Joint Community and Structural Hole Spanner Detection via Harmonic Modularity

Lifang He*
Shenzhen University
Shenzhen, China
lifanghescut@gmail.com

Chun-Ta Lu
University of Illinois at Chicago
Chicago, IL, USA
clu29@uic.edu

Jiaqi Ma
Tsinghua University
Beijing, China
mjq12@mails.tsinghua.edu.cn

Jianping Cao
National University of Defense Technology
Changsha, China
caojianping@nudt.edu.cn

Linlin Shen†
Shenzhen University
Shenzhen, China
llshen@szu.edu.cn

Philip S. Yu
University of Illinois at Chicago
Chicago, IL, USA
Tsinghua University
Beijing, China
psyu@uic.edu

ABSTRACT

Detecting communities (or modular structures) and structural hole spanners, the nodes bridging different communities in a network, are two essential tasks in the realm of network analytics. Due to the topological nature of communities and structural hole spanners, these two tasks are naturally tangled with each other, while there has been little synergy between them. In this paper, we propose a novel harmonic modularity method to tackle both tasks simultaneously. Specifically, we apply a harmonic function to measure the smoothness of community structure and to obtain the community indicator. We then investigate the sparsity level of the interactions between communities, with particular emphasis on the nodes connecting to multiple communities, to discriminate the indicator of SH spanners and assist the community guidance. Extensive experiments on real-world networks demonstrate that our proposed method outperforms several state-of-the-art methods in the community detection task and also in the SH spanner identification task (even the methods that require the supervised community information). Furthermore, by removing the SH spanners spotted by our method, we show that the quality of other community detection methods can be further improved.

Keywords

Community detection; structural hole; harmonic function; modularity; social network

*This work was done while the first author was at the University of Illinois at Chicago.

†Corresponding author

KDD '16, August 13–17, 2016, San Francisco, CA, USA.

DOI: http://dx.doi.org/10.1145/2939672.2939807

1. INTRODUCTION

Detecting communities (or modular structures) has been one of the flourishing issues in understanding the characteristics of real-world networks (*e.g.*, computer networks, biological, semantic and social networks). Exemplar applications include recognizing functions of protein in bioinformatics networks [36], and to forecast the information diffusion process in social networks [19]. It is non-trivial mainly because some bridging nodes, which keep the communication between different communities, blur the boundary of communities. From another point of view, these bridging nodes, known as "hubs" in neurology and "structural hole (SH) spanners" in sociology, have more control over the information that is being transmitted among communities [1, 4, 20, 33]. In neurology, examining the function and role of these hubs is of special interest as they play a central role in establishing and maintaining efficient global brain communication, a crucial feature for healthy brain functioning [1, 33]. In sociology, the theory of structural holes [4] suggests that individuals would acquire more potential resources from filling the "holes" between communities that are otherwise disconnected.

Due to the topological nature of communities and SH spanners, both detection tasks are naturally intermingled with each other. To date, however, studies on these two tasks have been performed independently. Traditional community detection approaches focus on finding clusters such that nodes inside a cluster are tightly connected to each other than to nodes in other clusters. Division [11], agglomeration [22], label propagation [8], and optimization [7] which continuously update the network partition to minimize or maximize a given measure of the quality of the network partition (*e.g.*, spectral clustering and modularity) are typical examples for such approaches. However, SH spanners are inherently connected to multiple communities, effectively linking diverse communities into a weakly-knit network. It's not surprising that well defined communities in real world networks are hard to find without considering the existence of SH spanners.

For mining SH spanners, many approaches [5, 13, 18, 20, 28, 32] have been proposed. For example, the authors in [20]

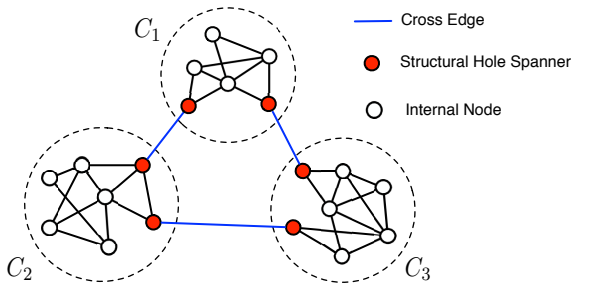


Figure 1: A simple network with three communities and six SH spanners

assume the communities are known in advance and formulate the SH spanners as the nodes such that after removing these nodes, the decrease of minimal cut for communities in network can be maximized. However, for most social networks, it is often difficult to determine the boundaries of communities, as mentioned above. Furthermore, the quality of the identified SH spanners are determined by the communities, which are usually hard to be discovered without removing the SH spanners.

Without given the community information, several importance measures, such as the PageRank [26] and degree centrality [16, 34] have been applied to identify nodes that are important in information diffusion. Some studies [5, 13, 32] utilize the betweenness measure to identify the SH spanners as nodes that have a large number of shortest paths that pass through them. Most recently, [28] proposes to identify the SH spanners by exploiting the bounded inverse closeness centralities of vertices and making use of articulation points of the network. However, the above methods fail to consider one of the most important properties of SH spanners: the information flows between communities are dominated by SH spanners.

Figure 1 illustrates an example of a network with three communities and six SH spanners, where the communities are enclosed by the dashed circles, SH spanners are represented as the red nodes, and the cross edges that connect different communities are marked in blue color. One can notice that these SH spanners are either not having the highest degree centrality or the highest betweenness centrality, but they are the only nodes that can spread information from one community to the other communities. When the community labels (*e.g.*, C_1, C_2 and C_3) are not available, though, it is hard to inspect the information flows between communities.

To unravel the tangled problems regarding community detection and SH spanner detection, we propose to tackle them simultaneously. For the sake of generality, we solely consider the topological structure of the given network. In this paper, we devise a HArmonic Modularity (HAM) scheme to formulate the interaction between communities. Specifically, we apply a harmonic function to measure the harmony between each node and its neighbors, and produce harmonic assignment in each detected community, so that the internal nodes and their intra-community neighbors are as harmonious as possible. By investigating the sparsity level of the interactions between communities, with particular emphasis on the nodes connecting to multiple communities, we are able to discriminate the indicator of SH spanners and assist the community guidance.

Our contributions can be summarized as follows:

- To the best of our knowledge, this is the first attempt to address the problem of joint community and structural hole spanner detection in real-world networks.
- We use the harmonic function to model the topological nature of community and SH spanners, and establish cooperation together. This provides an innovative perspective on the analysis of network structure.
- Through extensive experiments on three real-world network datasets, we present that the proposed HAM method outperforms several state-of-the-art methods in the community detection task and also in the SH spanner detection task (even the methods that require the supervised community information).
- We demonstrate HAM can capture the most intermediate nodes between communities. Furthermore, the SH spanners identified by HAM are more effective in spreading information to different communities than that identified by the alternative methods.
- By removing the SH spanners spotted by our method, we show that the quality of other community detection methods can be further improved.

