Social Network Analysis and Mining (2021) 11:7
https://doi.org/10.1007/s13278-020-00711-1

ORIGINAL ARTICLE

A Regularized Convex Nonnegative Matrix Factorization Model for signed network analysis

Jia Wang[1] · Rongjian Mu[2]

Received: 2 January 2020 / Revised: 23 November 2020 / Accepted: 28 November 2020 / Published online: 2 January 2021

Abstract

Community detection and link prediction are two basic tasks of complex network system analysis, which are widely used in the detection of telecom fraud organizations and recommendation systems in the real world. In ordinary unsigned networks, these two analyses have been developed for a long time. However, due to the existence of negative edges, the study of community detection and link prediction in signed networks is still limited now. Most existing methods have high computational complexity and ignore the generation of the networks based on heuristics. In this paper, we propose a regularized convex nonnegative matrix factorization model (RC-NMF) from the perspective of the generative model to detection communities in the signed network. This algorithm introduces graph regularization to constrain nodes with negative edges into different communities and nodes with positive edges into the same communities as much as possible. Experiments on synthetic signed networks and several real-world signed networks validate the effectiveness and accuracy of the proposed approach both in community detection and link prediction.

Keywords Signed network · Community detection · Link prediction · Nonnegative matrix factorization (NMF)

1 Introduction

Social networks in the real world can be modeled through a complex network (Ghoshal et al. 2014; Rossi and Ahmed 2019; Vasudevan and Deo 2012). In ordinary complex networks, the links between nodes represent a certain connection between individuals in social networks. Then, in the signed network, the links between nodes imply the positive and negative attribute relations among individuals in the social network. If there is a positive connection between nodes, it is shown as positive links, and vice versa. And in the signed network, there is a certain community structure; that is, the nodes within the community are mostly with positive links, and the nodes among different communities are mostly with negative links (Davis 1967). The above theory shows that in social network, individuals with a certain cooperative relationship are in the same cluster, and individuals in different clusters have a certain competitive relationship. Moreover, if the nodes in the signed network are positively or negatively connected with the other two nodes, then the two nodes are more inclined to produce positive links. Generally, in a signed network constructed by an online social network, positive links indicate "support", "like" or "cooperation", while negative links indicate "opposing", "dislike" or "competition". For example, users on the Slashdot website (Leskovec et al. 2010) can mark other users as friends or enemies based on other users' comments, and in consumer review site Epinions.com (Leskovec et al. 2010) which is a who-trusted-whom online social network, members of the site can decide whether to "trust" each other. Compared with the common complex network, the signed network with positive and negative attributes can contain more information when it represents the social network, so the analysis of the signed network has attracted more and more attention in the field of social network analysis.

Community detection and link prediction are two basic issues in signed network analysis. In a signed network, community detection is to find community structures that are represented as dense positive links in the same community and intensive negative links in different communities (Yang et al. 2007), and link prediction is to predict the

✉ Jia Wang
wangjia0313@126.com

[1] Department of Information Technology, Shanxi Professional College of Finance, Taiyuan, China

[2] Tianjin International Engineering Institute, Tianjin University, Tianjin, China

states of unknown links like positive or negative (Li et al. 2018a). Community detection can detect the community structure in social networks, which is helpful to analyze the grouping of individuals in social networks. Link prediction can predict the connection status of individuals in the social network in the next stage and can be used in the recommendation system. Therefore, the community detection algorithm and link prediction algorithm of the signed network are very helpful for the analysis of the social network.

Although some algorithms for community detection and link prediction in signed networks have been proposed in recent years, their development is still immature and not proven or still being developed. For example, some algorithms (Li et al. 2014; Anchuri and Magdon-Ismail 2012) based on optimization objective functions and heuristic-based algorithms (Yang et al. 2017; Zhao et al. 2017) have high computational complexity. Some model-based algorithms (Yang et al. 2007; Jiang 2015) have low accuracy in performance or need probabilistic statistical inference methods to select models, such as EM algorithm, resulting in a large computational burden. Some algorithms are based on deep learning (Wang et al. 2017, 2018) with high computational performance but poor interpretability. And most of the above algorithms can only be used for community detection or link prediction. Faced with these challenges, we propose a new RC-NMF model for community detection and link prediction.

In this paper, we introduce a graph regularization based on the convex nonnegative matrix factorization (Convex-NMF) algorithm. Convex-NMF is an improvement of the semi-NMF algorithm, which constrains the base matrix in the semi-NMF by adding a weight matrix. The introduced graph regularization can simultaneously constrain the nodes with positive links to enter the same community and the nodes with negative links to enter different communities. In addition to being used for community detection, our proposed RC-NMF can also be used for link prediction. We have compared experiments with other current advanced algorithms on artificially generated signed network datasets and real large-scale signed networks and proved the validity and accuracy of our proposed RC-NMF algorithms.

The structure of the paper is as follows: In Sect. 2, we introduce the related work on signed network community detection and link prediction algorithms in recent years. In Sect. 3, we introduce the RC-NMF algorithm we proposed in detail. Section 4 shows the comparison of other state-of-the-art algorithms on artificial signed network datasets and large-scale real-signed network datasets, which verifies the validity and accuracy of our proposed algorithm. Section 5 summarizes our contributions.

2 Related work

In recent years, a large number of algorithms have emerged for signed network community detection and link prediction. These algorithms can be broadly divided into the following categories: modularity optimization-based, balance theory-based, model-based and deep learning-based.

Modularity optimization-based methods The standard modularity is developed for unsigned networks, and it measures how far the real positive connections deviate from the expected random connections. And standard modularity optimization is essentially a discrete combination problem (Newman 2016). The communities in the network can be detected by optimizing the modularity objective function (Newman 2016). But this standard modularity optimization method was initially only applicable to unsigned networks. Li et al. (2014) defined signed modularity by improving standard modularity in the unsigned network and made it capable of handling negative links. Signed modularity balances the trend of users with positive links to forming community and the trend of users with negative links to destroying community by adding weights on positive and negative components in signed networks. Based on the above, some heuristics algorithms based on signed modularity optimization have been proposed. For example, Anchuri and Magdon-Ismail (2012) generalized spectral partitioning (SpePart) approach with iterative optimization to explore the community in the signed network, which is an extension about standard modularity optimization in the unsigned network.

