Frontiers of Information Technology & Electronic Engineering
www.jzus.zju.edu.cn; engineering.cae.cn; www.springerlink.com
ISSN 2095-9184 (print); ISSN 2095-9230 (online)
E-mail: jzus@zju.edu.cn

Science Letter:

Quantifying multiple social relationships based on a multiplex stochastic block model*

Mincheng WU, Zhen LI, Cunqi SHAO, Shibo HE†‡

College of Control Science and Engineering, Zhejiang University, Hangzhou 310027, China

†E-mail: s18he@zju.edu.cn

Received Nov. 8, 2020; Revision accepted Dec. 8, 2020; Crosschecked Jan. 14, 2021; Published online Jan. 30, 2021

Abstract: Online social networks have attracted great attention recently, because they make it easy to build social connections for people all over the world. However, the observed structure of an online social network is always the aggregation of multiple social relationships. Thus, it is of great importance for real-world networks to reconstruct the full network structure using limited observations. The multiplex stochastic block model is introduced to describe multiple social ties, where different layers correspond to different attributes (e.g., age and gender of users in a social network). In this letter, we aim to improve the model precision using maximum likelihood estimation, where the precision is defined by the cross entropy of parameters between the data and model. Within this framework, the layers and partitions of nodes in a multiplex network are determined by natural node annotations, and the aggregate of the multiplex network is available. Because the original multiplex network has a high degree of freedom, we add an independent functional layer to cover it, and theoretically provide the optimal block number of the added layer. Empirical results verify the effectiveness of the proposed method using four measures, i.e., error of link probability, cross entropy, area under the receiver operating characteristic curve, and Bayes factor.

Key words: Social network; Multiplex network; Stochastic block model
https://doi.org/10.1631/FITEE.2000617 **CLC number:** TP3-05

1 Introduction

The development of information technology has made it easy for people to make social connections all over the world using online social networks (OSNs) such as Twitter and Facebook. Notably, the structure of OSNs is always the OR-aggregation of various social ties. For example, an individual's friends in an OSN might be his/her club members, school mates, relatives, and so on, but these various relationships are often difficult to distinguish on the online platform. It is challenging to achieve a deep understanding of the mechanism of social networks without including more information. Fortunately, in addition to the edges (friendships) in OSNs, the annotations that describe the attributes (e.g., gender and occupation) of nodes are available. These natural attribute labels allow us to study the network structure using group-based network models, among which the stochastic block model (SBM) has been paid great attention (Holland et al., 1983). In SBM, nodes are clustered in groups (blocks) to provide a more general description of their preferential attachment in a network, which is encoded by link probability between groups.

Typically, a node in an OSN has various attributes, which means multiple roles. Newman and Clauset (2016) developed a method to improve the performance of community detection using the marginal information in an SBM. However, the complicated relationships in an OSN cannot be well described by a single-layer network. Instead, the

‡ Corresponding author
* Project supported by the National Natural Science Foundation of China (No. 61731004)
ORCID: Mincheng WU, https://orcid.org/0000-0002-9966-8427; Shibo HE, https://orcid.org/0000-0002-1505-6766

multiple social ties can be more accurately captured by a multiplex network. A multiplex network contains several layers, and each layer has the same node set. Edges in different layers represent different types of relationships, indicating that the multiplex network is a kind of heterogeneous information network (Sun and Han, 2012). Assuming that each layer of a multiplex network is described by a single SBM, a multiplex network forms a multiplex stochastic block model (MSBM). Studies have revealed that many real-world networks are more likely to have a multiplex structure, and that the MSBM outperforms the classical SBM in terms of the accuracy of link prediction (Vallès-Català et al., 2016; Lacasa et al., 2018).

