Mutual or Unrequited Love: Identifying Stable Clusters in Social Networks with Uni- and Bi-directional Links

Yanhua Li*, Zhi-Li Zhang, and Jie Bao

University of Minnesota, Twin Cities
{yanhua,zhzhang,baojie}@cs.umn.edu

Abstract. Many social networks, e.g., Slashdot and Twitter, can be represented as directed graphs (*digraphs*) with two types of links between entities: mutual (bi-directional) and one-way (uni-directional) connections. Social science theories reveal that mutual connections are more stable than one-way connections, and one-way connections exhibit various tendencies to become mutual connections. It is therefore important to take such tendencies into account when performing clustering of social networks with both mutual and one-way connections.

In this paper, we utilize the *dyadic* methods to analyze social networks, and develop a generalized mutuality tendency theory to capture the tendencies of those node pairs which tend to establish mutual connections more frequently than those occur by chance. Using these results, we develop a *mutuality-tendency-aware* spectral clustering algorithm to identify more stable clusters by maximizing the *within-cluster* mutuality tendency and minimizing the *cross-cluster* mutuality tendency. Extensive simulation results on synthetic datasets as well as real online social network datasets such as Slashdot, demonstrate that our proposed mutuality-tendency-aware spectral clustering algorithm extracts more stable social community structures than traditional spectral clustering methods.

1 Introduction

Graph models are widely utilized to represent relations among entities in social networks. Especially, many online social networks, e.g., Slashdot and Twitter, where the users' social relationships are represented as directed edges in directed graphs (or in short, *digraphs*). Entity connections in a digraph can be categorized into two types, namely, bi-directional links (mutual connections) and uni-directional links (*one-way* connections). Social theories [28] and online social network analysis [2, 6, 28] have revealed that various types of connections exhibit different stabilities, where mutual connections are more stable than one-way connections. In other words, mutual connections are the source of social cohesion [3, 4] that, if two individuals mutually attend to one another, then the bond is reinforced in each direction.

Studying the social network structure and properties of social ties have been an active area of research. Clustering and identifying social structures in social networks is an especially important problem [8, 17, 24] that has wide applications, for instance, community detection and friend recommendation in social networks. Existing clustering

* The work is supported in part by the NSF grants CNS-0905037, CNS-1017647 and the DTRA Grant HDTRA1-09-1-0050.

A. Bonato and J. Janssen (Eds.): WAW 2012, LNCS 7323, pp. 113–125, 2012.

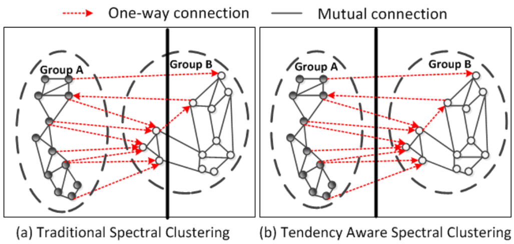

Fig. 1. An example network

methods [21, 29] are originally developed for *undirected* graphs, based on the classical *spectral clustering theory*. Several recent studies (see, e.g., [10, 21, 27, 29]) extend the spectral clustering method to digraphs, by first converting the underlying digraphs to undirected graphs via some form of *symmetrization*, and then apply spectral clustering to the resulting symmetrized (undirected) graphs. However, all these methods have two common drawbacks, which prevent them from obtaining *stable* clusters with *more mutual connections*. First, these methods do not explicitly distinguish between *mutual* and *one-way* connections commonly occurring in many social networks, treating them essentially as the same and therefore ignoring the different social relations and interpretations these two types of connections represent (see Section 2 for more details). Second, by simply minimizing the total cross-cluster links (that are symmetrized in some fashion), these methods do not explicitly account for the potential tendencies of node pairs to become mutually connected. As a simple example, Fig. 1 shows two groups of people in a network, where people in the same group tend to have more mutual (stable) connections, and people across two groups have more one-way (unstable) connections. When using the traditional spectral clustering method, as shown in Fig. 1(a), group B will be partitioned into two clusters, due to its strict rule of minimizing the total number of across cluster edges. On the other hand, the correct partition should be done as shown in Fig. 1(b), where the majority of mutual (stable) connections are placed within clusters, and one-way (unstable) connections are placed across clusters.

In this paper, we propose and develop a stable social cluster detection algorithm that takes into account the tendencies of node pairs whether to form mutual (thus stable) connections or not, which can result in more *stable* cluster structures. To tackle this clustering problem, we need to answer the following questions: 1) how to track and evaluate the tendencies of node pairs to become mutual (stable) relations? and 2) how to cluster the entities in social networks by accounting for their mutuality tendencies so as to extract more stable clustering structures?

To address these questions, we utilize dyadic methods to analyze social networks, and develop a generalized mutuality tendency theory which better captures the tendencies of node pairs that tend to establish mutual connections more frequently than those occur by chance. Using these results, we develop a *mutuality-tendency-aware* spectral clustering algorithm to detect more stable clusters by maximizing the *within-cluster* mutuality tendency and minimizing the *cross-cluster* mutuality tendency. Our contributions are summarized as follows.

