Block Modelling and Learning for Structure Analysis of Networks with Positive and Negative Links

Xuehua Zhao[1], Hua Chen[1], Xueyan Liu[2], Xu Tan[3(✉)], and Wenzhuo Song[2]

[1] School of Digital Media, Shenzhen Institute of Information Technology, Shenzhen 518172, China
lcrlc@sina.com

[2] College of Computer Science and Technology, Jilin University, Changchun 130012, China
dyyzlxy@163.com, wzsong17@gmail.com

[3] School of Software Engineering, Shenzhen Institute of Information Technology, Shenzhen 518172, China
tanx@sziit.edu.cn

Abstract. Currently, many community mining methods for signed networks with positive and negative links have been proposed, however, these methods can only efficiently find the community of signed networks and unable to find other structure, such as bipartite, multipartite and so on. In this study, we present a mathematically principled community mining method for signed networks. Firstly, a probabilistic model is proposed to model the signed networks. Secondly, a variational Bayesian approach is deduced to learn the proximation distribution of model parameters. In our experiments, the proposed method is validated in the synthetic and real-word signed networks. The experimental results show the proposed method not only can efficiently find communities of signed networks but also can find the other structure.

Keywords: Social networks · Community mining · Block modelling

1 Introduction

Signed networks usually are composed of the nodes, positive links and negative links, in which the nodes represent the individuals, the positive links represent like, trust or support relationship and the negative links represent dislike, distrust or oppose relationship [13,14]. In contrast to the unsigned networks [4,6,8], the signed networks may contain more information by extending the single relationship to the positive and negative relationships. Structure analysis is an important problem in the network studies since network structures are closely related to the functions and evolution of systems. Community structure, which is the dense subnetwork within a larger network, is the best-studied structure in networks.

W. Liu et al. (Eds.): KSEM 2018, LNAI 11062, pp. 396–402, 2018.
https://doi.org/10.1007/978-3-319-99247-1_35

In general, the links in communities are dense but the links between communities are sparse [9]. Because there are the negative links in the signed networks, the communities in the signed networks also show another characteristic that is most of the links in communities are the positive links and most of the links between communities are the negative links. To analyze the signed networks, until now, many methods have been proposed to find the communities in the signed networks. The representative methods is as follows: Doreian and Mrvar proposed a frustration-based method (referred to as DM) [3]. Traag et al. proposed a modularity-optimization-based algorithm for signed networks [11]. Yang et al. proposed a fast method based on Markov stochastic process (referred to as FEC) [12]. Anchuri et al. proposed a generalized spectral method for signed network partition [1]. For these methods, their common drawback is it is very difficult to design a good objective. To address the above problems, Zhao et al. proposed an EM-based community detection method for the signed networks [14]. Yang et al. proposed a signed stochastic block model and its variational Bayes learning algorithm for the signed networks [13]. However, these methods mainly focus on the community structure.

Block modelling is a form of statistical inference for the networks. The idea of block modelling for the network analysis is to find the structure of networks by fitting a specific block model to a network. Based on this idea, in this paper, we present a mathematically principled community mining method for signed networks. Firstly, a probabilistic model is proposed to model the signed networks, Secondly, a variational Bayesian approach is deduced to learn the proximation distribution of model parameters. In our experiments, the proposed method is validated in the synthetic and real-word signed networks. The experimental results show the proposed method not only can efficiently find community of signed networks but also can find the other structure.

2 Model and Method

Let $\boldsymbol{a}$ denote the adjacency matrix of the signed network N containing n nodes. The element a_{ij} is equal to 1, -1 or 0 if there is a positive, negative or no link between the node i and the node j. Suppose all of the nodes are divided into K groups and the nodes in the same group have the similar connection patten with the nodes of other groups. The proposed model is defined as follows

$$X = (K, \boldsymbol{z}, \boldsymbol{\omega}, \boldsymbol{\pi}) \tag{1}$$

where K is the number of groups. $\boldsymbol{\omega}$ is a K-dimension vector in which the element ω_k denotes the probability that a node is assigned to the group k, and $\sum_{k=1}^{K} \omega_k = 1$. $\boldsymbol{\pi}$ is a $K \times K \times 3$ matrix, where π_{lq1}, π_{lq2} and π_{lq3} denote the probability that there is a positive link, no link or negative link between a pair of nodes in the group l and q, respectively. In addition, the proposed model contains an indicating variable (or latent variable) $\boldsymbol{z}$, which is the $n \times K$ matrix containing the group information of nodes. $z_{ik} = 1$ if the node i is assigned to the group k, otherwise $z_{ik} = 0$.