2 Precision and loss function

Extending an SBM to an MSBM requires the definition of a joint distribution of variables that describe link probabilities among multiple groups (Barbillon et al., 2017). We denote the observed aggregate network by A^{O}, and let $\boldsymbol{X}^1, \boldsymbol{X}^2, \cdots, \boldsymbol{X}^K$ be K networks, indicating the relationships of K attributes. In network $\boldsymbol{X}^\alpha$ $(\alpha = 1, 2, \cdots, K)$, specifically, there are Q_α blocks (groups or clusters), where the partition is determined by data annotations. To avoid confusion, specifically, we use Latin letters $i, j, \cdots$ to indicate nodes and Greek letters $\alpha, \beta, \cdots$ to indicate layers. Assume that $\forall (q_\alpha, l_\alpha) \in \Psi_\alpha = \{1, 2, \cdots, Q_\alpha\}^2$, and for any node pair (i, j) that has the attribute (q_α, l_α) in the α^{th} layer, their relationship is denoted by $X_{ij}^\alpha(q_\alpha, l_\alpha) \in \{0, 1\}$. In the MSBM, every pair of nodes (i, j) has a joint partition $\boldsymbol{z} = ((q_1, l_1), (q_2, l_2), \cdots, (q_K, l_K))$, where $\boldsymbol{z} \in Z = \Psi_1 \times \Psi_2 \times \cdots \times \Psi_K$. For any multiplex link relationship $\boldsymbol{w} \in \{0, 1\}^K$, we define a joint distribution on $\boldsymbol{X}^{1:K} = (\boldsymbol{X}^1, \boldsymbol{X}^2, \cdots, \boldsymbol{X}^K)$ as

$$P(\boldsymbol{X}_{ij}^{1:K}(\boldsymbol{z}) = \boldsymbol{w}) = \pi_{\boldsymbol{z}}^{\boldsymbol{w}}, \tag{1}$$

satisfying

$$\sum_{\boldsymbol{w} \in \{0,1\}^K} \pi_{\boldsymbol{z}}^{\boldsymbol{w}} = 1. \tag{2}$$

After that, we denote

$$Z_{ql}^\alpha = \{\boldsymbol{z} | \boldsymbol{z} \in Z, z(\alpha) = (q, l)\} \tag{3}$$

as a set of partitions for blocks q and l in layer α and

$$W^\alpha = \{\boldsymbol{w} | \boldsymbol{w} \in \{0, 1\}^K, w(\alpha) = 1\}. \tag{4}$$

Thus, for any $\boldsymbol{z} \in Z_{ql}^\alpha$,

$$p_{ql}^\alpha = \sum_{\boldsymbol{w} \in W^\alpha} \pi_{\boldsymbol{z}}^{\boldsymbol{w}}, \tag{5}$$

which describes the link probability between any two blocks q and l in layer α.

The core assignment in this research on multiplex networks is to precisely estimate parameters in the MSBM, given the aggregation of a multiplex network and available annotations. Assume that there is a link probability p_{ij} that describes the existence of a link between any pair of two nodes i and j. Our goal is to estimate this parameter, denoted by $\hat{p}_{ij}$. Then we define the loss function ε by

$$\varepsilon = -\sum_{i,j} p_{ij} \ln \hat{p}_{ij}, \tag{6}$$

which is also called the cross entropy between p_{ij} and $\hat{p}_{ij}$. Then we use ε to measure the model precision by estimation methods. As we know, the maximum likelihood estimation of $\hat{p}_{ij}$ is equivalent to the process of minimizing the loss function ε (Burg, 1972).

3 Adding an independent layer

A direct way to improve model precision is to add an extra layer to the present structure that introduces joint variables simultaneously. Fig. 1 demonstrates intuitively why adding a layer can increase the precision. In Fig. 1, the observed network is a social network of students from two different schools. First, we construct the stochastic block model by the school partition, and the model successfully reveals that students are more likely to be connected with their school fellows. However, it fails to discover precisely the intra-school relationships. Then we add a layer where students are clustered by gender on the former layer. This new model not only detects intra-fellow preferences, but also reveals that friendships are more likely to form between students of the same gender, and reduces error links as well.

By Eq. (6), we can see that the block partitions are the critical factors that determine precision once the number of layers and the partitions are given. In this case, the loss function ε reaches its minimum if and only if the number of nodes in every block is the same. Intuitively, this suggests that we are supposed to cluster nodes independently, i.e., to minimize the correlations of blocks in different layers. Here, two layers are called independent if the

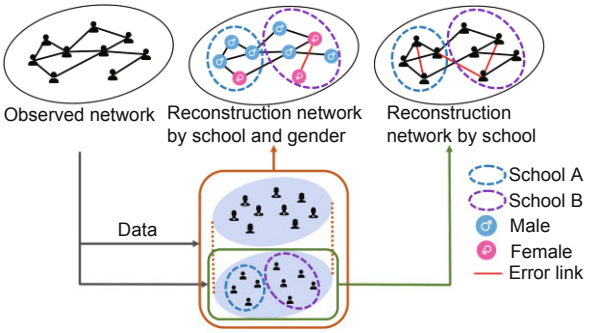

Fig. 1 An intuitive illustration of our method by adding a layer to improve model precision

proportion of nodes by division defined in one layer is the same as that in any block of another layer.

