Reliable Link Inference for Network Data with Community Structures

Lijia Ma, Jianqiang Li, *Member, IEEE,* Qiuzhen Lin, Maoguo Gong, *Senior Member, IEEE,* Carlos A. Coello Coello, *Fellow, IEEE,* and Zhong Ming

***Abstract*—Complex systems are often characterized by complex networks with links and entities. However, in many complex systems such as protein-protein interaction networks, recommender systems and online communities, their links are hard to be revealed directly, but they can be inaccurately observed by multiple data collection platforms or by a data collection platform at different times. Then, the links of the systems are inferred by the integration of the collected observations. As those data collection platforms are usually distributed over a large area and in different fields, their observations are unreliable and sensitive to the potential structures of the systems. In this paper, we consider the link inference problem in network data with community structures, in which the reliability of data collection platforms is unknown a priori and the link errors and reliability of platforms' observations are heterogeneous to the underlying community structures of the systems. We propose an Expectation Maximization algorithm for Link Inference in a network system with Community structures (short for EMLIC). The EMLIC algorithm is also used to infer the link errors and reliability of platforms' observations in different communities. Experimental results on both synthetic data and eight real-world network data demonstrate that our algorithm is able to achieve lower link errors than the existing reliable link inference algorithms when the network data have community structures.**

***Index Terms*—Link inference, network data, community structure, expectation maximization, reliability**

I. Introduction

The link structure in complex networks has become one of the most popular ways for interpreting the relationships between entities in many real-world complex systems [1], [2], including the social systems, biological systems, technological systems, economical systems, ecological systems, evolutionary systems, transportation systems, recommender systems and etc. For instances, in the social systems, links represent the interconnections, friendships, collaborations, votes and competitions among individuals [3]–[6], while in the biological networks they denote the transcription, signal transduction and metabolism processes among genes [7].

This work was supported by the Joint Funds of the National Natural Science Foundation of China under Key Program under Grant U1713212, the National Natural Science Foundation of China under Grants 61672358, 61572330 and 61772393, the Natural Science Foundation of Guangdong Province under grant 2017A030313338, and CONACyT under Grant 221551.

L. Ma, J. Li, Q. Lin and Z. Ming are with the College of Computer Science and Software Engineering, Shenzhen University, Shenzhen 518060, China.

M. Gong is with Key Laboratory of Intelligent Perception and Image Understanding of Ministry of Education, International Research Center for Intelligent Perception and Computation, Xidian University, Xi'an, Shaanxi Province 710071, China.

C.A. Coello Coello is with the Department of Computer Science, CINVESTAV-IPN, México, D.F., 07360, México.

*Corresponding author: J. Li (e-mail: lijq@szu.edu.cn).

It usually assumes that the links of a network are revealed directly [8], [9]. However, in some complex systems such as protein-protein interaction networks, recommender systems and online communities, their links cannot be revealed, but they can be inferred from other metadata, e.g., the properties and similarities of nodes, the extra observations, and etc [8], [10]–[14]. Generally, these metadata are collected by a set of platforms distributed in different fields. As they may suffer from measurement errors, including the recording errors, quantification errors, sampling bias and publication bias in the collection of metadata, it is difficult for these platforms to infer the real link structures from metadata [8], [11], [15], [16].

To detect the real link structures of systems, many studies for the link inference in network data have been presented in recent years [17]–[21]. One of the most common strategies is to measure the topological similarity of networks under the assumption of the reliable observations to links. A systematic review of these studies in inferring missing links and link recommendations could be found in [22], [23]. Note that, in some applications, it may be unreliable for the link observations of a network. For instance, in the link inference of protein-protein interaction networks, an observed link only represents the probability that a signaling pathway exists between the proteins across the link [8]. In the network crowdsourcing, the observations to a link may be varied at different times [11]. In these applications, those methods would result in the missing of real links and the adding of spurious links as they neglect the reliability of link observations.

