Non-backtracking Operators for Community Detection in Signed Networks

Zhaoyue Zhong†, Xiangrong Wang‡§, Cunquan Qu*, Guanghui Wang†
*Data Science Institute, Shandong University, China
†School of Mathematics, Shandong University, China
‡Institute of Future Networks, Southern University of Science and Technology, Shenzhen, China
§Research Center of Networks and Communications, Peng Cheng Laboratory, Shenzhen, China

Abstract—Community detection or clustering is crucial for understanding the structure of complex systems. In some networks, nodes are allowed to be linked by either 'positive' or 'negative' edges. Such networks are called signed networks. Discovering communities in signed networks is more challenging. In this article, we innovatively propose a non-backtracking matrix for signed networks, and theoretically derive a detectability threshold and prove the feasibility in community detection. Furthermore, we improve the operator by considering the balanced paths in the network (denoted as balanced non-backtracking operator). Simulation results demonstrate that the balanced non-backtracking matrix-based approach significantly outperforms the adjacency network (denoted as balanced non-backtracking operator). It shows great potential to detect communities with or without overlap.

Index Terms—Community detection, signed networks, non-backtracking operator, spectral analysis, detectability threshold.

I. INTRODUCTION

COMMUNITIES, also known as clusters or modules, are groups of nodes that may share common attributes or have similar properties in the graph. Community detection divides similar nodes or nodes with a large number of (positive, large weighted) connections into a group, providing people with a possible way to control the network. Since nodes with a large number of (positive) connected edges often have similar properties, in terms of graphs, community detection is also a process of finding cut edges. If a few edges are removed, the network can be divided into several parts, i.e. several connected components that are not connected with each other, then the division of these parts is on certain level equivalent to community partition.

Community detection is widely applied in biology, computer science, engineering, economics, political science, sociology and other fields [1]. For example, protein-protein interaction networks are a research hotspot in biology and bioinformatics [2]. The interaction between proteins is the basis of every process in the cell. Each interaction is observed by experiments and marked as a connection. Proteins with the same or similar functions are divided into one module. We expect them to participate in the same process. At this time, the community structure is associated with most of the immunohistochemistry as well as tumor and metastasis. The above is a classical application which abstracts the actual situation as an unsigned network. At the same time, social network is also a typical network with community structure. In general research, connection is regarded as positive, such as friends, likes and forwarding. However, there are often a lot of negative connections in social networks. Some websites, such as epinions.com and slashdot.com, allow users to identify friends and enemies [3]. The signed network introduced in this paper is a representation of this situation, such as the opposite opinions on the same topic [3], and the blackout and reporting among users.

Tracking back to the 1940s [4], Heider introduced the concept of signed networks and the well-known structural balance theory, which states that 'the friends of my friends, as well as the enemies of my enemies, are my friends', see in Fig. 1. As one of the most popular theories in social science, structural balance theory has been addressed increasing attention recently. One of the topics is to design algorithms for computing the structural balance of large-scale datasets [5], [6], [7]. Another question is studying the impact of structural balance on some concrete applications, such as recommender systems [8], dynamic process [9], and so forth.

Fig. 1. Triangles in signed networks. $T_1$ and $T_2$ are balanced and relatively stable. $T_3$ and $T_4$ are unbalanced and hence liable to break apart.

In social networks, user communities provide better services for websites, such as user recommendation of friends, etc. Clustering of web pages can be used to rank web pages and provide more relevant search results [11]. Furthermore, the application of community detection in social media can better explain the observed phenomena and provide benchmarks for social mechanism [10].

In general, we can classify the existing methods of community detection into the following classes: (1) traditional algorithms, such as graph partitioning [11], hierarchical clustering [12], partition clustering [13], spectral clustering [14]; (2) modular-based methods [15]; (3) dynamic algorithms [16].
we divide the node set into two groups, $A$ and $B$. Between the corresponding node pair, the positive and negative relationships are represented as the positive and negative edges. Each individual in the group members, a negative edge between in-group members, the conditional probability of a positive edge between out-group members, and a negative edge between out-group members, respectively. Thus, the conditional probabilities satisfy $\rho_{in}^+ + \rho_{in}^- = 1$ and $\rho_{out}^+ + \rho_{out}^- = 1$.

Generally speaking, when we say a network has a community structure, at least one of the following conditions hold:

Case 1: There are more intra-community positive links than negative links, i.e., $\rho_{in}^+ > \rho_{out}^+$ and $\rho_{in}^- < \rho_{out}^-$. Case 2: The density of intra-community links is more significant than inter-community relationships, i.e. $d_{in} > d_{out}$.

For the first case, the community structure is sign-sensitive, denoted as relationship dependent community. The second one is link density-sensitive, and we call it as density dependent community.

B. Definition of non-backtracking matrix

One of the main contributions of this work is to define non-backtracking operators in signed networks which shows great potential to detect communities. Though non-backtracking matrix is well defined on unsigned networks, which is presented for completeness, a proper definition of non-backtracking matrix is far from trivial as shown in following sections.