We denote the original model as M_{O}, which has K given layers $\boldsymbol{X}^1, \boldsymbol{X}^2, \cdots, \boldsymbol{X}^K$. Then we construct a new layer $\boldsymbol{X}^{\mathrm{f}}$ that is independent of all layers in M_{O}, forming a new model denoted by M_{F}. The constructed layer $\boldsymbol{X}^{\mathrm{f}}$ has Q_{f} blocks, and is called the functional layer. Mathematical analysis indicates that the optimal number of blocks Q_{f}^* of the constructed functional layer $\boldsymbol{X}^{\mathrm{f}}$ can be obtained by

$$Q_{\mathrm{f}}^* = \underset{Q_{\mathrm{f}}}{\operatorname{argmin}} \left| 2^K - 1 - \frac{Q_{\mathrm{f}}(Q_{\mathrm{f}}+1)}{2} \right|. \tag{7}$$

We briefly prove the result as follows: On one hand, recall that the model has $2^K \prod_{\alpha=1}^{K} \frac{Q_\alpha(Q_\alpha+1)}{2}$ variables (i.e., $\pi_{\boldsymbol{z}}^{\boldsymbol{w}}$) in Eq. (1). Furthermore, there are $\prod_{\alpha=1}^{K} \frac{Q_\alpha(Q_\alpha+1)}{2}$ constraints in Eq. (2), and $\sum_{\alpha=1}^{K} \frac{Q_\alpha(Q_\alpha+1)}{2} \prod_{\beta\neq\alpha} \left[\frac{Q_\beta(Q_\beta+1)}{2} - 1\right]$ constraints in Eq. (5). Thus, the degree of freedom n_{d} equals the number of variables minus the number of constraints, i.e.,

$$\begin{aligned} n_{\mathrm{d}} =& (2^K - 1) \prod_{\alpha=1}^{K} \frac{Q_\alpha(Q_\alpha+1)}{2} \\ & - \sum_{\alpha=1}^{K} \frac{Q_\alpha(Q_\alpha+1)}{2} \prod_{\beta\neq\alpha} \left[\frac{Q_\beta(Q_\beta+1)}{2} - 1 \right]. \end{aligned} \tag{8}$$

On the other hand, the number of parameters in the model M_{F} is

$$n_{\mathrm{p}} = \frac{Q_{\mathrm{f}}(Q_{\mathrm{f}}+1)}{2} \prod_{\alpha=1}^{K} \frac{Q_\alpha(Q_\alpha+1)}{2}. \tag{9}$$

The number of parameters is expected to be greater than or equal to the degree of freedom, i.e., $n_{\mathrm{p}} \geq n_{\mathrm{d}}$. Therefore, we have

$$2^K - 1 - \frac{Q_\alpha(Q_\alpha+1)}{2} \geq 0, \tag{10}$$

which finally leads to Eq. (7). In summary, this result gives the threshold for precision improvement of the method by extending the model M_{O} to M_{F}.

4 Empirical results

Next, we verify our results using simulations. In the following simulations, based on synthetic data, the undirected observed network has $N = 1000$ nodes, and is aggregated by six independent layers using the OR-aggregation mechanism. We assume that the first three layers have two blocks in each layer, which are consistent with the layers defined by annotations, i.e., the pre-determined layers composing the model M_{O}, where $K = 3$. The other three layers are assumed to be noisy networks.

In the model M_{F}, $\boldsymbol{X}^{\mathrm{f}}$ is independent of all the three two-block layers in M_{O}, and we increase Q_{f} from 2 to 4. According to Eq. (7), the optimal number of blocks Q_{f}^* in the functional layer is 3. The overall precision of model parameters is measured by the loss function ε defined in Eq. (6), and we consider the average loss by $\bar{\varepsilon} = \varepsilon/N^2$. We also measure the method's effectiveness in improving the precision of link probability. We denote $D = \{(0,0),(0,1),(1,1)\}$, $L = \{1,2,3\}$, and $P = \{p_d^\alpha | d \in D, \alpha \in L\}$, indicating the sets of link probabilities. Also, we define the mean error of link probability (ELP) by

$$\mathrm{ELP} = \frac{1}{|D||L|} \sum_{d\in D, \alpha\in L} \frac{|p_d^\alpha - \hat{p}_d^\alpha|}{p_d^\alpha}. \tag{11}$$

It is important to reconstruct the multiplex networks, where links disappear in the process of aggregation, using the estimated parameter $\hat{\theta}_{M_{\mathrm{F}}}$ (Here, the parameter $\hat{\theta}_{M_{\mathrm{F}}}$ indicates all the link probabilities in the model M_{F}). We measure the models' performances in network reconstruction with 30% spurious links by calculating the area under the receiver operating characteristic curve (i.e., AUC), which has been widely adopted to evaluate classification methods (Storey, 2003).