2. PRELIMINARIES

In this section we establish key definitions and notational conventions that simplify the exposition in later sections. Table 1 lists the important notations used in this paper.

Throughout this paper, matrices are written as boldface capital letters and vectors are denoted as boldface lowercase letters. For a matrix $\mathbf{M} \in \mathbb{R}^{n\times m}$, its elements are denoted by m_{ij}, and its i-th row, j-th column are denoted by $\mathbf{m}^i$, $\mathbf{m}_j$ respectively. The Frobenius norm of $\mathbf{M}$ is defined as $\|\mathbf{M}\|_F = \sqrt{\sum_{i=1}^{n} \|\mathbf{m}^i\|_2^2}$, the $\ell_{2,1}$ norm of $\mathbf{M}$ is defined as $\|\mathbf{M}\|_{2,1} = \sum_{i=1}^{n} \left\|\mathbf{m}^i\right\|_2$. For any vector $\mathbf{u} \in \mathbb{R}^n$, $Diag(\mathbf{u}) \in \mathbb{R}^{n\times n}$ is the diagonal matrix whose diagonal elements are u_i. $\mathbf{I}_n$ denotes an identity matrix with size n. $\|\mathbf{u}\|_0$ is the ℓ_0 norm, which counts the number of nonzero elements in the vector $\mathbf{u}$. We assume that there is a social network represented as an undirected graph $G = (V, E)$, where $V = \{v_1, \cdots, v_n\}$ is the set of nodes and $E \subset V \times V$ is the set of edges whose element $e_{ij} = (v_i, v_j)$ represents an interaction between the nodes v_i and v_j. We denote the adjacency matrix of G by $\mathbf{A} = [a_{ij}]$, where $1 \le i, j \le |V| = n$, $a_{ij} = 1$ if node v_i is connected by an edge to node v_j and 0 otherwise. We assume no self loops, thus $a_{ii} = 0$ for all i. In particular, assume that the nodes of the network can be grouped into m communities $C = \{C_1, \cdots, C_m\}$, with $V = C_1 \cup \cdots \cup C_m$ and $C_i \cap C_j = \varnothing$ for every pair i, j with $i \neq j$.

Next, we establish the definitions and main properties of nodes which will be used to formulate the problem.

Definition 1. (Internal Node) For any node $v_i \in C_p$, if all of its neighboring nodes belong to C_p, node v_i is called an internal node.

Table 1: Important Notations

Symbol	Definition
$V = \{v_i\}_{i=1}^n$	set of nodes
$C = \{C_i\}_{i=1}^m$	set of communities
n	total number of nodes ($\|V\|$)
m	number of communities ($\|C\|$)
k	number of top-ranked SH spanners
$\mathbf{A}, \mathbf{D}$	adjacency and degree matrices
$\mathbf{F}$	indicator matrix
d_i	degree of node v_i
$\|.\|_F$	Frobenius norm
$\|.\|_{2,1}$	$\ell_{2,1}$ norm
$\|.\|_0$	ℓ_0 norm

Definition 2. (Structural Hole Spanner) For any node $v_i \in C_p$, if there exists some neighboring nodes $v_j \in C_q (p \neq q)$, node v_i is called a structural hole spanner.

Definition 3. (Intra-Community Neighbor) For any node $v_i \in C_p$, if node v_i connects with node $v_j \in C_p (i \neq j)$, node v_j is called an intra-community neighbor of node v_i. All the intra-community neighbors of node v_i constitute its intra-community neighbor set.

Definition 4. (Inter-Community Neighbor) For any node $v_i \in C_p$, if node v_i connects with node $v_j \in C_q (p \neq q)$, node v_j is called an inter-community neighbor of node v_i. All the inter-community neighbors of node v_i constitute its inter-community neighbor set.

Definition 5. (Cross Edge) For any edge $e_{ij} = (v_i, v_j) \in E$, if v_i and v_j belong to different communities, edge e_{ij} is called a cross edge.

Definition 6. (Harmonic Function) Given a network $G = (V, E)$, a function $h : V \to \mathbb{R}$ defined on nodes of G is called harmonic if for every $v_i \in V$

$$h(v_i) \equiv \frac{1}{d_i} \sum_{(v_i, v_j) \in E} h(v_j) \tag{1}$$

where $d_i = \sum_j a_{ij}$ denotes the degree of node v_i. Intuitively, at every node $v_i \in V$, the value of a harmonic function is equal to the average of its values at the neighboring nodes.

3. HARMONIC MODULARITY

In this section, we first illustrate the formulation of our harmonic modularity (HAM) scheme. Then a detailed approach is rendered to solve the objective function of HAM. Further, we investigate its convergence and computational complexity.

3.1 Problem Formulation

Key intuitions: In graph theory, a community is described as a group of nodes more densely connected with each other than with the rest of the network. Intuitively, community structure characterizes the neighborhood relationships of the nodes, with nodes that are closer together in the graph having a similar community indicator. As such, the problem of community detection is much more related to the concept of *intra-community neighbor*. On the other hand, as SH spanners play a boundary-spanning role across communities, it is clear that the problem of top-k SH spanner detection is much more related to the concept of *inter-community neighbor*. Most of the existing studies focus on either one or the other of these two assignments. However, by bringing these two assignments together, we can see that although community detection and top-k SH spanner detection assign graph nodes from two different aspects, they both measure neighborhood relationships. Such neighborhood relationships would correspond to the *harmony* and *diversity* of nodes, respectively. From Definition 6, we know the harmonic property provides a systematic way to quantify the harmony and diversity between the indicator value at a given node and the average of its neighboring nodes. Thus, we propose to jointly detect community and top-k SH spanners by measuring the harmonic modularity of the given network. To overcome limitations in prior work, we state the following desiderata:

- **Nonparametric Guidance:** Utilize SH spanner information when inferring community assignment, and vice versa, so that assignment information is able to provide guidance to the detection process in a nonparametric fashion.
- **Harmony:** Produce a harmonic assignment in each community, so that the internal nodes and their intra-community neighbors are as harmonious as possible, even though they connect to the SH spanners.
- **Diversity:** Produce heterogeneous role assignments for internal nodes and SH spanners, so that community and top-k SH spanner assignments are as diverse from each other as possible.

Building upon these desiderata, we proceed to present HAM. We first present how to measure the harmonic modularity, which itself can be used to learn community assignment, of the given network. Since SH spanners involve different communities, we then model and analyze the topology of SH spanners to quantify the influence exerted on communities, which goes with the ability to identify SH spanners and improve community assignment. Let $\mathbf{F} \in \mathbb{R}^{n \times m}$ be the community indicator matrix, where $f_{ij} = 1$ if a node v_i is assigned to the j-th community, and 0 otherwise. The constraints on $\mathbf{F}$ can be written as

$$\mathbf{F} \in \{0, 1\}^{n \times m}, \quad \|\mathbf{f}^i\|_0 = 1, \quad \forall i,\ 1 \leq i \leq n, \tag{2}$$

where $\|\mathbf{f}^i\|_0 = 1$ is utilized to indicate the community that the node v_i most likely belongs to.