The reliable link inference, which considers the reliability of link observations, was firstly studied in [24]. This study presented a stochastic blockmodel for link inference under the assumption that the links of a network can be observed only once. In this model, the link observation A_o is the generation of an underlying probabilistic process p_M, and the link reliability R (i.e., the probability $p_M(\mathrm{A}_{ij}=1|\mathrm{A}_o)$ that edge e_{ij} is true under A_o) is used to distinguish the missing and spurious links of networks. This link inference model has been widely generalized to discover communities and estimate the number of communities in undirected networks, directed networks, signed networks and multiplex networks [24]–[27]. Following the degree-corrected stochastic block based inference model, Ball et al. [28] presented a principled statistical approach based on the generative network model to infer the link communities in networks. In this generative network model, the existence of edges between two vertices i and j is determined by $d_i \cdot d_j \cdot \theta_{rs}$, where d_i is the degree of node i, and r and s are the clusters of nodes i and j, respectively. This model was further extended

to infer the communities and semantics of attributed networks by simultaneously considering their link topologies, node and link semantics [29], [30]. Moreover, Martin et al. [8] proposed a reliable link inference model for uncertain networks, in which the observed links are the real ones with a certain probability. This model then uses a principled maximum-likelihood method to infer both the underlying community structures and the underlying truth of links.

Recently, Newman presented a novel reliable link inference model [11], in which each link of a network is independently observed more than once by a platform and the observations to a link may conflict with each other. This model is under the assumption that the observations A_o are captured by an error model $p(\mathrm{A}_o|\Theta, \mathrm{Z})$ with error parameter Θ and the real state Z of links, and it uses an expectation maximization (EM) method to infer Z and Θ by maximizing $p(\mathrm{A}_o|\Theta, \mathrm{Z})$. Note that, in social systems, the communications between individuals can be observed in multiple platforms (e.g., Fackbook, Webchat, Email, and transportation systems). Moveover, the social systems are composed of many groups (e.g., friends, classmates and colleagues) with different communication densities. In these systems, the data collection platforms may have heterogeneous reliability of observations to different groups. Actually, the heterogeneity of platforms' observations has been widely considered into classical crowdsourcing models. It has been demonstrated by one of its classical model EMLI [31] that the aggregation accuracy of systems can be effectively improved under the setting of heterogeneous reliability. Despite recent process in heterogeneous crowdsourcing, the integration of the structure heterogeneity of networks into reliable link inference remains to be an outstanding problem.

In this paper, following the work [11], we present a general link inference model for network data, which enables a deep understanding of the impacts of the heterogeneous reliability of platforms' observations and the community structures of networks on the link inference. Similar to the work [11], the presented link inference model is also under the assumption that the links of the systems can be observed multiple times. Different from the work [11], it considers that i) the observations can be collected by multiple platforms with heterogeneous reliability; and ii) the link inference may be influenced by the community structures of systems in which nodes are densely linked in the same community whereas they are sparsely connected across different communities [32]. Generally, the community structures of networks are unknown a priori, but they can be well detected by many classical methods, including the modularity optimization [32]–[35], statistical inference [28]–[30], [36], [37], Markov dynamics [38], multiobjective optimization [39], refinement heuristic [40], and etc. Here, we adopt the modularity-based optimization method BGLL to detect the underlying communities of networks due to its high detection performance and low computational complexity. Our main contributions are summarized as follows:

1) We develop a general link inference model that takes into account the heterogeneous reliability of platforms' observations, the potential community structures and the effects of those structures on the reliability and the parameter errors of observations to links.
2) We propose a likelihood based expectation maximization method (named as EMLIC) to infer the true state of links, the reliability of platform's observations to links in different communities and the parameter errors of links in different communities.
3) We verify the performances of EMLIC through extensive experiments on both simulations and eight real-world data sets. The results suggest that EMLIC has a lower link error and a higher community preservation than the existing reliable link inference algorithms when the network data have community structures.

The rest of the paper is organized as follows. In Section II, we present our link inference system model and problem formulation. In Section III, we propose an EM algorithm for the reliable link inference in the system. Extensive simulations and experiments are given in Section IV. Finally, we conclude and briefly discuss future work in Section V.

Notations: We use italic lower-case letters and block upper-case letters to represent scalars and matrices, respectively. The upper-case letters are used to denote vectors and sets and the bold faced symbols are adopted to represent random variables. Let A be a matrix, and A_{ij} be the (i, j)-th entry of A. The operator $\mathbb{E}$ represents mathematical expectation. The operators $|S|$ and $||S||$ denote the number of elements and the sum of the absolute value of elements in S, respectively.

