities S such that $S \subseteq \mathbf{C}$ and $|S| \geq 2$, we define $H(v, S) \in [0, 1]$ as the structural hole score of v in S, i.e., the likelihood of v spanning structural holes across all communities in S.

Here, each node has an importance score in each community and a structural hole score in every possible $S \subseteq \mathbf{C}$ ($|S| \geq 2$). The two types of scores are defined in terms of each other in a mutual recursion as follows:

$$I(v, C_i) = \max_{\substack{e_{uv} \in E, \\ S \subseteq \mathbf{C} \wedge C_i \in S}} \{I(v, C_i), \alpha_i I(u, C_i) + \beta_S H(u, S)\} \quad (2)$$

$$H(v, S) = \min_{C_i \in S} \{I(v, C_i)\} \quad (3)$$

where α_i and β_S are two tunable parameters. The importance score of user v is computed as the maximal value of the linear combination of v's friend's importance score and the structural hole score. The structural hole score is then defined as the minimal value of user v's importance scores in different communities in S. Essentially, in Eq. 2, the importance score can be explained as the maximal information flow a user can receive from one of her/his friends. Eq. 3 suggests that a structural hole spanner in S should be active in all the communities in S. The two update rules, Eqs. 2 and 3 are the basic approaches by which structural holes and importance (authority) reinforce each other. For initialization, we can use an algorithm such as PageRank [27] or HITS [17] to calculate the authority score $r(v)$ of each node v. Then we initialize the importance score $I(v, C_i)$ in the following ways:

$$\begin{aligned} I(v, C_i) &= r(v), \quad v \in C_i \\ I(v, C_i) &= 0, \qquad v \notin C_i \end{aligned} \quad (4)$$

The two update rules Eqs. 2 and 3 run in an alternating fashion until desired equilibrium values for the two scores are reached. Practically, the two scores $I(v, C_i)$ and $H(v, S)$ could be infinitely large without any constraints. We give the following theorem for the condition of the existence of a convergent solution.

Theorem 1. Given α_i and β_S, the two scores $I(v, C_i)$ and $H(v, S)$ always exists for any graph $G = (V, E)$, if and only if,

$$\max_{C_i \in \mathbf{C},\ C_i \in S} \{\alpha_i + \beta_S\} \leq 1 \quad (5)$$

PROOF. For the *only if* direction, suppose there exists $C_i \in \mathbf{C}$ and $C_i \in S$ such that $\alpha_i + \beta_S > 1$. We consider two connected nodes v_1 and v_2, with $r(v_1) = r(v_2) = 1$, $v_1 \in \cap_{C_j \in S} C_j$ and $v_2 \in C_i$. Thus, we have $I(v_1, C_i) = 1$. By Eq. 3, we get $H(v_1, S) = \min_{C_j \in S} I(v_1, C_j) = 1$. By Eq. 2, we get $I(v_2, i) \geq \alpha_i I(v_1, C_i) + \beta_S H(v_1, S) = \alpha_i + \beta_S > 1$, which is impossible.

Now we prove the *if* direction, if for each C_i and $C_i \in S$, we have $\alpha_i + \beta_S \leq 1$. We can use induction to prove that, after infinite number of iterations, it satisfies $I(v, C_i) \leq 1$. In the first iteration, we have $I^{(0)}(v, C_i) \leq r(v) \leq 1$. After the k-th iteration, we have $I^{(k)}(v, C_i) \leq r(v) \leq 1$. Hence, in the $(k+1)$-th iteration, for each $C_i \in S$, we have $I^{(k+1)}(v, C_i) \leq \alpha_i I^{(k)}(u, C_i) + \beta_S H^{(k)}(u, S) \leq (\alpha_i + \beta_S) I^{(k)}(u, C_i) \leq I^{(k)}(u, C_i) \leq 1$. □

Algorithm 1 gives the implementation to update Eqs. 2 and 3, which results in a complexity of $O(K2^l|E|)$, where K is the number of iterations. Let us first prove the ϵ-convergence of the algorithm and then discuss its efficiency.

Theorem 2. Algorithm 1 satisfies ϵ-convergence. Denote $\gamma = \max_{C_i \in \mathbf{C},\ C_i \in S} \{\alpha_i + \beta_S\}$, we have

$$\max_{v \in V,\ C_i \in \mathbf{C}} |I^{(k+1)}(v, C_i) - I^{(k)}(v, C_i)| \leq \gamma^k \quad (6)$$

PROOF. Firstly, parameters α_i and β_S satisfy $\gamma = \max_{C_i \in \mathbf{C},\ C_i \in S} \{\alpha_i + \beta_S\} \leq 1$. In addition, during the iterations, for any $v \in V$, $C_i \in \mathbf{C}$ and $S \subseteq \mathbf{C}$, the value of $I^{(k)}(v, C_i)$ and $H^{(k)}(v, S)$ are non-decreasing wrt the parameter k.

Now, we use induction to prove

$$\max_{v \in V, C_i \in \mathbf{C}} |I^{(k+1)}(v, C_i) - I^{(k)}(v, C_i)| \leq \gamma^k$$

and

$$\max_{v \in V, S \subseteq \mathbf{C}} |H^{(k+1)}(v, S) - H^{(k)}(v, S)| \leq \gamma^k.$$

When $k = 0$, for each $v \in V$ and $C_i \in \mathbf{C}$, we have $I^{(1)}(v, C_i) \leq 1$, thus $|I^{(1)}(v, C_i) - I^{(0)}(v, C_i)| \leq 1$. And for each $v \in V$ and $S \subseteq \mathbf{C}$, $H^{(1)}(v, S) \leq 1$, thus we also have $|H^{(1)}(v, S) - H^{(0)}(v, S)| \leq 1$.

