Link Mining in Online Social Networks with Directed Negative Relationships

Baofang Hu[1,2(✉)] and Hong Wang[2]

[1] School of Information Technology, Shandong Women's University, Jinan 250014, China
hbf0509@126.com

[2] School of Information Science and Engineering, Shandong Normal University, Jinan 250014, China
30900607@qq.com

Abstract. One of the most important work to analyse online social networks is link mining. A new type of social networks with positive and negative relationships are burgeoning. We present a link mining method based on random walk theory to mine the unknown relationships in directed social networks which have negative relationships. Firstly, we define an extended Laplacian matrix based on this type of social networks. Then, we prove the matrix can be used to compute the similarities of the node pairs. Finally, we propose a link mining method based on collaboration recommendation method. We apply our method in two real social networks. Experimental results show that our method do better in terms of sign accuracy and AUC for mining unknown links in the two real datasets.

Keywords: Social networks · Link mining · Collaborative filtering · Random walk

1 Introduction

In recent years, a new type of networks called signed social networks is burgeoning, such as Slashdot news review site, Epinions consumer review site and Wikipedia vote site. The relationships in these networks can be positive (friendly, like) or negative (hostile, dislike) and are more complicated than the relationships in traditional social networks whose links are all positive. So the research in social networks needs more comprehensive analysis of the two types of relationships and the research results in traditional unsigned networks are not applicable. Link mining in signed social networks is more complicated than that in general networks and can offer us more information. We can mine not only the possibilities of future links between unrelated nodes but also the future relationships (friendly or hostile).

A popular type of methods analysing the structure of general social networks is to calculate the commute distance based on random walk theory. These methods show good performance both in terms of accuracy and time complexity [1,2].

© Springer Science+Business Media Singapore 2016
W. Che et al. (Eds.): ICYCSEE 2016, Part I, CCIS 623, pp. 428–440, 2016.
DOI: 10.1007/978-981-10-2053-7_38

We aim at mining the sign and direction of future links in directed social networks using the relationships between Laplacian matrix and commute distance. However, the commute distance should be symmetric and traditional Laplacian matrix in directed graph is asymmetric. We define an extended Laplacian matrix and prove that it can be a legal similarity distance in directed signed networks. We also mine the sign and direction of links based on the idea of collaborative filtering which is usually used in recommendation systems.

The rest of the paper is organized as follows. We review related works about link mining in signed social networks in Sect. 2 and introduce some definitions and basic theories in Sect. 3. In Sect. 4, we present our exact definition of commute distance in signed social networks. And link mining process is proposed in Sect. 5. In Sect. 6, we design different experiments and show the experimental results. Finally, we provide the conclusions in Sect. 7.

2 Related Works

Link mining in signed social networks became popular through the work of Guha and Kumar [3]. They proposed a framework of trust propagation schemes to mine the sign of links in undirected signed networks. Kunigis et al. [4,5] studied the resistance distance in signed networks and mined the friend/foe relationship, however, they did not mine the directions of the links. Leskovec et al. [6] proposed status theory in signed networks and used it to mine positive and negative links. Chiang et al. [7] gave a definition of social imbalance (MOIs) based on l-cycles in signed social networks and proposed a link mining method.

Although there are large bodies of works involving negative relationships in on-line domains, they pursue directions different from our work focus here. In this paper, we focus on commute distance property in directed signed social networks to mine not only the sign but also the direction of the unknown relationships. It is well known that the commute distance is related to the spectrum of the graph Laplacian in general undirected social networks [8]. Our work focuses on the relationship between the commute distance and graph Laplacian in directed signed networks and using the relationship to mining the unknown links.

3 Preliminaries

3.1 Mathematical Model

We begin our work by describing the method for directed unweighted signed social networks, and then extend it to weighted networks. Given a directed graph $G = (V, E)$ with a sign (positive or negative) on each edge, we let adjacent matrix $A := (a_{ij})_{i:j=1,2,\cdots,n}$ denote the adjacent matrix of graph G. The element a_{ij} indicates the sign of the edge from node i to j. That is, $a_{ij} = 1$ when i marks j as a friend, -1 when i marks j as a foe, 0 when i doesn't mark j. Because G is a directed graph, A is asymmetric and $a_{ij} \neq a_{ji}$. And because there are too many users in social networks, the matrix A is a sparse matrix.