Balance theory-based methods In the 1940s, Heider (1946) introduced the balance theory that the two positively related individuals had the same attitude toward the third person by studying perceptions and attitudes of individuals, which generally implies that "the friend of my friend is my friend" and "the enemy of my enemy is my friend." In the 1950s, Cartwright and Harary (1956) further developed the theory in the graph theoretical at the group level and validated the Harary and Frank (1953) that a signed graph is balanced if and only if nodes can be divided into two mutually exclusive clusters such that intra-links are positive and inter-links are negative. Therefore, the theory was developed in more than two clusters (Kulakowski et al. 2019), which introduced that a weakly balanced graph exists a partition of the nodes into k clusters just as nodes with positive links are in the same cluster and nodes with negative links are between different clusters. Based on this theory, we can find the community structure of the signed social network by cutting off all negative links. However, the signed social networks in the real world have been normally unbalanced since the existence of frustration that presents itself as the positive

inter-links and the negative intra-links. To address this challenge, many algorithms for signed network analysis based on structural balance theory are proposed. For example, Chiang et al. (2014) extended the applicability of the balance theory from the local features to the global features in the signed network. Amelio and Pizzuti (2016) developed a correlation clustering method (CC) that maximizes positive links in a community and negative links between communities or minimizes frustration to detect community in signed networks. Li et al. (2018a) presented a novel framework including two implicit features and two latent features for predicting link, one of which is obtained by balance theory. Derr et al. (2020) used the theory of structural balance among individuals to predict link and interaction polarity in signed networks.

Model-based methods This type of methods focuses on modeling the generated mechanism which tends to apply to the network. For example, Yang et al. (2007) proposed an agent-based random walk model framework (FEC), which is the two-stage approach. First, FC (finding community) is conducted on the positive component of network based on a random walk model, and then, EC (extracting community) is conducted by minimizing predefined signed cut criteria according to the links of nodes obtained in the first stage (Yang et al. 2007). The algorithm is capable of giving nearly optimal solutions in linear time concerning the size of a network, but its performance is poor. Chen et al. (2014) proposed a novel approach named signed probabilistic mixture (SPM) model for overlapping community detection. Some of the above methods are based on optimization objectives or heuristic to detect a community structure in the signed network and do not care about the generation of the network. Jiang (2015) proposed a generalized stochastic block model that is the signed stochastic block model (SSBM) to explore the mesoscopic structures in signed networks from a node perspective where each node is assigned to a block or community and links are independently generated for pairs of nodes. Yang et al. (2017) adopted the variational Bayes EM algorithm to estimate the parameters and select model by approximate Bayesian model evidence based on the signed stochastic blockmodel (SSBM) that was proposed to characterize and generate the block structures of signed networks by explicitly formulating the link density and sign based on a stochastic perspective. Zhao et al. (2017) presented a statistical inference method in signed networks (SISN) for community detection, in which a probabilistic model is presented to model signed networks and an expectation–maximization (EM)-based parameter estimation method is deduced to find communities in signed networks. Li et al. (2018b) proposed a regularized semi-nonnegative matrix tri-factorization (Res-NMTF), which splits the matrix in the traditional semi-nonnegative matrix factorization into three terms and adds regularization on this basis to constrain nodes with negative links into different communities. However, this algorithm does not consider the nodes with positive links, so the accuracy is relatively low.

Deep learning-based methods With the rise of deep learning, some algorithms based on machine learning to discover community structure and link prediction emerged. Wang et al. (2017) proposed a novel framework SNEA (social network embedding with attributes), which exploits the network structure and user attributes simultaneously for network representation learning. Although the performance of deep learning is better than some traditional algorithms, the interpretation of these models is weak. Wang et al. (2018) proposed a novel and flexible end-to-end signed heterogeneous information network embedding (SHINE) framework to predict the sign of unobserved links. The SHINE framework gets the implicit low-dimensional vectors of nodes in the network through deep autoencoders and then does the similarity analysis of the nodes on this basis.

3 Our work

3.1 Convex Nonnegative Matrix Factorization (NMF)

Convex nonnegative matrix factorization (Convex-NMF) (Jordan 2009) is the improvement of semi-nonnegative matrix factorization (semi-NMF). Semi-NMF is one of the most popular methods for community detection in signed network, formulated as the following model:

$$\begin{aligned} &L = \left\| A^{\pm} - F^{\pm} H^{T} \right\|_{F}^{2} \\ &s.t \quad F \in R_{\pm}^{N \times C}, \quad H \in R_{+}^{N \times C} \end{aligned} \tag{1}$$

where A is the adjacency matrix of the signed network G with N nodes, F is the basis matrix, H is the community indicators matrix where element h_{jc} is the propensity of node j in community c, and C is the community amount.

Convex nonnegative matrix factorization (Convex-NMF) constrains the basis matrix F in semi-NMF based on the definition that F lies within the column space of adjacency matrix A, formulated as the following model:

$$\begin{aligned} &L = \left\| A^{\pm} - A^{\pm} W^{+} H^{T} \right\|_{F}^{2} \\ &s.t \quad F = AW, \quad W \in R_{+}^{N \times C}, \quad H \in R_{+}^{N \times C} \end{aligned} \tag{2}$$

where W is the weight matrix.

3.2 Regularized Convex-NMF model(RC-NMF)

Because of the influence of negative links in signed networks, we introduce graph regularization (Zheng and Skillicorn 2015) to minimize the positive ratio cut and negative ratio association

simultaneously. The regularization that constrains the nodes connected with negative links to be distributed into different communities and the nodes connected with positive links which are in the same communities simultaneously is defined as follows:

$$\min \sum_{j=1}^{k} \frac{h_j^{\top}(D^p - A)h_j}{h_j^{\top}h_j} \tag{3}$$

where A is the adjacency matrix of the signed network, D^p is a diagonal matrix with $D^p_{ii} = \sum_i A^p_{ij}$, and h_j is jth vector of the community indicator matrix H.

In addition, we can observe that $\mathrm{tr}(H^{\top}(D^p - A)H) = \sum_{j=1}^{k} \frac{h_j^{\top}(D^p - A)h_j}{h_j^{\top}h_j}$, and then, the regularization term optimization problem can be obtained by addressing the following optimization problem:

$$\min_{H \in R^{n \times k}} \mathrm{tr}\left(H^T(D^p - A)H\right) \text{ s.t. } H_{ij} \geq 0, \forall i, j \tag{4}$$

Furthermore, to make the node try its best to belong to only one community, we use $\|H\|_1^2$ to control the sparsity of the node probability matrix.

Combining the regularization term into Convex-NMF, the final RC-NMF model objective function is:

$$L = \left\|A - AWH^T\right\|_F^2 + \lambda \mathrm{tr}(H^{\top}(D^p - A)H) + \gamma \|H\|_1^2 \tag{5}$$

