2019 IEEE Intl Conf on Parallel & Distributed Processing with Applications, Big Data & Cloud Computing, Sustainable Computing & Communications, Social Computing & Networking (ISPA/BDCloud/SocialCom/SustainCom)

A Double Non-negative Matrix Factorization Model for Signed Network Analysis

1st Wei Yu
College of Intelligence and Computing
Tianjin University
Tianjin, China
weiyu@tju.edu.cn

2nd Rongjian Mu
College of Intelligence and Computing
Tianjin University
Tianjin, China
murj@tju.edu.cn

3rd Ying Sun
College of Intelligence and Computing
Tianjin University
Tianjin, China
yingsun@tju.edu.cn

4th Xue Chen
Law School
Tianjin University
Tianjin, China
xuechen@tju.edu.cn

5th Wenjun Wang
College of Intelligence and Computing
Tianjin University
Tianjin, China
wjwang@tju.edu.cn

6th Huaming Wu
Center for Applied Mathematics
Tianjin University
Tianjin, China
whming@tju.edu.cn

***Abstract*—As one of the most important characteristics of complex networks, community structure have been studied extensively in real cases. However, in signed networks that include both positive and negative edges, the development of community discovery is still limited now. Because the sign information on edges poses a challenge to modeling the signed network. Most existing methods are based on heuristics, so these methods tend to have high computational complexity and ignore the generation of the networks. Here, we propose a Double Non-negative Matrix Factorization (DouNMF) model from the perspective of generative model to detect communities in the signed network. This algorithm skillfully applies the Non-negative matrix Factorization algorithm to the signed network. In addition, the algorithm integrates indegree information into the process of matrix factorization. Large amounts of experiments on several artificial and real-world signed networks validate that the effectiveness and accuracy of our proposed approaches both in community discovery and link prediction.**

***Keywords*-signed network; community discovery; link prediction; non-negative matrix factorization.**

I. Introduction

Most of complex systems in the real world can be modelled with complex networks [1] [2]. Compared with the unsigned network, signed networks can model more information from real-world complex systems as they have both positive and negative edges. In fact, many complex relations in the real world can be denoted by positive and negative edges of signed networks. For example, a positive edge in online social network usually signifies 'support', 'like' or 'cooperation', while a negative edge means 'opposite', 'dislike' or 'hostility'. Therefore, the signed network analysis has been paid more and more attention in various fields [3].

Community discovery and link prediction are two basic problems in the signed network analysis. The mission of community discovery in signed networks is to discover the community structures that expressed as dense positive edges within the communities and dense negative edges among the communities [4]. In addition, link prediction is to predict the sign of unknown edge in signed network [5]. Although in recent years there have been some algorithms proposed for community discovery and link prediction in the signed network, but its development is still immature. For example, some algorithms [6] [7] based on optimization objective functions and heuristics have high computational complexity. Some model-based algorithms [4] [8] [9] [10] have low accuracy in performance or need probabilistic statistical inference methods to select model, such as EM algorithm, resulting in a large computational burden. Some algorithms are base on network embedding [11] [12] with high computational performance but poor interpretability. And most of the above algorithms can only be used for community discovery or link prediction. For the challenge, we propose a new model, Double Non-negative Matrix Factorization (DouNMF), for community discovery and link prediction.

In this work, we divide a signed network into positive and negative components and implement non-negative matrix factorization (NMF) [18] on the two components respectively. At the same time, we integrate the indegree information of nodes into the non-negative matrix factorization process to obtain the node probability matrix with higher accuracy. Except for community discovery, the proposed DouNMF can also be used for link prediction. We design experiments on both artificial signed network and real-world signed networks. Large amounts of experimental results demonstrate that our approach is more effective when compared with several state-of-the-art approaches.

We organize the other contents of the work as follows: the recent related work of community discovery in signed networks are introduced in section II, the details about our DouNMF approach are introduced in section III, experiments and related discussion on the artificial networks

978-1-7281-4328-6/19/$31.00 ©2019 IEEE

DOI 10.1109/ISPA-BDCloud-SustainCom-SocialCom48970.2019.00136

and real-world networks are showed in Section IV, which demonstrate the effective and accuracy of our proposed method, and at last our contributions are summarized in Section V.

II. Related work

Recently, large amounts of algorithms have emerged for community discovery and link prediction in signed work. These algorithms can be roughly divided into the following four categories: balance theory-based, modularity optimization-based, model-based, and network embedding-based.