◇ Motivated by the social science mutuality tendency theory, we establish a new *cluster-based* mutuality tendency theory. It yields a symmetrized mutuality tendency for each node pair, that measures the strength of social ties in (or across) clusters (Sec 3).
◇ Based on our theory, we develop a *mutuality-tendency-aware* spectral clustering algorithm that partitions the social graphs into stable clusters, by maximizing within-cluster mutuality tendencies and minimizing across-cluster mutuality tendencies (Sec 4).
◇ The experimental results – based on both social network structures of synthetical and real social network datasets – confirm that our clustering algorithm generates more stable clusters than the traditional spectral clustering algorithms (Sec 5).

2 Preliminaries, Related Work and Problem Definition

In this section, we first introduce the existing dyadic analysis methods in the social theory literature for analyzing and characterizing social network mutual connections and one-way connections. We then present the classic spectral clustering theory which was developed for *undirected* graphs, and briefly survey some related works which apply this theory to *digraphs* through *symmetrization*. We end the section with the problem definition, namely, how to identify *stable* clusters in social networks by taking into account mutuality tendencies of mutual and one-way connections.

2.1 Dyadic Analysis and Mutuality Tendency

Given a social network with both uni- and bi-directional links, such a network can be represented as a (simple) digraph $G = (V, E)$ with $|V| = n$ nodes. Let A be the standard adjacency matrix of the digraph, where $A_{ij} = 1$ if the directed edge $i \to j$ is present, and $A_{ij} = 0$ otherwise. Social scientists commonly view the social network G as a collection of dyads [28], where *a dyad is an* unordered *pair of nodes and* directed *edges between two nodes in the pair.* Denote a dyad as $Dy_{ij} = (A_{ij}, A_{ji})$, for $i < j$. Since dyad is an unordered notion, we have in total $N_d = n(n-1)/2$ dyads in G. Hence, there are only three possible isomorphism dyads. The first type of dyads is *mutual* relationship, where both directional edges $i \to j$ and $j \to i$ are present. The second type of dyads is *one-way* relationship, where either $i \to j$ or $j \to i$ is present, but not both. The last type of dyads is *null* relationship, where no edges show up between i and j. Let m, b, and u denote the number of mutual, one-way, and null dyads in the network. Clearly, $m + b + u = n(n-1)/2$.

Interpretations of Dyads. Social scientists have observed that mutual social relations and one-way relations in social networks typically exhibit different stabilities, namely, mutual relations are more stable than one-way relations [28]. Hence in the social science literature, one prevalent interpretation of dyadic relations in social networks are the following: mutual dyads are considered as stable connections between two nodes and null relation dyads represent no relations; the one-way dyads [1,5,16,18,20] are viewed as an *intermediate* state of relations, which are in transition to more stable equilibrium states of reciprocity (mutual or no relation). Several recent empirical studies [6, 9] of online social networks have further revealed and confirmed that mutual social relations are more stable relations than one-way connections.

Measuring Mutuality Tendency. The notion of mutuality tendency has been introduced in the social science literature (see, e.g., [7, 28]) to measure the tendency for a node pair to establish mutual connections. For any dyad between i and j in a digraph G, if i places a link to j, ρ_{ij} represents the tendency that j will reciprocate to i more frequently than would occur by chance. Let $\mathbf{X}_{ij}$ denote the random variable that represents whether or not node i places a directed edge to node j. There are only two possible events (i.e., $\mathbf{X}_{ij}$ takes two possible values): $\mathbf{X}_{ij} = 1$, representing the edge is present; or $\mathbf{X}_{ij} = 0$, the edge is not present. Let X_{ij} (resp. $\bar{X}_{ij}$) denote the event $\{\mathbf{X}_{ij} = 1\}$ (resp. $\{\mathbf{X}_{ij} = 0\}$). Then the probability of the event X_{ij} occurring is $P(X_{ij})$. The probability that i places a directed edge to j and j reciprocates back (i.e., node i and node j are mutually connected) is thus given by $P(X_{ij}, X_{ji}) = P(X_{ij})P(X_{ji}|X_{ij})$. Wofle [28] introduces the following measure of mutuality tendency in terms of the conditional probability $P(X_{ji}|X_{ij})$ as follows:

$$P(X_{ji}|X_{ij}) = P(X_{ji}) + \rho_{ij} P(\bar{X}_{ji}) = \frac{P(X_{ij}, X_{ji})}{P(X_{ij})}, \tag{1}$$

where $-\infty < \rho_{ij} \leq 1$ ensures $0 \leq P(X_{ji}) + \rho P(\bar{X}_{ji}) \leq 1$ to hold. Like many indices used in statistics, $-\infty < \rho \leq 1$ is dimensionless and easy to interpret, since it uses 0 and 1 as benchmarks, representing no tendency and maximum tendency for reciprocation. From eq.(1), the joint distribution $P(X_{ij}, X_{ji})$ in eq.(1) can be measured by the observed graph, namely, either $P(X_{ij}, X_{ji}) = P^{(\omega)}(X_{ij}, X_{ji}) = 1$, when i and j have mutual connection, or $P(X_{ij}, X_{ji}) = P^{(\omega)}(X_{ij}, X_{ji}) = 0$, otherwise, where the superscript ω indicates that the probability is obtained from the observed graph. On the other hand, the distribution for each individual edge is measured by $P(X_{ij}) = P^{(\mu)}(X_{ij}) = \frac{d_i}{|V|-1}$, where d_i is the out-going degree of node i. $P^{(\mu)}(X_{ij})$ represents the probability of edge $i \to j$ being generated under a random graph model, denoted by the superscript μ, with edges randomly generated while preserving the out-degrees. Hence, the tendency ρ is obtained by implicitly comparing the observed graph with a reference random digraph model.