For each node v_i, its community indicator $\mathbf{f}^i$ should be as harmonious with its neighbors as possible, *i.e.*, the difference between the value of $\mathbf{f}^i$ and the averaged value of its neighbors $\frac{1}{d_i} \sum_{(v_i, v_j) \in E} \mathbf{f}^j$ should be minimized. Hence, a harmonic function can be embedded to learn the community indicator matrix $\mathbf{F}$. On the basis of the Harmonic analysis, we formulate the following minimization problem

$$\begin{aligned} \min_{\mathbf{F}} \ & \left\|\mathbf{F} - \mathbf{D}^{-1}\mathbf{A}\mathbf{F}\right\|_F^2 \\ \text{s.t.} \ & \left\|\mathbf{f}^i\right\|_0 = 1, \quad \forall\, i, 1 \leq i \leq n \\ & \mathbf{F} \in \{0, 1\}^{n \times m} \end{aligned} \tag{3}$$

One can see if there are no cross edges or SH spanners in the network, the value of the objective function is essen-

tially zero. However, when a node connects with the inter-community neighbors, it will leads to a relatively large value. SH spanner identification is expected to moderate this influence, as the more influential SH spanner is more likely to get involved into interaction between communities. Moreover, to exploit the formulation of (3) on community detection more effectively, it is crucial for the community indicator matrix $\mathbf{F}$ to have discriminative ability for SH spanners, *i.e.*, promoting row-wise sparsity to discriminate relevant SH spanners. We introduce the $\ell_{2,1}$-norm penalty and orthogonality constraint to make it and thus solve the following optimization problem

$$\begin{aligned} &\min_{\mathbf{F}} \|\mathbf{F} - \mathbf{D}^{-1}\mathbf{A}\mathbf{F}\|_{2,1} \\ &\text{s.t.} \quad \mathbf{F}^{\mathrm{T}}\mathbf{F} = \mathbf{I}_m \end{aligned} \tag{4}$$

Note that the relaxation of orthogonality condition has two benefits: it not only avoids solving the NP-hard problem of ℓ_0 norm, but also allows the sparsity of the community indicator matrix $\mathbf{F}$ to be exploited. It can be seen the sparsity-inducing property of $\ell_{2,1}$ norm pushes $\mathbf{F}$ to be sparse in rows. More specifically, $\mathbf{f}^i$ shrinks to zero if the neighbors of node v_i belongs to different communities. In particular, the more inter-neighbors the node v_i connects to the more different communities, the larger $\|\mathbf{f}_i - \mathbf{D}^{-1}\mathbf{A}\mathbf{f}_i\|_2^2$ is, so the value of $\mathbf{f}^i$ gets penalized more harshly. Therefore, we can obtain the top-k SH spanners corresponding to the top-k smallest values of $\|\mathbf{f}^i\|_2$. Further, the shrinkage of the $\mathbf{f}^i$ diminishes the influence of the node v_i on its neighbors, making them to be more harmonious with their intra-community neighbors.

It is not difficult to see that the formulation of (4) characterizes graph nodes from two different aspects: harmony and diversity. The harmonic modularity provides a measure of the smoothness of $\mathbf{F}$ over the edges in G, and thus help produce harmonic community assignment. The $\ell_{2,1}$ norm provides an investigation on the sparsity level of the interactions between communities, and thus help discriminate SH spanners and assist the community guidance.

3.2 Solution

Directly minimizing Eq. (4) involving $\ell_{2,1}$ norm is non-trivial. Here we propose an iterative algorithm based on the half-quadratic minimization [25] to solve this problem. We start by introducing the following lemma [15].

Lemma 1. *Let $\phi(.)$ be a function satisfying the conditions: $x \to \phi(x)$ is convex on R; $x \to \phi(\sqrt{x})$ is convex on R_+; $\phi(x) = \phi(-x), \forall x \in R$; $\phi(x)$ is C^1 on R; $\phi''(0^+) \geq 0$, $\lim_{x\to\infty} \phi(x)/x^2 = 0$. Then for a fixed $\|\mathbf{u}^i\|_2$, there exists a dual potential function $\varphi(.)$, such that*

$$\phi(\|\mathbf{u}^i\|_2) = \inf_{p\in R}\{p\|\mathbf{u}^i\|_2^2 + \varphi(p)\} \tag{5}$$

where p is determined by the minimizer function $\varphi(.)$ with respect to $\phi(.)$.

Let $\mathbf{P} = \mathbf{F} - \mathbf{D}^{-1}\mathbf{A}\mathbf{F}$. According to the analysis for the $\ell_{2,1}$ norm in [15], if we define $\phi(x) = \sqrt{x^2 + \epsilon}$, we can replace $\|\mathbf{P}\|_{2,1}$ with $\sum_{i=1}^n \phi(\|\mathbf{p}^i\|_2)$. Thus, based on Lemma 1, the objective function of Eq. (4) can be reformulated as follows:

$$\begin{aligned} &\min_{\mathbf{F}} \operatorname{Tr}(\mathbf{P}^{\mathrm{T}}\mathbf{Q}\mathbf{P}) \\ &\text{s.t.} \quad \mathbf{F}^{\mathrm{T}}\mathbf{F} = \mathbf{I}_m \end{aligned} \tag{6}$$

Algorithm 1 Harmonic Modularity (HAM)

Input: $G = (V, E)$, m, k
Output: Assignments to m communities and the top-k SH spanners

1: Initialize $\mathbf{F}_0$ s.t. $\mathbf{F}_0^{\mathrm{T}}\mathbf{F}_0 = \mathbf{I}_m$, $t \leftarrow 0$;
2: **while** not converge **do**
3: Set $\mathbf{Q}_t \leftarrow Diag(\frac{1}{2\sqrt{\|\mathbf{p}_t^i\|_2^2+\epsilon}})$;
4: Compute $\mathbf{R}_t$ according to Eq. (8)
5: Compute $\mathbf{F}_{t+1}$ as the eigenvectors of $\mathbf{R}_t$ corresponding to the first m smallest eigenvalues;
6: $t \leftarrow t + 1$;
7: **end while**
8: Sort each node according to $\|\mathbf{f}\|_2$ in **ascending** order and select the top-k ranked ones as SH spanners;
9: Remove the top-k SH spanners from $\mathbf{F}$, and then cluster $\mathbf{F}$ by K-means to obtain clustering communities.

where $\mathbf{Q} = Diag(\mathbf{q})$, and $\mathbf{q}$ is an auxiliary vector of the $\ell_{2,1}$ norm. The elements of $\mathbf{q}$ are computed as follows.

$$q_i = \frac{1}{2\sqrt{\|\mathbf{p}^i\|_2^2 + \epsilon}} \tag{7}$$

where ϵ is a smoothing term that avoids division by zero, which is usually set to be a small constant value (we set $\epsilon = 10^{-4}$ in this paper).