Balance theory-based methods. This type of method usually is based on a heuristic approach with structural balance theory of sociology [13]. Generally, one could find the community structure in signed social networks by cutting off negative edges. Unfortunately, the structures of signed social networks in real case are normally unbalanced since the existence of frustration that presents 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 [14] extended the applicability of the balance theory from the local features of signed networks to the global features. Amelio et al. [15] developed a correlation clustering method (CC) that maximizes positive edges within communities and negative edges among communities or minimum frustration to detect community in signed networks. Li et al. [5] present a novel framework including two implicit features and two latent features for predicting link, one of which is obtained by balance theory.

Modularity optimization-based methods. Modularity optimization [16] essentially optimizes the modularity objective function to maximize the modular level based on the topological structure in the network. Li [6] defined signed modularity by improving standard modularity in the unsigned network, and made it capable of handling negative edges. Signed modularity balances the trend of entities with positive edges to forming community and the trend of entities with negative edges to destroying community by adding weights on positive and negative components. Then some heuristics algorithms based on signed modularity optimization have been proposed. For example, Anchuri et al. [7] generalized spectral partitioning (SpePart) approach with iterative optimization to mine the community structures, and this is an extension about standard modularity optimization in the unsigned network.

Model-based methods. Model-based methods focus on modeling the generated mechanism which tends to applicable to the network. For instance, Yang et al. [4] developed an agent-based random walk model framework (FEC) based on the assumption that an agent should have a higher probability of staying in the same community rather than going to another community after lots of walks, to mine the community structures. The algorithm is able to give the nearly optimal solutions in linear time obeying the size of networks, but its performance is poor. Chen et al. [17] proposed a novel approaches, which is called SPM, to discover overlapping community sturctures. Some of the above methods are based on optimization objectives or heuristic to mine community structures in the signed network and do not care about the generation of the network. Jiang et al. [8] proposed a generalized signed SBM (SSBM), to explore the mesoscopic structures in signed networks from a node perspective. Yang et al. [9] proposed SSBM Learning (SSL) algorithm that can learn SSBM with exploratory networks based on variational Bayesian techniques. We need to use some probability inference methods to fit the network observed in the model, such as expectation-maximization (EM) algorithm, which increases the computational burden. Zhao et al. [10] proposed a statistical inference approach in signed networks (shorthand for SISN), which model signed networks by a probabilistic model and find communities by EM algorithm in signed networks, is a mathematically principled method.

Network embedding-based methods. With the rise of deep learning, some algorithms based on network embedding to mine community structure and link prediction emerge. Wang et al. [12] proposed a framework named Signed Heterogeneous Information Network Embedding (SHINE) to predict the sign of unobserved edges. SHINE gets the implicit low-dimension vectors of nodes in the network through deep autoencoders, and then do the similarity analysis of the nodes on this basis. In addition, Wang [11] developed a new framework named Social Network Embedding with Attributes (SNEA), to exploit the network structures and user attributes simultaneously. Although the performance of deep learning is better than some traditional algorithms, the interpretation of these models is weak.

III. Our work

A. Non-negative Matrix Factorization (NMF)

NMF [18] is one of the most common unsupervised learning methods for community discovery in an unsigned network can be formulated as follows:

$$\min_{W,H} \sum_{i,j} \left(A_{i,j} - \left(WH^T\right)_{i,j} \right)^2 \tag{1}$$
$$s.t \quad W \in R_+^{N\times C}, \quad H \in R_+^{N\times C}$$

where A is the adjacency matrix of network G of n nodes, W is the basis matrix, and H is the community membership matrix where element h_{jc} is the propensity of node j in community c, and C is the community amount.

B. DouNMF model

Considering the existence of negative edges, we can not directly use the NMF on the signed network. Here, we firstly decompose a signed network into positive and negative components, then carry out NMF on the positive and negative components respectively. And the traditional NMF algorithm decomposes the initial matrix into two terms, the basis matrix and the weight matrix. In order to further suppress the randomness of matrix factorization, we introduce the indegree information matrix in the factorization process.

For a signed network G, the elements of adjacency matrix A include +1, -1 and 0, representing positive links, negative links, and un-links respectively. A can be separated into two matrices A^+ and A^-, which represent positive component and negative component of signed network G, respectively, and $A = A^+ - A^-$. By using non-negative matrix factorization into positive component A^+ and negative component A^- of the signed network we can derive the hidden features of the positive and negative components separately. Then our proposed DouNMF adds weight λ to balance the effect from the positive and negative components. Therefore, the objective function can be constructed as follows:

$$L = \left\|A^+ - W^+H^T\right\|_F^2 + \lambda\left\|A^- - W^-H^T\right\|_F^2, \tag{2}$$

where $W^+ \in R_+^{N\times C}$ and $W^- \in R_+^{N\times C}$ are the basis matrix of the positive component and the negative component respectively, and $H \in R_+^{N\times C}$ is the community membership matrix. The degree of nodes is only the implicit feature we can get to analyze the community structure in a given network G. Although the adjacency matrix of a signed network implicitly expressed node degrees in the above process, it is not subject to the control in matrix factorization. So we definitely apply indegree of nodes to the process of matrix factorization, maximally control the effect of degree information on node probability matrix H. Hence we introduce indegree information matrix K into the model (3) for better numerical results:

$$L = \left\|A^+ - W^+H^TK^+\right\|_F^2 + \lambda\left\|A^- - W^-H^TK^-\right\|_F^2, \tag{3}$$

where $K^+ \in R_+^{N\times N}$ and $K^+ \in R_+^{N\times N}$ are the indegree information matrices of the positive component and the negative component respectively, which are diagonal matrices where elements k_{ii} represent the possibility of the link from other nodes to node *i*. So the expectation of links between node *i* and node *j*:

$$\hat{a}_{ij} = w_{ik}h_{jk}k_{jj} \tag{4}$$

Furthermore, in order 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. The final objective function of the propoesd DouNMF model is as follows:

$$\begin{aligned} L = {} & \left\|A^+ - W^+H^TK^+\right\|_F^2 + \lambda\left\|A^- - W^-H^TK^+\right\|_F^2 + \\ & \gamma\|H\|_1^2 \\ & = Tr(A^+ - W^+H^TK^+)(A^{+T} - K^{+T}HW^{+T}) + \\ & \lambda Tr(A^- - W^-H^TK^-)(A^{-T} - K^{-T}HW^{-T}) + \\ & \gamma Tr(H\mathbf{1}H^T), \end{aligned} \tag{5}$$

where $\mathbf{1}$ is a matrix of size $k \times k$ whose elements are 1.

We develop multiplicative update rules based on gradient descent to optimize objective function (5), which is shown in Algorithm 1. Accordingly, We can easily deduce that the most time-consuming part of Algorithm 1 is the updating of H, of which the time complexity is $O(n_{iter}(N^2C + NC^2))$, where n_{iter} is the number of iterations. It's worth noting that the structures of the real-world signed networks are usually so sparse that N^2 can be approximately equal to the average number of links M. In addition, C can almost be ignored as it always be much less than N and M. Therefore, the time complexity of the optimization algorithm for the proposed DouNMF can degrade to $O(n_{iter}(M + N))$.

Algorithm 1 DouNMF Algorithm

Input:
Adjacency matrix A of signed network G ;
the node indegree matrix K ;
the number of communities C ;
the balance parameter λ and γ ;
Output:
Community membership matrix H ;
Initialize: H, W^+, W^-;
for $t = 1 : iter$ **do**
  $W^+ \leftarrow W^+ \odot \frac{A^+K^{+T}H}{W^+H^TK^+K^+H}$
  $W^- \leftarrow W^- \odot \frac{A^-K^{-T}H}{W^-H^TK^-K^-H}$
  $H \leftarrow H \odot \frac{K^+A^{+T}W^+ + \lambda K^-A^{-T}W^-}{K^+K^{+T}HW^{+T}W^+ + \lambda K^-K^{-T}HW^{-T}W^- + \gamma HN}$
end for
return H;

IV. Experiments

In this section, we designed large amounts of experiments on real-world signed networks and artificial signed networks to validate our proposed model including the convergence of the optimization algorithm.

A. Experiments on artificial signed networks

Here, we compared DouNMF with several state-of-the-art methods in artificial data to validate the accuracy and effectiveness of DouNMF for community discovery and link prediction.

1) Validation of community discovery:

Artificial signed networks. : We generate four kinds of signed networks by signed stochastic block model (SSBM) [9], which is $X = (K, \Pi, \Theta, \Omega)$, where K denotes the number of community, $\Pi(\pi_1, \pi_{-1}, \pi_0)$ indicates the prior probability of the intra-link, $\Theta(\theta_1, \theta_{-1}, \theta_0)$ indicates the prior probability of the inter-link and Ω indicates the prior probability that a node belongs to a community in the signed network. Here, we set $K = 4$, $\Omega = (\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4})$ and n = 128. By setting the parameters Π and Θ, we can get four kinds of signed networks:

Type I Weakly balanced signed network (BN)

$$X_{BN} = (K, (\pi_1, 0, \pi_0), (0, \theta_{-1}, \theta_0), (\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4}))$$

In this kind of signed network, we set $\pi_{-1} = 0$ and $\theta_1 = 0$, which controls the generated networks have no negative intra-links and positive inter-links. Here, we generate much more BN networks by setting the parameters π_1 and θ_{-1} range [0.1,1] with 0.1 steplength respectively.