Given the parameter $\boldsymbol{\omega}$, the probability distribution of $\boldsymbol{z}$ is as follows

$$p(\boldsymbol{z}|\boldsymbol{\omega}) = \prod_{i=1}^{n}\prod_{k=1}^{K}\omega_k^{z_{ik}} \tag{2}$$

Given $\boldsymbol{z}$, a_{ij} follows the following multinomial distribution with parameter $\boldsymbol{\pi}$:

$$p(a_{ij}|\boldsymbol{z},\boldsymbol{\pi}) = \prod_{l,q=1}^{K}\prod_{h=1}^{3}\pi_{lqh}^{z_{il}z_{jq}\delta(a_{ij},2-h)} \tag{3}$$

where $\delta(x,y)$ is Kronecker function, if $x=y$, the function value is 1, otherwise the value is zero.

When the priors of the model parameters $(\boldsymbol{\pi},\boldsymbol{\omega})$ are specified, we can describe the proposed model in a full Bayesian framework. Since $p(z_i|\boldsymbol{\omega})$ and $p(a_{ij}|\boldsymbol{z},\boldsymbol{\pi})$ satisfy the multinomial distribution, respectively, we can select the Dirichlet distribution as their conjugate prior distributions, as follows

$$p(\boldsymbol{\omega}|\boldsymbol{\rho}^0 = \{\rho_1^0,...,\rho_K^0\}) = \mathrm{Dir}(\boldsymbol{\omega};\boldsymbol{\rho}^0) \tag{4}$$

$$p(\pi_{lq}|\boldsymbol{\eta}_{lqh}^0 = \{\eta_{lq1}^0,\eta_{lq2}^0,\eta_{lq3}^0\}) = \mathrm{Dir}(\pi_{lq};\boldsymbol{\eta}_{lqh}^0) \tag{5}$$

where $\boldsymbol{\rho}_q^0$ and $\boldsymbol{\eta}_{lqh}^0$ are the hyperparameters. In the full Bayesian framework, the parameters $\boldsymbol{\pi}$ and $\boldsymbol{\omega}$ can be regarded as the random variables which follow the distributions with their respective hyperparameters.

To analyze the structure of network, we need to learn the parameters of model, then analyze the networks based on the learned values of parameters. Since the posterior distribution of $\boldsymbol{z}$, under the condition of data and model parameters, cannot be explicitly derived as an input required, we adopt the variational Bayesian approach [2,7] to learn the approximate distributions of parameters and variable.

The log-likelihood $\mathcal{L}(N)$ of the network N can be decomposed into two terms

$$\mathcal{L}(N) = \mathcal{L}(q(\cdot)) + KL(q(\cdot)||p(\cdot|N)) \tag{6}$$

In Eq. 6, $KL(q \parallel p)$ denotes the Kullback-Leibler divergence between the two distributions of $q(\boldsymbol{z},\boldsymbol{\pi},\boldsymbol{\omega})$ and $p(\boldsymbol{z},\boldsymbol{\pi},\boldsymbol{\omega}|N)$. $p(\boldsymbol{z},\boldsymbol{\pi},\boldsymbol{\omega}|N)$ is the true posterior distribution of the variables $\boldsymbol{z}$ and the parameters $(\boldsymbol{\pi},\boldsymbol{\omega})$ given the network N, $q(\boldsymbol{z},\boldsymbol{\pi},\boldsymbol{\omega})$ is an approximation of the true posterior distribution. $\mathcal{L}(q(\cdot))$ is called the lower bound of $\mathcal{L}(N)$. The Kullback-Leibler vergence satisfies $KL(q \parallel p) \geq 0$, with equality if, and only if, $q(\cdot) = p(\cdot)$.

The variational approach aims at optimizing a lower bound of $\mathcal{L}(N)$ by approximating the true distributions of the parameters and variable. To obtain a computationally tractable algorithm, we use mean field approximation, in which we assume the posterior $q(\boldsymbol{z},\boldsymbol{\pi},\boldsymbol{\omega})$ is a fully factorized approximation, which is written as follows

$$q(\boldsymbol{z},\boldsymbol{\pi},\boldsymbol{\omega}) = q(\boldsymbol{\pi})q(\boldsymbol{\omega})\prod_{i=1}^{n}q(z_i) \tag{7}$$

where $q(z_i)$, $q(\boldsymbol{\pi})$ and $q(\boldsymbol{\omega})$ denote the distributions of variables z_i, $\boldsymbol{\pi}$ and $\boldsymbol{\omega}$, respectively.

Next, we need to seek the distributions of $q(z_i)$, $q(\boldsymbol{\pi})$ and $q(\boldsymbol{\omega})$, which make the lower bound $\mathcal{L}(q(\cdot))$ largest. This requires us to deduce the expressions of the distributions $q(z_i)$, $q(\boldsymbol{\pi})$ and $q(\boldsymbol{\omega})$.