Limitations of Wolfe's Mutuality Tendency Measure for Stable Social Structure Clustering. Although the node pair in a dyad is unordered (i.e., the two nodes are treated "symmetrically" in terms of dyadic relations), Wolfe's measure of mutual tendency is in fact *asymmetric*. This can be easily seen through the following derivation. By definition,

$$\frac{\rho_{ji}}{\rho_{ij}} = \frac{P(X_{ji})P(\bar{X}_{ij})}{P(X_{ij})P(\bar{X}_{ji})} = \frac{P(X_{ji}) - P(X_{ij})P(X_{ji})}{P(X_{ij}) - P(X_{ij})P(X_{ji})}$$

We see that $\rho_{ij} = \rho_{ji}$ if and only if $P(X_{ij}) = P(X_{ji})$ holds. Hence, given an arbitrary dyad in a social network Wolfe's measure of mutuality tendency of the node pair is asymmetric – in a sense that it is a *node-specific* measure of mutuality tendency. It does not provide a measure of mutuality tendency of the (unordered) *node pair* viewed together. In Section 3, we will introduce a new measure of mutuality tendency that is *symmetric* and captures the tendency of a node pair in a dyadic relation to establish mutual connection. This measure of mutuality tendency can be applied to clusters and a whole network in a straightforward fashion, and leads us to develop a *mutuality-tendency-aware* spectral clustering algorithm.

2.2 Spectral Clustering Theory and Extensions to Digraphs via Symmetrization

Spectral clustering methods (see, e.g., [15, 22, 26, 27, 29]) are originally developed for clustering data with symmetric relations, namely, data that can be represented as *undirected* graphs, where each relation (edge) between two entities, $A_{ij} = A_{ji}$, represents their similarity. The goal is to partition the graph such that entities within each cluster are more similar to each other than those across clusters. This is done by minimizing the total weight of cross-cluster edges. Especially, [12] provides a systematic study on comparing a wide range of undirected graph based clustering algorithms using real large datasets, which gives a nice guideline of how to select clustering algorithms based on the underlying networks and the targeting objectives.

When relations between entities are *asymmetric*, or the underlying graph is *directed*, spectral clustering cannot be directly applied, as the notion of (semi-)definiteness is only defined for *symmetric* matrices. Several recent studies (see, e.g., [10, 21, 27, 29]) all attempt to circumvent this difficulty by first converting the underlying digraphs to undirected graphs via some form of *symmetrization*, and then apply spectral clustering to the resulting symmetrized (undirected) graphs. For example, the authors in [21] discuss several symmetrization methods, including the symmetrized adjacency matrix $\bar{A} = (A + A^T)/2$, the bibliographic coupling matrix AA^T and the co-citation strength matrix $A^T A$, and so forth. Symmetrization can also be done through a random walk on the underlying graph, where $P = D^{-1}A$ is the probability transition matrix and $D = diag[d_i^{out}]$ is a diagonal matrix of node out-degrees. For example, taking the objective function as the random walk flow circulation matrix $F_\pi = \Pi P$, where Π is the diagonal stationary distribution matrix, we have the symmetrized Laplacian of the circulation matrix as $\bar{\mathcal{L}} = (\tilde{\mathcal{L}} + \tilde{\mathcal{L}}^T)/2$, where $\tilde{\mathcal{L}}$ is the (asymmetric) digraph Laplacian matrix [13]. Then the classical spectral clustering algorithm can then be applied using $\bar{\mathcal{L}}$ which is symmetric and semi-definite. Zhou and et al [27, 29] use this type of symmetrization to perform clustering on digraphs. Moreover, Leicht and Newman [10] propose the digraph modularity matrix $Q = [Q_{ij}]$, which captures the difference between the observed digraph and the hypothetical random graph with edges randomly generated by preserving the in- and out-degrees of nodes, namely, $Q_{ij} = A_{ij} - d_i^{out} d_j^{in}/m$. Then, if the sum of edge modularities in a cluster S is large, nodes in S are well connected, since the edges in S tend to appear with higher probabilities than occur by chance. However, Q by definition is asymmetric, where [10] uses the symmetrized $\bar{Q} = (Q + Q^T)/2$ as objective to perform spectral clustering method. Essentially, the edge modularity captures how an individual edge appears more frequently than that happens by chance, thus the modularity based clustering method tends to group those nodes with more connections than expected together, which like all other clustering methods presented above completely ignores the distinction between mutual and one-way connections.

Problem Definition. In this paper we want to solve the following clustering problem in social networks with bi- and uni-directional links: Given a directed (social) graph where mutual connections represent more stable relations and one-way connections represent intermediate transferring states, *how can we account for mutual tendencies of dyadic* *relations and cluster the entities in such a way that nodes within each cluster have maximized mutuality tendencies to establish mutual connections, while across clusters, nodes have minimized tendencies to establish mutual connections?* The clusters (representing social structures or communities) identified and extracted thereof will hence likely be more stable.