Type II Unbalanced sigend network #1 (UN-I)

$$X_{UN-I} = (K, (\pi_1, \pi_{-1}, 0.2), (\theta_1, \theta_{-1}, 0.8), (\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4}))$$

In this kind of sigend network, we set $\pi_0 = 0.2$ and $\theta_0 = 0.8$ and make sure that the intra-links are dense and the inter-links are sparse. Here, we generate UN-I networks by setting the parameters $\pi_1 \in [0, 0.8]$ with 0.08 steplength and $\theta_{-1} \in [0, 0.2]$ with 0.02 steplength, respectively.

Type III Unbalanced sigend network #2 (UN-II)

$$X_{UN-II} = (K, (\pi_1, \pi_{-1}, 0.8), (\theta_1, \theta_{-1}, 0.2), (\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4}))$$

In this kind of sigend network, we set $\pi_0 = 0.8$ and $\theta_0 = 0.2$ and make sure that the intra-links are sparse and the inter-links are dense. Here, we generate UN-II networks by setting the parameters $\pi_1 \in [0, 0.2]$ with 0.02 steplength and $\theta_{-1} \in [0, 0.8]$ with 0.08 steplength, respectively.

Type IV Unbalanced sigend network #3 (UN-III)

$$X_{UN-III} = (K, (\pi_1, \pi_{-1}, 0.2), (\theta_1, \theta_{-1}, 0.2), (\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4}))$$

In this kind of sigend network, we set set $\pi_0 = 0.2$ and $\theta_0 = 0.2$ and make sure that the intra-links and the inter-links are both dense. Here, we generate UN-III networks by setting the parameters $\pi_1 \in [0.0.8]$ and $\theta_{-1} \in [0.0.8]$ with 0.08 steplength respectively.

Validation metrics: We choose normalized mutual information (NMI) [19], which is usually used to measure the similarity of two clusterings, to test the performance of the approaches for community discovery with ground-truth. In detail, NMI is computed as follows:

$$NMI(\mathcal{L}, \mathcal{L}') = \frac{\sum_{i=1}^{C} \sum_{j=1}^{C} n_{ij} log \frac{n_{ij} n}{n_i^{(1)} n_j^{(2)}}}{\sqrt{\sum_{i=1}^{C} n_i^{(1)} log \frac{n_i^{(1)}}{n} \sum_{j=1}^{C} n_j^{(2)} log \frac{n_j^{(2)}}{n}}} \quad (6)$$

Where $\mathcal{L}$ and $\mathcal{L}'$ denote ground-truth and detected community partition by algorithm respectively, C denotes the number of communities, n denotes the number of nodes, n_{ij} denotes the number of nodes of ground community i that are divided in community j in detected community partition, $n_i^{(1)}$ denotes the number of nodes in knowed community i, and $n_j^{(2)}$ denotes the number of nodes in detected community j.

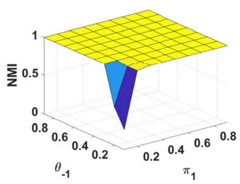
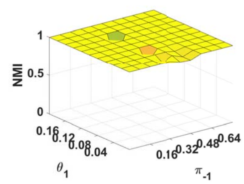
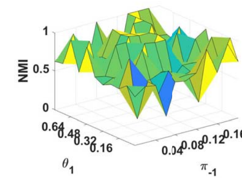
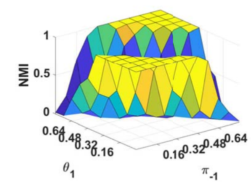
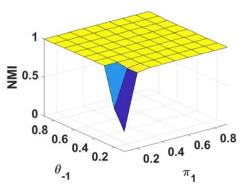
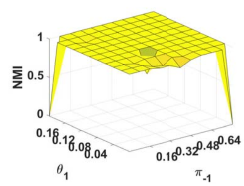
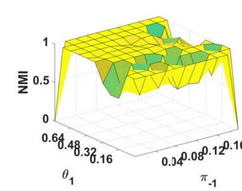
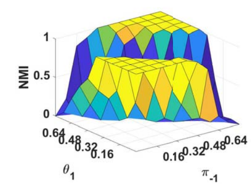
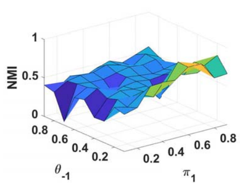
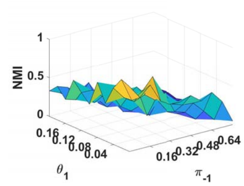
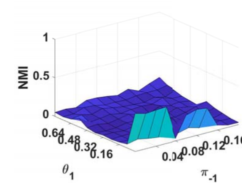
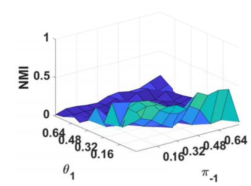
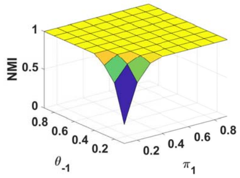
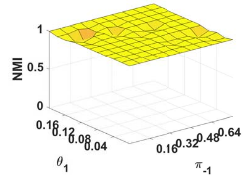
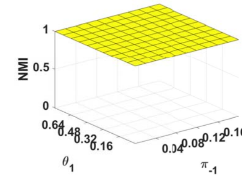
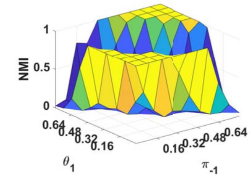
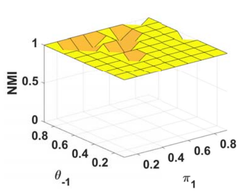
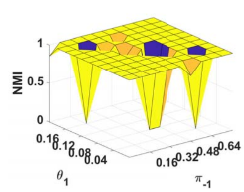
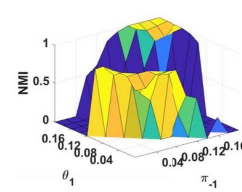
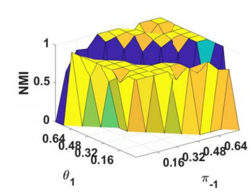
Comparison methods: To test the performance of DouNMF, we design comparisons experiments with other three state-of-the-art approaches for community discovery. The three state-of-the-art methods are FEC [4], SISN [10] and SSL [9].