Prior to the formal definition of non-backtracking matrix in signed networks, we present first the definition in unsigned or general networks. The non-backtracking matrix $H$, often called Hashimoto matrix in mathematics, are defined as follows:

$$\tilde{H}_{e,f} = \tilde{A}_e \tilde{A}_f 1(e_2 = f_1) 1(e_1 \neq f_2),$$ (1)

where $\tilde{A}$ is the adjacency matrix of unsigned networks, and $e = (e_1, e_2)$ and $f = (f_1, f_2)$ are two directed edges. Note that if the network is undirected, then we treat each undirected edge as two directed edges. Hence, the matrix $\tilde{H}$ is of dimension $2m \times 2m$, where $m$ is the number of edges in the network. Actually, the non-backtracking matrix $H_{2m \times 2m}$ can be written in the format below,

$$\tilde{H}(e_1 \rightarrow e_2).(f_1 \rightarrow f_2) = \begin{cases} 1 & \text{if } e_2 = f_1 \text{ and } e_1 \neq f_2, \\ 0 & \text{otherwise.} \end{cases}$$ (2)

Similar with Eq. (1), the non-backtracking matrix, denoted as $H$, of signed networks can be directly derived as follows,

$$H_{e,f} = A_e A_f 1(e_2 = f_1) 1(e_1 \neq f_2).$$ (3)

We can write it as

$$H(e_1 \rightarrow e_2).(f_1 \rightarrow f_2) = \begin{cases} 1 & \text{if } e_2 = f_1, e_1 \neq f_2 \text{ and } \\
\text{sign}(e_1 \rightarrow e_2) = \text{sign}(f_1 \rightarrow f_2), \\ -1 & \text{if } e_2 = f_1, e_1 \neq f_2 \text{ and } \\
\text{sign}(e_1 \rightarrow e_2) \neq \text{sign}(f_1 \rightarrow f_2), \\ 0 & \text{otherwise,} \end{cases}$$ (4)

where $\text{sign}(e_1 \rightarrow e_2)$ denotes the sign of a directed edge $e_1 \rightarrow e_2$ which takes value of either 1 or $-1$. The significance of the defined non-backtracking matrix $H$ is that real information can be transferred between the edge pairs of two identical signs and false information can be transferred between two edges with different signs, which accurately encodes the theory of structural balance (a triple with either one or three negative signs is unstable).
C. Alternative definition of non-backtracking matrix based on linearized belief propagation

Belief propagation (BP) is a kind of acyclic message passing algorithm, which calculates the exact marginal distribution of each vertex in the network. Although BP is designed to work correctly on trees, it is usually applied to general graphs that are sparse and may contain loops [21], [24], [25].

This algorithm starts from the appropriate initial assignment and performs iteration for some “messages”. Specifically, for each edge \((v, w)\) in a graph \(G = (V, E)\), the message \(\eta_{v \to w}^+\) indicates the conditional probability that \(v\) belongs to community \(a\) when \(w\) does not, and the message \(\eta_{w \to v}^-\) indicates the probability that \(w\) belongs to community \(a\) when \(v\) does not. Usually \(\eta_{v \to w}^+ \neq \eta_{w \to v}^-\). As you can see, although the original graph is undirected, these messages are passed on the directed edge, with each message between 0 and 1. Based on information transfer, it can be calculated iteratively.

BP algorithm has a good consistency with the actual grouping of the underlying network. Although BP is designed to work correctly on trees, it is usually applied to general graphs that are sparse and may contain loops [21], [24], [25].

In the directed edge sign, we generalize the existing BP updating equation of the BP algorithm. Due to the appearance of the edges, we define the non-backtracking matrix from the following form, for \(u \in \mathcal{N}(v)\),

\[
\frac{\eta_{v \to w}^+}{\eta_{v \to w}^-} := e^{-h} \prod_{\text{sign}(u \to v) = \text{sign}(v \to w)} \eta_{u \to v}^+ c_{in} + \eta_{u \to v}^- c_{out} \prod_{\text{sign}(u \to v) = \neg \text{sign}(v \to w)} \eta_{u \to v}^+ c_{out} + \eta_{u \to v}^- c_{in} \prod_{\text{sign}(u \to v) \neq \text{sign}(v \to w)} \left(1 - \eta_{u \to v}^+ \right) c_{in} + \left(1 - \eta_{u \to v}^- \right) c_{out} \prod_{\text{sign}(u \to v) \neq \neg \text{sign}(v \to w)} \left(1 - \eta_{u \to v}^+ \right) c_{out} + \left(1 - \eta_{u \to v}^- \right) c_{in},
\]

(5)

where \(\eta_{v \to w}^+\) represents the probability that \(v\) belongs to a community when \(u\) does not belong to the network, \(\eta_{v \to w}^-\) represents two communities respectively. Note that \(e^{-h}\) indicates the information passed in from non-edges(points not adjacent to \(v\)), where \(h = (c_{in} - c_{out})(n_V^{BP} - n_E^{BP})\), and \(n_V^{BP}\) refers to the ratio of the current number of points in two communities to the total number of nodes estimated according to BP algorithm.

It should be noted that when \(u \to v\) and \(v \to w\) have different signs, the information passed in \(v \to w\) is not \(\eta_{v \to w}^+\), but the \((1 - \eta_{v \to w}^-)\). That means if these two edges have different signs, the fault information will be passed into \(v \to w\). Similarly, the trivial fixed point of the above updated equation is still \(\eta_{v \to w} = 1/2\), that is, the probability that each vertex is divided into two communities is equal.

Next, we consider the information update equation near the trivial fixed point. Writing \(\eta_{v \to w}^+ = 1/2 \pm \delta_{u \to v}\), and linearize around this fixed point (for more details, see Appendix A). We get an updating rule of \(\delta\)

\[
\delta := \frac{(c_{in} - c_{out})}{(c_{in} + c_{out})} H^T \delta.
\]

That is, \(H\) can also be obtained by BP algorithm.

In other words, we define the non-backtracking matrix from two different perspectives, deduces its role in community detection by theory, and proves its feasibility in basic stochastic block model by the linearization of BP updating equation around the fixed point as well.

III. COMMUNITY DETECTION

A. Analytical community detection threshold and detection vector

To demonstrate the applicability of the signed non-backtracking matrix in community detection, we derive the community detection threshold and a detection vector for an arbitrary signed network. Generalized by the conclusion in unsigned networks [21], [24], [26], [28], we define \(g^{out}\) and \(g^{in}\) as the \(N\)-dimensional vectors,

\[
g^{out}_u = \sum_{v \in \mathcal{N}(u)} g_{u \to v} \cdot \text{sign}(u \to v),
\]

\[
g^{in}_u = \sum_{v \in \mathcal{N}(u)} g_{v \to u} \cdot \text{sign}(v \to u),
\]

where \(\mathcal{N}(u)\) represents the neighbor set of node \(u\) and vector \(g\) in \(2m\)-dimension is a given vector. Different from the unsigned network, we not only sum over incoming and outgoing edges but also take the sign of edges into consideration.