3.2 Commute Distance in General Social Networks

Fouss et al. [8] present a method to compute the similarities of node pairs based on a Markov-chain model of random walk through the undirected general graph. They prove that the square root of average commute distance is an Euclidean distance and provide similarities between any pair of nodes, having the nice property of increasing when the number of paths connecting those elements increases and when the 'length' of paths decreases.

- **The average first hitting time** $h(i,j)$is defined as the expect time that a random walker, starting in state i hits the state j for the first time. It can be computed as shown in formula (1).

$$h(i,j) = \sum_{k \in nbs(i)}^{\mathrm{n}} p_{ik} + \sum_{k \in nbs(i)}^{\mathrm{n}} p_{ik}\, h(k,j) \tag{1}$$

  In general networks, p_{ik} means the transition probability from state i to state k. It can be expressed as shown in formula (2).

$$p_{ik} = \frac{a_{ik}}{\sum\limits_{k \in nbs(i)}^{\mathrm{n}} a_{ik}} = \frac{a_{ik}}{\sum\limits_{k=1}^{\mathrm{n}} a_{ik}} \tag{2}$$

  Where, a_{ik} is the element of the adjacency matrix A of the graph which is defined as usual as: $a_{ik} = w_{ik}$ if node i is connected to node k and $a_{ik} = 0$ otherwise.
- **The average commute distance** $n(i,j)$ is defined as the expect time that a random walker, starting in state i, enters state j for the first time and goes back to i.

Fouss et al. [8] prove the average commute distance can be computed by the Moore-Penrose pseudoinverse of the Laplacian matrix in undirected general networks as shown in formula (3).

$$n(i,j) = V_G(l^+_{ii} + l^+_{jj} - 2l^+_{ij}) \tag{3}$$

l^+_{ij} is the element of the Moore-Penrose pseudoinverse $(L^+)$of the Laplacian matrix (L) of the graph. $L = D - A$, D is the degree matrix of the graph and A is the adjacent matrix of the graph. V_G means the volume of the graph. $V_G = \sum\limits_i d_{ii}$, d_{ii} is the diagonal element of the degree matrix D.

4 Commute Distance in Directed Signed Social Networks

In this section we define the commute distance in directed signed networks and prove the distance is a legal kernel to compute the similarities of node pairs.

In signed networks, the weight a_{ik} may be negative and the transition probability p_{ik} may be negative by the previous definition in formula (2). It conflicts

with the traditional non-negative probability. However, if we simply see the negative weight a_{ik} as zero, the network would degenerate into an general network and lose a lot of important information.

Feynman in [9] proposed the definition of negative probability. In his work, probabilities may be negative under certain assumed conditions. In recent years, negative probabilities theory has been widely used in quantum field. Here, if we consider probabilities as an intermediary probability from one state to another state and showing some posture (hostile or friend), the transition probability between states can be negative.

In signed networks, transition probabilities between states in Markov chain can be involved in three situations as shown in Fig. 1. By definition, the transition probability from state S_0 to state S_n is equal to the product of every transition probability. There is no difference between the first situation and traditional general graph. The transition probability from state S_0 to state S_n in the second situation and the third may be negative or positive. It depends on the number of negative edges in these situations. The final probability is negative when the number of negative edges is $(-1)^{2k+1}(k \geq 0)$, but positive when the number is $(-1)^{2k}(k \geq 0)$. The product coincides with the structure balance theory that the foe of my foe is more likely to be my friend.

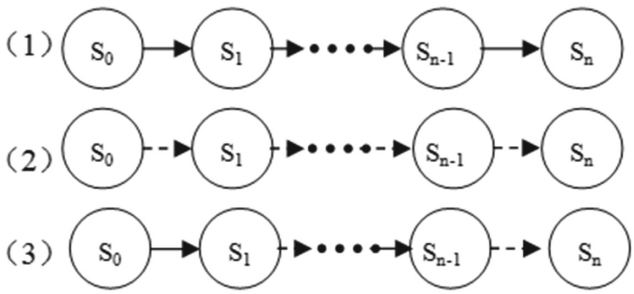

Fig. 1. Transition probabilities in signed graph(dotted line means -1, solid line means $+1$, (1)all edges are positive (2)all edges are negative (3)some edges are positive and some edges are negative)