Input: $G = (V, E)$, parameters α_i, β_S, and convergence threshold ϵ
Output: Importance I and structural hole score H

Initialize $I(v, C_i)$ according to Eq. 4 ;
repeat
  foreach $v \in V$ **do**
    foreach $C_i \in \mathbf{C}$ **do**
      $P(v, C_i) =$
      $\max_{S \subseteq \mathbf{C} \wedge C_i \in S} \{\alpha_i I(v, C_i) + \beta_S H(v, S)\}$;
    end
  end
  foreach $v \in V$ **do**
    foreach $C_i \in \mathbf{C}$ **do**
      $I'(v, C_i) = \max\{I(v, C_i), \max_{e_{uv} \in E} P(u, C_i)\}$;
    end
    foreach $S \subseteq \mathbf{C}$ **do**
      $H'(v, S) = \min_{C_i \in S} I'(v, C_i)$;
    end
  end
  Check the ϵ-convergence condition by

$$\max_{v \in V, C_i \in \mathbf{C}} |I'(v, C_i) - I(v, C_i)| \leq \epsilon$$

  Update $I = I'$ and $H = H'$;
until *Convergence*;

Algorithm 1: HIS-algorithm.

Suppose after k iterations, for each $v \in V$ and $C_i \in \mathbf{C}$, we have $|I^{(k+1)}(v, C_i) - I^{(k)}(v, C_i)| \leq \gamma^k$, and for each $v \in V$ and $S \subseteq \mathbf{C}$, $|H^{(k+1)}(v, S) - H^{(k)}(v, S)| \leq \gamma^k$. Hence, in the $(k+1)$-th iteration, for each $v \in V$ and $C_i \in \mathbf{C}$, if $I^{(k+2)}(v, C_i) = I^{(k+1)}(v, C_i)$, then $|I^{(k+2)}(v, C_i) - I^{(k+1)}(v, C_i)| = 0$, otherwise, there exists u, s.t. $e_{uv} \in E$ and $S \subseteq \mathbf{C}$, such that $C_i \in S$ and

$$\begin{aligned} I^{(k+2)}(v, C_i) &= \alpha_i I^{(k+1)}(u, C_i) + \beta_S H^{(k+1)}(u, S) \\ &\leq \alpha_i (I^{(k)}(u, C_i) + \gamma^k) + \beta_S (H^{(k)}(u, S) + \gamma^k) \\ &\leq \alpha_i I^{(k)}(u, C_i) + \beta_S H^{(k)}(u, S) + \gamma^{k+1} \\ &\leq I^{(k+1)}(v, C_i) + \gamma^{k+1} \end{aligned}$$

Thus, we have $|I^{(k+2)}(v, C_i) - I^{(k+1)}(v, C_i)| \leq \gamma^{k+1}$. For each $v \in V$ and $S \subseteq \mathbf{C}$, there exists $C_i \in S$, such that

$$\begin{aligned} H^{(k+1)}(v, S) &= I^{(k+1)}(v, C_i) \\ &\geq I^{(k+2)}(v, C_i) - \gamma^{k+1} \\ &\geq H^{(k+2)}(v, S) - \gamma^{k+1} \end{aligned}$$

Hence $|H^{(k+2)}(v, S) - H^{(k+1)}(v, S)| \leq \gamma^{k+1}$. □

Therefore, when $\gamma = \max_{C_i \in \mathbf{C},\ C_i \in S} \{\alpha_i + \beta_S\} < 1 - \delta$, where δ is a small constant, the algorithm is guaranteed to be convergent after a finite number of iterations.

We now discuss the efficiency of the algorithm. The time complexity of the algorithm is $O(2^l |E| / \log \gamma)$, which is insufficient for large networks. We here introduce an improved algorithm. Notice that in the $(k+1)$-th iteration of Algorithm 1, we only need to recompute the values of $I(v, *)$, when one of v's neighbors changes its value in the k-th iteration. We can record the change status of each node in the k-th iteration, and broadcast its change to all neighbors in the next iteration. In this way, we can update Eqs. 2-3 in linear-time on the degree of v and 2^l. In each iteration, we select the node v with the largest updated $I(v, *)$, then broadcast the value to its neighbors. The selection of the node v with the largest updated $I(v, *)$ can be done in time $O(\log |E|) = O(\log |V|)$, by using a priority queue. In § 5, we will give the efficiency performance of the improved algorithm. After running the algorithm, we select k nodes with the highest $\max_{|S| \geq 2} \{\beta_S H(v, S)\}$ as the top-k structural hole spanners.

3.2 Model Two: MaxD

The second model is based on the idea that users who span structural holes play an important role in information diffusion between different communities (confirmed in the observable analysis in § 2). Following this, we formalize the problem of structural hole spanner detection by *minimal cut*.

Definition 3. **Minimal Cut.** Given a network $G = (V, E)$ and l communities $\mathbf{C} = \{C_1, \cdots, C_l\}$, we call $D \subseteq E$ as the minimal cut (denoted as $\mathrm{MC}(G, \mathbf{C})$) of communities $\mathbf{C}$ in G, if D is the minimal number of edges to separate those nodes in each community C_i from the others $\{C_j | j \neq i\}$.

Given this, the structural hole spanner detection problem can be cast as finding top-k nodes such that after removing these nodes, the minimal cut of $\mathbf{C}$ in G will be significantly reduced, i.e., the decrease of the minimal cut after removing will be maximized. The idea is natural, as structural hole spanners play bridging roles between communities. Without these structural hole nodes, the connections between different communities would be minimized. To make the idea precise, we propose the following problem definition:

Definition 4. **Detecting top-k structural hole spanners by minimal cut.** Given a graph $G = (V, E)$, and l communities $\mathbf{C} = \{C_1, \cdots, C_l\}$, the task of detecting top-k structural hole spanners by minimal cut is to find $|V_{SH}| = k$ nodes such that after removing these k nodes, the *decrease* of minimal cut of $\mathbf{C}$ in G should be maximized, i.e.,

$$Q(V_{SH}, \mathbf{C}) = \mathrm{MC}(G, \mathbf{C}) - \mathrm{MC}(G \backslash V_{SH}, \mathbf{C}), \ \ |V_{SH}| = k \quad (7)$$

The following theorem shows the hardness of this definition.

Theorem 3. The problem of detecting top-k structural hole spanners by minimal cut is NP-hard, even in the case of $l = 2$.

PROOF. In the case of $l = 2$, the problem can be reduced from the k-DENSEST SUBGRAPH problem, which tries to find a k-node subgraph with the maximum number of edges from a given graph. For the decision version of the problem, it asks whether there exists a k-node subgraph containing at least d edges.