3 Cluster-Based Mutuality Tendency Theory

Inspired by Wolfe's study in [28], we propose a new measure of mutuality tendency for dyads that can be generalized to groups of nodes (clusters), and develop a *mutuality tendency theory* for characterizing the strength of social ties within a cluster (network structure) as well as across clusters in an asymmetric social graph. This theory lays the theoretical foundation for the network structure classification and community detection algorithms we will develop in section 4.

Let $\mathbf{X}_{ij}$ denote the random variable that represents whether or not node i places a directed edge to node j. There are only two possible events (i.e., $\mathbf{X}_{ij}$ takes two possible values): $\mathbf{X}_{ij} = 1$, representing the edge is present; or $\mathbf{X}_{ij} = 0$, the edge is not present. Let X_{ij} (resp. $\bar{X}_{ij}$) denote the event $\{\mathbf{X}_{ij} = 1\}$ (resp. $\{\mathbf{X}_{ij} = 0\}$). Given an *observed* (asymmetric) social graph G, to capture the *mutuality tendency* of dyads in this graph, we compare it with a *hypothetical*, *random* (social) graph, denoted as $G^{(\mu)}$, where links (dyadic relations) are generated randomly (i.e., by chance) in such a manner that the (out-)degree d_i of each node i in $G^{(\mu)}$ is the same as that in the observed social graph G. Under this random social graph model, the probability of the event X_{ij} occurring is $P^{(\mu)}(X_{ij}) = \frac{d_i}{|V|-1}$; namely, i places a (directed) link to node j randomly or by chance (the superscript μ indicates the probability distribution of link generations under the random social graph model). The probability that i places a directed edge to j and j reciprocates back (i.e., node i and node j are mutually connected) is thus given by $P^{(\mu)}(X_{ij}, X_{ji}) = P^{(\mu)}(X_{ij})P^{(\mu)}(X_{ji}|X_{ij}) = P^{(\mu)}(X_{ij})P^{(\mu)}(X_{ji})$, since $\mathbf{X}_{ij}$ and $\mathbf{X}_{ji}$ are independent under the random social graph model. On the observed social graph, denote $P^{(\omega)}(X_{ij}, X_{ji})$ to represent the event whether there is a mutual connection (symmetric link) between node i and node j, i.e., $P^{(\omega)}(X_{ij}, X_{ji}) = 1$, if the dyad Dy_{ij} is a mutual dyad in the *observed* social graph, and $P^{(\omega)}(X_{ij}, X_{ji}) = 0$, otherwise. We define the *mutuality tendency* of dyad Dy_{ij} as follows:

$$\theta_{ij} := P^{(\omega)}(X_{ij}, X_{ji}) - P^{(\mu)}(X_{ij}, X_{ji}) = P^{(\omega)}(X_{ij}, X_{ji}) - P^{(\mu)}(X_{ij})P^{(\mu)}(X_{ji}), \tag{2}$$

which captures how the node pair i and j establish a mutual dyad more frequently than would occur by chance.

This definition of mutuality tendency is a symmetric measure for dyad Dy_{ij}, i.e., $\theta_{ij} = \theta_{ji}$. In addition, it is shown that $\theta_{ij} \in [-1, 1]$. We remark that $\theta_{ij} = 0$ indicates that if node i places a directed link to node j, the tendency that node j will reciprocate back to node i is no more likely than would occur by chance; the same holds true if node j places a directed link to node i instead. On the other hand, $\theta_{ij} > 0$ indicates that if node i (resp. node j) places a directed link to node j (resp. node i), node j (resp. node i) will more likely than by chance to reciprocate. In particular, with $\theta_{ij} = 1$, node

j (resp. node i) will almost surely reciprocate. In contrast, $\theta_{ij} < 0$ indicates that if node i (resp. node j) places a directed link to node j (resp. node i), node j (resp. node i) will tend not to reciprocate back to node i (resp. node j). In particular, with $\theta_{ij} = -1$, node j (resp. node i) will almost surely not reciprocate back. Hence θ_{ij} provides a measure of strength of social ties between node i and j: $\theta_{ij} > 0$ suggests that the dyadic relation between node i and j is stronger, having a higher tendency (than by chance) to become mutual; whereas $\theta_{ij} < 0$ suggests that node i and j have weaker social ties, and their dyadic relation is likely to remain asymmetric or eventually disappear.