In some special cases, the denominator in formula (2) ($\sum_{j \in nbs(i)}^{n} a_{ij}$) may be zero when the number of positive edges is equal to that of negative edges. To avoid this meaningless situation, we use the absolute value of the out-degree to calculate the sum of the weights. We denote the extended transition probability $\tilde{P}^{(1)}$ in signed networks as shown in formula (4).

$$\tilde{p}_{ik} = \frac{a_{ik}}{\sum_{j \in nbs(i)}^{n} |a_{ij}|} \ (-1 \leq \tilde{p}_{ik} \leq 1) \tag{4}$$

We also extend the definition of the diagonal degree matrix D denoted as $\tilde{D}$, and $\tilde{d}_{ii} = \sum_{(i,j)\in E} |a_{ij}|$. According to the definition of transition probability matrix, $\tilde{P}$ is equal to $\tilde{D}^{-1}A$. In most cases, $\sum_i \tilde{p}_{ik}$ in formula (4) is not equal to 1 and clashes with the property of transition probability matrix. Hence, we should normalize the matrix $\tilde{P}$.

The corresponding Laplacian matrix is asymmetric because matrix A is asymmetric. In this paper, we use the normalize Laplacian matrix which is proposed by Chung [11], $\tilde{L} = (L + L^T)/2$, to define Laplacian matrix in directed signed network as shown in formula (5).

$$\tilde{L} = \frac{L + L^T}{2} = \frac{(\tilde{D} - A) + (\tilde{D}^T - A^T)}{2} = \frac{\tilde{D} + \tilde{D}^T}{2} - \frac{A + A^T}{2} = \tilde{D}' - B \quad (5)$$

Note: $\tilde{D}^T$ is not the transposed matrix of $\tilde{D}$ and $\tilde{d}_{ii}^T = \sum_{(j,i)\in E} |a_{ji}|$.

In the paper of Fouss [8], the derivation process of the relationship between commute distance and Laplacian matrix doesn't make any requirements for the value of transition possibility matrix P. The relationship between commute distance and Laplacian matrix $\tilde{L}$ shown in formula (6) is still proper when the transition probability is negative.

$$n(i,j) = V_G(\tilde{l}_{ii}^+ + \tilde{l}_{jj}^+ - 2\tilde{l}_{ij}^+) \quad (6)$$

Now, we should prove the Laplacian matrix $\tilde{L}$ is a legal kernel and it can express the commute distance. A legal kernel should meet the Mercer's theorem and be symmetric and positive semidefinite. Undoubtedly, $\tilde{L}$ is symmetric by definition. Let's prove $\tilde{L}$ is positive semidefinite.

Theorem 1. *$\tilde{L}$ defined in this paper is positive semidefinite in any graph G.*

Proof. Let $\tilde{L}$ be the sum over the edges of graph G.

$\tilde{L} = \sum_i \sum_j \tilde{L}^{(i,j)}$. Where, $\tilde{L}^{(i,j)} \in R^{V*V}$ has four non-zero elements:

$\tilde{l}_{ii}^{(i,j)} = \tilde{l}_{jj}^{(i,j)} = \frac{|a_{ij}|+|a_{ji}|}{2} \geq \frac{|a_{ij}+a_{ji}|}{2}$, $\tilde{l}_{ij}^{(i,j)} = \tilde{l}_{ji}^{(i,j)} = -b_{ij} = -\frac{a_{ij}+a_{ji}}{2}$

Let $x \in R^V$ be a vertex column vector. Considering the bilinear of $\tilde{L}^{(i,j)}$, we find $\tilde{L}^{(i,j)}$ is positive semidefinite.