Applying \(H\) to \(g\), we get that

\[
(Hg)^{out}_u = \sum_{v \in \mathcal{N}(u)} g^{out}_u - g^{in}_u,
\]

\[
(Hg)^{in}_u = (d_u - 1) \sum_{v \in \mathcal{N}(u)} g^{out}_u,
\]

where \(d_u\) means the degree of node \(u\) (regardless of the sign of edges).

By rewriting the above two equations in a matrix form, we get that

\[
\begin{bmatrix} (Hg)^{in}_u \\ (Hg)^{out}_u \end{bmatrix} = H^\prime \begin{bmatrix} g^{in}_u \\ g^{out}_u \end{bmatrix},
\]

\[
H^\prime = \begin{bmatrix} 0 & D - I \\ -I & A \end{bmatrix},
\]

(7)

where \(I\) is the identity matrix, \(D\) is the diagonal matrix of vertex degrees, and \(A\) is the adjacency matrix of the underlying unsigned structure corresponding to the signed network.

Suppose that \(Hg = \mu g\), we have

\[
\mu \begin{bmatrix} g^{in}_u \\ g^{out}_u \end{bmatrix} = H^\prime \begin{bmatrix} g^{in}_u \\ g^{out}_u \end{bmatrix}.
\]

If \(g^{in}\) and \(g^{out}\) are nonzero, then \(g^{in}\) is an eigenvector of \(H^\prime\) with the same eigenvalue \(\mu\). Hence,

\[
\mu g^{out} = \tilde{A} \cdot g^{out} - g^{in} = [A - \mu^{-1}(D - I)]g^{out}.
\]

So \(\mu\) is a root of the quadratic eigenvalue equation

\[
det \left[ \mu^{2}I - \mu \tilde{A} + (D - I) \right] = 0.
\]

(8)

Compared with the original non-backtracking matrix \(H\), the complexity to calculate eigenvalues of \(H^\prime\) will be greatly
reduced. This equation is well known in the theory of graph zeta functions [28]. It accounts for 2n of H’s eigenvalues, and the other 2m−2n are ±1.

Actually, we directly and simply prove that the spectrum of H is the same as that of H̃, which regards the network as an unsigned network. Because H can also be derived by multiplying all the elements of some rows and their symmetrical columns of H by −1, that is,
\[ |I−H| = (−1)^{2n−1} \cdot |I−H̃| = |I−H̃|, \]
where |ɛ| is the number of negative edges in the network.

Therefore, according to the previous conclusion, the bulk of the spectrum of B are also confined to the disk of radius \( \sqrt{c} \) in signed networks. Note that \( c = d \cdot n \) is the average degree of the network. Similarly, we can define \( c_{in} \) and \( c_{out} \) respectively.

Further, we can get the first and second eigenvalues of H,
\[ \begin{align*}
\mu_1 &\approx c \\
\mu_2 &\approx \mu_c = \frac{c_{in}−c_{out}}{2} = \frac{d_{in}−d_{out}}{2n}.
\end{align*} \tag{10} \]

In the unsigned network, the second eigenvector of the non-backtracking matrix is a community-correlated eigenvector. If the second eigenvalue of H̃ is separated from the bulk of the spectrum, then the eigenvector corresponding to the second eigenvalue can be used in the community detection (label vertices according to the sign of the sum of all incoming edges at each vertex) [29]. Similar conclusions in signed networks are verified in the following.

Now, we first attempt to construct a vector g which is correlated with the communities and is an approximate eigenvector with eigenvalue \( \mu_c \). We assume that \( c = O(1) \), so the graph is sparse and locally tree-like. For any positive integer \( r \), and any directed edge \((u,v)\), we define that,
\[ (g^{(r)})_{u→v} = \mu_c^{−r} \cdot \sum_{(w,x):d(u→v,w→x)=r} \sigma_x \cdot \sigma_{u→v}, \]
where \( \sigma_x = ±1 \) denotes \( x \)'s community, \( \sigma_{u→v} = ±1 \) denotes the sign of edge \((u,v)\), \( d(u→v,w→x) \) denotes the number of steps required to go from \( u → v \) to \( w → x \) in the graph of directed edges, as shown in Figure 2.

Fig. 2. An illustration for calculating \( d(u→v,w→x) \). Going from edge \( x → y_1 \) to edge \( z_1 → w \) needs to transverse two edges \( y_1 → z_1 \) and \( z_1 → w \). Thus, \( d(x→y_1,z_1→w) = 2 \).

Applying H to \( g^{(r)} \), we have,
\[ (Hg^{(r)})_{u→v} = \mu_c^{−r} \cdot \sum_{(w,x):d(u→v,w→x)=r+1} \sigma_x \cdot \sigma_{u→v}, \]
which can be simplified as
\[ (Hg^{(r)})_{u→v} = \mu_c \cdot g^{(r+1)}_{u→v}. \]

We may write \( g^{(r)}_{u→v} \) as
\[ \mu_c^{−r} \cdot \sigma_{u→v} \cdot \sum_{(w,x):d(u→v,w→x)=r} (\sigma_x − \mu_c^{−1} \sum_{y \in N(x) \setminus u} \sigma_y). \]

Now, there are (in expectation) \( c^r \) terms in this sum, each of which, conditioned on the \( \sigma_x \)'s, has an expected value of zero and a constant variance. Hence,
\[ E[(g^{(r)})_{u→v} − g^{(r+1)}_{u→v}]^2 = O(c^r \mu_c^{−2r}). \]

Summing over all the edges, we have,
\[ E[(g^{(r)}) − g^{(r+1)}]^2 = O(c^r \mu_c^{−2r} |E|). \]

Therefore, when the community-correlated eigenvalue (the second eigenvalue) satisfies
\[ \mu_c > \sqrt{c}. \tag{11} \]

When it is separated from the bulk spectrum, it can be naturally considered that the error is small and approaches zero for large \( r \).