Mutuality Tendency of Clusters. The mutuality tendency measure for dyads defined in eq.(2) can be easily generalized for an arbitrary cluster (a subgraph) in an observed social graph, $S \subseteq G$. We define the mutuality tendency of a cluster S, Θ_S, as follows:

$$\Theta_S := \sum_{i\sim j; i,j\in S} P^{(\omega)}(X_{ij}, X_{ji}) - \sum_{i\sim j; i,j\in S} P^{(\mu)}(X_{ij}) P^{(\mu)}(X_{ji}), \tag{3}$$

where the subscript $i \sim j : i,j \in S$ means that the summation accounts for all (unordered) dyads, and i,j are both in S. Denote the second term in eq.(3) as $m_S^{(\mu)}$, and the (out-degree) volume of the cluster S as $d_S := \sum_{i\in S} d_i$. As $P^{(\mu)}(X_{ij}) = d_i/(|V|-1)$ and $P^{(\mu)}(X_{ji}) = d_j/(|V|-1)$,

$$m_S^{(\mu)} = \sum_{i\sim j; i,j\in S} \frac{d_i d_j}{(|V|-1)^2} = \frac{d_S^2 - \sum_{i\in S} d_i^2}{2(|V|-1)^2}, \tag{4}$$

which represents the expected number of mutual connections among nodes in S *under the random social graph model*. Given the cluster S in the observed social graph G, define $m_S^{(\omega)} := \sum_{i\sim j; i,j\in S} P^{(\omega)}(X_{ij}, X_{ji})$, namely, $m_S^{(\omega)}$ represents the number of (observed) mutual connections among nodes in the cluster S in the observed social graph G. The mutual tendency of cluster S defined in eq.(3) is therefore exactly $\Theta_S = m_S^{(\omega)} - m_S^{(\mu)}$.

Hence Θ_S provides a measure of strength of (likely mutual) social ties among nodes in a cluster: $\Theta_S > 0$ suggests that there are more mutual connections among nodes in S than would occur by chance; whereas $\Theta_S < 0$ suggests that there are fewer mutual connections among nodes in S than would occur by chance. Using Θ_S, we can therefore quantify and detect clusters of nodes (network structures or communities) that have strong social ties. In particular, when $S = G$, Θ_G characterizes the mutuality tendency for the entire digraph G, i.e., $\Theta_G = m_G^{(\omega)} - m_G^{(\mu)} = \sum_{i\sim j} \theta_{ij}$, where $m_G^{(\omega)} := \sum_{i\sim j} P^{(\omega)}(X_{ij}, X_{ji})$ represents the number of (observed) mutual dyads among nodes in the observed social graph G, and

$$m_G^{(\mu)} = \sum_{i\sim j} \frac{d_i d_j}{(|V|-1)^2} = \frac{d^2 - \sum_{i\in V} d_i^2}{2(|V|-1)^2}, \tag{5}$$

represents the expected number of mutual dyads among nodes in G *under the random social graph model*. Likewise, given a bipartition $(S, \bar{S})$ of G, we define the cross-cluster mutuality tendency as

$$\Theta_{\partial S} := \sum_{i \in S \sim j \in \bar{S}} (P^{(\omega)}(X_{ij}X_{ji}) - P^{(\mu)}(X_{ij})P^{(\mu)}(X_{ji})) \tag{6}$$

Denote the second quantity in eq.(6) as $m_S^{(\mu)}$,

$$m_{\partial S}^{(\mu)} = \sum_{i \in S \sim j \in \bar{S}} \frac{d_i d_j}{(|V|-1)^2} = \frac{d_S d_{\bar{S}}}{(|V|-1)^2} \tag{7}$$

which represents the expected number of mutual connections among nodes across S and $\bar{S}$ under the random social graph model. Define $m_{\partial S}^{(\omega)} := \sum_{i \in S \sim j \in \bar{S}} P^{(\omega)}(X_{ij}, X_{ji})$ representing the number of (observed) mutual connections among nodes across clusters S and $\bar{S}$ in the observed social graph G. The mutuality tendency across cluster S and $\bar{S}$ defined in eq.(6) is therefore exactly $\Theta_{\partial S} = m_{\partial S}^{(\omega)} - m_{\partial S}^{(\mu)}$.

The mutuality tendency theory outlined above accounts for different interpretations and roles mutual and one-way connections represent and play in asymmetric social graphs, with the emphasis in particular on the importance of mutual connections in forming and developing stable social structures/communities with strong social ties. In the next section, we will show how we can apply this mutuality tendency theory for detecting and clustering stable network structures and communities in asymmetric social graphs.

4 Mutuality-Tendency-Aware Spectral Clustering Algorithm

In this section, we establish the basic theory and algorithm for solving the mutuality-tendency-aware clustering problem. Due to the space limitation, some proofs are delegated to the technical report [14].

Without loss of generality, we consider only simple (unweighted) digraphs $G = (V, E)$ (i.e., the adjacency matrix A is a 0-1 matrix). Define the mutual connection matrix $M := \min(A, A^T)$, which expresses all the mutual connections with unit weight 1. In other words, if node i and node j are mutually connected (with bidirectional links), $M_{ij} = M_{ji} = 1$, otherwise, $M_{ij} = M_{ji} = 0$. Hence, we have $M_{ij} = P^{(\omega)}(X_{ij}, X_{ji})$, representing the event whether there is a mutual connection (symmetric link) between node i and node j, i.e., in the dyad Dy_{ij} in the observed social graph. In addition, let δ_{ij} be the Kronecker delta symbol, i.e., $\delta_{ij} = 1$ if $i = j$, and $\delta_{ij} = 0$ otherwise. Then, we define matrix