3.3 Update rules for RC-NMF model

We design multiplicative update rules to solve (5). The object function can be written as:

$$\begin{aligned} L &= \left\|A - AWH^T\right\|_F^2 + \lambda \mathrm{tr}\left(H^T(D^p - A)H\right) + \gamma \|H\|_1^2 \\ &= \mathrm{tr}\left(\left(A - AWH^T\right)\left(A - AWH^T\right)^T\right) \\ &\quad + \lambda \mathrm{tr}\left(H^T(D^p - A)H\right) + \gamma \mathrm{tr}\left(HNH^T\right) \\ &= \mathrm{tr}\left(A^TA - 2H^TA^TAW + W^TA^TXWH^TH\right) \\ &\quad + \lambda \mathrm{tr}\left(H^T(D^p - A)H\right) + \gamma \mathrm{tr}\left(HNH^T\right) \\ &\quad s.t \quad W \geq 0 \quad \& \quad H \geq 0 \end{aligned} \tag{6}$$

where N is a matrix of size $k \times k$ whose elements are 1. In the face of constrained optimization problems, we construct Lagrange functions of W and H,

$$\begin{aligned} J(W) =& \left(-2H^TA^TAW + W^TA^TXWH^TH\right) \\ &- \beta W \end{aligned} \tag{7}$$

$$\begin{aligned} J(H) =& \mathrm{tr}\left(-2H^TA^TAW + W^TA^TXWH^TH\right) \\ &+ \lambda \mathrm{tr}\left(H^T(D^p - A)H\right) + \gamma \mathrm{tr}\left(HNH^T\right) \\ &- \beta H \end{aligned} \tag{8}$$

where β is Lagrange multiplier for $W \geq 0$ and $H \geq 0$. First, update W(fixing H), take the derivative of W's Lagrange function $J(W)$ and set it at zero, and we get the following:

$$\frac{\partial J}{\partial W} = -2A^TAH + 2A^TAWH^TH - \beta = 0 \tag{9}$$

because of $W \geq 0$ and using the KKT conditions $\beta W = 0$, we can get the following:

$$\left(-2A^TAH + 2A^TAWH^TH\right)W = \beta W = 0 \tag{10}$$

$$\left(-A^TAH + A^TAWH^TH\right)W^2 = \beta W^2 = 0 \tag{11}$$

where $A^TA = (A^TA)^+ - (A^TA)^-$, and we can get the following:

$$\begin{aligned} &\left[(-((A^TA)^+H - (A^TA)^-H)) + ((A^TA)^+WH^TH\right. \\ &\quad \left.-(A^TA)^-WH^TH)\right]W^2 = 0 \end{aligned} \tag{12}$$

$$\begin{aligned} &\left[(A^TA)^-H + (A^TA)^+WH^TH\right]W^2 \\ &\quad = \left[(A^TA)^+H + (A^TA)^-WH^TH\right]W^2 \end{aligned} \tag{13}$$

$$W_{ik}^2 = W_{ik}^2 \frac{(A^TA)^+H + (A^TA)^-WH^TH}{(A^TA)^-H + (A^TA)^+WH^TH} \tag{14}$$

Similarly, we update H(fixing W), take the derivative of H's Lagrange function $J(H)$ and set it at zero, and we get the following:

$$\begin{aligned} \frac{\partial L}{\partial H} =& -2A^TAW + 2HW^TA^TAW \\ &+ 2\lambda(D^p - A)H + 2\gamma HN - \beta = 0 \end{aligned} \tag{15}$$

because of $H \geq 0$ and using the KKT conditions $\beta H = 0$, we can get the following:

$$\begin{aligned} &\left(-2A^TAW + 2HW^TA^TAW + 2\lambda(D^p - A)H + 2\gamma HN\right) \\ &\quad H_{ik} = \beta H_{ik} = 0 \end{aligned} \tag{16}$$

$$\begin{aligned} &\left(-A^TAW + HW^TA^TAW + \lambda(D^p - A)H + \gamma HN\right) \\ &\quad H_{ik}^2 = \beta H_{ik}^2 = 0 \end{aligned} \tag{17}$$

where $A^TA = (A^TA)^+ - (A^TA)^-$ and $D^p - A = ((D^p - A)^+ - (D^p - A)^-)$, and we can get the following:

$$\begin{aligned}&\big[-((A^TA)^+W-(A^TA)^-W)+(HW^T(A^TA)^+W\\&\quad-HW^T(A^TA)^-W)\\&\quad+(\lambda(D^p-A)^+H-\lambda(D^p-A)^-H)\\&\quad+\gamma HN]H_{ik}^2=0\end{aligned}\tag{18}$$

$$\begin{aligned}&\big[(A^TA)^-W)+HW^T(A^TA)^+W\\&\quad+\lambda(D^p-A)^+H+\gamma HN\big]H_{ik}^2\\&=\big[(A^TA)^+W)+HW^T(A^TA)^-W\\&\quad+\lambda(D^p-A)H^-\big]H_{ik}^2\end{aligned}\tag{19}$$

$$H_{ik}^2=H_{ik}^2\frac{(A^TA)^+W+HW^T(A^TA)^-W+\lambda(D^p-A)^-H}{(A^TA)^-W+HW^T(A^TA)^+W+\lambda(D^p-A)^+H+\gamma HN}\tag{20}$$

Finally, we can get the updating rules of W and H as follows by formula (14) and formula (20):

$$W_{ik}\leftarrow W_{ik}\sqrt{\frac{\left[\left(A^TA\right)^+H\right]_{ik}+\left[\Psi_2H^TH\right]_{ik}}{\left[\left(A^TA\right)^-H\right]_{ik}+\left[\Psi_1H^TH\right]_{ik}}}\tag{21}$$

$$H_{ik}\leftarrow H_{ik}\sqrt{\frac{\left[\Psi_1\right]_{ik}+\left[HW^T\Psi_2\right]_{ik}+\left[\lambda\Psi_3^-H\right]_{ik}}{\left[\Psi_2\right]_{ik}+\left[HW^T\Psi_1\right]_{ik}+\left[\lambda\Psi_3^+H\right]_{ik}+[\gamma HN]}}\tag{22}$$

where $\Psi_1=\left(A^TA\right)^+W,\Psi_2=\left(A^TA\right)^-W$ and $\Psi_3=(D^p-A)$. And the iterative update strategy for the model is shown in Algorithm 1.

Algorithm 1 RC-NMF Algorithm

Require:
Adjacency matrix A of signed network G;
The number of communities k;
The parameters λ and γ;
Ensure:
Community membership matrix H;
Initialize: H, W;
1: **for** $t=1:iter$ **do**
2: update W according to (21)
3: update H according to (22)
4: **end for**
5: **return** H;

3.4 Computational complexity

The computational complexity of updating W of the proposed RC-NMF algorithm is $O(niter(N^2K+NK^2))$, and that of updating H is $O(niter(N^2K+NK^2))$, where niter is the number of iterations and K is the number of community. It is worth noting that in the real world, the signed network is sparse, so N^2 can be represented as the number of links M in the network. Moreover, the number of community K is far less than the number of nodes N, so K^2 can be ignored. Therefore, the computer complexity of the optimization algorithm for the proposed RC-NMF can degrade to $O(niter(M+N))$.

4 Experiments

In this section, we designed a series of experiments in synthetic data and real-world signed networks to validate our model including the convergence of our algorithm. In order to ensure that our RC-NMF algorithm yields the best results, we made relevant parameter sensitivity experiments in Sect. 4.4 and determined the optimal parameter as: $\lambda=3$ and $\gamma=7$.