And according to the conclusion in unsigned networks [30], [31], it can be inferred that, under the condition of the threshold and \( n \to \infty \), for every \( u → v \),
\[ < g^{(r)}_{u→v}, \sigma_u \cdot \sigma_{u→v} > \neq 0. \]

Thus, we can draw a conclusion that
\[ |Hg^{(r)} − \mu_c g^{(r)}| = o(1). \]

So \( g^{(r)} \) is indeed an approximate eigenvector for \( H \) with eigenvalue \( \mu_c \), which may be used to detect the community structure of the signed networks. And Ineq. (11) is the detection threshold finally deduced in this paper, which shows agreement with the threshold in unsigned networks.

B. Beyond two communities

Fig. 3. Community detection results when the number of communities is greater than 2. \( q = 3, N = 120, d_{in} = 0.6, d_{out} = 0.2, \rho_{in} = \rho_{out} = 0.6 \) (a) non-backtracking matrix, \( \Omega_{bl} = 1 \); (b) adjacency matrix, \( \Omega_{bl} = 0.85 \)

The above analysis is demonstrated on stochastic block models with two communities (\( q = 2 \)). In fact, according to the above derivation process of this article, the non-backtracking matrix is also well applied in the model with community number greater than 2 (\( q > 2 \)). Its detectable threshold should be similar to the conclusion in the unsigned network, that is,
the community-correlated eigenvalue satisfies Ineq. (11), i.e.,

\[ \mu_c > \sqrt{c}. \]

In this case, the second eigenvalue is,

\[ \mu_2 \approx \frac{c_{in} - c_{out}}{q} = \frac{d_{in} - d_{out}}{q} n. \]

In general, we can take the first \( k \) eigenvectors and use \( k \)-means algorithm to determine the grouping of nodes.

Given the case that \( q = 3, N = 120, d_{in} = 0.6, d_{out} = 0.2, p_{in} = p_{out} = 0.6 \), apply \( k \)-means algorithm to partition the network. In Fig. 3 different colors represent different groups, and non-backtracking matrix and adjacency matrix detection are used to obtain overlaps of 1 and 0.86 respectively.

It can be seen that the algorithm based on non-backtracking matrix still has a good application. However, it should be noted that when the number of communities is greater than two, the threshold value of detecting network structure by using adjacency matrix is not known yet, so we do not carry out a deeper comparative evaluation.

C. An improved non-backtracking operator for community detection in signed networks

According to the analysis above, we know that the proposed non-backtracking operator is density sensitive rather than sign sensitive. Even in the case that the negative edges of inter-community connections and the positive edges of intra-community connections account for the majority, the final results are not ideal as long as the density of connected edges within and between groups does not meet the threshold value we derived. The insensitivity to edge signs is essentially against our original intention to explore the community structure of the signed networks.

In fact, according to the threshold of the adjacency matrix operator mentioned below, we can know that this operator is only sign sensitive and also does not meet our expectations for community detection. The ideal tool should have good performance in both aspects.

In the construction of non-backtracking matrix, we have considered the structure balance theory and the belief propagation theory. Why does this sign insensitive occur? We can see that in Sec. IIIA, the approximation vector of the second eigenvector is actually only related to the \( k \)-order neighbors of a node \( x \), which is just a simple sum of its belonging communities, without considering whether the path between two nodes is balanced, that is, whether the internal relationship is friendly or hostile.

Hence, we improve the operator under the following assumption. We assume that there should be expecting a larger number of balanced or stable paths of a given length \( k \) between two vertices \( u \) and \( v \) if they belong to the same community. So the improved matrix, denoted by \( H^b \), is defined as follows,

\[ H^b_{e,f} = 1(e_2 = f_1)1(e_1 \neq f_2)1(A_e \cdot A_f = 1). \]

Or we can write it as

\[ H^b_{(e_1 \rightarrow e_2),(f_1 \rightarrow f_2)} = \begin{cases} 1 & \text{if } e_2 = f_1, e_1 \neq f_2 \text{ and } \\
\text{sign}(e_1 \rightarrow e_2) = \text{sign}(f_1 \rightarrow f_2), & \text{otherwise}. \end{cases} \]

So corresponding to the BP algorithm, during the propagation process, we only pass the real information to the edges with same signs, not the wrong message to the edges with different signs. For simplicity, we call this matrix as balanced non-backtracking matrix. Actually, \( H^b \) is an approximation of \( H \), and the detection threshold for community detection is still unclear. But through the experiment, we can see that it does show an ideal role in community detection.

In the following section, we will compare the performance of these three matrices in community detection.

IV. RESULTS

To evaluate the performance of the signed non-backtracking matrix \( H \) and balanced non-backtracking matrix \( H^b \) in community detection, we carry out extensive simulations on signed networks. The accuracy of community detection is quantified by the concept of overlap [21] which is defined as the proportion of correctly predicted nodes to all nodes. The overlap can be expressed as

\[ ovl = \frac{1}{N} \sum u \delta_{g_u, \tilde{g}_u}, \]

where \( g_u \) is the true group label of vertex \( u \), and \( \tilde{g}_u \) is the label found by the algorithm. When \( g_u = \tilde{g}_u \) for every node \( u \), we have \( ovl = 1 \) and the detection accuracy achieves 100%.