$$\bar{M} = \frac{dd^T - diag[d^2]}{(|V|-1)^2}$$

with d as the out-going degree vector, where each entry

$$\bar{M}_{ij} = \frac{d_i d_j - \delta_{ij} d_i^2}{(|V|-1)^2} = \begin{cases} \frac{d_i d_j}{(|V|-1)^2} & \text{if } i \neq j \\ 0 & \text{if } i = j \end{cases} \tag{8}$$

represents the probability that two nodes i and j independently place two unidirectional links to each other to form a mutual dyad. Hence, $\bar{M}_{ij} = P^{(\mu)}(X_{ij})P^{(\mu)}(X_{ji})$ represents the probability of node pair i and j to establish a mutual connection under random graph model with edges randomly generated by preserving the node out-degrees. We denote $T = M - \bar{M}$ as the mutuality tendency matrix, with each entry

$$T_{ij} = P^{(\omega)}(X_{ij}, X_{ji}) - P^{(\mu)}(X_{ij})P^{(\mu)}(X_{ji}) = \theta_{ij} \tag{9}$$

as the individual dyad mutuality tendency.

Mutuality Tendency Lapacian. T is symmetric and those entries associated with non-mutual dyads are negative, representing less mutuality tendencies to establish mutual connections than those occur by chance. Define the mutuality tendency Laplacian matrix as

$$L_T = D_T - T \tag{10}$$

where $D_T = diag[d_T(i)]$ is the diagonal degree matrix of T, with $d_T(i) = \sum_j T_{ij}$. We have the following theorem presenting several properties of L_T.

Theorem 1. *The mutuality tendency Laplacian matrix L_T as defined in eq.(10) has the following properties*

- *Given a column vector $x \in \mathbb{R}^{|V|}$, the bilinear form $x^T L_T x$ satisfies*

$$x^T L_T x = \sum_{i \sim j} T_{ij}(x_i - x_j)^2. \tag{11}$$

- *L_T is symmetric and in general indefinite. In addition, L_T has one eigenvalue equal to 0, with corresponding eigenvector as $\mathbf{1} = [1, \cdots, 1]^T$.*

Mutuality Tendency Ratio Cut Function. For a digraph $G = (V, E)$, and a partition $V = (S, \bar{S})$ on G, we define the *mutuality tendency ratio cut function* as follows.

$$TRCut(S, \bar{S}) = \Theta_{\partial S}\left(\frac{1}{|S|} + \frac{1}{|\bar{S}|}\right), \tag{12}$$

which represents the overall mutuality tendency across clusters balanced by the "sizes" of the clusters. Then, the clustering problem is formulated as a minimization problem with $K = 2$ clusters. (More general cases with $|V| \geq K > 2$ will be discussed in the next subsection.)

$$\min_S TRCut(S, \bar{S}) \tag{13}$$

Since $\Theta_{\partial S} = \Theta_G - (\Theta_S + \Theta_{\bar{S}})$ holds true, we have

$$TRCut(S, \bar{S}) = (\Theta_G - (\Theta_S + \Theta_{\bar{S}}))\left(\frac{1}{|S|} + \frac{1}{|\bar{S}|}\right).$$

For a given graph G, the graph mutuality tendency Θ_G is a constant, the minimization problem in eq.(13) is equivalent to the following maximization problem:

$$\max_S \left\{(\Theta_S + \Theta_{\bar{S}} - \Theta_G)\left(\frac{1}{|S|} + \frac{1}{|\bar{S}|}\right)\right\}. \tag{14}$$

Hence, minimizing the cross-cluster mutuality tendency is equivalent to maximize the within-cluster mutuality tendency. Using the results presented in Theorem 1, we prove the following theorem which provides the solution to the above mutuality tendency optimization problem.

Theorem 2. *Given the tendency Laplacian matrix $L_T = D_T - T$, the signs of the eigenvector of L_T corresponding to the smallest non-zero eigenvalue indicate the optimal solution $(S, \bar{S})$ to the optimization problem eq.(13).*

Moreover, the mutuality-tendency-aware spectral clustering can be easily generalized for the case of $K > 2$ (See more details in [14]).

Choice of K. We choose K, i.e., the total number of clusters, using the eigengap heuristic [25]. Theorem 1 shows that L_T has all real eigenvalues. Denote the eigenvalues of L_T in an increasing order, i.e., $\lambda_1 \leq \cdots \leq \lambda_n$, The index of the largest eigengap, namely, $K := \text{argmax}_{2 \leq K \leq n}(g(K))$, where $g(K) = \lambda_K - \lambda_{K-1}$, $K = 2, \cdots, n$, indicates how many clusters there are in the network.

5 Evaluations

In this section, we evaluate the performance of the *mutuality-tendency-aware* spectral clustering method by comparing it with various symmetrization methods based digraph spectral clustering algorithms. We only present the comparison results for the adjacency matrix symmetrization method, with objective matrix as $\bar{A} = (A + A^T)/2$. For other settings, we obtained similar results and omit them here, due to the space limitation. We will 1) first test the performances using synthetic datasets, and then 2) apply our method to real online network datasets, e.g., Slashdot social network, and discover stable clusters with respect to mutual and one-way connections.