Given an instance $\phi = \{G^* = (V, E), k, d\}$ of the k-DENSEST SUBGRAPH problem, we denote $n = |V|$ and $m = |E|$. We build a graph G consisting of $(n + (n^2 + 1)m + 2)$ nodes, denoted as $\{s, t, x_1, \cdots, x_n, y_{i,1}, \cdots, y_{i,m}\}$, $1 \leq i \leq n^2 + 1$, where $\mathbf{C} = \{C_1, C_2\}$, $C_1 = \{s\}$ and $C_2 = \{t\}$. In the following, we use graphs with polynomially bounded weight (weighted graph for short), and it is straightforward to construct an equivalent un-weighted graph of polynomial size. Graph G contains $(n + (n^2 + 1)(n - 1)m)$ edges. For each $1 \leq j \leq n$, we add one edge between s and x_j with capacity $(n^2 + 1)m$. For each $1 \leq j \leq n$, $1 \leq i \leq n^2 + 1$ and $1 \leq k \leq m$, if node x_j does not appear on the k-th edge (each edge could be regarded as a set of two nodes), we add one edge between x_j and $y_{i,k}$ with capacity

1. For each $1 \leq i \leq n^2+1$ and $1 \leq k \leq m$, we add one edge between $y_{i,k}$ and t with capacity 1.

According to the max-flow and min-cut theory, it is easy to see that, $\mathrm{MC}(G, \mathbf{C}) = (n^2+1)m$. Now we want to prove that, the k-DENSEST SUBGRAPH instance ϕ is satisfiable, if and only if, there exists a subset $|V_{SH}| = n-k$, such that $\mathrm{MC}(G \setminus V_{SH}, \mathbf{C}) \leq (n^2+1)(m-d)$.

For the *only if* direction, suppose the k-DENSEST SUBGRAPH instance ϕ is satisfiable, then we have the subgraph consists of nodes $\{x_{s_j}\}$ and at least d edges. Thus, we can choose $V_{SH} = \{x_j\} \setminus \{x_{s_j}\}$. For the k-th edge e_k in graph G, if e_k exists in the subgraph, for all $1 \leq i \leq n^2+1$, node $y_{i,k}$ cannot be reached. Hence, we have $\mathrm{MC}(G \setminus V_{SH}, \mathbf{C}) \leq (n^2+1)(m-d)$.

For the *if* direction, if there exists a subset $|V_{SH}| = n-k$ such that $\mathrm{MC}(G \setminus V_{SH}, \mathbf{C}) \leq (n^2+1)(m-d)$. Denote $V^*_{SH} = V_{SH} \bigcap \{x_j\}$, we have $|V^*_{SH}| \leq n-k$, and $\mathrm{MC}(G \setminus V^*_{SH}, \mathbf{C}) \leq (n^2+1)(m-d)$. Thus, let the subgraph be the set of corresponding nodes in $\{x_j\} \setminus V^*_{SH}$, there are at least d edges in the graph whose both endpoints are contained in the subgraph. Therefore, the k-DENSEST SUBGRAPH instance ϕ is satisfiable.

Based on the above, we establish the theorem. □

The k-DENSEST SUBGRAPH is hard to approximate, the best known approximation algorithm is $O(n^{1/4+\epsilon})$ [3]. The results in literature [2] indicated the hardness of approximating k-DENSEST SUBGRAPH within $n^{\Omega(1)}$ factors.

Theorem 4. Suppose the k-DENSEST SUBGRAPH is hard to approximate within $n^{\Omega(1)}$, then the problem of detecting top-k structural hole spanners by minimal cut is hard to approximate within $n^{\Omega(1)}$ as well.

PROOF. Suppose there is an approximation algorithm $\mathcal{A}$ for the problem of detecting top-k structural holes by minimal cut with an approximation ratio of $O(f(|G|))$. Given an instance $\phi = \{G^* = (V, E), k, d\}$ of the k-DENSEST SUBGRAPH problem, again we denote $n = |V|$ and $m = |E|$. We continue using the construction in the proof of Theorem 3. Suppose the optimal solution $\{x^*_{s_j}\}$ of ϕ contains d^* edges, then there exists a subset $V^{'}_{SH} = \{x_j\} \setminus \{v^*_{x_j}\}$ such that $|V^{'}_{SH}| = n-k$ and $\mathrm{MC}(G \setminus V^{'}_{SH}, \mathbf{C}) \leq (n^2+1)(m-d^*)$. We call algorithm $\mathcal{A}$ to compute a subset $|V_{SH}| = n-k$ such that

$$\begin{aligned} \mathrm{MC}(G \setminus V_{SH}, \mathbf{C}) &\leq m(n^2+1) - d^*(n^2+1)/O(f(n^{C_0})) \\ &= (n^2+1)(m - d^*/O(f(n^{C_0}))) \end{aligned} \tag{8}$$

where C_0 is a constant. Then denote $V^*_{SH} = V_{SH} \bigcap \{x_j\}$, we have $|V^*_{SH}| \leq n-k$, and $\mathrm{MC}(G \setminus V^*_{SH}, \mathbf{C}) \leq (n^2+1)\lceil m - d^*/O(f(n^{C_0}))\rceil$. Thus, let the subgraph be the set of corresponding nodes in $\{x_j\} \setminus V^*_{SH}$, which contains at least $\lfloor d/O(f(n^{C_0}))\rfloor$ edges. Therefore, the problem of detecting top-k structural holes by minimal cut is also hard to approximate within $n^{\Omega(1)}$. □

Approximate algorithms. Now, we present a polynomial-time algorithm to approximate the problem of structural hole spanners detection by minimal cut.

For any pair of communities, we select $k/\binom{l}{2}$ nodes between them as structural hole spanners using a greedy strategy (referred as MaxD-AL1). In each round, we choose the node which will result in a maximal decrease of the minimal cut when removed it from the network.

Theorem 5. The greedy algorithm can achieve an approximation ratio of $n^{O(1)}$.

Input: $G = (V, E), k, l, \mathbf{C} = \{C_i\})$
Output: Top-k structural hole nodes V_{SH}

```
Initialize V_SH = ∅ ;
while |V_SH| < k do
    Initialize f(v) = 0, for each v ∈ V ;
    foreach non empty S ⊂ {1, ··· , l} do
        E_S = ∪_{i∈S} C_i and E_T = ∪_{i∉S} C_i ;
        Compute the maximal flow with source E_S and sink E_T on
        the induced graph G \ V_SH ;
        foreach v ∈ V do
            Add f(v) by the flow though node v ;
        end
    end
    Choose O(k) nodes with the largest f as candidates D;
    Compute p* = arg max_{p∈D} MC(G \ (V_SH ⋃ {p}), C);
    Update V_SH = V_SH ⋃ {p*}
end
return I and H ;
```

Algorithm 2: MaxD-AL2 Algorithm.