We break symmetry by maximizing overall \( q! \) permutations of the groups, where the nodes are divided into \( q \) groups. The prediction is totally exact when overlap equals to 1 and under this definition, the minimum value of overlap can be taken as \( 1/q \). The overlap is normalized as

\[ ovl = \frac{1}{N} \sum u \delta_{g_u, \tilde{g}_u} - \frac{1}{q} \right) / (1 - \frac{1}{q}). \]

The overlap is ranging from 0 to 1. Here, 0 means that the prediction is inaccurate due to random grouping. For the sake of visualization, we still use the unnormalized overlap in the following numerical simulation.

A. Analysis of detection accuracy

According to the results of numerical simulation, the following conclusions are obtained.

First of all, it is feasible and accurate to use signed non-backtracking matrix \( H \) to detect communities in signed networks if the bulk of spectrum is within \( \sqrt{c} \) and the second real eigenvalue exceeds \( \sqrt{c} \), the overlap obtained by using the
corresponding eigenvector is close to (sometimes equal to) 1. Fig. 4(a) shows a typical example of accurate detection.

Second, as for balanced non-backtracking matrix $H^b$, although the intrinsic principle has not been studied yet, we can still see that the bulk of its eigenvalues are within a certain radius range. As shown in Fig. 4(b), unlike before, there are three real eigenvalues outside the circle. It is gratifying that BNBT also leads to satisfactory results in community detection. We find that the calculated overlap is close to 1, which means that the optimized BNBT algorithm can indeed be used to detect community structure. In the following sections, we will provide a detailed description and analysis of its performance.

Besides, the threshold of community detection using non-backtracking matrix is closely related to $d_{in}$ and $d_{out}$, but not to $p^+_in$ and $p^-_{out}$. Meanwhile, it is noted that the community structure can not be detected near the threshold on certain conditions. It can be seen from the Section III that, since $r$ does not really tend to infinity (so the error will not be infinitely small), $g^{(r)}$ can not be simply regarded as the approximation of the second eigenvector. Thus is only marginally effective on the detection of community structure. For instance, in the case that $N = 100$, $p^+_in = p^-_{out} = 0.6$, $d_{out} = 0.2$ and $d_{in} = 0.3$ (or 0.35), all the cases meet the threshold conditions in theory, but there is only one real eigenvalue out of the bulk of the spectrum and the overlap equals to 0.5350 (or 0.5050).

B. Comparison with adjacency-matrix-based detection

The leading eigenvector of adjacency matrix has shown the community structure, and the number of eigenvalues beyond the threshold is no longer expressed as the number of communities (should be the number of communities minus 1).

When there are only two communities, as long as the following conditions are met, the sign of the main eigenvector can be used to detect communities,

$$p_{out} > \frac{1}{2} - \frac{d_{in}}{d_{out}} \left( p^+_in - \frac{1}{2} \right) + \frac{1}{d_{out}} \sqrt{d_{in} + d_{out} - 8d^2_{in} \left( p^+_in - \frac{1}{2} \right)^2}$$  \hspace{1cm} (14)

Considering the average degree $c$ of network, we use the signed non-backtracking matrix as long as the following inequality is satisfied,

$$\left\{ \begin{array}{l} \frac{1}{2} < p^+_in < \frac{1}{2} + \frac{1}{2} \sqrt{\frac{c}{c^2 + N d^2_{in}}} \\ d_{in} > \frac{c + \sqrt{c}}{N} \end{array} \right.$$  \hspace{1cm} (15)

When applying NBT matrix, we can get better results in some cases than using the adjacency matrix. Here, we give a simple but general example to compare the two methods (see Fig. 5). In all the comparisons, the paper takes $p^+_in = p^-_{out}$ for convenience. We only consider the case $p^+_in > 0.5$ here, because when $p^+_in < 0.5$, the community structure can be represented by $u_N$ (the last eigenvector of adjacency matrix).

However, considering the dynamic evolution process of the network, only the driving role played by the leading eigenvector of the initial network in the structural balance evolution gives rise to a dynamical manifestation of the detectability transition[22].

1. When $(d_{in}, d_{out}, p^+_in)$ belongs to region 3 and region 4, in other words, it does not meet the threshold value Ineq. (11), and the non-backtracking matrix method fails, so the adjacency matrix method should be considered.

2. When $(d_{in}, d_{out}, p^+_in)$ belongs to region 1 and region 2, compared with the method based on adjacency matrix, the method based on non-backtracking matrix has less correlation with $p^+_in$ and $p^-_{out}$.

From the aspect of computation, adjacent matrix needs to calculate the first eigenvalue and eigenvector of a $n \times n$ matrix, and non-backtracking matrix needs to calculate the second eigenvalue and eigenvector of a $2n \times 2n$ matrix (in the actual numerical simulation process, the program will calculate multiple eigenvalues for comparison).

(i) When $(d_{in}, d_{out}, p^+_in)$ is in region 1, i.e. meeting the threshold value of the algorithm based on the adjacency matrix in Ineq. (14), the adjacency matrix method should be used for community detection considering the computational complexity.

(ii) When $(d_{in}, d_{out}, p^+_in)$ is in region 2, the non-backtracking matrix can be considered. That is to say, our algorithm has a better indication for clustering when the algorithm based on adjacency matrix does not meet the threshold value.

In fact, Fig. 5 is an abstract representation of the community related parameters in the real network. Region 4 represents the situation that $d_{in} < d_{out}$, $p^+_in \approx p^-_{in}$ and $p^-_{out} \approx p^-_{out}$. In
other words, the random blocks generated by the parameters in region 4 are not the two communities mentioned above. For the case that the parameters meet region 2 (or region 3), we actually hope that the algorithm is not sign sensitive (or density sensitive), and then the balanced non-backtracking operator works.

It is worth mentioning that since the theoretical detection threshold of the balanced non-backtracking matrix is still unknown, we will not discuss it furtherly.