Synthetic Datasets. We first consider synthetic datasets designed specifically to test the performance of our mutuality-tendency-aware spectral clustering method. We randomly generate a network with 1200 nodes and $K = 3$ clusters, that contain 500, 400 and 300 nodes, respectively. There are 54675 directional edges, among which 27336 edges are bidirectional and 27339 edges are unidirectional. We are randomly placed 90.02% of the bidirectional edges *in* clusters, and 89.6% of the unidirectional edges *across* clusters. Fig. 2(i)-(iii) show that traditional spectral clustering algorithm detects clusters with 661, 538 and 1 entities, respectively, while our method identify correct clusters (See Fig. 3(i)-(iii)).

Real Social Networks. In the second set of simulations, we applied our *mutuality-tendency-aware* spectral clustering algorithm to several real social network datasets, e.g., Slashdot [23], Epinions [19], and email communication network [11] datasets, and compare with various symmetrization methods based digraph clustering algorithms, such as $A = (A + A^T)/2$, AA^T and $F_\pi = \Pi P$. Here we only show the comparison results with adjacency matrix symmetrization based digraph spectral clustering on Slashdot dataset. All other settings lead to similar results and we omit them here.

Slashdot is a technology-related news website founded in 1997. Users can submit stories and it allows other users to comment on them. In 2002, Slashdot introduced

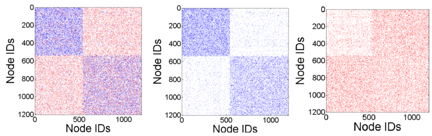

Fig. 2. (i)All edges, (ii)Bidirectional edges, (iii)Unidirectional edges (Traditional spectral clustering method in synthetic dataset)

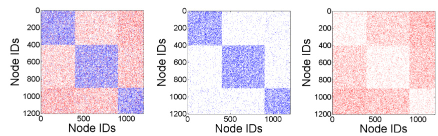

Fig. 3. (i)All edges, (ii)Bidirectional edges, (iii)Unidirectional edges (Tendency aware spectral clustering in synthetic dataset)

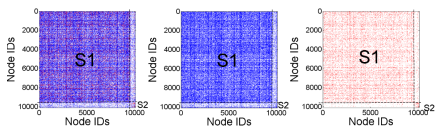

Fig. 4. (i)All edges, (ii)Bidirectional edges, (iii)Unidirectional edges (Traditional spectral clustering method in Slashdot dataset)

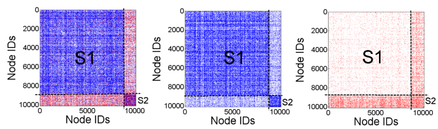

Fig. 5. (i)All edges, (ii)Bidirectional edges, (iii)Unidirectional edges (Tendency aware spectral clustering in Slashdot dataset)

the Slashdot Zoo feature which allows users to tag each other as friends or foes. The network data we used is the Slashdot social relation network, where a directed edge from i to j indicates an interest from i to j's stories (or topics). Hence, two people with mutual connections thus share some common interests, while one-way connections infer that one is interested in the other's posts, but the interests are not reciprocated back. The Slashdot social network data was collected and released by Leskovec [23] in November 2008.

The statistics[1] are shown in Table 1. It shows that the largest strongly connected component (SCC) include about 70355 nodes. Then, we remove those nodes with very low in-degrees and out-degrees, say no more than or equal to 2. By finding the largest strongly connected component of the remaining graph, we extract a "core" of the network with 10131 nodes and 197378 edges, among which there are 21404 unidirectional edges and 175974 bidirectional edges, respectively. In our evaluations, we observe that there is a large "core" of the network, and all other users are attached to this core network. In our study, we are interested in extracting the community structure from the "core" network. When applying our spectral clustering algorithm to the "core" network, two clusters with 8892 and 1239 nodes are detected (shown in Fig.5(i)-(iii)). In our result, a large portion (about 35.04%) of cross-cluster edges are unidirectional edges which in turn yield lower mutuality tendency across clusters. On the other hand, when using the traditional symmetrized $\bar{A} = (A + A^T)/2$, two clusters with 9640 and 491 nodes are extracted instead (shown in Fig.4(i)-(iii)). We can see that the clustering result obtained using the traditional spectral clustering method has only around 5.75% of the total edges across clusters as unidirectional edges, which boost up the mutuality tendency across clusters. However, in our clustering result, we have more unidirectional edges placed across clusters, which decreases the mutuality tendency

[1] Here, the total number of edges is smaller than that is shown on the website [23], because we do not count for those selfloops.

Table 1. Statistics of Slashdot Dataset (U-edge: Unidirectional edge, B-edge: Bidirectional edge)

Nodes	77360	Nodes in largest SCC	70355	Nodes in the "core" component	10131
Edges	828161	Edges in largest SCC	818310	Edges in the "core" component	197378
U-edges	110199	U-edges in largest SCC	100930	U-edges in the "core" component	21404
B-edges	717962	B-edges in largest SCC	717380	B-edges in the "core" component	175974

Table 2. Ave. mutuality tendency comparison on Slashdot dataset

.	θ_G	θ_{S1}	θ_{S2}	$\theta_{\partial S}$
Mutuality tendency aware clustering	0.0017	**0.0049**	**0.0028**	**0.00033**
Traditional clustering	0.0017	0.0018	0.0021	0.00070

across clusters. From Fig. 5(i), we can clearly see that we have unidirectional (red) edges dominating the cross-cluster parts. Moreover, Table 2 shows the average mutuality tendency comparison between different clustering methods, where we can see that the mutuality-tendency-aware spectral clustering algorithm can group nodes together with higher within-cluster tendencies than that of traditional spectral clustering.