PROOF. Based on the fact that the minimal cut of the graph is bounded by $n^{O(1)}$, the theorem is proved. □

Suppose the time to compute the minimal cut of all communities $\mathbf{C} = \{C_1, \cdots, C_l\}$ in G is $O(T_l(n))$. Thus, the time-complexity of the greedy algorithm is $O(nkT_l(n))$. To scale up the algorithm to large networks, we consider two strategies to improve the efficiency of the algorithm. One idea is to restrict the number of candidates in the greedy algorithm. The first algorithm (MaxD-AL1) only considers $O(k)$ high-degree nodes as candidates, which improves the time-complexity to $O(k^2T_l(n))$. In the second algorithm (called MaxD-AL2), for each partition E_S and E_T, we call the network-flow algorithm [8, 11] to compute the minimal cut of E_S and E_T. We consider top $O(k)$ nodes with maximal sum of flows through them as candidates. Details can be found in Algorithm 2.

In Algorithm 2, one challenge is to estimate $\mathrm{MC}(G, \mathbf{C})$. As introduced in [10], by a reduction from the 3-DIMENSIONAL MATCHING problem, it is NP-hard to compute the minimal cut between multiple-sets(when $l > 2$). We develop the following algorithm to estimate the minimal cut of communities $\mathbf{C} = \{C_i\}$ in G. The approximation ratio of the algorithm is $O(\log l)$. The idea of the approximation algorithm is as follows. To find the minimal cut of all communities $\mathbf{C} = \{C_i\}$, we try all possible partitions of $\mathbf{C} = \mathbf{C_1} \cup \mathbf{C}_2$ and find the minimal cut (denoted as D) between $\bigcup_{C_i \in \mathbf{C_1}} C_i$ and $\bigcup_{C_i \in \mathbf{C}_2} C_i$. Then we remove D from the graph G and call sub-tasks on $\mathbf{C}_1$ and $\mathbf{C}_2$ recursively. The time-complexity of the algorithm for computing $\mathrm{MC}(G, \mathbf{C})$ is $O(2^{2l}T_2(n))$.

Theorem 6. The above algorithm for computing $\mathrm{MC}(G, \mathbf{C})$ provides an $O(\log l)$ approximation.

PROOF. The approximation ratio is bounded by the depth of the partition process. There is always a partition whose depth is at most $O(\log l)$. Thus, the approximation ratio is $O(\log l)$. □

4. MODEL APPLICATIONS

Now, we turn to discuss how structural holes can help real social applications. Specifically, we consider detecting community kernels [32] and inferring social ties [29]. The former aims to detect the community structure among influential (kernel) users and the latter is to predict the types of social relationships (can be generally considered as a link prediction task).

4.1 Community Kernel Detection

The community kernel detection problem is defined as: [32] Given a graph $G = (V, E)$, a weight vector $\overrightarrow{w}(v) = \{w_1(v), \cdots, w_l(v)\}$ is defined for each node v, with each $w_i(v)$ representing the relative importance of the node wrt the i-th community. Denote s as the size of community kernels. One goal of community kernel detection is to obtain the importance of each node wrt a community. Then those nodes with the highest importance scores are selected as the kernel members. The algorithm proposed in [32] is called WEBA.

Now, we study how to leverage structural hole to help community kernel detection. Our idea is based on the intuition that structural hole spanners may be connected with kernel members in different communities. Following this, we incorporate the output of structural hole analysis into the objective function of WEBA. For HIS, we define

$$p(v) = \max_{S \subseteq \{1, \cdots, l\}} \{\beta_S H(v, S)\} \tag{9}$$

For MaxD, we first calculate the top-k structural hole spanners V_{SH}. Then we define $p(v) = 1$ if $v \in V_{SH}$, and $p(v) = 0$ otherwise. Given this, we extend the objective function of WEBA, and define the following optimization problem:

$$\begin{aligned} \text{max} \quad & \mathcal{L}(\vec{w}) = \sum_{(u,v)\in E} \vec{w}(u) \cdot \vec{w}(v) + s \sum_{v \in V} \sum_{1 \leq i \leq l} p(v) w_i(v) \\ \text{subject to} \quad & \sum_{v \in V} w_i(v) = s, \ \forall i \in \{1, \cdots, \ell\}; \\ & \sum_{1 \leqslant i \leqslant \ell} w_i(v) \leqslant 1, \ \forall v \in V; \\ & w_i(v) \geqslant 0, \ \forall v \in V, \ \forall i \in \{1, \cdots, \ell\}. \end{aligned} \tag{10}$$

We use a similar algorithm as that in [32] to solve the optimization problem in Eq. 10. We still use coauthor data set in [32] to evaluate the performance of community kernel detection in terms of precision, recall, and F1-score.

4.2 Link Prediction

We also apply the results of structural hole analysis to help predict the types of social relationships, an important link prediction task. Specifically, we consider the following data sets used in [29].

Slashdot is a network of friends. Slashdot is a site for sharing technology related news. In Slashdot, users can tag each other as "friends" (like) or "foes" (dislike). The data set is comprised of 77,357 users and 516,575 edges.Our goal is to predict the "friend" relationships between users.

Mobile is a network of mobile users. The data set is from [9]. It consists of the logs of calls, blue-tooth scanning data and cell tower IDs of 107 users during about ten months. If two users communicated (by making a call and sending a text message) with each other or co-occurred in the same place, we create an edge between them. In total, the data contains 5,436 edges. Our goal is to predict whether two users have a friend relationship. For evaluation, all users are required to complete an online survey, in which 157 pairs of users are labeled as friends.

For predicting the types of social relationships on the above data sets, [29] presents a number of algorithms, among which the graphical model PFG (Partially Labeled Factor Graph) achieves the best performance in one single network. We first perform the proposed models to mine structural hole spanners on the two data sets. As for the communities, we use the Newman's algorithm [26]. Then we use the identified structural hole spanners to define correlation features in the PFG algorithm. Specifically, given a structural hole spanner, for any two users who have relationships with the spanner, we create a binary correlation feature. For example, if both users v_i and v_j have a friend relationship with a spanner v_k, then a correlation feature $h(y_{ik} = 1, y_{jk} = 1)$ is defined. For more details about the feature definition, please refer to [29].