$$\begin{aligned}
&x^T\tilde{L}^{(i,j)}x \\
&= x_i^2 * \tfrac{|a_{ij}|+|a_{ji}|}{2} - 2x_ix_j * \tfrac{a_{ij}+a_{ji}}{2} + x_j^2 * \tfrac{|a_{ij}|+|a_{ji}|}{2} \\
&\geq x_i^2 * \left|\tfrac{a_{ij}+a_{ji}}{2}\right| - 2x_ix_j * \tfrac{a_{ij}+a_{ji}}{2} + x_j^2 * \left|\tfrac{a_{ij}+\tilde{a}_{ji}}{2}\right| \\
&= \left|\tfrac{a_{ij}+a_{ji}}{2}\right| (x_i - \operatorname{sgn}\left(\tfrac{a_{ij}+a_{ji}}{2}\right) x_j)^2 \\
&\geq 0 \\
&x^T\tilde{L}x = \sum_i \sum_j x^T\tilde{L}^{(i,j)}x \geq 0
\end{aligned}$$

Hence $\tilde{L}$ is positive semidefinite.

The extended Laplacian matrix $\tilde{L}$ can be used to calculate the similarity of node pairs. The time complexity in calculating the inverse of Laplacian matrix is $O(n^2)$. The time cost of commute distance method is huge for social networks which have millions of users. Matrix factorization method such as singular value decomposition (SVD) [12] is one of the effective ways to reduce the computational cost.

5 Link Mining Based on Collaborative Recommendation

Our proposed link mining method is based on the idea of collaborative filtering. We regard the target node as an item and the edges from the top-k nodes to the target node as the ratings given by the users. The top-k similarities of node i are used to predict the edge form i to j as shown in formula (7).

$$r(i,j) = \frac{\sum\limits_{m \in i'stop-knodes} sim(i,m) * a_{mj}}{\sum\limits_{m \in i'stop-knodes} sim(i,m)} \tag{7}$$

Where, $r(i,j)$ is denoted as average attitude (friendly/hostile) from the top-k nodes of i to the node j. $sim(i,m)$ is nodes similarity which can be calculated with the commute time proposed in formula (6). $sim(i,j) = n(i,j)$. a_{mj} is the element of the adjacent matrix A of the directed signed network.

6 Experiments

6.1 Experiment Process

The link mining algorithm proposed in this paper consists of three steps:

(1) Compute the extended Laplacian matrix $\tilde{L}$ as shown in formula (5) and it's Moore-Penrose pseudo-inverse.
(2) Compute the commute distance as shown in formula (6).
(3) Mine the direction and sign of the node pair in test set.

We compare experimentally our algorithm with four existing link mining algorithms, the low rank modelling with matrix factorization algorithm [7], transitive node similarity algorithm [13] and resistance distance method [4] denoted as LR-ALS, FriendTNS+ and A-sym, respectively. Henceforth, our proposed link mining approach based on commute distance similarity is denoted as CSLP.

6.2 Datasets and Metrics

To evaluate the performance of the algorithms, we adopt two real social signed networks: the Slashdot Zoo dataset and the Epinions dataset (downloaded from snap.stanford.edu). The two real datasets show high local clustering coefficients

and low average shortest path lengths. These features can be mainly discovered in small-world networks.

In the experiment process, we adopt rand walk methodology to select 4000 nodes from each dataset and divide each dataset into two sets: (i) the training set E^T is regarded as known information and, (ii) the test set E^P is used for testing and no information in the test set is allowed to be used for link mining. We evaluate and compare these algorithms using a 10-fold cross-validation methodology.

In this paper, we use sign accuracy, AUC, precision and recall as evaluation metrics.

- **Sign accuracy.** Sign accuracy is the ratio of the number of right edges in the predicting results, which have the same signs as the corresponding edges in the test set, to the number of edges in the test set E^P.
- **AUC.** AUC is equivalent to the area under the receiver-operating characteristic (ROC) curve. It is the probability that a randomly chosen missing edge (an edge in E^P) is given a higher similarity value than a randomly chosen non-existent edge (an edge in $E - E^T$, where E denotes the universal set). In the implement process, among n times of independent experiments, if there are m times the missing edge having higher similarity value and p times the missing edge and non-existent edge having the same similarity value, we define AUC, as follows: $AUC = (m + 0.5 * p)/n$.
- **Precision.** Precision is the ratio of the number of right edges in the predicting results, having the right signs and direction, to the number of edges in the predicting results. As there is no existing method for predicting links' direction, the precision measure is only for the method proposed in this paper.
- **Recall.** Recall is the ratio of the number of right edges (both sign and direction) in the predicting results to the number of edges in the test set E^P. This measure is also only for the method proposed in this paper.