In addition to the above three methods, we also did an ablation study to verify the effectiveness of adding indegree matrix to the non-negative matrix factorization for the constraint community indicator vector. In the DouNMF (without K) method, we simply decompose the adjacency matrix into basis matrix and weight matrix by NMF, without adding indegree information matrix, and set the optimal values of parameters λ and γ to be 1.5 and 6.4, respectively.

Comparison experiments were designed on four signed network models, each with 121 signed networks. First we generate signed networks on each network model, and then calculate the NMI value of each algorithm on each network to form a three-dimensional (3D) graph. Each point of 3D surfaces represents a NMI value acquired in a fixed network with a given value of Π and Θ. In the BN model, the x-axis and y-axis of the 3D graph represent the prior possibility of positive intra-link and negative inter-link respectively. In the other three UN models, the x-axis and y-axis of the 3D graph represent the prior possibility of positive inter-link and negative intra-link. These two kinds of links represent the frustration of the signed network.

The results of the comparison experiment of community discovery method on the artificial signed networks are shown in table I. In the BN, which have no positive inter-link and negative intra-link, our DubleNMF algorithm, DouNMF (without K) without indegree information matrix K and SISN performance in detecting community are weak in the initial state of most connectionless networks. As the number of links increases, community discovery performance is always the best. SSL algorithm performance fluctuates several times in networks with multiple connections, but is generally excellent. In the UN-I, which is intra-link dense and inter-link sparse, our algorithm performs better than others. In the UN-II, which is intra-link sparse and inter-link dense, SISN algorithm performance is the best, our DouNMF algorithm appears a slide in networks with proportion of negative intra-link tend to the middle which is the community structure in the singed network that tends to the chaotic state but there is a big improvement over the original DouNMF (without K) method by adding indegree information matrix. And the SSL algorithm performance decreases greatly when the

Table I
COMPARISON EXPERIMENT RESULTS OF FOUR ALGORITHMS IN COMMUNITY DISCOVERY IN DIFFERENT ARTIFICIAL DATASET.

networks / methods	BN	UN-I	UN-II	UN-III
DouNMF (without K)				
DouNMF				
FEC				
SISN				
SSL				

frustration in the network increases. In the UN-III, intra-links and inter-links are both dense, the performance of all algorithms is not good in the networks with the highest frustration, because the dense noise connection in these networks has a much greater impact on network balance than the previous UN network models. But the performance of these algorithms is superior in networks that are not the noisiest in this kind of network model.

Finally, We can find that FEC algorithm performance is not good in all these signed network model due to the random selection of initial node and the uncertain length of the local random walk in the first step of FEC. We find that our DouNMF algorithm is better than others in signed networks that are more real-world like UN-I. In addition, through the comparison with the DouNMF (without K) method, we prove that adding the indegree information matrix into the non-negative matrix factorization improves the accuracy of community discovery.