Clearly, the optimal solution of (6) can be computed via solving the eigenvector problem for the matrix:

$$\mathbf{R} = (\mathbf{I}_n - \mathbf{D}^{-1}\mathbf{A})^{\mathrm{T}}\mathbf{Q}(\mathbf{I}_n - \mathbf{D}^{-1}\mathbf{A}) \tag{8}$$

Based on the above analysis, we summarize the detailed optimization algorithm in Algorithm 1.

3.3 Convergence and Complexity

The Algorithm 1 to optimize Eq. (4) is presented from line 2 to line 7. We prove that it converges to the optimal solution $\mathbf{F}$. We begin with the following Lemma [24].

Lemma 2. *For any nonzero vectors $\mathbf{v}_t^i \in \mathbb{R}^c, 1 \leq i \leq r$, where r is an arbitrary number. The following inequality holds:*

$$\sum_i \|\mathbf{v}_{t+1}^i\|_2 - \sum_i \frac{\|\mathbf{v}_{t+1}^i\|_2^2}{2\|\mathbf{v}_t^i\|_2} \leq \sum_i \|\mathbf{v}_t^i\|_2 - \sum_i \frac{\|\mathbf{v}_t^i\|_2^2}{2\|\mathbf{v}_t^i\|_2} \tag{9}$$

Proof. The detailed proof can be found in the work [24]. □

Next, we show that the iterative algorithm shown in Algorithm 1 converges by the following theorem.

Theorem 3. *The iterative approach in Algorithm 1 (line 2 to line 7) monotonically decreases the objective function value of $\min_{\mathbf{F}^{\mathrm{T}}\mathbf{F}=\mathbf{I}_m} \|\mathbf{F} - \mathbf{D}^{-1}\mathbf{A}\mathbf{F}\|_{2,1}$ in each iteration.*

Proof. Let $\mathbf{\Delta} = \mathbf{I}_n - \mathbf{D}^{-1}\mathbf{A}$, then $\mathbf{P} = \mathbf{\Delta}\mathbf{F}$. In line 5 of Algorithm 1, we can see that

$$\mathbf{F}_{t+1} = \arg\min_{\mathbf{F}^{\mathrm{T}}\mathbf{F}=\mathbf{I}_m} \operatorname{Tr}(\mathbf{F}^{\mathrm{T}}\mathbf{\Delta}^{\mathrm{T}}\mathbf{Q}_t\mathbf{\Delta}\mathbf{F}) \tag{10}$$

Therefore, we have

$$\begin{aligned} &\operatorname{Tr}(\mathbf{F}_{t+1}^{\mathrm{T}}\mathbf{\Delta}^{\mathrm{T}}\mathbf{Q}_t\mathbf{\Delta}\mathbf{F}_{t+1}) \leq \operatorname{Tr}(\mathbf{F}_t^{\mathrm{T}}\mathbf{\Delta}^{\mathrm{T}}\mathbf{Q}_t\mathbf{\Delta}\mathbf{F}_t) \\ \Rightarrow &\sum_i \frac{\|\mathbf{p}_{t+1}^i\|_2^2}{2\|\mathbf{p}_t^i\|_2} \leq \sum_i \frac{\|\mathbf{p}_t^i\|_2^2}{2\|\mathbf{p}_t^i\|_2} \end{aligned}$$

Then according to Lemma 2, $\sum_i \|\mathbf{p}_{t+1}^i\|_2 - \sum_i \frac{\|\mathbf{p}_{t+1}^i\|_2^2}{2\|\mathbf{p}_t^i\|_2} \leq \sum_i \|\mathbf{p}_t^i\|_2 - \sum_i \frac{\|\mathbf{p}_t^i\|_2^2}{2\|\mathbf{p}_t^i\|_2}$, we have the following inequality

$$\sum_i \|\mathbf{p}_{t+1}^i\|_2 \leq \sum_i \|\mathbf{p}_t^i\|_2 \tag{11}$$

Based on the definition of $\ell_{2,1}$ norm and $\mathbf{P} = \mathbf{F} - \mathbf{D}^{-1}\mathbf{A}\mathbf{F}$, we can obtain

$$\|\mathbf{F}_{t+1} - \mathbf{D}^{-1}\mathbf{A}\mathbf{F}_{t+1}\|_{2,1} \leq \|\mathbf{F}_t - \mathbf{D}^{-1}\mathbf{A}\mathbf{F}_t\|_{2,1} \tag{12}$$

which indicates that the value of $\min_{\mathbf{F}^{\mathrm{T}}\mathbf{F}=\mathbf{I}_m} \|\mathbf{F} - \mathbf{D}^{-1}\mathbf{A}\mathbf{F}\|_{2,1}$ monotonically decreases using the updating rule in Algorithm 1. □

According to Theorem 3, we can see that the iterative approach in Algorithm 1 converges to local optimal $\mathbf{F}$ corresponding to Eq. (4). The proposed optimization algorithm is efficient. In the experiment, we observe that our algorithm usually converges around only 20 iterations.

The complexity is briefly discussed as follows. The complexity of computing $\mathbf{Q}$ is $\mathcal{O}(n^2)$. To obtain $\mathbf{F}$, we need to conduct eigendecomposition of $\mathbf{R}$, which is $\mathcal{O}(n^3)$ in complexity. It can be reduced to $\mathcal{O}(n^{2.376})$ using the Coppersmith-Winograd algorithm. The complexity of identifying communities by K-means is $\mathcal{O}(m^2nt)$, where t is the iterative index required for K-means to converge. The complexity of top-k SH spanner selection is $\mathcal{O}(n\log(n) + nm)$.

4. INTERPRETATION AND CONNECTION

Here we present the special case of our HAM method and relate it to the existing works. These different viewpoints provide a rich and complementary set of techniques for reasoning about this approach to the joint community and SH spanner detection problem.

Perhaps the most interesting and significant aspect of HAM is that it can be viewed as a direct response to the harmonic property. As we use the $\ell_{2,1}$ norm to push the community indicators of SH spanners shrink to zero, this equals to remove SH spanners from the graph. Thus, in our method, the value of the objective function is close to zero, which is consistent with the notion of *harmony*. Moreover, one can see if there are no cross edges or SH spanners in the network, then the auxiliary matrix $\mathbf{Q}$ is always constant, regardless of any changes of community indicator $\mathbf{F}$. In this case, we will solve the minimum Frobenius norm residual problem (*i.e.*, $\min \|\mathbf{F} - \mathbf{D}^{-1}\mathbf{A}\mathbf{F}\|_F$, s.t. $\mathbf{F}^{\mathrm{T}}\mathbf{F} = \mathbf{I}_m$), and HAM degenerates into the random walk approach (also called harmonic function learning) [31].