II. System model and problem formulation

We consider a reliable link inference system model with a certain n-node network G and q platforms (or a platform at q different times), in which this network has m uncertain links, each of which has an associated real state $\mathbf{Z}_{ij} \in \{0, 1\}$ representing the presence ($\mathbf{Z}_{ij} = 1$) or absence ($\mathbf{Z}_{ij} = 0$) of the link e_{ij} between nodes i and j, and the real link structures are hard to be revealed directly, but they can be inaccurately observed by multiple platforms or a platform at different times. For each link e_{ij}, a platform α gives an observation $\mathrm{A}_{ij}^{[\alpha]}$ indicating whether this link exists in reality. Fig. 1 gives a schematic illustration of a toy link inference system with $n = 8$ nodes and $m = 10$ links observed by q platforms. As shown in Fig. 1, the observed edges consist of two types of links: i) the real links that exist in reality and ii) the spurious links that are absent actually but they are observed by some platforms. For all observed edges, the q platforms may give conflicting observations. As the reliability of platforms is unknown a priori, it is hard to determine the real link structures of the network. Our link inference system is to estimate the real link structures Z of the network system G from the unreliable observations A collected by the q platforms.

The link inference system with an n-node network and a set of platforms corresponds to a graph $G_S = (V, E, I, \mathrm{A})$ with $|V| = n$, $|E| = m$ and $|I| = q$, where V and I are the sets of nodes $V = \{1, 2, \ldots, n\}$ and platforms $I = \{1, 2, \ldots, q\}$, respectively, and A denotes the observations of platforms to the link structures of a network. For an edge e_{ij} and a platform α, let $\mathrm{A}_{ij}^{[\alpha]} = 1$ denote that the link e_{ij} is collected or observed by the platform α and let $\mathrm{A}_{ij}^{[\alpha]} = 0$ indicate otherwise. E is the

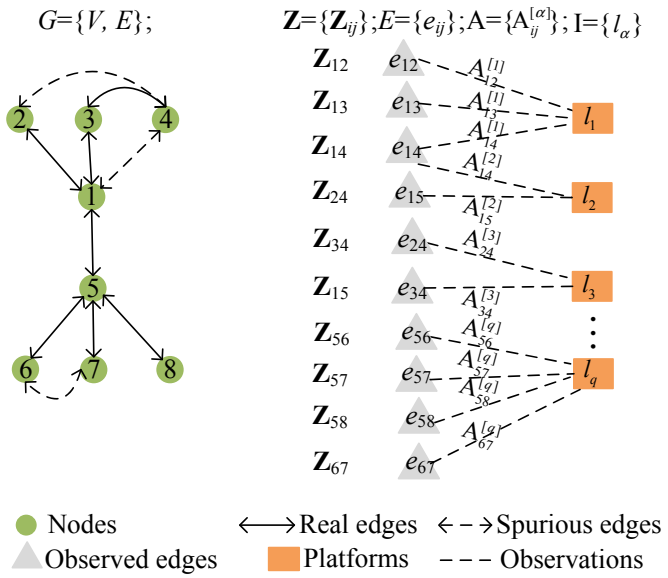


Fig. 1: Illustration example of the link inference system with $n = 8$ nodes, $m = 10$ links and q platforms. Left panel: The certain n-node network with uncertain edges, including real edges and spurious edges. Right panel: The observations A of platforms to observed edges. Each observation $\mathrm{A}_{ij}^{[\alpha]}$ denotes whether the edge e_{ij} is collected by the platform α, and each edge has an unknown real state $\mathbf{Z}_{ij}$.

set of observed edges, and each edge $e_{ij} \in E$ is observed by at least one platform, i.e., $E = \{e_{ij} | \sum_{\alpha=1}^{q} \mathrm{A}_{ij}^{[\alpha]} > 0, i, j \in V\}$.

Many real-world networks have community structures, i.e., nodes and links in the same community have similar properties whereas these in different communities show different properties. Moreover, links densely exist in the same community whereas they are sparse across different communities [32]. For instance, in protein-protein interaction networks, they have many functional modules (e.g. co-regulation, co-expression, signaling pathway, aggregating cellular, and etc), each of which corresponds to a community. It has been demonstrated that i) the links in some protein-protein interaction networks cannot be exactly extracted, but they can be collected by a set of gene methods or platforms; ii) each community has its own link structures and loss ratio of links; and iii) the difficulties to observe links may vary with their community locations. Here, we call the links in the same community and the links across different communities as the intra-community links and inter-community links, respectively.