5. EXPERIMENTS

In this section, we evaluate the effectiveness and efficiency of our algorithms proposed in Section 3. All data sets and codes used in this work are publicly available.[4]

5.1 Experimental Setup

To quantitatively evaluate the proposed models, we consider the following performance metrics:

- **Accuracy.** In the Coauthor network, we evaluate the proposed models in terms of Precision, Recall, and F1-Measure. In both Twitter and Coauthor networks, we use the maximization of information diffusion to evaluate the proposed models.
- **Case study.** We use several case studies as the anecdotal evidence to further demonstrate the effectiveness of the proposed models.
- **Application improvement.** We apply the detected structural hole nodes to help community kernel detection and link prediction. This will demonstrate how the quantitative measurement of structural holes can benefit other social networking applications.

Comparison Methods We compare the following methods for detecting top-k structural hole spanners.

Pathcount [12]: for each node, the algorithm counts the average number of shortest paths (between each pair of nodes) it lies on, and then selects those nodes with the highest numbers as structural hole spanners.

2-Step Connectivity [29]: for each node, it counts the number of pairs of neighbors who are not directly connected. And then those nodes with the highest numbers are viewed as structural hole spanners.

PageRank: it uses PageRank [27] to estimate the importance of each node and then selects those nodes with the highest PageRank scores as structural hole spanners.

PageRank+: it selects those nodes who have the highest Pagerank scores and appear in more than one communities as structural hole spanners.

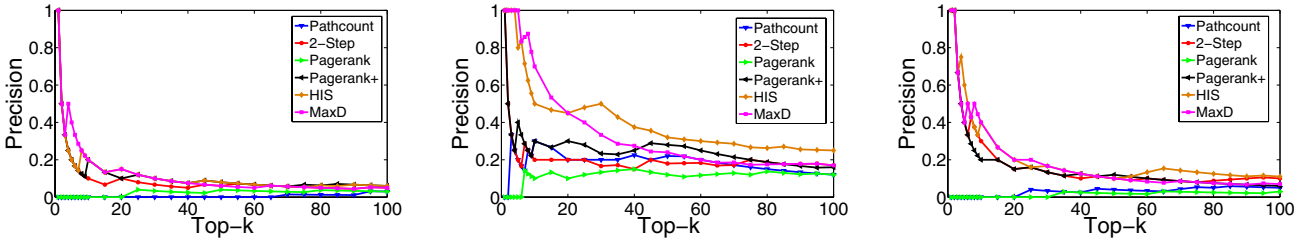
HIS: the first proposed model. We empirically set $\alpha_i = 0.3$ and $\beta_S = 0.5 - 0.5^{|S|}$.

MaxD: the second proposed model. By default, we use the MaxD-AL2 algorithm to approximate the model.

In Coauthor, we consider each subject area as a community; in Twitter, we use the $(\alpha-\beta)$ algorithm [24] to find overlapping communities; and in Inventor, we take each company as a community.

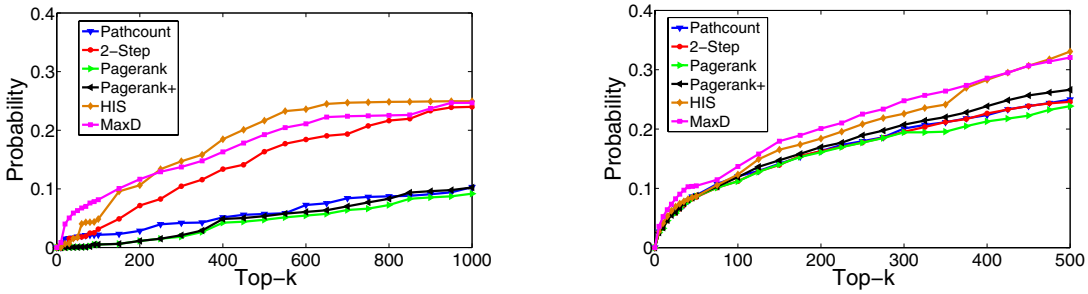
All codes are implemented in C++, and all experiments are performed on a PC running Windows 7 with Intel (R) Core (TM) 2 CPU 6600 (2.4GHz) and 4GB memory. Table 2 lists the running time of the comparison algorithms. In general, HIS has a very good efficiency performance and can perform the detection on large network (Inventor) with millions of nodes in 26 seconds. MaxD results in a bit lower efficiency, but is comparable with Pathcount.

[4]http://arnetminer.org/structural-hole/



(a) AI-DM (b) DB-DM (c) DP-NC

Figure 5: Accuracy performance of different algorithms for detecting top-k structural hole nodes on Coauthor.

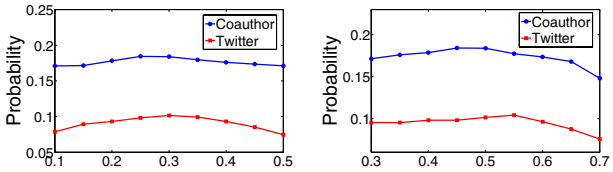


(a) Twitter (b) Coauthor

Figure 6: Results of maximization of information diffusion by different algorithms. (a) probability of the detected structural hole nodes appearing on a tweet-forwarding path across different communities; (b) probability of the detected structural hole nodes receiving cross-domain citations.

Table 2: Running time of different algorithms.