In this section we compare CLSP with other methods in terms of sign accuracy and AUC. Sign prediction accuracies of various methods are calculated with different values of k and the accuracies are averaged by 10-fold cross validation. The detailed results are shown in Fig. 2. In our algorithm, k is the top-k similarities. In FriendTNS+, k means the top-k transitive similarities. k is equal to the reduced dimension in LR-ALS and A-sym algorithm. The line charts in Fig. 2 show the accuracy for signs with different values of k. We can see that CLSP consistently achieves the highest accuracy for most of thresholds T in two real datasets, and the CLSP algorithm gets obviously higher sign accuracy in the Slashdot dataset.

We also compare CLSP with other algorithms in terms of AUC as shown in Fig. 3. We use a pure chance predictor as baseline algorithm which simply randomly selects pairs of nodes to be friends. The AUC value of pure chance predictor is 0.5. We use the AUC metric, which pays attention to an algorithm's overall ability to rank all the missing links over non-existent ones. We plot a

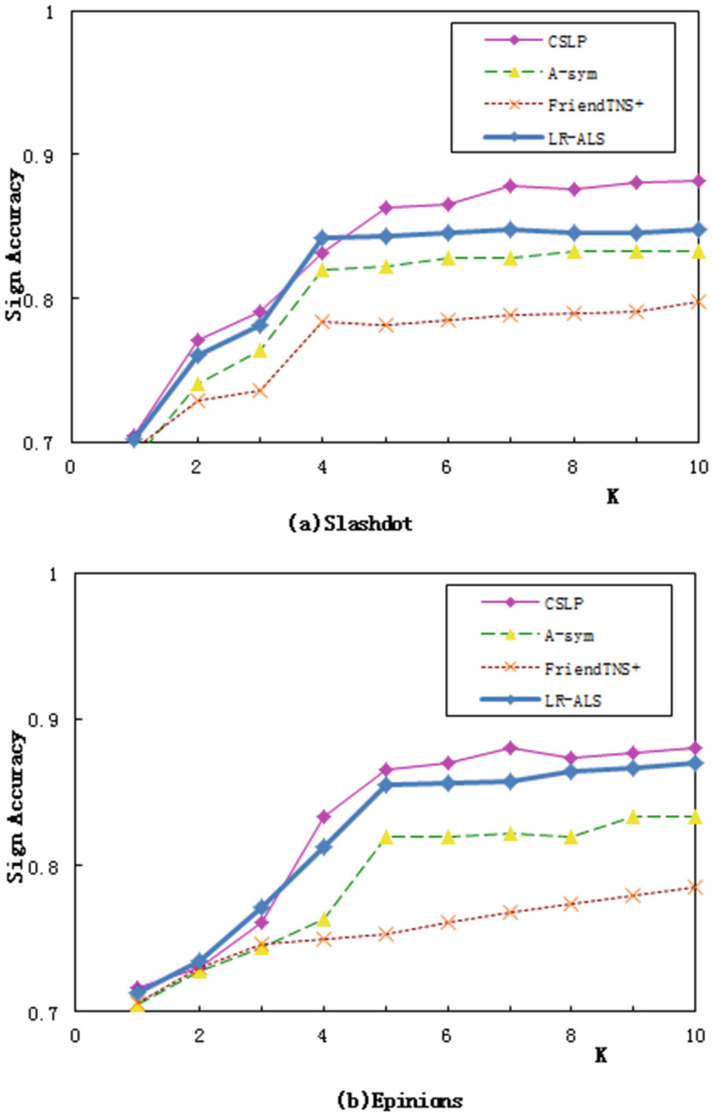

Fig. 2. Sign accuracy comparison of CSLP, A-sym, FriendTNS+ and the LR-ALS algorithm for Slashdot (b) Epinions datasets.

curve for AUC vs. the fraction of observed edges used in the training set. As shown, CLSP does better than pure chance and other algorithms, indicating that it is a strong predictor of missing structure. The main reason is the method seizes the edges' sign and direction and the link predicting process is based on the idea of collaborative filtering.