4.1 Experiments on synthetic signed networks

In this section, we design a series of control experiments of our RC-NMF model with other algorithms on the artificial signed network dataset and the real large-scale signed network dataset to verify the performance of our model in community detection and link prediction. Finally, we analyze the convergence of our RC-NMF algorithm.

4.1.1 Validation of community detection

Synthetic signed networks SG benchmark network (Yang et al. 2017): The SG benchmark network (Yang et al. 2017) is evolved from the GN benchmark network (Yang et al. 2007). The GN benchmark network includes four parameters c, n, k and p_{in}, which can only generate ordinary complex networks. Where c is the community number, n is the number of nodes in each community, k is the average degree in the network, and p_{in} denotes the probability of internal links. On this basis, the SG benchmark network can generate signed network, which adds two noise-level parameters p_+ and p_- that represent the prior probability of positive inter-links and negative intra-links, respectively. As p_{in} decrease or noise level increases, the community structures will become less clear and more difficult to be detected, we set the parameters as follows: c = 4; n = 32; k = 16 and generate two kinds of SG networks:

Type I Weakly balanced signed network(SG-BN)

There is no noise in SG-BN, i.e., $p_+=0$ and $p_-=0$. The parameter p_{in} is from 0.1 to 0.9. The larger the p_{in} value is, the clearer the community structure of the signed network tends to be.

Type II Unbalanced signed network(SG-UN)

Noise is added with different levels, i.e., p_+ from 0 to 0.5 in 0.05 steps and p_- from 0 to 0.5 in 0.05 steps, the parameter p_{in} is set to 0.8. We set the threshold of noise level to 0.5 because if the noise is too large, the generated network model does not satisfy the characteristics of the signed network that the links in the same community are mostly positive and the links in different communities are mostly negative.

SLFR benchmark network (Yang et al. 2017): The signed Lancichinetti–Fortunato–Radicchi (SLFR) benchmark network (Yang et al. 2017) is derived from Lancichinetti–Fortunato–Radicchi (LFR) (Lancichinetti et al. 2008). Compared with GN benchmark network above, LFR benchmark network considers the non-uniform distribution of node degree and community number and has eight parameters: n, k_{avg}, k_{max}, λ_1, λ_2, c_{min}, c_{max} and μ, where n is number of nodes, k_{avg} and k_{max} represent average degree and maximum degree, respectively, λ_1 and λ_2 mean minus exponent for the degree sequence and the community size distribution, respectively, c_{min} and c_{max} mean the minimum and maximum of community number, respectively, and μ is the mixing parameter that indicates the fraction of edges connecting the neighbors in other communities. On this basis, the SLFR benchmark network can generate signed network, which adds two noise-level parameters $p+$ and p_- that represent the fraction of positive external links and negative internal links, respectively. As p_{in} decrease or noise level increases, the community structures will become less clear and more difficult to be detected. In this paper, we set the parameters as follows: $n = 128$, $k_{avg} = 16$, $k_{max} = 20$, $\lambda_1 = 2$, $\lambda_2 = 1$, $c_{min} = 20$, $c_{max} = 40$, and generate two kinds of SLFR networks::

Type I Weakly balanced signed network(SLFR-BN)

There is no noise in SLFR-BN, i.e., $p_+ = 0$ and $p_- = 0$. The parameter μ is from 0.1 to 0.9. The μ value represents the degree of confusion of the community in the signed network. The larger the μ value is, the more likely the signed network tends to be in a state of no community structure.

Type II Unbalanced signed network(SLFR-UN)

Noise is added with different levels, i.e., the parameter μ is set to 0.2, and $p_+ \in [0, 0.5]$ in 0.05 steps and $p_- \in [0, 0.5]$ in 0.05 steps.

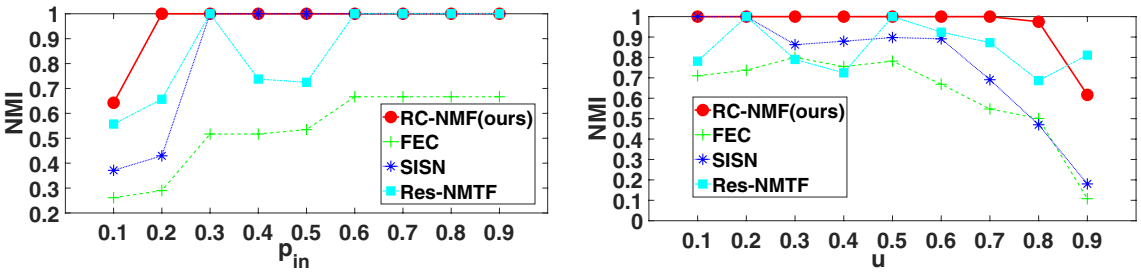
Validation metrics As for the accuracy measurement of community detection, we use normalized mutual information(NMI) (Jiao et al. 2018), which is widely used to measure the degree of similarity between predicted community structure and real community structure. The larger the NMI value is, the higher the accuracy of the community division gets. The NMI can be expressed as follows:

$$NMI(C, C') = \frac{\sum_{i=1}^{k}\sum_{j=1}^{k} n_{ij} log\frac{n_{ij}n}{n_i^{(1)}n_j^{(2)}}}{\sqrt{\sum_{i=1}^{k} n_i^{(1)} log\frac{n_i^{(1)}}{n}\sum_{j=1}^{k} n_j^{(2)} log\frac{n_j^{(2)}}{n}}} \tag{23}$$

where C and C' denote ground truth and detected community partition by algorithm, respectively, k is the number of the communities, n is the number of nodes, n_{ij} denotes the number of nodes in ground community i that are assigned in community j in detected community partition, $n_i^{(1)}$ is the number of nodes in true community i, and $n_j^{(2)}$ is the number of nodes in detected community j.

Comparison methods The RC-NMF performance was first tested with respect to the community detection with signed networks. Comparisons were made using three state-of-the-art methods: FEC (Yang et al. 2007), SISN (Zhao et al. 2017) and Res-NMTF (Li et al. 2018b).

Analysis of experimental results In order to more accurately measure the performance of community detection, the comparison experiments were designed in the four different types of signed networks as above 4.1.1. In the two noise-free signed network datasets, SG-BN and SLFR-BN, the x-axis represents the internal connection ratio of the signed network and the confusion of the community structure. In the two unbalanced signed network datasets SG-UN and SLFR-UN, the x- and y-axes represent the internal negative links and external positive links of the signed network, respectively, which are also represented as noise in the signed network.