Data Set	Pathcount	2-Step	PageRank	HIS	MaxD
Coauthor	350.66s	4.71s	0.20s	0.60s	189.78m
Twitter	32.03m	12.09s	0.67s	3.87s	602.37m
Inventor	494.3 hr	98.96s	3.61s	26.11s	370.8hr



(a) α (b) β

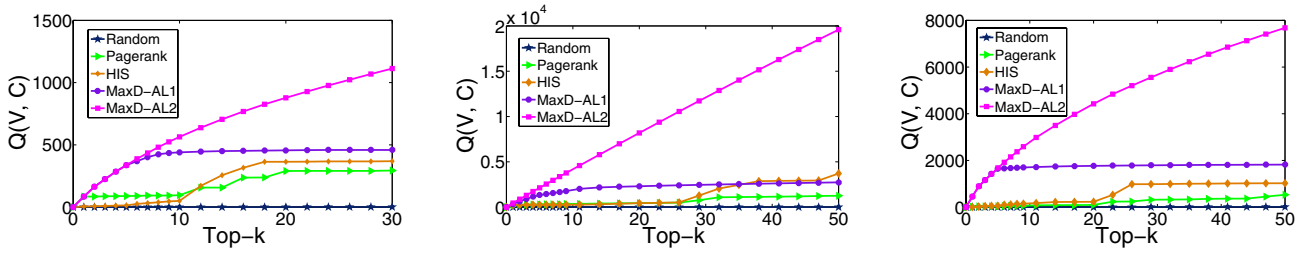
Figure 8: Performance of information spread by HIS with the two parameters α and β varied ($k = 200$).

5.2 Performance Analysis

Accuracy. We first use Coauthor as the benchmark data set to evaluate the proposed models. Figure 5 shows the performance of different algorithms. Both of the proposed models clearly outperform the comparison algorithms by +20-40% at top 20. As expected, choosing important nodes (by PageRank) only is not a good strategy. 2-Step Connectivity and Pathcount achieve a better performance than PageRank. This is because that the objective of PageRank, to find authority nodes, is different from that of finding structural hole spanners. In our first model, HIS, structural hole nodes are determined not only by the bridging positions, but also by the status (e.g., opinion leaders or not) of people connected by the bridging positions. We also note that the two proposed models present different behaviors. Roughly, MaxD performs a bit better at top 20, while HIS outperforms when the number increases to 40-100.

Maximization of information diffusion. Employing Twitter and Coauthor as the basis, we study how the detected structural hole spanners govern the diffusion of information. Specifically, we apply the different algorithms to the Twitter (or the Coauthor) network to detect top-k structural hole spanners. Then we use the tweet-forwarding (or the citation) information to verify the detected results. Figure 6 shows the performance of different algorithms on the two networks. We can see in Twitter the proposed models significantly outperform the comparison algorithms. In Coauthor, the improvements of our models over the comparison algorithms is still clear. We produce sign tests for each result, which confirms that all the improvements of our proposed models over the four methods are statistically significant ($p \ll 0.01$). We can also see that top 100 (0.2%) structural hole users (detected by MaxD) in Twitter influence almost 10% of the forwarding behaviors between different communities. Notice the striking patterns between the two models on the Twitter network. Although MaxD directly models the information diffusion process, HIS clearly outperforms MaxD when the number of k increases up to 200, and by over 20% when $k = 500$. This suggests that there is big a difference between the information network structure and the social network structure. How to combine the two network structures for mining structural hole spanners would be an interesting future work.

Model analysis. We now analyze several properties of the two models. For HIS, we compare the two algorithms described in §3.1



(a) Coauthor (b) Twitter (c) Inventor

Figure 7: Model analysis for MaxD on the three networks.

on the three data-sets. The two algorithms produce the same result, but the improved algorithm achieves an 25× speedup, comparing with the basic algorithm (as shown in Table 2). We further examine how the tunable parameters α and β influence the performance of information spread on the Coauthor and Twitter networks. Figure 8 shows the performance of information diffusion of HIS with the two parameters varied (by fixing $k = 200$). The performance is insensitive to the different parameter settings.

For the MaxD model, we compare different algorithms by the minimal cut described in § 3.2. Figure 7 shows the performance of the different algorithms in Model 2. The MaxD-AL2 algorithm outperforms the other algorithms in terms of $Q(V_{SH}, \mathbf{C})$. This is consistent with the theoretical analysis in §3.2. MaxD-AL1 is close to MaxD-AL2 for a small k ($k < 10$), but the difference quickly becomes wider when increasing the value of k.

5.3 Case study

Now we present a case study on the Inventor network to qualitatively demonstrate the effectiveness of the proposed models. Table 3 lists top-5 structural hole spanners detected by our proposed algorithms from the Inventor network. We find that most of the detected structural hole spanners have been working in more than one job. The exception is T. Kondo and S. Yamazaki. The former is the senior vice president of Sony and holds patents on semiconductor, image processing, and mobile devices. On each topic, he collaborated with people from different companies/universities. S. Yamazaki is the president of SEL (Semiconductor Energy Laboratory). He is a Japanese inventor in the field of computer science and solid-state physics. He holds over 2,680 U.S. utility patents. Part of his patents are in relation to SEL and many others are named individually. Another phenomenon worth mentioning is that HIS seems to select people with the highest PageRank scores, while MaxD tends to select people who have been working on more jobs. This result is consistent with the intuitions behind the two models.

Table 3: Top-5 structural hole nodes discovered by our algorithms on the Inventor network. Names with $*$ are inventors with the highest (top-5) PageRank scores.

Inventor	HIS	MaxD	Title
E. Boyden		√	Professor (MIT Media Lab)
			Associate Professor (MIT McGovern Inst.)
			Group Leader (Synthetic Neurobiology)
A. Czarnik		√	Founder and Manager (Protia, LLC)
			Visiting Professor (University of Nevada)
			Co-Founder (Chief Scientific Officer)
T. Kondo*	√	√	Senior vice president (Sony Corporation)
A. Nishio		√	Director of Operations (WBI)
			Director of Department Responsible (IDA)
E. Nowak*	√		Senior vice President (Walt Disney)
			Secretary of Trustees (The New York Eye)
A. Rofougaran	√		Consultant (various wireless companies)
			Co-founder (Innovent System Corp.)
			Leader (RF-CMOS).
M. Rofougaran	√	√	Engineering Director (Broadcom Corp.)
			Co-founder(Iran Today Publications)
S. Yamazaki*	√		President and majority shareholder (SEL)

6. APPLICATION IMPROVEMENT

We now turn to evaluate the performance improvement when applying the output of mining structural hole spanners to the two social applications: community kernel detection and link prediction.