Figs. 2a–2c show the ELP, AUC, and $\bar{\varepsilon}$, where the number of blocks Q_{f} in model M_{F} is denoted by the superscript. According to ELP and $\bar{\varepsilon}$, model M_{F}

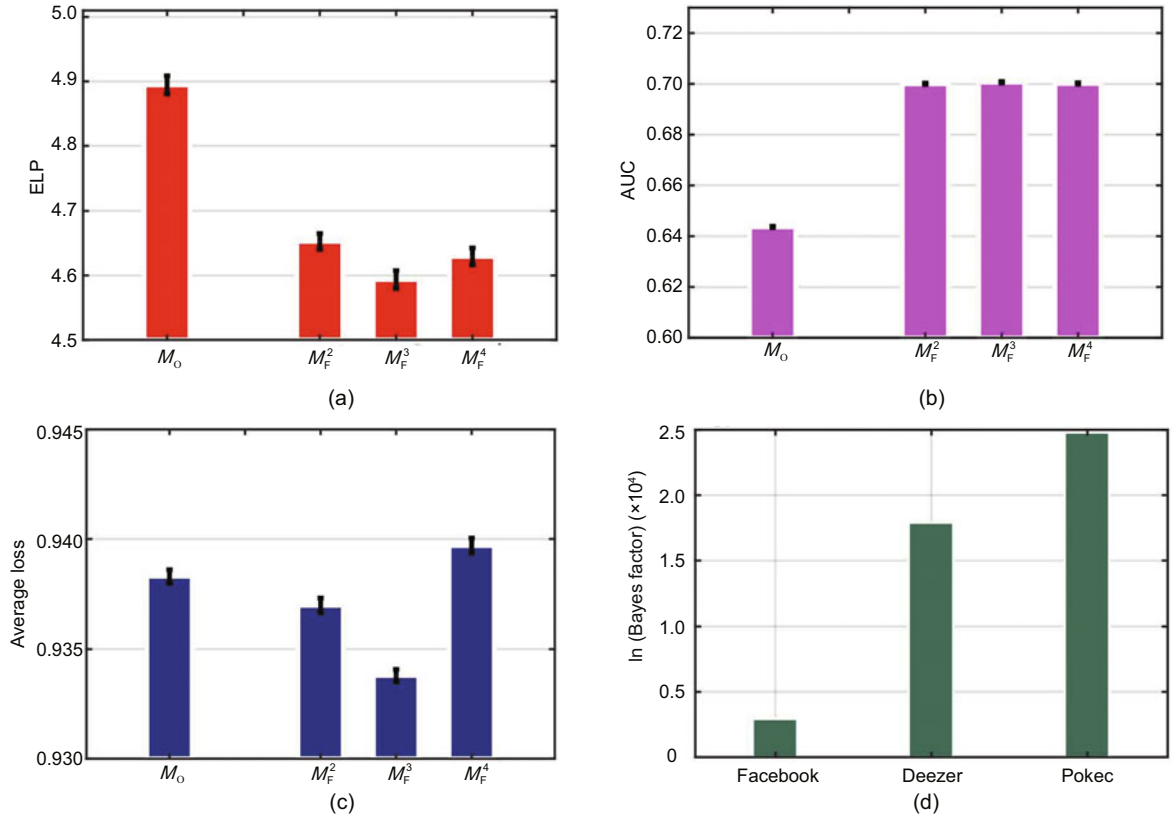

Fig. 2 Simulation results of the error of link probability (ELP) (a), AUC (b), average loss (c), and Bayes factors of different models (d)

with $Q_\mathrm{f} = 3$ has the best overall performance, which verifies the optimal Q_f^* in Eq. (7). The results also show that the AUCs of the model M_F with different numbers of blocks are comparatively stable, while the original model M_O has a significant defect.