(a) SG-BN **(b)** SLFR-BN

Fig. 1 NMI of community detection in SG-BN and SLFR-BN

The community detection comparison results obtained in the two noise-free signed network datasets of SG-BN and SLFR-BN are shown in Fig. 1. We can observe that in SG-BN 1(a), as the p_{in} increases, the community structure of the signed network tends to be clearer, and the NMI value also increases. The RC-NMF algorithm we proposed is represented in red. When $p_{in} > 0.1$, the NMI value is always 1, which means that the community divided by the algorithm is the same as the real community. The performance of the other three algorithms is inferior to the RC-NMF algorithm. The SISN algorithm in blue indicates that when $p_{in} > 0.2$, the NMI value is always 1, and the Res-NMTF algorithm is represented in cyan, when $p_{in} = 0.4$ and $p_{in} = 0.5$, the NMI value decreases, indicating that the performance of the algorithm is poor when the number of links in the community and between the communities tends to be equal. The FEC algorithm is represented by green. Although the NMI value shows an upward trend as the p_{in} increases, the overall NMI has always been at a low value, and its algorithm performance is poor. In SLFR-BN 1(b), as the μ value increases, the community structure of the signed network tends to be blurred, and the NMI value also decreases. Before $\mu = 0.9$, the NMI value obtained by our RC-NMF algorithm is always 1 showing the best accuracy. The other three algorithms show a downward trend in the process of increasing μ, but the accuracy is lower than our algorithm. And the community detection comparison results obtained on the unbalanced signed network dataset of SG-UN are shown in Fig. 2. We can observe that with the increase in external positive link p_+ and internal negative link p_-, which represents the increase in noise level in the signed network, the performance in community detection of RC-NMF algorithm and Res-NMTF algorithm has decreased to some extent. The accuracy of the RC-NMF algorithm we proposed decreased significantly when the internal negative link p_- was increased, and the external positive link p_+ had no significant impact on the performance of the algorithm. Moreover, when the internal negative link p_- was less than 0.25, our algorithm result obtained that the NMI was always 1, and its community detection performance was better than other algorithms. Moreover, in the real-world social network, there are fewer negative relationships among individuals in the same cluster, which, as shown in the signed network, there are fewer internal negative edges, and our algorithm has better performance in this case, so it can be well applied to social network analysis. However, in the case of high noise, the performance of our RC-NMF algorithm is worse than that of FEC and SISN, which reflects the poor robustness of our algorithm. Finally, the community detection comparison results obtained on the unbalanced signed network dataset of SLFR-UN are shown in Fig. 3. We can observe that with the increase in external positive link p_+ and internal negative link p_-, which represents the increase in noise level in signed network, the increase in the detection performance of all algorithms has been reduced to varying degrees. The accuracy of our RC-NMF algorithm is greatly affected by the internal negative links p_-, when the internal negative link $p_- < 0.2$, the NMI value is 1, and the performance of the algorithm is better than all other algorithms. With internal negative link $p_- > 0.2$, the accuracy of the algorithm decreases, which is inferior to the accuracy of the FEC algorithm and the SISN algorithm. And the accuracy

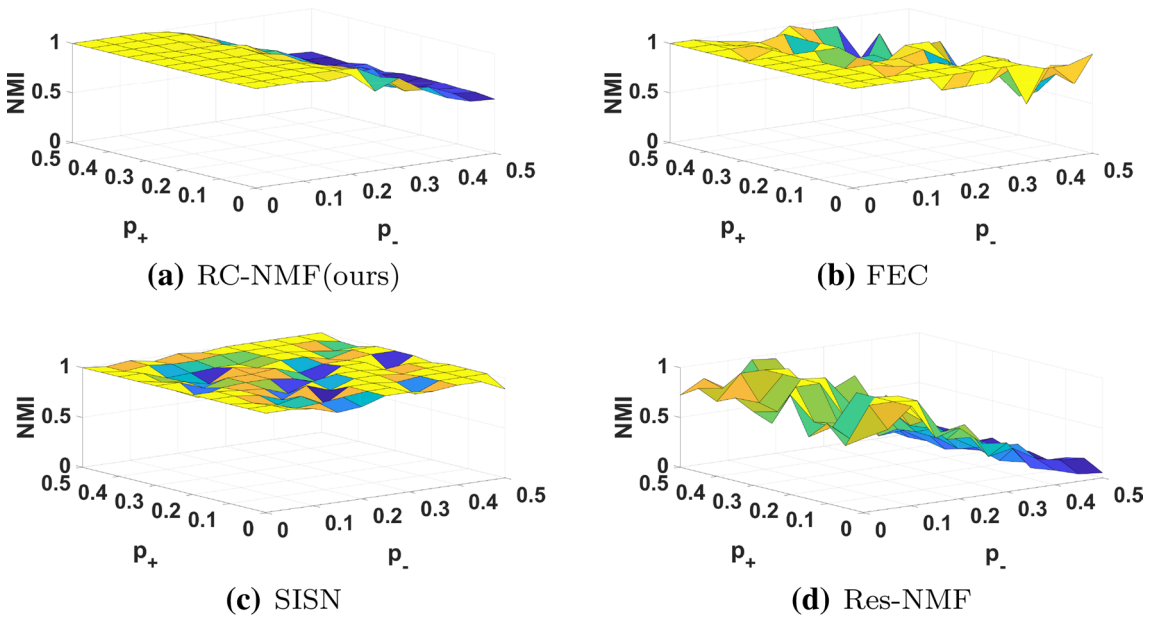


Fig. 2 NMI of community detection in SG-UN

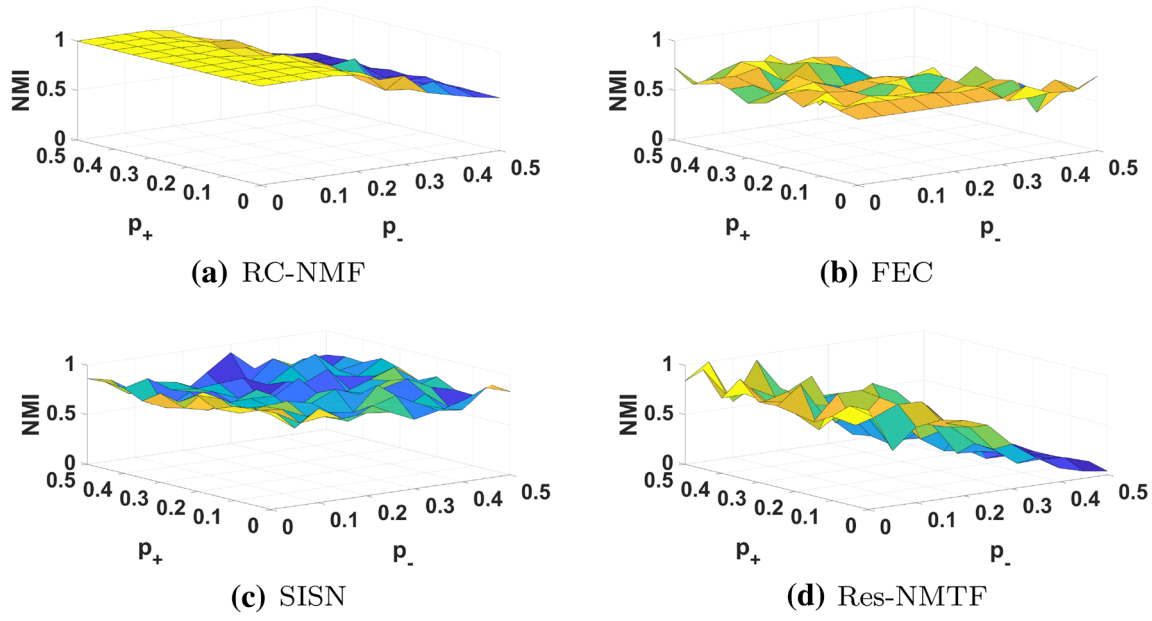


Fig. 3 NMI of community detection in SLFR-BN

of the Res-NMTF algorithm decreased significantly with the increase in internal negative link $p\_$.