References

1. Berscheid, E., Regan, P.: The psychology of interpersonal relationships. Pearson Prentice Hall (2005)
2. DeScioli, P., Kurzban, R., Koch, E., Liben-Nowell, D.: Best friends: Alliances, friend ranking, and the myspace social network. Perspectives on Psychological Science 6(1), 6–8 (2011)
3. Golder, S., Yardi, S., Marwick, A.: A structural approach to contact recommendations in online social networks. In: Workshop on Search in Social Media, SSM (2009)
4. Gouldner, A.: The norm of reciprocity: A preliminary statement. American Sociological Review, 161–178 (1960)
5. Heider, F.: Attitudes and cognitive organization. Journal of Psychology 21(1), 107–112 (1946)
6. Jamali, M., Haffari, G., Ester, M.: Modeling the temporal dynamics of social rating networks using bidirectional effects of social relations and rating patterns. In: WWW (2011)
7. Katz, L., Powell, J.: Measurement of the tendency toward reciprocation of choice. Sociometry 18(4), 403–409 (1955)
8. Kurucz, M., Benczur, A., Csalogany, K., Lukacs, L.: Spectral clustering in telephone call graphs. In: WebKDD (2007)
9. Kwak, H., Chun, H., Moon, S.: Fragile online relationship: A first look at unfollow dynamics in twitter. In: CHI (2011)
10. Leicht, E., Newman, M.: Community structure in directed networks. Physical Review Letters 100(11), 118703 (2008)
11. Leskovec, J., Kleinberg, J., Faloutsos, C.: Graph evolution: Densification and shrinking diameters. ACM Transactions on Knowledge Discovery from Data (TKDD) 1(1), 1–41 (2007)
12. Leskovec, J., Lang, K., Mahoney, M.: Empirical comparison of algorithms for network community detection. In: WWW (2010)
13. Li, Y., Zhang, Z.-L.: Random Walks on Digraphs, the Generalized Digraph Laplacian and the Degree of Asymmetry. In: Kumar, R., Sivakumar, D. (eds.) WAW 2010. LNCS, vol. 6516, pp. 74–85. Springer, Heidelberg (2010)

14. Li, Y., Zhang, Z.-L., Bao, J.: Mutual or unrequited love: Identifying stable clusters in social networks with uni- and bi-directional links. Arxiv preprint arXiv:1203.5474 (March 2012)
15. Luxburg, U.: A tutorial on spectral clustering. Statistics and Computing 17, 395–416 (2007)
16. Miller, H., Geller, D.: Structural balance in dyads. Journal of Personality and Social Psychology 21(2), 135 (1972)
17. Mishra, N., Schreiber, R., Stanton, I., Tarjan, R.E.: Clustering Social Networks. In: Bonato, A., Chung, F.R.K. (eds.) WAW 2007. LNCS, vol. 4863, pp. 56–67. Springer, Heidelberg (2007)
18. Price, K., Harburg, E., Newcomb, T.: Psychological balance in situations of negative interpersonal attitudes. Journal of Personality and Social Psychology 3(3), 265 (1966)
19. Richardson, M., Agrawal, R., Domingos, P.: Trust Management for the Semantic Web. In: Fensel, D., Sycara, K., Mylopoulos, J. (eds.) ISWC 2003. LNCS, vol. 2870, pp. 351–368. Springer, Heidelberg (2003)
20. Roudrigues, A.: Effects of balance, positivity, and agreement in triadic social relations. Journal of Personality and Social Psychology 5(4), 472 (1967)
21. Satuluri, V., Parthasarathy, S.: Symmetrizations for clustering directed graphs. In: EDBT/ICDT (2011)
22. Shi, J., Malik, J.: Normalized cuts and image segmentation. IEEE Trans. Pattern Anal. Mach. Intell. 22(8), 888–905 (2000)
23. Slashdot. dataset, `http://snap.stanford.edu/data/soc-Slashdot0811.html`
24. Smyth, S.: A spectral clustering approach to finding communities in graphs. In: SDM (2005)
25. von Luxburg, U.: A tutorial on spectral clustering. Technical Report No.TR-149, Max Planck Institute for Biological Cybernetics (2006)
26. Wang, X., Davidson, I.: Flexible constrained spectral clustering. In: KDD 2010, pp. 563–572 (2010)
27. Weston, J., Leslie, C.S., Ie, E., Zhou, D., Elisseeff, A., Noble, W.S.: Semi-supervised protein classification using cluster kernels. Bioinformatics 21(15), 3241–3247 (2005)
28. Wolfe, A.: Social network analysis: Methods and applications. American Ethnologist 24(1), 219–220 (1997)
29. Zhou, D., Huang, J., Schölkopf, B.: Learning from labeled and unlabeled data on a directed graph. In: ICML (2005)