2) Validation of link prediction :

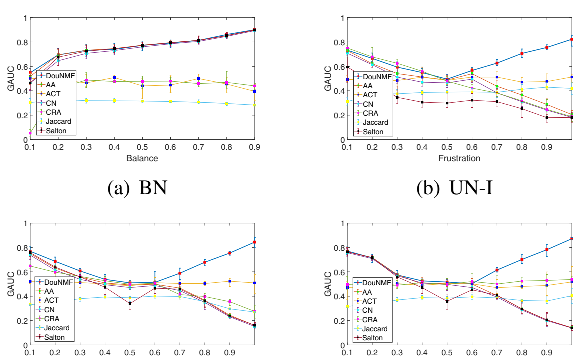
Validation metrics: We use GAUC (Generalized AUC over +1, 0 and 1) [20] to measure the overall ranking performance, formulated as:

$$\frac{1}{|P|+|N|}\left(\frac{1}{|U|+|N|}\sum_{a_i\in P}\sum_{a_j\in U\cup N} I\left(L\left(a_i\right)>L\left(a_j\right)\right)+\right.$$
$$\left.\frac{1}{|U|+|P|}\sum_{a_i\in N}\sum_{a_j\in U\cup P} I\left(L\left(a_i\right)<L\left(a_j\right)\right)\right) \tag{7}$$

where $|P|$, $|N|$ and, $|U|$ represent the number of positive edges, negative edges and un-edges 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. As the extension of AUC, GAUC defines a ranking score considering the three kinds of link status.

Comparison methods : The performance of DouNMF was validated with respect to the link prediction in signed networks. We choose six well-known index methods for computing node similarity as Comparison methods: AA, ACT, CN, CRA, Jaccard and Salton [21].

We use the standard 5-fold cross validation for the experiments of link prediction. Fig. 1 shows the performance of our DouNMF with the other six link prediction algorithms on the four network models respectively. Fig. 1(a) represents the BN model, where the x-aixs represents the balance level of the signed network, which is the proportion of positive intra-links and negative inter-links. We find that the network balance level is positively related to the performance of most tested link prediction algorithms. Fig. 1(b) - (d) represent the UN models, where the x-aixs represents the frustration in the signed network, which is the proportion of positive inter-links and negative intra-links. It can be found that when the frustration in these three kinds of the singed network tends to be intermediate, that is, when the proportion of positive and negative intra-links and inter-links tend to be the same, the performance of most link prediction algorithms that be tested are the lowest. This is because the structure of the signed network, in this case, is characterized by a chaotic state, where links of nodes in these signed networks tend to random at this time. In other states, our algorithm is superior to other comparison methods.



(a) BN (b) UN-I

(c) UN-II (d) UN-III

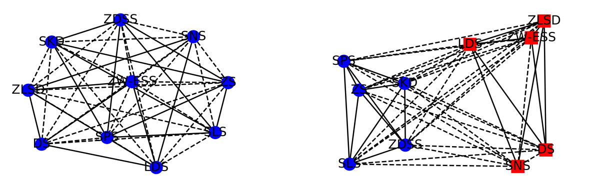
Figure 1. GAUC of link prediction in different artificial datasets.

B. Experiments on real-world signed networks

We compared DouNMF with other six approaches in real-world signed networks to validate the accuracy and effectiveness in community discovery and link prediction.

1) Validation of community discovery:

Slovene parliamentary party network [22]: This is a relational network about ten parties in Slovene parliament in 1994, which has 2 communities. In the community discovery, we only retain the sign of link in the network and ignore the weight of links. Fig. 2(a) shows the connection state between nodes in Slovene parliamentary party network. In the figure, the solid links represent the positive relationships, and the dash-dot links represent the negative relationship. Fig. 2(b) shows the community partition made by our DouNMF 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).



(a) (b)

Figure 2. Slovene parliamentary party network. (a) The connection state between nodes; (b) The community partition made by DouNMF.

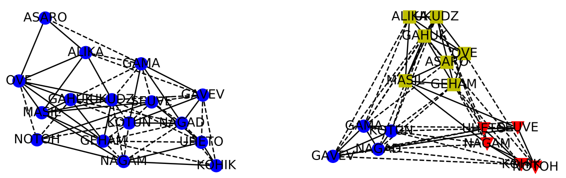
Gahuku-Gama subtribes network [23]: This network represents the culture relationship amongs subtribes of New Guinea Highland. It includes 16 subtribes from three communities. Fig. 3(a) shows the connection state between nodes in the Gahuku-Gama subtribes network, where solid edges represent the political alliance relationship and dash-dot edges represent the enmities relationship respectively. Fig. 3(b) shows the community partition made by our DouNMF 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).

Table II
LARGE SCALE SIGNED NETWORK DATASET STATISTICS .

Datasets	nodes	pos-links	neg-links	un-links
Epinions	131,828	717,667	123,705	1.73×10^{10}
Slashdot	77,357	396,378	120,197	5.98×10^{9}
Wiki	138,592	1,294,214	172,396	1.92×10^{10}
Bitcoinotc	5,881	31,714	3,547	3.46×10^{7}
Epinions@50	6,109	379,830	42,494	3.69×10^{7}
Slashdot@50	4,303	130,680	40,539	1.83×10^{7}
Wiki@50	11,047	573,423	69,012	1.21×10^{8}
Bitcoinotc@50	263	6,476	454	6,339

Table III
COMPARISON EXPERIMENT RESULTS OF FOUR ALGORITHMS IN LINK PREDICTION .