4.1.2 Validation of link prediction

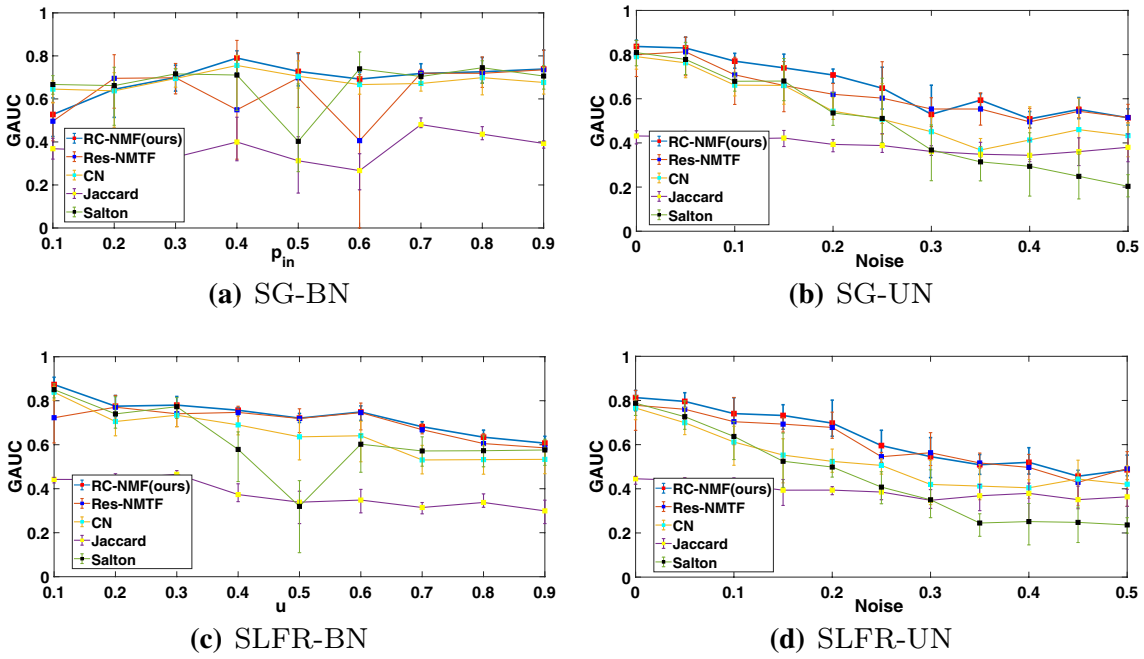
Validation metrics To measure the performance of algorithms about link prediction in signed network, we use GAUC (generalized AUC over +1, 0 and − 1) (Song and Meyer 2015), which is an extension of AUC measurement index and can be used to measure the accuracy of the three states of positive links, negative links and un-links. It is very suitable for measuring the accuracy of signed network link prediction, formulated as:

$$\frac{1}{|P|+|N|}\left(\frac{1}{|U|+|N|}\sum_{a_i\in P}\sum_{a_j\in U\cup N} I\big(L(a_i)>L(a_j)\big) + \frac{1}{|U|+|P|}\sum_{a_i\in N}\sum_{a_j\in U\cup P} I\big(L(a_i)<L(a_j)\big)\right) \tag{24}$$

where $|P|$, $|N|$ and $|U|$ represent the number of positive links, negative links and un-links in signed networks, respectively. $L(\cdot)$ is the link score function, and $I(\cdot)$ is the 0/1 indicator function that if the condition in $(\cdot)$ comes true, we get 0 loss, otherwise 1 loss. The larger the GAUC, the higher the accuracy of the link prediction algorithm.

Comparison methods Because the above-mentioned algorithm can only be used to detect the community and cannot perform link prediction experiments, we have selected several node similarity indicators and the Res-NMTF algorithm as the comparison algorithm of our RC-NMF algorithm about measuring the performance of link prediction. Several indicators include: CN, Jaccard and Salton.

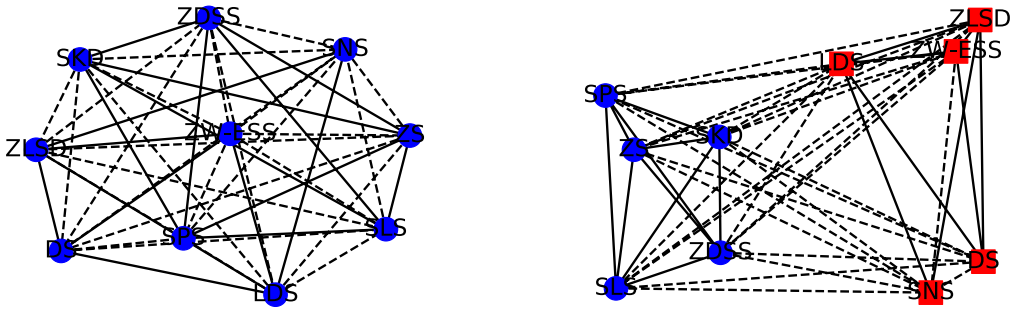
Analysis of experimental results In order to make the experimental results more accurate and comprehensive, we used the standard fivefold cross-validation for training and testing. Figure 4 shows the performance of our RC-NMF algorithm with the Res-NMTF algorithm and several indicators on the four network models, respectively. In the SG-BN signed network dataset shown in Fig. 4a, the x-axis represents the proportional of the internal links p_{in}. In the SLFR-BN dataset shown in Fig. 4c, the x-axis represents the degree of confusion of the community structure. In the SG-UN data set and the SLFR-UN data set shown in Fig. 4b, d, respectively, the x-axis represents the ratio of the internal negative links and the external positive link, which denote the noise level of the signed network. It can be observed that in SG-BN, as the internal links p_{in} increase, the community structure tends to be obvious, and the performance of the link prediction algorithms is improved. In SLFR-BN, as the degree of confusion in the community structure μ increases, the performance of the link prediction algorithm decreases to varying degrees. In SG-UN and SLFR-UN, the performance of the link prediction algorithm decreases with the increase in noise. The accuracy of our RC-NMF algorithm is always the second best or the best among the comparison results of various algorithms, which indicates that our algorithm is superior in link prediction.



(a) SG-BN **(b)** SG-UN

(c) SLFR-BN **(d)** SLFR-UN

Fig. 4 GAUC of link prediction in different synthetic datasets



(a) The connection state between nodes in Slovene parliamentary party network.