C. Comprehensive comparison among three operators in community detection

![Figure 6: The performance in community detection when applying different matrices. N = 10000, c = 10, d_{in} + d_{out} = 1, and p_{in}^+ = p_{out}. d_{in} and p_{in}^+ vary from 0.5 to 1 with an interval of 0.05 respectively.](image)

In the following, we give examples to show the better performance of the improved algorithm. The results are shown in Fig. 6. The x-axis represents the $p_{in}^+$, and the y-axis represents the $d_{in}$. The z-axis as well as the color represents the overlap. The experiments are performed on signed networks with size of $10^4$ and average degree of 10. We set $p_{in}^+ = p_{out}$ for convenience, which varies from 0.5 to 1 with an interval of 0.05. It is worth noting that $d_{in}$s in the figure are normalized, that is $d_{in} + d_{out} = 1$, which also varies from 0.5 to 1 with an interval of 0.05. As we can see, the signed non-backtracking matrix’s performance is only sensitive to the link density of inter- and intra-community connections, while the adjacency-based algorithm is more sensitive to the link signs. The balanced non-backtracking matrix-based approach takes both advantage of the above matrices, and its performance depends on the link density and link signs. Moreover, the area of undetectable filed (in which the overlap is about 0.5) is smallest, which indicates the best performance of this operator.

In Fig. 7, we show four concrete examples by taking $d_{in} \in \{0.5, 0.70, 0.75, 0.80\}$ and varying $p_{in}^+$ from 0.5 to 1. Actually, each line in this figure is a slice of Fig. 6 by taking $d_{in}$ as the corresponding values. The black line indicates the results detected using the adjacency matrix, the red line means the results detected using the non-backtracking matrix, and the blue line represents the results detected using the balanced non-backtracking matrix. For each case, we conduct 10 experiments and calculate the average to connect them with the black/red/blue broken lines. Except for the first graph, which is the case of $d_{in} = 0.5$, we can draw the following conclusion. The overlap is close to 0.5 (which means the nodes are labelled almost randomly) when the detection threshold based on adjacency matrix is not satisfied, while the method based on non-backtracking matrix still has a good performance. As is discussed above, the overlap does not change with $p_{in}^+$ increases. As $p_{in}^+$ increases, the adjacency matrix-based algorithm outperforms the signed non-backtracking matrix-based one. However, the balanced non-backtracking-based method always performs better than the adjacency matrix-based approach. All the above analysis can lead us to a satisfactory conclusion. The first graph is the only exception. In this case, the density is useless to divide the community structure ($d_{in} = d_{out} = 0.5$), so the signed non-backtracking matrix-based algorithm is invalid, and the performance of the balanced non-backtracking-based method is weaker than that of sign sensitive algorithm.

In Fig. 8, we discuss furtherly from the other perspective, that is we show three concrete examples by taking $p_{in}^+ \in \{0.50, 0.60, 0.70\}$ and varying $d_{out}$ from 0.5 to 1. These are actually slices from another angle of Fig. 6. The performance of adjacency matrix-based approach is not completely independent with $d_{out}$. When $p_{in}^+ = 0.7$, the overlap increases when $d_{out}$ is small. However, after the threshold is satisfied,
the performance do not increase as $d_{out}$ decreases. It is further proved that the original non-backtracking operator $H$ is sign insensitive. As before, in the case of $p_{in}^+ = 0.5$, density sensitivity is the only factor we need to consider, so $H$ is superior, while in other cases, $H^b$ shows its advantages.

In addition, we know that the community structure may be difficult to detect when the graph is sparse. For example, in unsigned networks, when $c$ is constant and $n$ is large, the network is decomposed for many reasons. Most importantly, the leading eigenvalues of $A$ are indicated by the vertices of the highest degree, and the corresponding eigenvectors are localized around these vertices [21], [32]. At the same time, the non-backtracking matrix has better performance in sparse case. We expect to get the same conclusion in the signed network.

If the right side of the first inequality of (15) is regarded as a function of $d_{in}$, we get a lower bound according to the monotonicity of the function as below,

$$
\begin{align*}
\frac{1}{2} < & \frac{1}{2} + \frac{1}{2} \sqrt{\frac{1}{2c}} = d_{in} < \frac{c + \sqrt{c}}{N} .
\end{align*}
$$

Therefore we prove theoretically that when the non-backtracking matrix detection is feasible, the smaller $c$ is, the better the result based on the non-backtracking matrices is compared with the adjacent matrix. Note that what we say is better performance in the case of sparse is relative. In fact, as the network becomes sparse, our detection accuracy will certainly decrease correspondingly. The result is also confirmed by numerical simulations performed on stochastic block models with $N = 10^4$, $p_{in}^+ = 0.7$, $d_{in} = 0.75$, and the average degree $c$ varies from 10 to 60 with an interval of 5. As is shown in Fig. 9 the performance of the non-backtracking matrix-based algorithm is more stable and robust compared with the adjacency matrix regarding to the average degree.

Based on the above comparison, we are also interested in the newly proposed $H^b$, but these analyses are only based on numerical simulation, and the underlying theory remains to be studied. However, we can draw a conclusion that $H^b$ is sensitive to both the sign of the edge and the connection density. In most cases, it has good performance. On the contrary, $A$ and $H$ have their own limitations. Although they may be the best choice in some very special situations, generally speaking, the balanced non-backtracking-based algorithm is the most reliable and effective method.

V. CONCLUSIONS AND DISCUSSION

This paper investigates an efficient community detection in signed network by demonstrating the feasibility to define a non-backtracking matrix for signed networks. We provide the definition of a proper non-backtracking matrix from perspectives of both structural balance theory and belief propagation. Based on the proposed non-backtracking matrix, we analytically determine the community detectability and propose the most efficient operator, the balanced non-backtracking-based operator $H^b$, which significantly outperforms the adjacency matrix based detection algorithms.