Community kernel detection. For fair comparison, we still use the benchmark coauthor network used in [32] to evaluate the performance of community kernel detection in terms of precision, recall, and F1-score. The benchmark network is comprised of authors who have published papers on top conferences in five research areas:[5] Artificial Intelligence (AI), Databases (DB), Distributed and Parallel Computing (DP), Graphics, Vision and HCI (GV), and Networks, Communications and Performance (NC). For example, for DB, the conferences include VLDB, SIGMOD, PODS, ICDE, ICDT, and EDBT. The community of program committee (PC) members of those conferences in each area is viewed as the ground truth for quantitatively evaluating the performance of community kernel detection. We empirically set the value of k as 100.

By solving Eq. 10 with the similar algorithms as that in [32], we compare the community detection performance of WEBA with and without the help of structural hole. The performance is shown in Figure 9. Clear improvements on the coauthor data set can be obtained. In terms of F1-score, the average improvement is about 4.5%. HIS performs a bit better than MaxD.

Link prediction. We also apply the discovered structural holes spanners to help link prediction in the two networks: Slashdot and Mobile. Specifically, we first use the Newman's algorithm [26] to discover communities in each network and then use the proposed models to mine top-k structural holes spanners. Then we define correlation features based on the discovered structural hole spanners and add those defined features into the prediction algorithm PFG [29]. We use half of the data for training the PFG algorithm and the rest for testing its prediction performance. Table 4 lists the prediction performance of PFG before and after adding the structural holes-based features. It can be seen that by incorporating the structural holes-based features, the performance of predicting the

[5]The benchmark network is similar to the Coauthor data set introduced in §2.

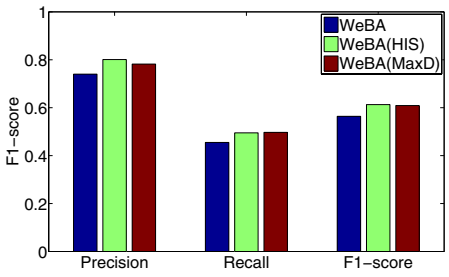


Figure 9: Performance of WEBA for detecting kernel communities with and without the help of structural holes mining.

Table 4: Performance of the PFG algorithm for predicting the type of social relationships before and after combining the structural hole-based features.

Dataset	Algorithm	K	Precision	Recall	F1-score
Mobile	PFG	-	0.9111	0.5694	0.7008
	PFG(HIS)	5	0.8958	0.5972	0.7166
	PFG(HIS)	15	0.8491	0.6250	0.7200
	PFG(HIS)	25	0.8519	**0.6389**	**0.7302**
	PFG(MaxD)	5	**0.9130**	0.5833	0.7118
	PFG(MaxD)	15	0.8776	0.5972	0.7107
	PFG(MaxD)	25	0.8723	0.5972	0.7090
Slashdot	PFG	-	0.6619	0.7281	0.6934
	PFG(HIS)	100	0.6562	0.7965	0.7196
	PFG(HIS)	150	0.6615	**0.8241**	**0.7339**
	PFG(HIS)	200	**0.6788**	0.7886	0.7296
	PFG(MaxD)	100	0.6602	0.7542	0.7041
	PFG(MaxD)	150	0.6667	0.7532	0.7073
	PFG(MaxD)	200	0.6619	0.7775	0.7151

type of social relationships by PFG is clearly improved (+1.4-5.8% by F1-score; t-test, $p < 0.01$). We also evaluate how the performance is affected by the number of k. On the Mobile data, as the network is relatively small, we set k as 5, 15, 25. On the Slashdot data, we set k as 100, 150, 200. The results show that the performance is indeed influenced by different settings of the value for k, but with all the different settings, the performance of link prediction by PFG can be improved. This confirms the effectiveness of the proposed structural hole mining models.

7. RELATED WORK

Structural holes. The concept of structural hole is first introduced in [4] and further elaborated in literature [1, 5, 6]. There have been a few works on mining structural holes from social networks. Goyal and Vega-Redondo [12] propose a model of network formation to study how structural holes are formed in social network. They consider a model in which a node u potentially benefits from serving as an intermediary between nodes v and w even when it resides on an arbitrarily long v-w path. Based on the model, they obtain a star network. However, in real world, many networks are not necessarily of the star topology. Buskens and van de Rijt [7] uses the game theory to model the network formation with structural holes. Kleinberg et al. [18] study the strategic and dynamic aspect to the theory of structural hole. They extend Burt's work [5] by modeling how social networks change over time if everyone is vying for those bridging positions. In this work, we study a novel problem of mining structural hole spanners in social networks, which can be considered as the reverse process of the study of strategic network formation with structural holes [7, 12, 18].

Information diffusion. Our work is also related to a growing body of research on information diffusion. For example, Gruhl et al. [13] study the dynamics of information propagation in environments of low-overhead personal publishing on a web blog data. They apply the theory of infectious diseases to model the "topic" flow on web blogs. Kumar et al. [19] explore the formation of the structure of conversations in social networks and propose a mathematical model to generate the basic structure underlying conversation behaviors. Liben-Nowell and Kleinberg [21] investigate the information spreading processes at a person-to-person level using methods to reconstruct the propagation of massively circulated Internet chain letters. They find the progress of the chain letters proceeds in a narrow but very deep tree-like pattern and propose a probabilistic model based on network clustering and asynchronous response times to produce the tree. Yang et al. [33] analyze how information spread on Twitter via the retweeting behavior and propose a semi-supervised framework to predict users' retweet behaviors. Myers et al. [25] study how the process of information diffusion is influenced by external sources. Matsubara et al. [23] propose a model called SPIKEM to model the rise and fall patterns of influence propagation. However, all these works do not consider how structural holes influence the procedure of information diffusion. To the best of our knowledge, this is the first work to systematically study the problem of mining structural hole spanners in social networks.

8. CONCLUSION

In this paper, we study the novel problem of mining structural hole spanners in large networks. We precisely define the problem of top-k structural hole detection and provide an objective (quality) function to formalize the problem. We develop two instantiation models for the objective function based on the principles of information flow. For both models, we provide theoretical analysis and proofs for their hardness, and develop efficient algorithms to solve with provable approximation guarantees. We validate the effectiveness and efficiency of the proposed models on three different types of networks. We also apply the detected structural hole spanners by the proposed models to help several social networking applications, which further demonstrate its effectiveness.