(b) The community partition made by our RC-NMF algorithm.

Fig. 5 Slovene parliamentary party network

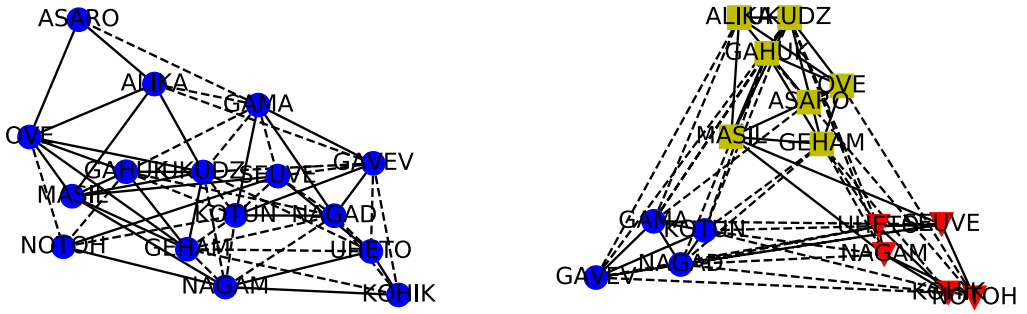
4.2 Experiments on real-world signed networks

In this section, we compared RC-NMF with the other above-approaches in real-world signed networks to validate the accuracy and effectiveness of our proposed algorithm in community detection and link prediction.

4.2.1 Validation of community detection

Slovene parliamentary party network The network is about the relations among the ten parties in Slovene parliament, *1994*, which has two communities (Ferligoj and Kramberger 1996). In the community detection, we only retain the sign of link in the network and ignore the weight of links. Figure 5a shows the connection state between nodes in the Slovene parliamentary party network, the solid edges represent the positive relationship, and the dash-dot edges represent the negative relationship. Figure 5b shows the community partition made by our RC-NMF algorithm, and the result is the same as the real situation, which is divided into two communities: (SKD, ZDSS, ZS, SLS, SPS) and (ZLSD, LDS, ZW-ESS, DS, SNS).

Gahuku-Gama subtribes network The network is about the culture of New Guinea Highland (Read 1954). There are



(a) The connection state between nodes in Gahuku-Gama subtribes network.

(b) The community partition made by our RC-NMF algorithm.

Fig. 6 Gahuku-Gama subtribes network

Table 1 Large-scale signed network dataset statistics

Datasets	Nodes	Pos-links	Neg-links	Un-links
Epinions@50	6109	379,830	42,494	3.69×10^7
Slashdot@50	4303	130,680	40,539	1.83×10^7
Wiki@50	11,047	573,423	69,012	1.21×10^8
Bitcoinotc@50	263	6476	454	6339

Table 2 Comparison experiment results of four algorithms in link prediction

	CN	Jaccard	Salton	Res-NMTF	RC-NMF
Epinions@50	0.8166	0.4328	0.7654	0.7954	**0.8837**
Slashdot@50	0.6671	0.3563	0.5888	0.7704	**0.7973**
Wiki@50	0.6832	0.4648	0.6369	0.745	**0.7219**
Bitcoins@50	0.6743	0.4176	0.6565	0.7397	**0.7143**

16 subtribes in this network falling into three communities. Figure 6a shows the connection state between nodes in the Gahuku-Gama subtribes network, where solid edges represent the political alliance relationship and dash-dot edges represent the enmities relationship, respectively. Figure 6b shows the community partition made by our RC-NMF algorithm, and the result is the same as the real situation, which is divided into three communities: (UKUNZ, GEHAM, MASIL, OVE, ASARO, ALIKA), (SEUVE, UHETO, NAGAM, NOTOH, KOHIK) and (KOTUN, GAMA, NAGAM, GAVEV).

4.2.2 Validation of link prediction

In order to detect the accuracy of link prediction in real-world signed network, we used four large-scale real network datasets in experiments of link prediction, i.e., Slashdot

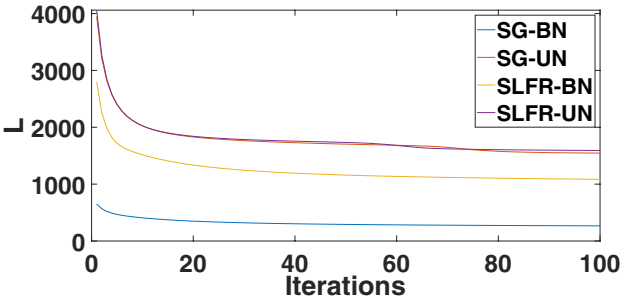


Fig. 7 Convergence of our gradient descent update rules

(Leskovec et al. 2010), Wiki (Maniu et al. 2011), Epinions (Leskovec et al. 2010) and Bitcoinotc (Kumar et al. 2018). And in the real world, a person has an average of 40 friends offline and 338 friends online. Therefore, it is more realistic to check users with a high degree (Li et al. 2018a). In the contrast experiment of link prediction in the large-scale real-signed network, we select nodes that the degree threshold is set at 50, and the network statistics after setting are shown in Table 1 where we used 'name@degree' to represent a specific dataset, e.g., Epinions@50 is the dataset about Epinions with $d \geq 50$. Table 2 shows the comparison results of our algorithm with other methods. Each number represents the GAUC value obtained by the link prediction experiment of the corresponding algorithm on the corresponding large-scale signed network dataset, where the experimental result value of our algorithm is shown in bold. We can observe that compared to other algorithms, our algorithm performs better than other methods in real-scale large-scale signed networks.

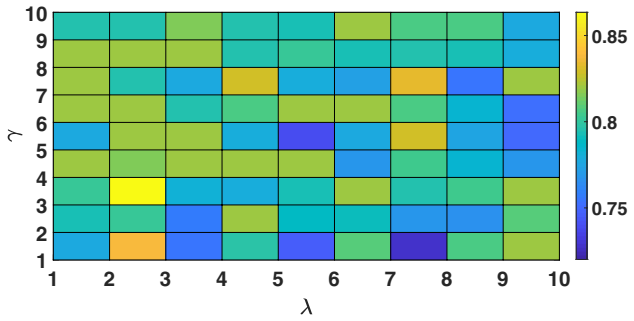


Fig. 8 Experimental results of parameter sensitivity

4.3 Algorithm convergence

To test the convergence of the algorithm, we perform experiments on four kinds of synthetic datasets and deriving networks from each kind of network model. As shown in Fig. 7, when the number of iterations is greater than 50, the value of our objective function tends to be stable and will not change, which indicates that the proposed algorithm satisfies the convergence condition.