It is worth noting that we certainly hope that the algorithm is sensitive and adaptable to both sign and density. Our algorithm (as well as the previous algorithms) is absolutely effective under the most standard community partition, but for communities with different characteristics and for more complex networks, we’d better have a simple understanding of it before selecting the appropriate algorithm. Considering this, The balanced non-backtracking matrix is most universal. The exception to this is that when the community in the actual network is not relationship dependent community or density dependent community, the above algorithms may not achieve satisfactory results.
The proposed framework shows great potential to detect communities with or without overlap and paves the way to understand the collective behaviors of systems where positive and negative relationships coexist.

APPENDIX
THE DERIVATION PROCESS OF THE UPDATING RULES OF $\delta$

In order to simplify the updating equation of $\delta$, let's consider Eq. (5) carefully.

First of all, we notice that $n_{BP}^{+} = n_{BP}^{-}$ holds near the trivial fixed point, that is, $e^{-h}$ can be written as 1. Suppose an unknown constant $Z$, we split the Eq. (5) into two equations,

$$\eta_{u \rightarrow v}^{+} = Z \times \prod_{u \in N(v)} \left( \eta_{u \rightarrow v}^{+}, \eta_{u \rightarrow v}^{-} \right) \left[ (1 - \eta_{u \rightarrow v}^{+}) c_{in} + (1 - \eta_{u \rightarrow v}^{-}) c_{out} \right],$$

for $\eta_{u \rightarrow v}^{+} = 1/2 \pm \delta_{u \rightarrow v}$,

$$\eta_{u \rightarrow v}^{-} = Z \times \prod_{u \in N(v)} \left( \eta_{u \rightarrow v}^{+}, \eta_{u \rightarrow v}^{-} \right) \left[ (1 - \eta_{u \rightarrow v}^{+}) c_{out} + (1 - \eta_{u \rightarrow v}^{-}) c_{in} \right],$$

for $\eta_{u \rightarrow v}^{-} = 1/2 \pm \delta_{u \rightarrow v}$,

Rewriting Eq. (17) near the trivial fixed point $\eta_{u \rightarrow v}^{+} = 1/2 \pm \delta_{u \rightarrow v},$

$$\frac{1}{2} + \delta_{u \rightarrow v} = Z \times \prod_{u \in N(v)} \left( \eta_{u \rightarrow v}^{+}, \eta_{u \rightarrow v}^{-} \right) \left[ \left( \frac{1}{2} + \delta_{u \rightarrow v} \right) c_{in} + \left( \frac{1}{2} - \delta_{u \rightarrow v} \right) c_{out} \right],$$

$$\frac{1}{2} + \delta_{u \rightarrow v} = Z \times \prod_{u \in N(v)} \left( \eta_{u \rightarrow v}^{+}, \eta_{u \rightarrow v}^{-} \right) \left[ \left( \frac{1}{2} - \delta_{u \rightarrow v} \right) c_{in} + \left( \frac{1}{2} + \delta_{u \rightarrow v} \right) c_{out} \right].$$

By merging the similar items, we get,

$$\frac{1}{2} + \delta_{u \rightarrow v} = Z \times \prod_{u \in N(v)} \left( \eta_{u \rightarrow v}^{+}, \eta_{u \rightarrow v}^{-} \right) \left[ \frac{1}{2} (c_{in} + c_{out}) + \delta_{u \rightarrow v} (c_{in} - c_{out}) \right],$$

$$\times \prod_{u \in N(v)} \left( \eta_{u \rightarrow v}^{+}, \eta_{u \rightarrow v}^{-} \right) \left[ \frac{1}{2} (c_{in} + c_{out}) - \delta_{u \rightarrow v} (c_{in} - c_{out}) \right].$$

And Eq. (18) is simplified to

$$\frac{1}{2} - \delta_{u \rightarrow v} = Z \times \prod_{u \in N(v)} \left( \eta_{u \rightarrow v}^{+}, \eta_{u \rightarrow v}^{-} \right) \left[ \frac{1}{2} (c_{in} + c_{out}) - \delta_{u \rightarrow v} (c_{in} - c_{out}) \right] \times \prod_{u \in N(v)} \left( \eta_{u \rightarrow v}^{+}, \eta_{u \rightarrow v}^{-} \right) \left[ \frac{1}{2} (c_{in} + c_{out}) + \delta_{u \rightarrow v} (c_{in} - c_{out}) \right].$$

Linearizing Eq. (20) and Eq. (21), it follows that

$$\frac{1}{2} \pm \delta_{u \rightarrow v} \approx Z \times \left\{ \left( \frac{1}{2} (c_{in} + c_{out}) \right)^{|N(v)|} \right\},$$

In order to eliminate the constant $Z$, we calculate Eq. (22) plus Eq. (23) and Eq. (22) minus Eq. (23) respectively,

$$1 = 2Z \left\{ \left( \frac{1}{2} (c_{in} + c_{out}) \right)^{|N(v)|} \right\},$$

$$2\delta_{u \rightarrow v} = 2Z \left\{ \left( \frac{1}{2} (c_{in} + c_{out}) \right)^{|N(v)|} \right\} \times (c_{in} - c_{out})$$

$$\times \left\{ \sum_{u \in N(v)} \frac{c_{in} - c_{out}}{c_{in} + c_{out}} \right\} \delta_{u \rightarrow v}.$$

After eliminating the constant $Z$, we have

$$\delta_{u \rightarrow v} = \frac{c_{in} - c_{out}}{c_{in} + c_{out}} \times \left\{ \sum_{u \in N(v)} \frac{c_{in} - c_{out}}{c_{in} + c_{out}} \right\} \delta_{u \rightarrow v},$$

Thus, we get the updating rule of $\delta$ in signed networks,

$$\delta = \frac{(c_{in} - c_{out})}{(c_{in} + c_{out})} H^T \delta.$$
ACKNOWLEDGMENT

Z-YZ, C-QQ, and G-HW were supported in part by National Natural Science Foundation of China under Grant 11631014, Grant 11871311, in part by China Postdoctoral Science Foundation under Grant 2019TQ0188, Grant 2019M662315, in part by Shandong University multidisciplinary research and innovation team of young scholars under Grant 2020QNQT017. X-RW acknowledges the partial support of the project "PCL Future Greater-Bay Area Network Facilities for Large-scale Experiments and Applications (LZC0019)".