In our link inference system, observation data are often collected by multiple platforms from different domains and the difficulties of collecting links in different communities are also different. Hence, it is typically unknown a priori for the reliability of platforms' observations depending on the two factors: whether these platforms are observing a real or spurious link and which function module the observed link comes from. As the inter-community links are sparse generally, we assume that the reliability of a platform' observations to inter-community links is only determined by the former factor. Here, we model the reliability of a platform α as

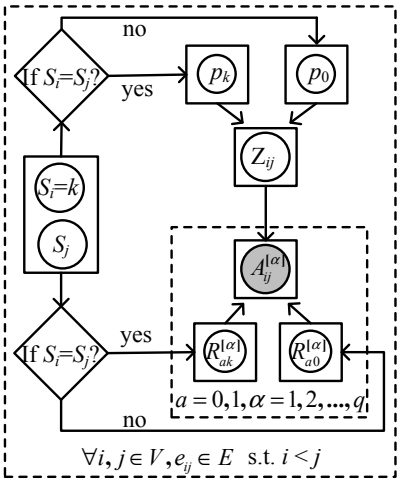


Fig. 2: A sketch of the mathematical model of reliable link inference for network data with community structures.

$R^{[\alpha]} = \{R_{1k}^{[\alpha]}, R_{10}^{[\alpha]}, R_{0k}^{[\alpha]}, R_{00}^{[\alpha]}\}$, where $k = 1, \ldots, c$, c is the number of communities of the network. Letting S_i be the community label of node i, for each $a \in \{0, 1\}$, $R^{[\alpha]}$ is represented as follows:

- $R_{ak}^{[\alpha]}$: the probability that platform α gives a true observation to a link (e.g., link e_{ij}) conditioned on $Z_{ij} = a$, $S_i = S_j$ and $S_i = k$,
- $R_{a0}^{[\alpha]}$: the probability that platform α gives a true observation to a link (e.g., link e_{ij}) conditioned on $Z_{ij} = a$ and $S_i \neq S_j$.

Here, it is unknown a priori for the $R^{[\alpha]}$ values that need to be estimated by our EMLIC method.

In our link inference system, we consider a general case of real applications, in which it may be different for the probability that an intra-community link actually exists in different communities. Moreover, as the inter-community links are sparse, it is assumed that the probability of the presence of an inter-community link is the same. Here, we model the prior probability of the existence of real links in a network G as $p(G) = \{p_k, p_0\}$, $k = 1, 2, \ldots, c$, where

- p_k: the prior probability that an intra-community link (e.g., e_{ij}) exists actually conditioned on $S_i = k$,
- p_0: the prior probability that an inter-community link (e.g., e_{ij}) exists actually.

Here, the $p(G) = \{p_k, p_0\}$ values are unknown a priori and they are estimated by our EMLIC method.

Problem Formulation. Given a link inference system model with q platforms and a network G with n nodes and m uncertain links, our problem is to find the optimal estimators of unknown parameters $\Theta = \{R, p(G)\}$, including the reliability of platforms $R = \{R^{[1]}, R^{[2]}, \ldots, R^{[q]}\}$, and the expected probability of the existence of real links $p(G)$ in the system, together with the most likely link structures Z, so as to maximize the marginal likelihood or probability of observations A.

Fig. 2 gives a detailed sketch of the mathematical model of the reliable link inference for network data with community structures. Given the potential communities S of the network G, the marginal likelihood $p(\mathrm{A}; S, \Theta)$ of observations A is

computed as follows:

$$\begin{aligned} p(\mathrm{A};S,\Theta) &= \sum_{\mathrm{Z}} p(\mathrm{A},\mathrm{Z};S,\Theta) \\ &= \sum_{\mathrm{Z}} p(\mathrm{Z};S)\cdot p(\mathrm{A}|\mathrm{Z};S,\Theta) \\ &= \sum_{\mathrm{Z}} \prod_{e_{ij}\in E} \Big[p(\mathrm{Z}_{ij};S_i,S_j)\cdot\Big(\prod_{\alpha=1}^{q} p(\mathrm{A}_{ij}^{[\alpha]}|\mathrm{Z}_{ij};S_i,S_j,\Theta)\Big)\Big] \end{aligned},$$