Structural hole is an important concept in social theory and it relates to a wide range of indicators of social success. As for the future work, it would be intriguing to combine the content information with the user network information and design a unified model for mining structural hole spanners. It is also interesting to further improve the proposed algorithms. For example the MaxD-AL2 algorithm still suffers from the high computational cost (as shown in Table 2) and this HIS model is still lack of a theoretical guarantee. In addition, though the MaxD model uses the information diffusion in the evaluation, but not really uses the process to identify the structural hole spanners. How to elegantly incorporate the information diffusion process into the MaxD model would be a very interesting research topic. Another potential issue is to systematically study how structural holes can help the other social networking applications (e.g., recommendation).

Acknowledgements. We thank Jon Kleinberg for insightful discussions. The work is supported by the Natural Science Foundation of China (No. 61222212, 61073073).

9. REFERENCES

[1] G. Ahuja. Collaboration networks, structural holes, and innovation: A longitudinal study. *Administrative Science Quarterly*, 45(3):425–455, 2000.

[2] A. Bhaskara, M. Charikar, E. Chlamtac, U. Feige, and A. Vijayaraghavan. Detecting high log-densities: an $o(n^{1/4})$ approximation for densest k-subgraph. In *STOC*, pages 201–210, 2010.

[3] A. Bhaskara, M. Charikar, V. Guruswami, A. Vijayaraghavan, and Y. Zhou. Polynomial integrality gaps for strong sdp relaxations of densest k-subgraph. In *SODA*, pages 1395–1408, 2012.

[4] R. S. Burt. *Structural Holes: The Social Structure of Competition*. Harvard University Press, 1992.

[5] R. S. Burt. Structural holes and good ideas. *American Journal of Sociology*, 110:349–399, 2004.

[6] R. S. Burt. Secondhand brokerage: Evidence on the importance of local structure for managers, bankers, and analysts. *Academy of Management Journal*, 50:119–148, 2007.

[7] V. Buskens and A. van de Rijt. Dynamics of networks if everyone strives for structural hole. *American Journal of Sociology*, 114(2):371–407, 2008.

[8] Y. Dinitz. Dinitz' algorithm: The original version and even's version. In *Essays in Memory of Shimon Even*, pages 218–240, 2006.

[9] N. Eagle, A. S. Pentland, and D. Lazer. Inferring social network structure using mobile phone data. *PNAS*, 106(36), 2009.

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

[11] A. V. Goldberg and S. Rao. Flows in undirected unit capacity networks. *SIAM J. Discrete Math.*, 12(1):1–5, 1999.

[12] S. Goyal and F. Vega-Redondo. Structural holes in social networks. *Journal of Economic Theory*, 137(1):460–492, 2007.

[13] D. Gruhl, R. Guha, D. Liben-Nowell, and A. Tomkins. Information diffusion through blogspace. In *WWW'04*, pages 491–501, 2004.

[14] J. He, J. Hopcroft, H. Liang, S. Suwajanakorn, and L. Wang. Detecting the structure of social networks using (α, β)-communities. In *WAW'11*, 2011.

[15] J. Hopcroft, T. Lou, and J. Tang. Who will follow you back? reciprocal relationship prediction. In *CIKM'11*, pages 1137–1146, 2011.

[16] E. Katz. The two-step flow of communication: an up-to-date report of an hypothesis. *In Enis and Cox(eds.), Marketing Classics*, pages 175–193, 1973.

[17] J. Kleinberg. Authoritative sources in a hyperlinked environment. *Journal of the ACM*, 46(5):604–632, 1999.

[18] J. Kleinberg, S. Suri, E. Tardos, and T. Wexler. Strategic network formation with structural holes. In *EC'08*, pages 284–293, 2008.

[19] R. Kumar, M. Mahdian, and M. McGlohon. Dynamics of conversations. In *KDD'10*, pages 553–562, 2010.

[20] P. F. Lazarsfeld, B. Berelson, and H. Gaudet. *The people's choice: How the voter makes up his mind in a presidential campaign*. Columbia University Press, New York, USA, 1944.

[21] D. Liben-Nowell and J. Kleinberg. Tracing information flow on a global scale using internet chain-letter data. *PNAS*, 105(12):4633–4638, Mar. 2008.

[22] T. Lou, J. Tang, J. Hopcroft, Z. Fang, and X. Ding. Learning to predict reciprocity and triadic closure in social networks. *TKDD*, 2013, (accepted).

[23] Y. Matsubara, Y. Sakurai, B. A. Prakash, L. Li, and C. Faloutsos. Rise and fall patterns of information diffusion: model and implications. In *KDD'12*, pages 6–14, 2012.

[24] N. Mishra, R. Schreiber, I. Stanton, and R. E. Tarjan. Finding strongly-knit clusters in social networks. *Internet Mathematics*, 5(1–2), 2009.

[25] S. A. Myers, C. Zhu, and J. Leskovec. Information diffusion and external influence in networks. In *KDD'12*, pages 33–41, 2012.

[26] M. E. J. Newman. Fast algorithm for detecting community structure in networks. *Phys. Rev. E*, 69(066133), 2004.

[27] L. Page, S. Brin, R. Motwani, and T. Winograd. The pagerank citation ranking: Bringing order to the web. Technical Report SIDL-WP-1999-0120, Stanford University, 1999.

[28] J. M. Podolny and J. N. Baron. Resources and relationships: Social networks and mobility in the workplace. *American Sociological Review*, 62(5):673, 1997.

[29] J. Tang, T. Lou, and J. Kleinberg. Inferring social ties across heterogeneous networks. In *WSDM'12*, pages 743–752, 2012.

[30] J. Tang, B. Wang, Y. Yang, P. Hu, Y. Zhao, X. Yan, B. Gao, M. Huang, P. Xu, W. Li, and A. K. Usadi. Patentminer: Topic-driven patent analysis and mining. In *KDD'2012*, pages 1366–1375, 2012.

[31] J. Tang, J. Zhang, L. Yao, J. Li, L. Zhang, and Z. Su. Arnetminer: Extraction and mining of academic social networks. In *KDD'08*, pages 990–998, 2008.

[32] L. Wang, T. Lou, J. Tang, and J. Hopcroft. Detecting community kernels in large social networks. In *ICDM'11*, pages 784–793, 2011.

[33] Z. Yang, J. Guo, K. Cai, J. Tang, J. Li, L. Zhang, and Z. Su. Understanding retweeting behaviors in social networks. In *CIKM'10*, pages 1633–1636, 2010.