According to variational Bayes, the optimal distribution $q(z_i)$, $q(\boldsymbol{\omega})$ and $q(\boldsymbol{\pi})$ are the following multinomial or Dirichlet distribution, respectively.

$$q(z_i) = M(z_i; 1, \tau_{i1}, ..., \tau_{iK}) \tag{8}$$

where τ_{ik} is the probability of node i belonging to group k, and satisfies:

$$\tau_{il} \propto e^{\psi(\rho_l)-\psi(\sum_{l=1}^{K}\rho_l)} \times \prod_{j\neq i}^{n}\prod_{q=1}^{K}\left(e^{\tau_{jq}\sum_{h=1}^{3}\delta(a_{ij},2-h)(\psi(\eta_{lqh})-\psi(\sum_{h=1}^{3}\eta_{lqh}))}\right) \tag{9}$$

where $\psi(\cdot)$ is *digamma* function.

$$q(\boldsymbol{\omega}) = \mathrm{Dir}(\boldsymbol{\omega}; \boldsymbol{\rho}), \qquad \rho_q = \rho_q^0 + \sum_{i=1}^{n}\tau_{iq} \tag{10}$$

$$q(\boldsymbol{\pi}) = \prod_{l,q}\mathrm{Dir}(\boldsymbol{\pi}_{lq}; \boldsymbol{\eta}_{lq}) \tag{11}$$

For $q \neq l$, the hyperparameter η_{qlh} $(h = \{1, 2, 3\})$ is given by

$$\eta_{lqh} = \eta_{lqh}^0 + \sum_{i\neq j}^{n}\tau_{il}\tau_{jq}\delta(a_{ij}, 2-h) \tag{12}$$

For $\forall q$, the hyperparameter η_{qqh} $(h = \{1, 2, 3\})$ is given by

$$\eta_{qqh} = \eta_{qqh}^0 + \sum_{i<j}^{n}\tau_{iq}\tau_{jq}\delta(a_{ij}, 2-h) \tag{13}$$

The Eqs. 9, 10, 12 and 13, build the main steps of our algorithm. We iterate to update these equations to convergence. Finally, the values of the learned hyperparameters $(\boldsymbol{\tau}, \boldsymbol{\eta})$ can be used to analyze the structure of signed networks. Since the time complexity of the proposed algorithm is mainly determined by calculating Eqs. 9, 10, 12 and 13, the total time complexity of the algorithm is $O(K^2n^2)$.

3 Experiments

The proposed algorithm is called SASN here. SASN is validated in the synthetic and real-world networks in our experiments. We also make comparisons with other four algorithms which are respectively DM [3], SSL [13], FEC [12] and

SISN [14]. In our experiments, the normalized mutual information (NMI) [14] is used to evaluate the performance of the algorithms.

Firstly, we generate synthetic networks by the generate model in Ref. [12]. The model parameters are set as follows: $(4, 32, 32, 0.5, 0.05 * p-, 0)$ and varying $p-$ from 0 to 0.5 with the interval 0.05. The larger the value of $p-$ is, the more the negative links in the communities are. The results of five algorithms running this type of signed networks are shown in Fig. 1. We can see that, all the NMI values of the SASN and SSL are 1 when $p-$ varies from 0 to 0.5. This indicates our method and SSL can correctly find the communities in the networks.

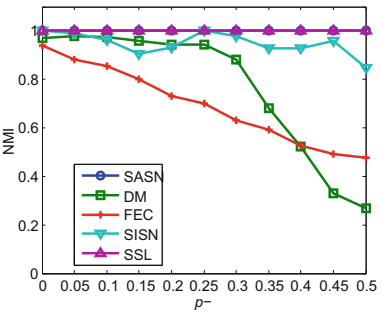

Fig. 1. Results of five algorithms.

Secondly, we generate the networks with coexisting structure. The type of networks is generated according to the following way. First, all the nodes are divided into four groups, each of which includes 32 nodes. Then, the links within or between the groups are generated according to the following $\pi_{a,b}$ value, where a and b denote the labels of groups. $\pi_{11} = \{0.6, 0.1, 0.3\}$, $\pi_{12} = \{0.1, 0.2, 0.7\}$, $\pi_{13} = \{0.1, 0.2, 0.7\}$, $\pi_{14} = \{0.1, 0.2, 0.7\}$, $\pi_{22} = \{0.2, 0.1, 0.7\}$, $\pi_{23} = \{0.01, 0.4, 0.59\}$, $\pi_{24} = \{0.01, 0.4, 0.59\}$, $\pi_{33} = \{0.01, 0.01, 0.98\}$, $\pi_{34} = \{0.01, 0.4, 0.59\}$, $\pi_{44} = \{0.01, 0.01, 0.98\}$, and other π is zero. The positive, no and negative links between two nodes within or between groups follow the multinomial distribution with parameter π. Figure 2 illustrates the adjacency matrix of randomly generated network according to the above parameter set.