Our HAM uses the eigenvectors of a matrix ($\mathbf{R}$ in Eq. (8)) to reveal the community structure in the graph, therefore it can be regarded as belonging to the category of spectral clustering approaches. However, there are two major differences. The matrix whose eigenvectors are used for clustering plays the key role in spectral clustering. In HAM, this matrix is computed based on the harmonic function learning idea in conjunction with the sparsity-inducing $\ell_{2,1}$-norm, while conventional spectral clustering methods are often based on the graph Laplacian matrices. Second, spectral clustering gives a closed-form solution, and HAM needs to be optimized in a half-quadratic way. In practice, the main computational load of HAM and spectral clustering is to compute the eigenvectors, therefore they have the same order of time complexity.

5. EXPERIMENTS

To evaluate the effectiveness of HAM, we conduct extensive experiments on real-world social networks, and compare them with various baseline methods. As there hardly exist comparative methods that can simultaneously detect communities and SH spanners, we compare our approach to community detection and SH spanner detection methods, respectively. Since the quality of detected communities can be well affected by the spotted SH spanners, we first investigate the effectiveness of HAM on SH spanner detection. We then study the performance of HAM on community detection. Finally, we discuss the results and the reasons behind them.

5.1 Dataset Description

We adopt three real-world social network datasets from different contexts.

- *Karate Club* [38] is the network of friendships between members of a karate club that splits into two clubs due to a dispute between the coach and administrator.
- *DBLP* is a co-authorship network where two authors are connected if they publish at least one paper together. Publication venue defines an individual ground-truth community; authors who published to the same journal or conference form a community.
- *YouTube* is a video-based social network, where users form friendship with each other based on their interactions over videos and users can create groups which other users can join. Such user-defined groups are considered as ground-truth communities.

The *DBLP* and *YouTube* data were obtained from [37]. We sample 5 datasets from each and measure the average performance. Since we focus on detecting distinct communities, the communities of interest should not have too much overlaps with other communities. We compute the number of cross edges between each pair of communities, and divide it by the number of nodes in the smaller community as the cross ratio between the community pair. The sampling rule is: randomly pick a community, and then take all the nodes within the community if the cross ratios between the picked community and the existing selected communities is less than a threshold (set as 0.3 in the experiment). We repeat the sampling process until the number of nodes or the number of communities in the sampled network reaches a predefined limit (set as $2,000$ and 20, respectively, in the experiment) or no more community to be picked. Table 2 summarizes the original datasets and also the sampled datasets for the large networks.

5.2 Structural Hole Spanner Detection

5.2.1 Compared Methods

We compare HAM with the following seven state-of-the-art methods, each of which represents a different strategy for detecting the top-k SH spanners.

- HAM is our proposed method, which selects the nodes that have neighbors belonging to more different communities as the SH spanners.

Table 2: Summary of experimental datasets. Sampled datasets are listed by the mean ± standard deviation.

		Characteristics				
Types of data	Datasets	# Nodes	# Edges	# Cross edges	# SH spanners	# Communities
Original data	Karate Club	34	78	11	13	2
	DBLP	1557.6±362.19	4915.6±451.95	127.4±41.60	189±49.44	15±4.24
Sampled data	YouTube	1310±133.67	2853.5±289.69	82.5±15.18	91.3±13.57	15.25±2.06

Table 3: Structural hole spanner detection results (average *SHII*). Column 2 indicates the used Information Diffusion model

		Comparative Methonds							
Datasets	ID model	HAM	Constraint	PageRank	BC	2-Step	MaxD	HIS	AP_BICC
Karate Club	LT	**0.343**	0.295	0.159	0.159	0.159	0.159	0.132	0.295
	IC	**0.002**	**0.002**	0.001	0.001	0.001	0.001	0.001	**0.002**
	SH spanners	[3,20,9]	[1,34,3]	[34,1,33]	[1,34,33]	[34,1,33]	[34,1,33]	[32,9,14]	[1,3,34]
DBLP	LT	**5.384**	0.404	0.357	0.958	0.394	0.272	0.718	0.550
	IC	**3.578**	0.229	0.190	0.821	0.203	0.135	0.304	0.495
YouTube	LT	**3.951**	2.447	1.236	1.226	1.935	1.674	3.198	1.630
	IC	**2.452**	1.254	0.662	0.791	1.014	0.798	2.148	0.799

- Constraint [4] uses constraint to estimate the importance of each node and select the top-k nodes with the lowest constraint scores as the SH spanners.

- PageRank [26] is the traditional node ranking algorithm, which returns the nodes with the top-k page rank scores as the SH spanners.

- Betweenness Centrality (BC) [2] assigns each node a score that is the number of shortest paths (between all pairs of nodes) on which the node lies, then selects the top-k nodes with the highest scores as the SH spanners.

- 2-Step [32] assigns each node a score that is the number of pairs of its neighbors without edges between them, and then selects the top-k highest scores as the SH spanners.

- MaxD [20] is a supervised learning strategy, which focuses on predefined communities and minimal cut theory to select the top-k nodes, such that after removing these nodes the decrease of the minimal cut will be maximized.

- HIS [20] is also a supervised method. It assigns each node v a score that simulates the likelihood of v as a structural hole spanner across the given subset of communities, and then selects the top-k nodes with the highest scores as the SH spanners.

- AP_BICC [28] is a recently proposed SH spanner detection method, which selects the top-k SH spanners based on articulation points (AP) (that is, nodes of a graph that connect two or more otherwise unconnected parts of the graph) and bounded inverse closeness centrality (BICC), such that after removing these nodes the increase of the mean distance of the network will be maximized. We set the bounded parameter to $l = 2$, and the number of nodes used in the BICC to $K = 50$, as they suggested.

5.2.2 Evaluation criteria

Currently there is no standard criteria available for evaluating the performance of the top-k SH spanners. Here we base on simulating the information diffusion process [12] in the given network to evaluate the performance. Since a SH spanner usually dominates the spread of information across communities, when using the more effective SH spanner as a seed (source) node, the faster the information would be diffused to different communities. Hence, the number of the influenced outsiders, which are the ones reside in the communities different from the SH spanner's community, should be larger. However, the number of outsiders is related to the size of the given community. Further, considering a well connected node in the center of a community can also propagate its influence to other communities through its neighbors, simply relying on the number of influenced outsiders may not be able to discriminate SH spanners from center nodes. Thus, to give a evaluation criterion that is suitable for a variety of SH spanners and can distinguish SH spanners from center nodes, we consider the proposition of the number of the influenced outsiders to the total number of all the influenced nodes. Formally, we define the *structural hole influence index* under a certain diffusion model as the evaluation criterion.

- *Structural Hole Influence Index (SHII)*: Let s be the given seed, C_p be the community that the given seed belongs to, and I_v be an indicator of whether a node v is influenced by using a certain information diffusion model. We define *SHII* as follows:

$$SHII(s) = \frac{\sum_{C_i \in C \setminus C_p} \sum_{v \in C_i} I_v}{\sum_{C_i \in C} \sum_{v \in C_i} I_v}$$

  where C is the set of all the communities. Generally, higher *SHII* corresponds to better performance.