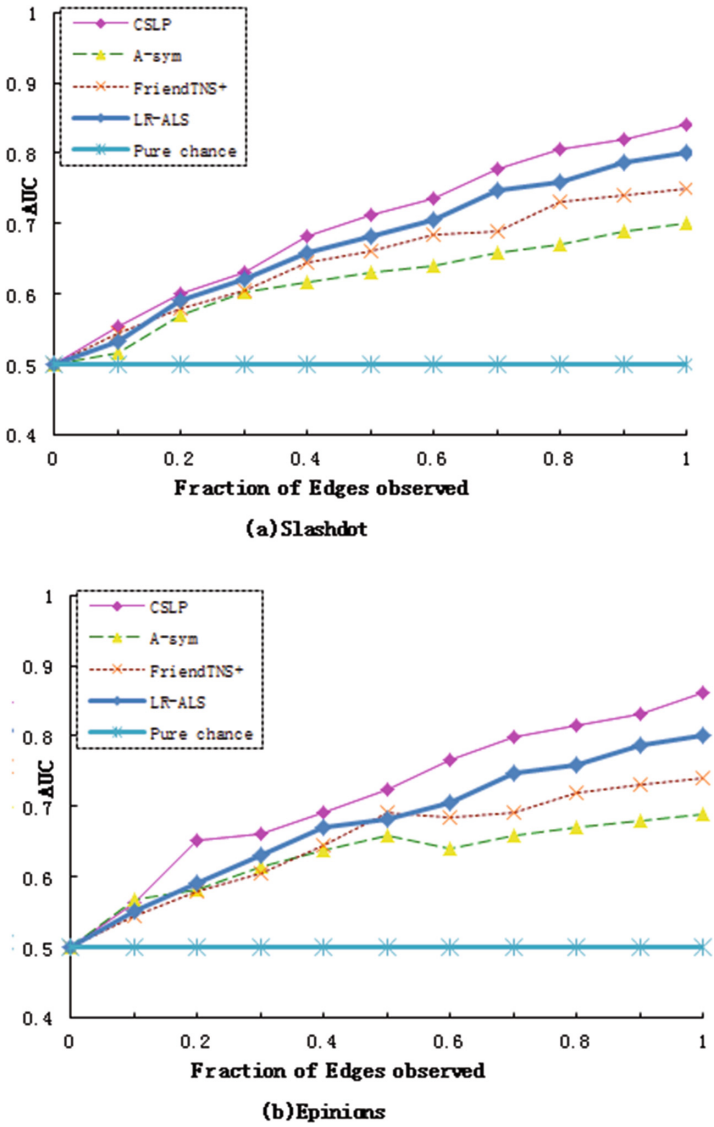

Fig. 3. AUC comparison of CSLP, A-sym, FriendTNS+, LR-ALS algorithm and pure-chance for (a)Slashdot (b) Epinions datasets.

We present the precision performance of CLSP when we take into account both right direction and sign in predicting results. As shown in Fig. 4, the precision is lower than sign accuracy shown in Fig. 2 at corresponding different levels of k. However, the precision is close to 80 percent in the two real datasets. In the Slashdot dataset, the precision is high to 0.7813 while it is 0.8010 in Epinions dataset.