where $p(\mathrm{Z}_{ij};S_i,S_j)$ is the probability of the existence of a link e_{ij} connecting the nodes i and j, in which node i is in the community S_i while node j is in the community S_j, and it is computed as follows:

$$p(\mathrm{Z}_{ij};S_i,S_j) = \begin{cases} (p_k)^{\mathrm{Z}_{ij}}\cdot(1-p_k)^{(1-\mathrm{Z}_{ij})} & \text{if } S_i=S_j,\ S_i=k, \\ (p_0)^{\mathrm{Z}_{ij}}\cdot(1-p_0)^{(1-\mathrm{Z}_{ij})} & \text{if } S_i\neq S_j. \end{cases}$$

$p(\mathrm{A}_{ij}^{[\alpha]}|\mathrm{Z}_{ij};S_i,S_j,\Theta)$ is the probability of the observation of platform α to link e_{ij} conditioned on the true state of the edge e_{ij} being Z_{ij} given S_i, S_j and Θ, and it is computed as follows:

$$p(\mathrm{A}_{ij}^{[\alpha]}|\mathrm{Z}_{ij};S_i,S_j,\Theta) = \begin{cases} (R_{1k}^{[\alpha]})^{\mathrm{A}_{ij}^{[\alpha]}}\cdot(1-R_{1k}^{[\alpha]})^{(1-\mathrm{A}_{ij}^{[\alpha]})} & \text{if } \mathrm{Z}_{ij}=1,\ S_i=S_j,\ S_i=k, \\ (1-R_{0k}^{[\alpha]})^{\mathrm{A}_{ij}^{[\alpha]}}\cdot(R_{0k}^{[\alpha]})^{(1-\mathrm{A}_{ij}^{[\alpha]})} & \text{if } \mathrm{Z}_{ij}=0,\ S_i=S_j,\ S_i=k, \\ (R_{10}^{[\alpha]})^{\mathrm{A}_{ij}^{[\alpha]}}\cdot(1-R_{10}^{[\alpha]})^{(1-\mathrm{A}_{ij}^{[\alpha]})} & \text{if } \mathrm{Z}_{ij}=1,\ S_i\neq S_j, \\ (1-R_{00}^{[\alpha]})^{\mathrm{A}_{ij}^{[\alpha]}}\cdot(R_{00}^{[\alpha]})^{(1-\mathrm{A}_{ij}^{[\alpha]})} & \text{if } \mathrm{Z}_{ij}=0,\ S_i\neq S_j. \end{cases}$$

Based on the descriptions above, our link inference problem can be modeled as follows:

$$\begin{aligned} \Theta^* &\leftarrow \arg\max_{\Theta} p(\mathrm{A};S,\Theta), \\ \mathrm{Z}^* &\leftarrow \arg\max_{\mathrm{Z}} p(\mathrm{Z}|\mathrm{A};S,\Theta^*). \end{aligned} \tag{1}$$

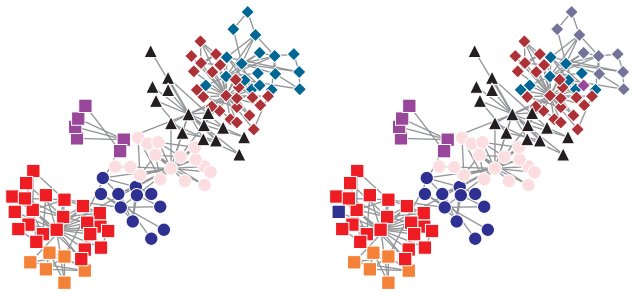
Next, we will present an EM algorithm to find the optimal values of Z and Θ.

III. Expectation Maximization for Link Inference in a Network with Communities: EMLIC

As shown in Eq. (1), the optimal estimate of Θ depends on the marginal likelihood $p(\mathrm{A};S,\Theta)$ of A. As we known, the computation of $p(\mathrm{A};S,\Theta)$ is often intractable as the latent variables in $\mathrm{Z}=\{\mathrm{Z}_{ij}\}$, $e_{ij}\in E$, can take any possible value. In this section, an EM algorithm (called as EMLIC) is presented to find the maximum likelihood estimate (MLE) of this statistic marginal likelihood model with observed data A, unknown parameter Θ and latent variable Z. The EMLIC finds the MLE of $p(\mathrm{A};S,\Theta)$ by iteratively maximizing the expectation of its log likelihood function. The likelihood $p(\mathrm{A},\mathrm{Z};S,\Theta)$ is computed as follows:

$$\begin{aligned} p(\mathrm{A},\mathrm{Z};S,\Theta) &= p(\mathrm{A}|\mathrm{Z};S,\Theta)\cdot p(\mathrm{Z};S) \\ &= \prod_{e_{ij}\in E}\Big[p(\mathrm{Z}_{ij};S_i,S_j)\cdot\Big(\prod_{\alpha=1}^{q} p(\mathrm{A}_{ij}^{[\alpha]}|\mathrm{Z}_{ij};S_i,S_j,\Theta)\Big)\Big]. \end{aligned} \tag{2}$$

As known from Eq. (2), the computation of $p(\mathrm{A},\mathrm{Z};S,\Theta)$ needs to know the community structure S of the network G. For some networks with known community structures, we can directly use their community information. However, for most complex systems, their communities cannot be observed a priori, but they can be detected based on the link structures of networks. Recently, many community detection algorithms have been proposed, and they can be used here to detect the communities of the network G. Here, we choose a classical algorithm BGLL to detect the community structure of the network G with observed link structures E, and its systematic descriptions are detailed in [33], [34]. The algorithm BGLL is chosen as i) it has a good performance on detecting real communities without knowing the number of communities a priori; ii) it generates the same community division at different trials, which removes the impacts of the instability of detected communities on the link inference in EMLIC; and iii) it has a low computational complexity, which makes it possible to detect communities in large-scale networks [41].



(a) (b)

Fig. 3: (Color online) The community structures of the SFI network detected by the BGLL algorithm. (a) the community structure of the original SFI network and (b) the community structure of spurious SFI network with 10% randomly spurious links. Nodes in different communities are plotted with different colors and shapes.

In our link inference system, the observed link structures include both real links and spurious links, and their true states are unknown a priori. Here, we detect the community structure of a network by using the BGLL algorithm [33] on its observed link structures. Recent studies have demonstrated that the existence of spurious links would not change the original community structures of a network obviously. In order to demonstrate it, we use the BGLL algorithm [33] to detect communities in both the real SFI network with 118 nodes and 200 edges and its spurious version with 10% spurious links. The detected communities are shown in Fig. 3. As shown in Fig. 3, the spurious SFI network has a similar community structure with the real one. More specifically, the real and spurious SFI networks have 8 and 9 communities, respectively, and the normalized mutual information [42] which evaluates the community similarity between the two network partitions is 0.9361. The main difference between the two network partitions is that the community (drawn by the blue diamond in Fig. 3(a)) of the real SFI network is further divided into two small communities (drawn by the purple diamond and blue diamond in Fig. 3(b)) of the spurious network. The small difference between the communities in a real network

$$\begin{aligned}\mathbf{Q}(\Theta|\Theta^{(n)}) &= \mathbf{E}_{\mathrm{Z}|\mathrm{A};S,\Theta^{(n)}}\Big[\ln p(\mathrm{A},\mathrm{Z};S,\Theta)\Big] \\ &= \sum_{e_{ij}\in E}\Big\{p(\mathrm{Z}_{ij}=1|\mathrm{A}_{ij};S_i,S_j,\Theta^{(n)})\cdot\Big[\ln\Big(p(\mathrm{Z}_{ij}=1;S_i,S_j)\cdot(\prod_{\alpha=1}^{q}p(\mathrm{A}_{ij}^{[\alpha]}|\mathrm{Z}_{ij}=1;S_i,S_j,\Theta))\Big)\Big] \\ &+ p(\mathrm{Z}_{ij}=0|\mathrm{A}_{ij};S_i,S_j,\Theta^{(n)})\cdot\Big[\ln\Big(p(\mathrm{Z}_{ij}=0;S_i,S_j)\cdot(\prod_{\alpha=1}^{q}p(\mathrm{A}_{ij}^{[\alpha]}|\mathrm{Z}_{ij}=0;S_i,S_j,\Theta))\Big)\Big]\Big\}.\end{aligned} \tag{3}$$

and a spurious network would not affect the link inference performances of EMLIC.

Based on the detected community structure S, our EMLIC algorithm finds the MLE of unknown parameter Θ by iteratively executing the following two steps: **E-step** and **M-step**.