Real network \ Method	AA	ACT	CN	CRA	Jaccard	Salton	DouNMF (without K)	DouNMF
Epinions@50	0.8424	0.6629	0.8166	0.7987	0.4328	0.7654	0.8312	**0.8427**
Slashdot@50	0.4966	0.5954	0.6712	0.5454	0.3563	0.5888	0.7594	**0.7714**
Wiki@50	0.7096	0.6364	0.6714	0.6276	0.4648	0.6369	0.7022	**0.7156**
Bitcoins@50	0.7089	0.7041	0.6652	0.6265	0.4176	0.6565	0.7084	**0.7212**



(a) (b)

Figure 3. Gahuku-Gama subtribes network. (a) The connection state between nodes; (b) The community partition made by DouNMF.

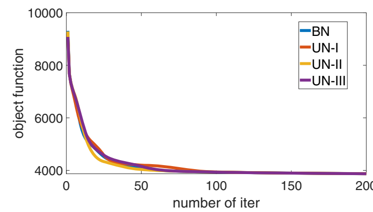


Figure 4. Convergence of our update rules

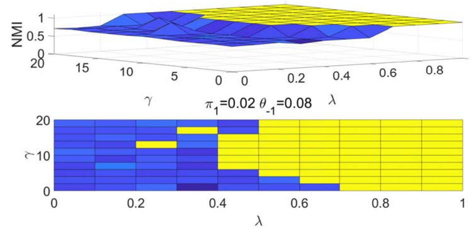


Figure 5. Sensitivity of our model to parameters

2) Validation of link prediction: As shown in Table II, we used four large scale real network datasets in experiments of link prediction, i,e., Epinions [24], Bitcoinotc [25], Slashdot [24] and Wiki [26]. 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 high degree [5]. In the experiment of link prediction in large-scale real network, we select nodes with high degree. The threshold of degree we set is 50, and the network statistics after setting are also shown in Table II where we use name@degree to represent a specific dataset e.g., Epinions@50 is the dataset about Epinions with $d \geq 50$. Table III shows the comparison results of DouNMF with other methods. We can see that the performance of our algorithm is improved to different degrees compared with other methods. And the results of comparison with DouNMF (without K) method also prove that the introduction of indegree information matrix in the process of NMF improves the performance of our algorithm in link prediction.

C. Algorithm Convergence and Parameters Sensitivity

To test the convergence of the algorithm, we perform experiments on four kinds of artificial datasets and deriving networks from each kind of network model. As shown in Fig. 4, when the number of iterations is bigger than 50, our objective function tends to converge.

In addition, we evaluate our DouNMF algorithm in terms of its sensitivity to the weight parameter λ and regularization parameter γ. Fig. 5 shows the sensitivity of our model, in which our DouNMF algorithm tends to be stable when $\lambda > 0.9$ and γ in 8−15. We choose $\lambda = 3$ and $\gamma = 10$ in our subsequent series of experiments.

V. CONCLUSIONS

Community discovery and link prediction are two of basic tasks in signed network analysis. Some of the previous algorithms rely on predefined optimization objective functions or heuristics, while others are not very explanatory. In the face of these challenges, we propose DouNMF method that converges in a reasonable number of times and has good interpretability. We decompose the signed network into positive and negative components. In order to constrain the influence of negative connection, weight parameters are added in the negative component, and the node indegree information is explicitly incorporated into the matrix factorization process so that a more accurate community matrix of the node can be obtained. Then in order to constrain the situation of overlapping communities, we add sparse constraints to our model. And we conduct a series of the experiment for community discovery and link prediction in artificial data and real-world signed network to validate the effectiveness and accuracy of our DouNMF algorithm.

ACKNOWLEDGMENT

This work was supported by the National Key R & D Program of China (2018YFC0809800) and the National Public Complaints and Proposals Administration (2019G0704).

REFERENCES

[1] M. Zanin, D. Papo, P. A. Sousa, E. Menasalvas, A. Nicchi, E. Kubik, and S. Boccaletti, "Combining complex networks and data mining: why and how," *Physics Reports*, vol. 635, pp. 1–44, 2016.

[2] L. C. Freeman, *Research methods in social network analysis*. Routledge, 2017.

[3] J. Tang, Y. Chang, C. Aggarwal, and H. Liu, "A survey of signed network mining in social media," *ACM Computing Surveys (CSUR)*, vol. 49, no. 3, p. 42, 2016.