The results of five algorithms running in this type of signed networks are shown in Fig. 3. The NMI values of the results for the SASN, DM, FEC, SISN and SSL are 1, 0.8641, 0, 0.8827 and 0.8338, respectively. This indicates the SASN has more excellent performance in such networks with the coexisting structure than other four algorithms.

For the real-world signed networks, we select two real-world networks with ground truth community structure to validate the proposed algorithm. The selected signed network are Slovene parliamentary party network (SPPN) [5] and Gahuku-Gama subtribes network (GGSN) [10], respectively. For the SPPN, the nodes in the SPPN are divided into two communities and the results of our

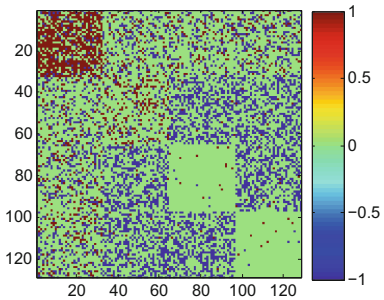

Fig. 2. Adjacency matrix of network.

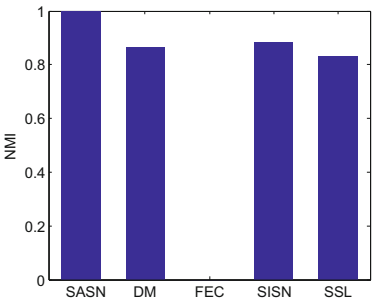

Fig. 3. Results of five algorithms.

algorithm are consistent with their ground truth. For the GGSN, the nodes in the GGSN are divided into three communities and the results of our algorithm are consistent with their ground truth.

4 Conclusions

In this paper, we present a mathematically principled community mining method for the signed network. Firstly, based on block modelling idea, we propose an probability model for the signed network, which can efficiently model the well-known structure. Secondly, we deduce the specific equations of parameters in the variational Bayesian framework. The proposed method is validated in the synthetic and real-world signed networks. The experimental results show the proposed method not only can efficiently find community of signed networks but also can find the other structure.

Acknowledgments. This work is funded by the MOE (Ministry of Education in China) Project of Humanities and Social Sciences (17YJCZH261, 17YJCZH157), National Science Foundation of China (61571444), Guangdong Province Natural Science Foundation (2016A030310072), Special Innovation Project of Guangdong Education Department (2017GKTSCX063), and Special Funds for the Cultivation of Scientific and Technological Innovation for College Students in Guangdong (pdjh2018b0862).

References

1. Anchuri, P., Magdon-Ismail, M.: Communities and balance in signed networks: a spectral approach. In: 2012 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM), pp. 235–242. IEEE (2012)
2. Blei, D.M., Kucukelbir, A., McAuliffe, J.D.: Variational Inference: A Review for Statisticians. ArXiv e-prints, January 2016
3. Doreian, P., Mrvar, A.: A partitioning approach to structural balance. Soc. Netw. **18**(2), 149–168 (1996)
4. Ghoshal, G., Mangioni, G., Menezes, R., Poncela-Casanovas, J.: Social system as complex networks. Soc. Netw. Anal. Min. **4**(1), 1–2 (2014)

5. Kropivnik, S., Mrvar, A.: An analysis of the slovene parliamentary parties network. In: Developments in Statistics and Methodology, pp. 209–216 (1996)
6. Liu, X., Wang, W., He, D., Jiao, P., Jin, D., Cannistraci, C.V.: Semi-supervised community detection based on non-negative matrix factorization with node popularity. Inf. Sci. **381**, 304–321 (2017)
7. Murphy, K.P.: Machine Learning: A Probabilistic Perspective. MIT Press, Cambridge (2012)
8. Newman, M.: Networks: An Introduction. Oxford University Press, Oxford (2010)
9. Newman, M.E.: Communities, modules and large-scale structure in networks. Nat. Phys. **8**(1), 25–31 (2012)
10. Read, K.E.: Cultures of the central highlands, New Guinea. Southwest. J. Anthropol. **10**(1), 1–43 (1954)
11. Traag, V.A., Bruggeman, J.: Community detection in networks with positive and negative links. Phys. Rev. E **80**(3), 036115 (2009)
12. Yang, B., Cheung, W., Liu, J.: Community mining from signed social networks. IEEE Trans. Knowl. Data Eng. **19**(10), 1333–1348 (2007)
13. Yang, B., Liu, X., Li, Y., Zhao, X.: Stochastic blockmodeling and variational bayes learning for signed network analysis. IEEE Trans. Knowl. Data Eng. **PP**(99), 1 (2017)
14. Zhao, X., Yang, B., Liu, X., Chen, H.: Statistical inference for community detection in signed networks. Phys. Rev. E **95**(4), 042313 (2017)