4.4 Parameter sensitivity

In this section, we study the parameter sensitivity in our proposed algorithm. We selected one of the SLFR-UN datasets, in which $p_+ = 0.25$ and $p_- = 0.25$ represent the signed network with certain noise but not the maximum, which is similar to the signed network in the real world. The experimental results of parameter sensitivity are shown in Fig. 8. The parameters have some influence on the accuracy of the algorithm, but not very much. When parameter λ is greater than 5, the blue color increases, and the accuracy of the algorithm decreases to a certain extent. Moreover, parameter γ has little influence on the accuracy of the algorithm, and the better results are concentrated between 4 and 9. Therefore, we select parameter values: $\lambda = 3$ and $\gamma = 7$.

5 Conclusion and future work

Community detection and link prediction are the basic tasks of signed network analysis. Many of the previous algorithms rely on pre-defined optimization objective functions or heuristic algorithms with high computational complexity, and most of the algorithms cannot simultaneously perform community detection and link prediction. In response to these challenges, we propose the RC-NMF method that converges in a reasonable number of times and can be used to complete community detection and link prediction at the same time. In this paper, in order to constrain the influence of negative links, we introduce a method of graph regularization to constrain nodes with positive links being assigned to the same community and nodes with negative links being assigned to different communities. Then, in order to constrain the situation of overlapping communities, we added sparse constraints to our model. After that, we conducted a series of the experiments for community detection and link prediction in synthetic data and real-world signed network to validate the effectiveness and accuracy of our RC-NMF algorithm.

In the future, we will work on using our proposed algorithm in real social network analysis, using crawler technology to crawl the text data of comments between users on social media such as Facebook or Weibo. Then, we use some emotion analysis methods to analyze the emotional tendency in the text, such as friendliness or hostility, and build a signed network based on this. Furthermore, we use the algorithm we proposed to find some community structures and special societies in the social network such as fraud groups. Finally, we use it to predict the emotional tendency of users, such as positive or negative signed network, which can serve as a foundation for the friend recommendation system.

References

Amelio A, Pizzuti C (2016) An evolutionary and local refinement approach for community detection in signed networks. Int J Artif Intell Tools 25(04):1650021

Anchuri P, Magdon-Ismail M (2012) Communities and balance in signed networks: a spectral approach. In: Proceedings of the 2012 international conference on advances in social networks analysis and mining (ASONAM 2012), IEEE Computer Society, pp 235–242

Cartwright D, Harary F (1956) Structural balance: a generalization of heider's theory. Psychol Rev 63(5):277

Chen Y, Wang X, Yuan B, Tang B (2014) Overlapping community detection in networks with positive and negative links. J Stat Mech Theory Exp 3:P03021

Chiang KY, Hsieh CJ, Natarajan N, Dhillon IS, Tewari A (2014) Prediction and clustering in signed networks: a local to global perspective. J Mach Learn Res 15(1):1177–1213

Davis JA (1967) Clustering and structural balance in graphs. Hum Relat 20(2):181–187

Derr T, Wang Z, Dacon J, Tang J (2020) Link and interaction polarity predictions in signed networks. Soc Netw Anal Min 10

Ferligoj A, Kramberger A (1996) An analysis of the slovene parliamentary parties network. Dev Stat Methodol 12:209–216

Ghoshal G, Mangioni G, Menezes R, Poncela-Casanovas J (2014) Social system as complex networks. Soc Netw Anal Min 4(1)

Harary Frank (1953) On the notion of balance of a signed graph. Mich Math J 2(2):143–146

Heider F (1946) Attitudes and cognitive organization. J Psychol 21(1):107–112

Jiang JQ (2015) Stochastic block model and exploratory analysis in signed networks. Phys Rev E 91:062805

Jiao P, Yu W, Wang W, Li X, Sun Y (2018) Exploring temporal community structure and constant evolutionary pattern hiding in dynamic networks. Neurocomputing 314:224–233

Jordan MI (2009) Convex and semi-nonnegative matrix factorizations. IEEE Trans Pattern Anal Mach Intell 32(1):45–55

Kulakowski K, Stojkow M, Zuchowska-Skiba D (2019) Heider balance, prejudices and size effect. J Math Sociol 1–9

Kumar S, Hooi B, Makhija D, Kumar M, Faloutsos C, Subrahmanian V (2018) Rev2: fraudulent user prediction in rating platforms. In: Proceedings of the eleventh ACM international conference on web search and data mining, ACM, pp 333–341

Lancichinetti A, Fortunato S, Radicchi F (2008) Benchmark graphs for testing community detection algorithms. Phys Rev E Stat Nonlinear Soft Matter Phys 78(4 Pt 2):046110

Leskovec J, Huttenlocher D, Kleinberg J (2010) Signed networks in social media. In: Proceedings of the SIGCHI conference on human factors in computing systems, ACM, pp 1361–1370

Li X, Fang H, Zhang J (2018a) File: a novel framework for predicting social status in signed networks. In: Thirty-second AAAI conference on artificial intelligence

Li Y, Liu J, Liu C (2014) A comparative analysis of evolutionary and memetic algorithms for community detection from signed social networks. Soft Comput 18(2):329–348

Li Z, Chen J, Fu Y, Hu G, Pan Z, Zhang L (2018b) Community detection based on regularized semi-nonnegative matrix tri-factorization in signed networks. Mob Netw Appl 23(1):71–79

Maniu S, Abdessalem T, Cautis B (2011) Casting a web of trust over wikipedia: an interaction-based approach. In: Proceedings of the 20th international conference companion on world wide web, ACM, pp 87–88

Newman M (2016) Community detection in networks: Modularity optimization and maximum likelihood are equivalent. arXiv preprint arXiv:160602319

Read KE (1954) Cultures of the central highlands, new guinea. Southwest J Anthropol 10(1):1–43

Rossi RA, Ahmed NK (2019) Complex networks are structurally distinguishable by domain. Soc Netw Anal Min

Song D, Meyer DA (2015) Recommending positive links in signed social networks by optimizing a generalized auc. In: Twenty-ninth AAAI conference on artificial intelligence

Vasudevan M, Deo N (2012) Efficient community identification in complex networks. Soc Netw Anal Min 2(4):345–359

Wang H, Zhang F, Hou M, Xie X, Guo M, Liu Q (2018) Shine: signed heterogeneous information network embedding for sentiment link prediction. In: Proceedings of the eleventh ACM international conference on web search and data mining, ACM, pp 592–600

Wang S, Aggarwal C, Tang J, Liu H (2017) Attributed signed network embedding. In: Proceedings of the 2017 ACM on conference on information and knowledge management, ACM, pp 137–146

Yang B, Cheung W, Liu J (2007) Community mining from signed social networks. IEEE Trans Knowl Data Eng 19(10):1333–1348

Yang B, Liu X, Li Y, Zhao X (2017) Stochastic blockmodeling and variational bayes learning for signed network analysis. IEEE Trans Knowl Data Eng 29(9):2026–2039

Zhao X, Yang B, Liu X, Chen H (2017) Statistical inference for community detection in signed networks. Phys Rev E 95:042313

Zheng Q, Skillicorn DB (2015) Spectral embedding of signed networks. In: Proceedings of the 2015 SIAM international conference on data mining, SIAM, pp 55–63