[4] B. Yang, W. Cheung, and J. Liu, "Community mining from signed social networks," *IEEE transactions on knowledge and data engineering*, vol. 19, no. 10, pp. 1333–1348, 2007.

[5] X. Li, H. Fang, and J. Zhang, "File: A novel framework for predicting social status in signed networks," in *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

[6] Y. Li, J. Liu, and C. Liu, "A comparative analysis of evolutionary and memetic algorithms for community detection from signed social networks," *Soft Computing*, vol. 18, no. 2, pp. 329–348, 2014.

[7] P. Anchuri and M. Magdon-Ismail, "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, 2012, pp. 235–242.

[8] J. Q. Jiang, "Stochastic block model and exploratory analysis in signed networks," *Phys. Rev. E*, vol. 91, p. 062805, Jun 2015. [Online]. Available: https://link.aps.org/doi/10.1103/PhysRevE.91.062805

[9] B. Yang, X. Liu, Y. Li, and X. Zhao, "Stochastic blockmodeling and variational bayes learning for signed network analysis," *IEEE Transactions on Knowledge and Data Engineering*, vol. 29, no. 9, pp. 2026–2039, 2017.

[10] X. Zhao, B. Yang, X. Liu, and H. Chen, "Statistical inference for community detection in signed networks," *Phys. Rev. E*, vol. 95, p. 042313, Apr 2017. [Online]. Available: https://link.aps.org/doi/10.1103/PhysRevE.95.042313

[11] S. Wang, C. Aggarwal, J. Tang, and H. Liu, "Attributed signed network embedding," in *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*. ACM, 2017, pp. 137–146.

[12] H. Wang, F. Zhang, M. Hou, X. Xie, M. Guo, and Q. Liu, "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, 2018, pp. 592–600.

[13] F. Heider, "Attitudes and cognitive organization," *The Journal of Psychology*, vol. 21, no. 1, pp. 107–112, 1946, pMID: 21010780. [Online]. Available: https://doi.org/10.1080/00223980.1946.9917275

[14] K.-Y. Chiang, C.-J. Hsieh, N. Natarajan, I. S. Dhillon, and A. Tewari, "Prediction and clustering in signed networks: a local to global perspective," *The Journal of Machine Learning Research*, vol. 15, no. 1, pp. 1177–1213, 2014.

[15] A. Amelio and C. Pizzuti, "An evolutionary and local refinement approach for community detection in signed networks," *International Journal on Artificial Intelligence Tools*, vol. 25, no. 04, p. 1650021, 2016.

[16] M. Newman, "Community detection in networks: Modularity optimization and maximum likelihood are equivalent," *arXiv preprint arXiv:1606.02319*, 2016.

[17] Y. Chen, X. Wang, B. Yuan, and B. Tang, "Overlapping community detection in networks with positive and negative links," *Journal of Statistical Mechanics: Theory and Experiment*, vol. 2014, no. 3, p. P03021, 2014.

[18] D. D. Lee and H. S. Seung, "Algorithms for non-negative matrix factorization," in *Advances in neural information processing systems*, 2001, pp. 556–562.

[19] P. Jiao, W. Yu, W. Wang, X. Li, and Y. Sun, "Exploring temporal community structure and constant evolutionary pattern hiding in dynamic networks," *Neurocomputing*, vol. 314, pp. 224–233, 2018.

[20] D. Song and D. A. Meyer, "Recommending positive links in signed social networks by optimizing a generalized auc," in *Twenty-ninth AAAI conference on artificial intelligence*, 2015.

[21] P. Jiao, F. Cai, Y. Feng, and W. Wang, "Link predication based on matrix factorization by fusion of multi class organizations of the network," *Scientific reports*, vol. 7, no. 1, p. 8937, 2017.

[22] A. Ferligoj and A. Kramberger, "An analysis of the slovene parliamentary parties network," *Developments in statistics and methodology*, pp. 209–216, 1996.

[23] K. E. Read, "Cultures of the central highlands, new guinea," *Southwestern Journal of Anthropology*, vol. 10, no. 1, pp. 1–43, 1954.

[24] J. Leskovec, D. Huttenlocher, and J. Kleinberg, "Signed networks in social media," in *Proceedings of the SIGCHI conference on human factors in computing systems*. ACM, 2010, pp. 1361–1370.

[25] S. Kumar, B. Hooi, D. Makhija, M. Kumar, C. Faloutsos, and V. Subrahmanian, "Rev2: Fraudulent user prediction in rating platforms," in *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining*. ACM, 2018, pp. 333–341.

[26] S. Maniu, T. Abdessalem, and B. Cautis, "Casting a web of trust over wikipedia: an interaction-based approach," in *Proceedings of the 20th international conference companion on World wide web*. ACM, 2011, pp. 87–88.