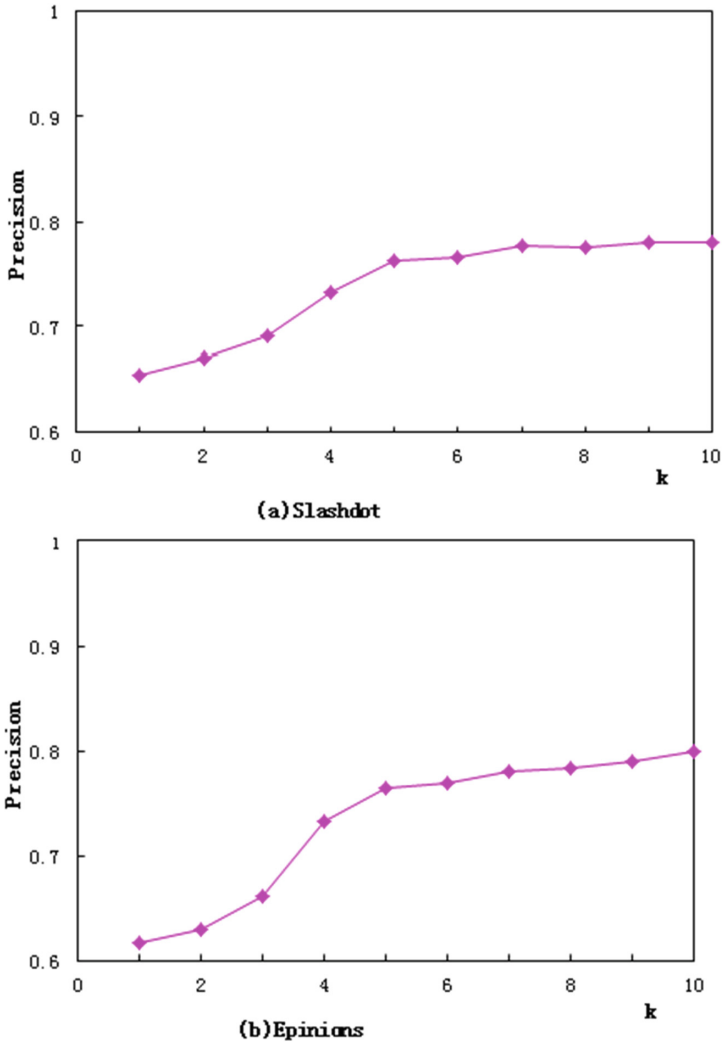

Fig. 4. Precision (right sign and direction) in CLSP proposed in this paper for (a)Slashdot (b) Epinions datasets.

Next, we proceed to examine the performance of our algorithm in terms of recall. The experiment result is shown in Fig. 5. The maximum value of recall is 0.4332 in the Slashdot dataset while it is 0.430 in the Epinions dataset.

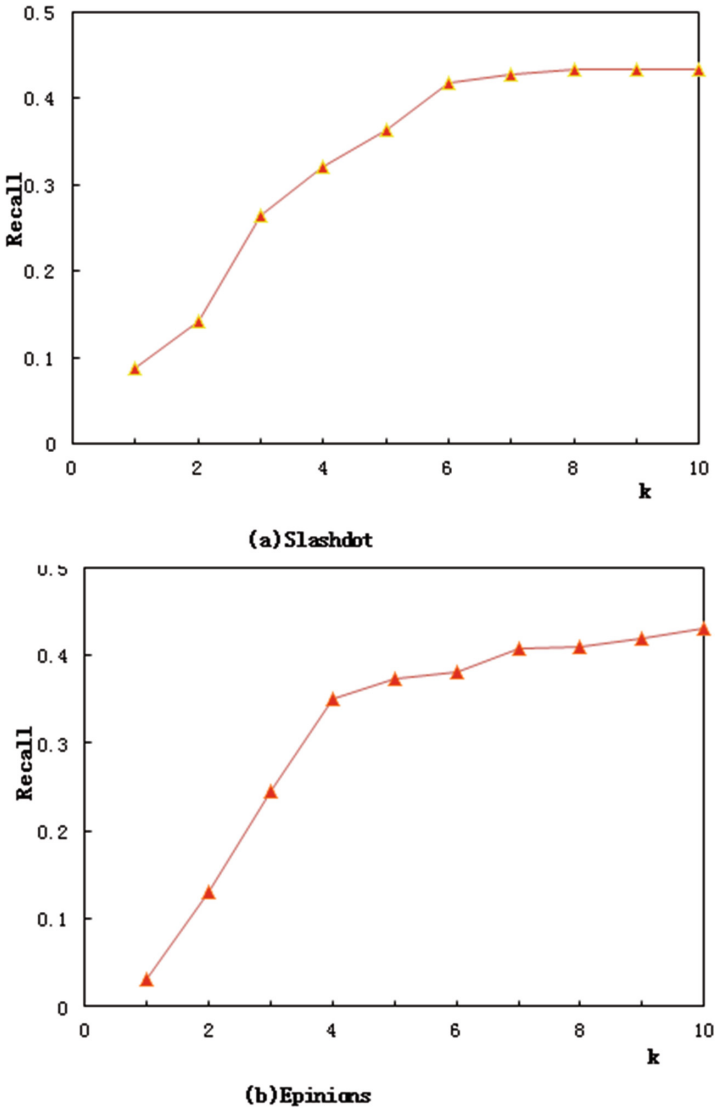

Fig. 5. Recall (right sign and direction) in CLSP proposed in this paper for (a)Slashdot (b) Epinions datasets.