Finally, we use the Bayes factor b as the criterion for comparing different models' performances:

$$b = P(A^\mathrm{O}|M_\mathrm{F})/P(A^\mathrm{O}|M_\mathrm{O}). \tag{12}$$

Thus, the observed network A^O is more likely to be generated by the model M_F if b is larger than one. The datasets are from three frequently used social APPs, Facebook, Deezer, and Pokec (Leskovec and Krevl, 2016), where the social relationships are naturally assumed to have an OR-aggregation structure and have 4039, 41 773, and 50 000 nodes, respectively. Based on the data, the original model M_O has three layers, and in each layer, there are two groups defined by user annotations, i.e., user attributes (such as gender and age). Then we construct an independent two-block layer on M_O, forming the model M_F. The Bayes factors are shown in Fig. 2d, where all factors are much larger than one. Thus, the real-world networks are more likely to be generated by the proposed model M_F.

5 Conclusions

Quantifying multiple social relationships of real-world networks via limited observations is currently of great interest. In this letter, we considered a multiplex stochastic block model by observing the OR-aggregation of a multiplex network, where blocks are determined by natural annotations of nodes (e.g., age and gender). Within this framework, the model precision was defined by the loss function ε, indicating the cross entropy of parameters between the data and model. Thus, we proposed a method to improve precision when estimating parameters in the multiplex stochastic block model with its aggregation. We analyzed the number of variables and the degree of freedom to derive the optimal number of blocks in the functional layer, which is deeply related to the number of layers K. Finally, our theoretical analysis for

adding an independent functional layer was verified using empirical results, having broad applications in social and engineering systems (Chen et al., 2017; Zhou et al., 2018).

Contributors

Mincheng WU and Shibo HE designed the research. Zhen LI and Cunqi SHAO processed the data. Mincheng WU and Zhen LI drafted the manuscript. Shibo HE revised and finalized the paper.

Compliance with ethics guidelines

Mincheng WU, Zhen LI, Cunqi SHAO, and Shibo HE declare that they have no conflict of interest.

References

Barbillon P, Donnet S, Lazega E, et al., 2017. Stochastic block models for multiplex networks: an application to a multilevel network of researchers. *J R Stat Soc Ser A*, 180(1):295-314. https://doi.org/10.1111/rssa.12193

Burg JP, 1972. The relationship between maximum entropy spectra and maximum likelihood spectra. *Geophysics*, 37(2):375-376. https://doi.org/10.1190/1.1440265

Chen JM, Hu K, Wang Q, et al., 2017. Narrowband Internet of Things: implementations and applications. *IEEE Intern Things J*, 4(6):2309-2314. https://doi.org/10.1109/JIOT.2017.2764475

Holland PW, Laskey KB, Leinhardt S, 1983. Stochastic blockmodels: first steps. *Soc Netw*, 5(2):109-137. https://doi.org/10.1016/0378-8733(83)90021-7

Lacasa L, Mariño IP, Miguez J, et al., 2018. Multiplex decomposition of non-Markovian dynamics and the hidden layer reconstruction problem. *Phys Rev X*, 8(3):031038. https://doi.org/10.1103/PhysRevX.8.031038

Leskovec J, Krevl A, 2016. SNAP datasets: Stanford large network dataset collection. http://snap.stanford.edu/data

Newman ME, Clauset A, 2016. Structure and inference in annotated networks. *Nat Commun*, 7:11863. https://doi.org/10.1038/ncomms11863

Storey JD, 2003. The positive false discovery rate: a Bayesian interpretation and the q-value. *Ann Stat*, 31(6):2013-2035. https://doi.org/10.1214/aos/1074290335

Sun YZ, Han JW, 2012. Mining heterogeneous information networks: principles and methodologies. *Synth Lect Data Min Knowl Discov*, 3(2):1-159. https://doi.org/10.2200/S00433ED1V01Y201207DMK005

Vallès-Català T, Massucci FA, Guimerà R, et al., 2016. Multilayer stochastic block models reveal the multilayer structure of complex networks. *Phys Rev X*, 6(1):011036. https://doi.org/10.1103/PhysRevX.6.011036

Zhou CW, Gu YJ, He SB, et al., 2018. A robust and efficient algorithm for coprime array adaptive beamforming. *IEEE Trans Veh Technol*, 67(2):1099-1112. https://doi.org/10.1109/TVT.2017.2704610