REFERENCES

[1] S. Fortunato, “Community detection in graphs,” Physics reports, vol. 486, no. 3-5, pp. 75–174, 2010.
[2] P. F. Jonsson, T. Cavanna, D. Zicha, and P. A. Bates, “Cluster analysis of networks generated through homology: automatic identification of important protein communities involved in cancer metastasis,” BMC bioinformatics, vol. 7, no. 1, p. 2, 2006.
[3] P. Anchur and M. Magdon-Ismail, “Communities and balance in signed networks: A spectral approach,” in 2012 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining. IEEE, 2012, pp. 235–242.
[4] F. Heider, “Attitudes and cognitive organization,” The Journal of psychology, vol. 21, no. 1, pp. 107–112, 1946.
[5] G. Facchetti, G. Iacono, and C. Altafini, “Computing global structural balance in large-scale signed social networks,” Proceedings of the National Academy of Sciences, vol. 108, no. 52, pp. 20953–20958, 2011.
[6] S. A. Marvel, J. Kleinberg, R. D. Kleinberg, and S. H. Strogatz, “Continuous-time model of structural balance,” Proceedings of the National Academy of Sciences, vol. 108, no. 5, pp. 1771–1776, 2011.
[7] A. Kirkley, G. T. Cantwell, and M. Newman, “Balance in signed networks,” Physical Review E, vol. 99, no. 1, p. 012320, 2019.
[8] N. Monika, G. Lavenya, K. Vaishnavi, M. Kavya, and V. Kaniya, “Structural balance theory based recommendation,” International Journal of Advanced Research in Computer Science, vol. 9, no. Special Issue 3, p. 45, 2018.
[9] C. Qu and H. Wang, “Impact of structural balance on self-avoiding pruning walk,” Physica A: Statistical Mechanics and its Applications, vol. 524, pp. 362–374, 2019.
[10] J. Leskovec, D. Huttenlocher, and J. Kleinberg, “Signed networks in social media,” in Proceedings of the SIGCHI conference on human factors in computing systems, 2010, pp. 1361–1370.
[11] B. W. Kernighan and S. Lin, “An efficient heuristic procedure for partitioning graphs,” The Bell system technical journal, vol. 49, no. 2, pp. 291–307, 1970.
[12] J. Friedman, T. Hastie, and R. Tibshirani, The elements of statistical learning. Springer series in statistics New York, 2001.
[13] J. MacQueen et al., “Some methods for classification and analysis of multivariate observations,” in Proceedings of the fifth Berkeley symposium on mathematical statistics and probability. Oakland, CA, USA, 1967, pp. 281–297.
[14] U. Von Luxburg, “A tutorial on spectral clustering,” Statistics and computing, vol. 17, no. 4, pp. 395–416, 2007.
[15] L. Yang, X. Cao, D. He, C. Wang, X. Wang, and W. Zhang, “Modularity based community detection with deep learning,” in IJCAI, vol. 16, 2016, pp. 2252–2258.
[16] M. Girvan and M. E. Newman, “Community structure in social and biological networks,” Proceedings of the national academy of sciences, vol. 99, no. 12, pp. 7821–7826, 2002.
[17] M. E. Newman and M. Girvan, “Finding and evaluating community structure in networks,” Physical review E, vol. 69, no. 2, p. 026113, 2004.
[18] D. J. MacKay and D. J. Mac Kay, Information theory, inference and learning algorithms. Cambridge university press, 2003.
[19] S. Chauhan, M. Girvan, and E. Ott, “Spectral properties of networks with community structure,” Physical Review E, vol. 80, no. 5, p. 056114, 2009.
[20] A. Pothen, “Graph partitioning algorithms with applications to scientific computing,” in Parallel Numerical Algorithms. Springer, 1997, pp. 323–368.

Zhaoyue Zhong received the B.S. degree from School of Mathematics, Shandong University, China in 2020. She is currently pursuing the M.S. degree in School of Mathematical Sciences, Fudan University, China. Her current research interests include complex networks and mathematical methods in neural networks.

Xiangrong Wang is currently a research assistant professor at Southern University of Science and Technology, Shenzhen, China. She received her Ph.D. degree from the Delft University of Technology, the Netherlands in 2016. Before joining Shenzhen, she was a postdoctoral researcher at the Delft University of Technology from 2017 to 2018. In 2017 and 2018, she was a visiting scholar at Zaragoza University, Zaragoza, Spain and ISI Foundation, Torino, Italy. Her research focuses on modeling and analysis of complex networks, nonlinear dynamics and graph spectral analysis.
Cunquan Qu is currently working as a Post-doctor researcher at Data Science Institute, Shandong University, Jinan, China. He did a joint Ph.D. project in the Multimedia Computing Group at the Delft University of Technology for two years. He received his Ph.D. degree from Shandong University in 2019. His research focuses on analyzing network structure, modeling the dynamics process, graph neural networks.

Guanghui Wang received a B.Sc. degree in Mathematics from Shandong University, Jinan, China, in 2001. He got the doctors degree from Paris SDU University and worked in Ecole Centrale Paris as a Post-doctor. Now he is a professor in School of Mathematics, Shandong University. His current interests include graph theory, combinatorics, complex networks and bioinformatics.