In this study, we use two different and widely used information diffusion models [17]: Linear Threshold (LT) model and Independent Cascade (IC) model to find the set of influenced nodes. Since it is unlikely to simulate the information diffusion procedure using only one seed, for the given SH spanner

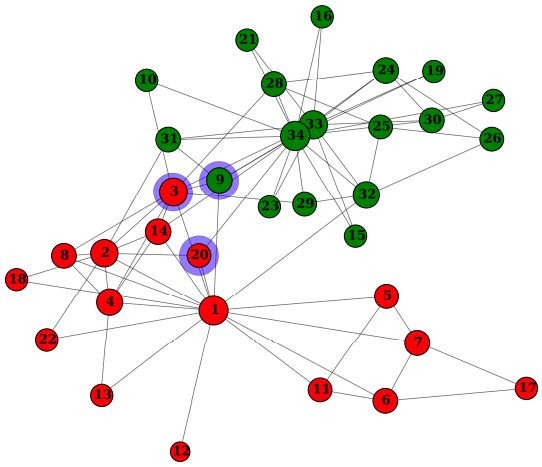


Figure 2: Visualization of the identified SH spanners in Karate Club, where the large blue shadowed nodes represent SH Spanners discovered by HAM.

candidate, we randomly select a certain number (5% of the number of the nodes that are in the same community as the candidate) of the candidate's neighbors along with the candidate as the seed set in the LT/IC model. For each SH spanner candidate, we repeat the random selection for 1,000 times and run the LT and IC models for 100,000 times for each selected seed set. Finally, we report the average *SHII* for each method.

5.2.3 Experiment Result

Table 3 reports the performance of each method in terms of the proposed evaluation criterion. From the results, we can observe that HAM significantly outperforms all the other baseline methods even the supervised methods (*i.e.*, MaxD and HIS), which are aware of the ground truth of the communities. Specifically, it can be seen in the DBLP data that the *SHII* achieved by HAM under the LT model is 5.384, which is more than 5 times higher than that obtained by BC and HIS, about 10 times larger than that achieved by Constraint, PageRank, 2-Step, MaxD, and AP_BICC.

For qualitative analysis, in Figure 2, we visualize the network of Karate Club using the tool *Networkx* [14] and mark the identified SH spanners by HAM to see the topological position of them. The nodes with large blue shadow are the top-3 SH spanners spotted and the colors of nodes indicate the different communities they belong to. The results of the other methods are listed in Table 3. We can find that HAM captures the most intermediate nodes between communities. In contrast, most of the compared methods take the nodes with the highest degree as SH spanners, but actually these nodes are usually the center of the communities. It obviously contradicts to the definition of SH spanners. Although HIS does not focus on finding the nodes with the highest degree, its identified SH spanners have less impact on the information diffusion across different communities, as can be seen that HIS has smaller values of *SHII* than all the other methods.

5.3 Community Detection

We apply the results of identified top-k SH spanners by HAM to help predict community structure. To further show the effectiveness of the identified SH spanners, we evaluate the community detection performance before and after removing the top-k SH spanners for each comparative methods. When considering all the nodes in the network as the prediction object, in our HAM method, for each SH spanner we determine its community label via majority voting using its connected nodes' community labels.

5.3.1 Compared Methods

We compare HAM with the following commonly used algorithms for community detection.

- Spectral Clustering [23] refers to a class of techniques which rely on the eigenstructure of the graph Laplacian to partition points into disjoint clusters with points in the same cluster having high similarity and points in different clusters having low similarity. Typical graph Laplacians are: the unnormalized Laplacian defined as $\mathbf{L} = \mathbf{D} - \mathbf{A}$, the symmetric normalized Laplacian $\mathbf{L}_{sym} = \mathbf{D}^{-\frac{1}{2}}\mathbf{L}\mathbf{D}^{-\frac{1}{2}}$ and the asymmetric normalized Laplacian $\mathbf{L}_{asym} = \mathbf{D}^{-1}\mathbf{L}$. We use these three matrices based on spectral clustering, and denote them as SP, SP_{sym} and SP_{asym}, respectively.
- Random Walk (RW) [31] is a special case of our approach, when the norm of the objective function is Frobenius norm.
- Q-modularity (Q) [22] is a recently introduced community detection method that greedily computes partitions of networks into communities by maximizing the so-called 'modularity'. In particular, the author has shown in [21] that it can use the largest eigenvectors of the 'modularity' matrix to characterize the community structure in the graph.

5.3.2 Evaluation criteria

In order to evaluate the quality of the communities generated by different community detection approaches, we adopt the following measures.

a) Accuracy (ACC) is one of the most frequently used criterion, which evaluates the quality of similarity between resulted cluster labels and ground truth labels. Let c_i represent the clustering label result from a community detection algorithm and y_i represent the corresponding ground truth label of an arbitrary node v_i. Then ACC is defined as follows:

$$ACC = \frac{\sum_{i=1}^{n} \delta(y_i, map(c_i))}{n}$$

where δ is the Kronecker delta function, and $map(c_i)$ is the best mapping function that permutes clustering labels to match the ground truth labels using the KuhnMunkres algorithm. A larger ACC indicates better clustering performance.

b) Normalized Mutual Information (NMI) measures the mutual information entropy between resulted cluster labels and ground truth labels. For any two arbitrary variables P and Q, NMI is defined as follows:

$$NMI = \frac{I(P,Q)}{\sqrt{H(P)H(Q)}}$$

where $I(P,Q)$ computes the mutual information between P and Q, and $H(P)$ and $H(Q)$ are the entropies

Table 4: Community detection results "average score" on three datasets. "↑" indicates the larger the value the better the performance; "↓" indicates the smaller the value the better the performance.

Datasets	Evaluations with top-k SH spanners				Evaluations without top-k SH spanners			
	Methods	ACC ↑	NMI ↑	ACE ↓	Methods	ACC ↑	NMI ↑	ACE ↓
Karate Club	HAM	**1.000**	**1.000**	**0.000**	HAM	1.000	1.000	0.000
	SP	0.912	0.646	0.253	SP	1.000	1.000	0.000
	SP_{sym}	0.971	0.837	0.115	SP_{sym}	1.000	1.000	0.000
	SP_{asym}	0.824	0.363	0.448	SP_{asym}	0.968	0.824	0.125
	RW	0.971	0.837	0.115	RW	1.000	1.000	0.000
	Q	1.000	1.000	0.000	Q	1.000	1.000	0.000
DBLP	HAM	**0.879**	**0.885**	**0.114**	HAM	0.879	0.886	0.112
	SP	0.693	0.790	0.162	SP	0.872	0.876	0.120
	SP_{sym}	0.680	0.756	0.179	SP_{sym}	0.821	0.788	0.255
	SP_{asym}	0.307	0.472	0.687	SP_{asym}	0.331	0.497	0.665
	RW	0.790	0.810	0.154	RW	0.916	0.877	0.136
	Q	0.429	0.634	0.373	Q	0.468	0.663	0.342
YouTube	HAM	**0.953**	**0.938**	**0.100**	HAM	0.954	0.940	0.100
	SP	0.888	0.881	0.176	SP	0.909	0.908	0.160
	SP_{sym}	0.802	0.815	0.212	SP_{sym}	0.826	0.857	0.158
	SP_{asym}	0.433	0.583	0.513	SP_{asym}	0.545	0.630	0.499
	RW	0.908	0.905	0.163	RW	0.933	0.922	0.143
	Q	0.492	0.626	0.395	Q	0.515	0.721	0.294

of P and Q. An advantage of *NMI* is that it does not necessarily increase when the number of clusters increase. The larger the value, the better the performance.

c) Average Cluster Entropy (ACE) is based on the impurity of a cluster given the true classes in the data. Let p_{ij} be the fraction of class j in obtained cluster i, and n_i be the size of cluster i, then ACE is defined as:

$$ACE = \sum_{i=1}^{m} \frac{n_i(-\sum_j p_{ij} \log(p_{ij}))}{n}$$

The smallest the value, the better the performance. Particularly, the low values of ACE indicate homogeneous distribution of the nodes within each group [6].