7 Conclusion

We introduce a definition of the commute distance similarity in directed signed networks and use the similarity to mine the direction and sign of links. Our core mining process is based on the idea of collaborative filtering between friendships and links. The study extends the research approach of link mining and can be

the necessary supplement of link mining in directed networks. The experiment results show the method gaining better performance in terms of sign accuracy and AUC measures than several existing algorithms.

In the future, we will continue our research on link mining in directed signed networks. With further study, we will explore the refinement of corresponding parameters and evaluation measure [15]. In addition, we will extend the predicted information by considering the bi-direction link prediction [14].

Acknowledgement. This work is supported by National Natural Science Foundation under Grant (No. 61373149, 61472233, 61572300), Technology Program of Shandong Province under Grant (No.2014GGB01617, ZR2014FM001), Taishan Scholar Program of Shandong Procince(No.TSHW201502038), Exquisite course project of Shandong Province (No. 2012BK294, 2013BK399, and 2013BK402), and Education scientific planning project of Shandong province (No. ZK1437B010).

References

1. Leicht, E.A., Holme, P., Newman, M.E.J.: Vertex similarity in networks. Phys. Rev. E **73**(3), 026120–026130 (2006)
2. Sarukkai, R.R.: Link prediction and path analysis using markov chains. Comput. Netw. **33**(1–6), 377–386 (2000)
3. Guha, R., Kumar, R., Raghavan, P., Tomkins, A.: Propagation of trust and distrust. In: Proceedings of the 13th International Conference on World Wide Web, pp. 403–412. ACM Press (2004) (doi:10.1145/988672.988727)
4. Kunegis, J., Lommatzsch, A., Bauckhage, C.: The Slashdot Zoo: mining a social network with negative edes. In: Proceedings of the 18th International Conference on World Wide Web, ESP, pp. 741–750. ACM Press (2009) (doi:10.1145/1526709.1526809)
5. Kunegis, J., Preusse, J., Schwagereit, F.: What is the added value of negative links in online social networks? In: Proceedings of the 22nd International Conference on World Wide Web, pp. 727–736 (2013)
6. Leskovec, J., Huttenlocher, D., Kleinberg, J.: Predicting positive and negative links in online social networks. In: Proceedings of the 19th International Conference on World Wide Web, pp. 641–650 (2010)
7. Chiang, K.-Y., Hsieh, C.-J., Natarajan, N., Dhillon, I.S., Tewari, I.S.: Prediction and clustering in signed networks: a local to global Perspective. J. Mach. Learn. Res. **15**(1), 1177–1213 (2014)
8. Fouss, F., Pirotie, A., Renders, J.M., et al.: Random-walk computation of similarities between nodes of a graph with application to collaborative recommendation. IEEE Trans. Knowl. Data Eng. **19**(3), 355–369 (2007)
9. Cartwright, D., Harary, F.: Structure balance: a generalization of Heiders theory. Psych. Rev. **63**(5), 277–293 (1956)
10. Hiley, B.J., Peat, F.D.: Quantum Implications: Essays in Honour of David Bohm. Psychology Press, London (1991)
11. Chung, F.: Laplacians and the Cheeger inequality for directed graphs. Ann. Comb. **9**(1), 1–19 (2005)
12. Koren, Y., Bell, R., Volinsky, C.: Matrix factorization techniques for recommender systems. IEEE Comput. **42**(1), 30–37 (2009)

13. Symeonidis, P., Tiakas, E., Manolopoulos, Y.: Transitive node similarity for link prediction in social networks with positive and negative links. In: Proceedings of the 2010 ACM Conference on Recommender Systems, pp. 183–190 (2010)
14. Wang, H., Yuan, W., Yu, X.: Bi-direction link prediction in dynamic multi-dimension networks. J. Comput. Inform. Syst. **10**(3), 1333–1340 (2014)
15. Zheng, Q., Skillicorn, D.B.: Spectral embedding of signed networks. In: SDM (2015)