5.3.3 Experiment Result

Table 4 summarizes the performance of different methods according to three evaluation criteria. The performance before removing the top-k SH spanners (with top-k SH spanners) are shown in the left column and without top-k SH spanners in the right column. From the left column of Table 4 it can be seen that the performance of each method on different datasets can be quite different. However, the best method that outperforms other methods in all datasets is HAM, especially for DBLP and YouTube datasets. We can see the accuracies achieved by HAM (0.879 and 0.953, respectively) are considerably larger than the second best method (0.790 and 0.888, respectively). Moreover, HAM significantly outperforms RW, which means that the sparsity-inducing $\ell_{2,1}$-norm is effective to increase the harmony within the community. From the right column of Table 4, we can see the performance of each comparative method substantially benefits from the removal of top-k SH spanners. For example, the accuracy performance of SP is improved from 0.693 to 0.872 on the DBLP dataset. These results suggest that our proposed method can guarantee to find more positive structural hole spanners connecting to different communities, and removing them makes the community structure more apparent, thus facilitating better community detection performance.

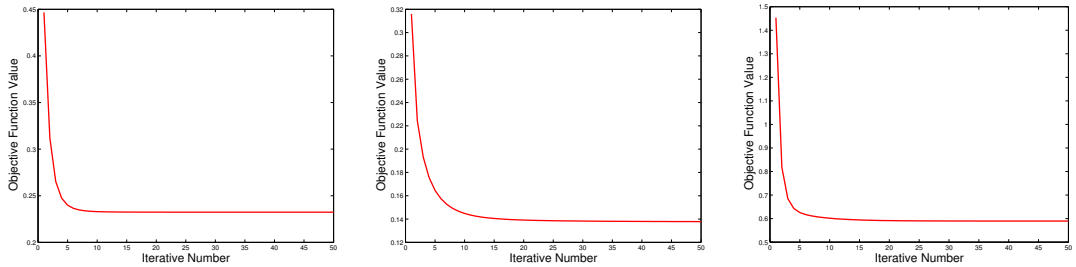
5.4 Discussion

We now turn to the discussion of the experimental results. Table 3 shows HAM is substantially better than other SH spanner detection methods. The reason is that HAM finds the SH spanners who bridge different communities, while most other methods detect the SH spanners with higher degree, which collaborate with many nodes in their own community. In other words, structural hole spanners are more likely to connect the nodes between communities than higher-degree nodes. Thus they have great potentials to control information flow between communities. Moreover, the left column of Table 4 shows the community detection performance gain of HAM over other methods is significant. This result is due to the absence of SH spanners in our original community detection procedure. Removing SH spanners helps make the community structure more tangible such that different communities are easier to be separated. As can be seen in the right column of Table 4, the performance of each comparative method is improved after removing the identified top-k SH spanners detected by HAM.

In addition, Figure 3 shows convergence curves of HAM. From this figure, we can see that the proposed optimization algorithm converges quickly in the vicinity of the minimum, *i.e.*, only around 20 iterations. At the same time, it can also be seen from Figure 3 that the objective function converges to a very small value, which is consistent with intuitive interpretation of SH spanner selection process. Furthermore, our method needs no parameter tuning, which makes it even more appealing.

6. RELATED WORK

To the best of our knowledge, this is the first work to simultaneously address the problems of community detection and SH spanner detection. Our work is related to both com-



(a) Karate Club (b) DBLP (c) YouTube

Figure 3: : Convergence curve of HAM over Karate Club, DBLP and YouTube datasets

munity detection techniques and SH spanners mining. We briefly discuss both of them.

6.1 Community Detection

Community detection, with its root in graph partitioning and graph clustering, has been pivotal to network science. A plethora of algorithms have been presented to address this task over the years, be it cut-based [9], spectrum-based [23], modularity-based [22], or information theoretic [29]. To cover all community detection algorithms is beyond the scope of this paper, and interested readers can refer to the survey papers such as [10].

In recent years, a variety of methods inspired by different paradigms are put forward for community detection [35]. A prominent one is to consider the structural roles of individual nodes. This desideratum is motivated by the observation on many real-world networks that, by nature, community and structural role discovery are interdependent and complementary to each other. Real-world communities often contain nodes with various roles for it to function, such as ones that interface with other communities and ones that are peripheral to community cores. On the other hand, the role assignment of a node also depends on the communities that the node itself, its neighbors and beyond belong to. Therefore, there exists a strong and crucial need to detect communities and roles jointly. Recent work has leveraged role detection techniques for community detection [30]. None of those methods, however, consider the SH spanner detection.

6.2 Structural Hole Spanner Detection

Structural Hole theory is first introduced by [4] to find the key employees in organizations for integrating operations across functional and business boundaries. A series of empirical studies [3, 27] have demonstrated that advantages accrue to SH spanners who occupy bridging positions between different communities. In the literature, many strategies [13, 18] have been devised to model the property of structural holes in a network. [13] proposed a network formation model that a vertex serves as an intermediary between many vertices. The strategic link formation in their model leads to a star network, while real-world networks are not necessary of the star topology.

In recent years, several approaches have been proposed to find the top-k SH spanners. [13] formulated SH spanners as nodes that reside on large number of shortest paths between different pairs of nodes. Because counting all the shortest paths is time-consuming, [32] proposed a 2-Step approach that only counts the number of shortest paths with length two. Most recently, [28] viewed the SH spanners as a set of vertices whose removal will result in the maximum increase on the mean distance of the network, which is the average of the lengths of all pairs of vertices in the network. They then proposed the AP_BICC model, by exploiting the bounded inverse closeness centrality (BICC) of vertices and making use of articulation points (AP) of the network.

To our best knowledge, there is only one paper that utilized the community information for mining the top-k SH spanners [20], which assumes the communities are given. One instantiation of their proposed model is to find a set of vertices whose removal leads to the maximum decrease in the minimum cut in the given set of communities. However, communities usually are not known in most scenarios, thus the quality of the solution relies on the quality of communities found. In contrast, given only the topological structure of the network, our proposed HAM can detect the communities and the top-k SH spanners simultaneously. Furthermore, as demonstrated in the experiment, HAM captures the most influential intermediate nodes between communities, while most of the previous methods take the nodes in the center of the communities as SH spanners.

7. CONCLUSION

In this work, we proposed a novel Harmonic Modularity (HAM) method for simultaneously detecting the potential communities and the top-k SH spanners, using only the topological structure of the network. Specifically, we applied the harmonic function analysis to measure the harmonic modularity and to obtain the community indicator. We further investigated the sparsity level of the interactions between communities, with particular emphasis on the nodes connecting to multiple communities, to discriminate the indicator of SH spanners and assist the community guidance. Extensive experiments conducted on three real-world social networks demonstrated that HAM can capture the characteristics of structural hole spanners, and the proposed algorithm significantly outperform several comparative methods (even the methods using the supervised community information) in the top-k SH spanner identification problem and also the community detection problem, respectively.

8. ACKNOWLEDGMENTS

The work is supported in part by NSF (III-1526499), NSFC (61272050, 61472089, 61503253), NSFC-Guangdong Joint Found(U1501254), and the Science Foundation of Guangdong Province (2014A030313556).

9. REFERENCES

[1] D. S. Bassett, E. T. Bullmore, A. Meyer-Lindenberg, J. A. Apud, D. R. Weinberger, and R. Coppola. Cognitive fitness of cost-efficient brain functional networks. *PNAS*, 106(28):11747–11752, 2009.

[2] U. Brandes. A faster algorithm for betweenness centrality*. *Journal of Mathematical Sociology*, 25(2):163–177, 2001.

[3] 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.

[4] R. S. Burt. *Structural holes: The social structure of competition*. Harvard university press, 2009.

[5] V. Buskens and A. Van de Rijt. Dynamics of networks if everyone strives for structural holes1. *American Journal of Sociology*, 114(2):371–407, 2008.

[6] K. Ciesielski, D. Czerski, M. Dramiński, M. A. Kłopotek, and S. T. Wierzchoń. Semantic information within the beatca framework. *Control and Cybernetics*, 39(2):377–400, 2010.

[7] A. Clauset, M. E. Newman, and C. Moore. Finding community structure in very large networks. *Physical review E*, 70(6):066111, 2004.

[8] G. Cordasco and L. Gargano. Community detection via semi-synchronous label propagation algorithms. In *BASNA*, pages 1–8, 2010.

[9] I. S. Dhillon, Y. Guan, and B. Kulis. Weighted graph cuts without eigenvectors a multilevel approach. *TPAMI*, 29(11):1944–1957, 2007.

[10] S. Fortunato. Community detection in graphs. *Physics reports*, 486(3):75–174, 2010.

[11] M. Girvan and M. E. Newman. Community structure in social and biological networks. *PNAS*, 99(12):7821–7826, 2002.

[12] A. Goyal, W. Lu, and L. V. Lakshmanan. Celf++: optimizing the greedy algorithm for influence maximization in social networks. In *WWW*, pages 47–48, 2011.

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

[14] A. A. Hagberg, D. A. Schult, and P. J. Swart. Exploring network structure, dynamics, and function using NetworkX. In *SciPy*, pages 11–15, 2008.

[15] R. He, T. Tan, L. Wang, and W.-S. Zheng. $\ell_{2,1}$ regularized correntropy for robust feature selection. In *CVPR*, pages 2504–2511, 2012.

[16] U. Kang and C. Faloutsos. Beyond'caveman communities': Hubs and spokes for graph compression and mining. In *ICDM*, pages 300–309, 2011.

[17] D. Kempe, J. Kleinberg, and É. Tardos. Maximizing the spread of influence through a social network. In *KDD*, pages 137–146, 2003.

[18] J. Kleinberg, S. Suri, É. Tardos, and T. Wexler. Strategic network formation with structural holes. In *EC*, pages 284–293, 2008.

[19] S. Lin, Q. Hu, G. Wang, and S. Y. Philip. Understanding community effects on information diffusion. In *PAKDD*, pages 82–95, 2015.

[20] T. Lou and J. Tang. Mining structural hole spanners through information diffusion in social networks. In *WWW*, pages 825–836, 2013.

[21] M. E. Newman. Modularity and community structure in networks. *PNAS*, 103(23):8577–8582, 2006.

[22] M. E. Newman and M. Girvan. Finding and evaluating community structure in networks. *Physical review E*, 69(2):026113, 2004.

[23] A. Y. Ng, M. I. Jordan, Y. Weiss, et al. On spectral clustering: Analysis and an algorithm. *NIPS*, 2:849–856, 2002.

[24] F. Nie, H. Huang, X. Cai, and C. H. Ding. Efficient and robust feature selection via joint $\ell_{2,1}$-norms minimization. In *NIPS*, pages 1813–1821, 2010.

[25] M. Nikolova and M. K. Ng. Analysis of half-quadratic minimization methods for signal and image recovery. *SISC*, 27(3):937–966, 2005.

[26] L. Page, S. Brin, R. Motwani, and T. Winograd. The pagerank citation ranking: bringing order to the web. 1999.

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

[28] M. Rezvani, W. Liang, W. Xu, and C. Liu. Identifying top-k structural hole spanners in large-scale social networks. In *CIKM*, pages 263–272, 2015.

[29] M. Rosvall and C. T. Bergstrom. Multilevel compression of random walks on networks reveals hierarchical organization in large integrated systems. *PloS one*, 6(4):e18209, 2011.

[30] Y. Ruan and S. Parthasarathy. Simultaneous detection of communities and roles from large networks. In *COSN*, pages 203–214, 2014.

[31] D. A. Spielman. Algorithms, graph theory, and linear equations in laplacian matrices. In *ICM*, volume 4, pages 2698–2722, 2010.

[32] J. Tang, T. Lou, and J. Kleinberg. Inferring social ties across heterogenous networks. In *WSDM*, pages 743–752, 2012.

[33] M. P. van den Heuvel and O. Sporns. Rich-club organization of the human connectome. *The Journal of neuroscience*, 31(44):15775–15786, 2011.

[34] L. Wang, T. Lou, J. Tang, and J. E. Hopcroft. Detecting community kernels in large social networks. In *ICDM*, pages 784–793, 2011.

[35] Z. Wang, Z. Chen, Y. Zhao, and S. Chen. A community detection algorithm based on topology potential and spectral clustering. *The Scientific World Journal*, 2014.

[36] W. Winterbach, P. Van Mieghem, M. Reinders, H. Wang, and D. de Ridder. Topology of molecular interaction networks. *BMC systems biology*, 7(1):90, 2013.

[37] J. Yang and J. Leskovec. Defining and evaluating network communities based on ground-truth. *KAIS*, 42(1):181–213, 2015.

[38] W. W. Zachary. An information flow model for conflict and fission in small groups. *Journal of anthropological research*, pages 452–473, 1977.