E-step: We compute the expected log likelihood function $\mathbf{Q}(\Theta|\Theta^{(n)})$ of $p(\mathrm{A},\mathrm{Z};S,\Theta)$ as Eq. (3) in the top of this page.

Here, $p(\mathrm{Z}_{ij}=1|\mathrm{A}_{ij};S_i,S_j,\Theta^{(n)})$ is the probability of the presence of the link e_{ij} conditioned on the observation being A_{ij} given the community assignments S_i and S_j and the current estimation $\Theta^{(n)}$ of parameters, and it can be computed as follows:

$$\begin{aligned}&p(\mathrm{Z}_{ij}=1|\mathrm{A}_{ij};S_i,S_j,\Theta^{(n)}) \\ &= \frac{p(\mathrm{Z}_{ij}=1|\mathrm{A}_{ij};S_i,S_j,\Theta^{(n)})}{\sum_{\mathrm{Z}_{ij}\in\{0,\ 1\}}p(\mathrm{Z}_{ij}|\mathrm{A}_{ij};S_i,S_j,\Theta^{(n)})} \\ &= \frac{p(\mathrm{Z}_{ij}=1;S_i,S_j)\cdot p(\mathrm{A}_{ij}|\mathrm{Z}_{ij}=1;S_i,S_j,\Theta^{(n)})}{\sum_{\mathrm{Z}_{ij}\in\{0,\ 1\}}p(\mathrm{Z}_{ij};S_i,S_j)\cdot p(\mathrm{A}_{ij}|\mathrm{Z}_{ij};S_i,S_j,\Theta^{(n)})}\end{aligned} \tag{4}$$

and $p(\mathrm{Z}_{ij}=0|\mathrm{A}_{ij};S_i,S_j,\Theta^{(n)}) = 1-p(\mathrm{Z}_{ij}=1|\mathrm{A}_{ij};S_i,S_j,\Theta^{(n)})$.

M-step: We maximize the expected log function $\mathbf{Q}(\Theta|\Theta^{(n)})$, and then get Θ of the next iteration. In order to find the maximal value of $\mathbf{Q}(\Theta|\Theta^{(n)})$, we take the derivative of $\mathbf{Q}(\Theta|\Theta^{(n)})$ to each parameter in Θ and let their derivatives be equal to 0. Thereafter, for each α, k and $a\in\{0,1\}$, we get

$$\begin{aligned}(R_{ak}^{[\alpha]})^{(n+1)} &= \frac{\sum_{e_{ij}\in\Upsilon_k,\mathrm{A}_{ij}^{[\alpha]}=a}p(\mathrm{Z}_{ij}=a|\mathrm{A}_{ij};S_i,S_j,\Theta^{(n)})}{\sum_{e_{ij}\in\Upsilon_k}p(\mathrm{Z}_{ij}=a|\mathrm{A}_{ij};S_i,S_j,\Theta^{(n)})}, \\ (R_{a0}^{[\alpha]})^{(n+1)} &= \frac{\sum_{e_{ij}\in\Gamma,\mathrm{A}_{ij}^{[\alpha]}=a}p(\mathrm{Z}_{ij}=a|\mathrm{A}_{ij};S_i,S_j,\Theta^{(n)})}{\sum_{e_{ij}\in\Gamma}p(\mathrm{Z}_{ij}=a|\mathrm{A}_{ij};S_i,S_j,\Theta^{(n)})}, \\ p_k^{(n+1)} &= \frac{\sum_{e_{ij}\in\Upsilon_k}p(\mathrm{Z}_{ij}=1|\mathrm{A}_{ij};S_i,S_j,\Theta^{(n)})}{|\Upsilon_k|}, \\ p_0^{(n+1)} &= \frac{\sum_{e_{ij}\in\Gamma}p(\mathrm{Z}_{ij}=1|\mathrm{A}_{ij};S_i,S_j,\Theta^{(n)})}{|\Gamma|},\end{aligned} \tag{5}$$

where $\Upsilon_k=\{e_{ij}\in E|S_i=k,S_j=k\}$ and $\Gamma=\{e_{ij}\in E|S_i\neq S_j\}$.

The two steps, i.e., E-step and M-step, are iteratively executed until the estimated Θ value converges (i.e., the deviation of Θ is smaller than a predefined value. Here, we set it as 0.01) or the iteration of EMLIC reaches its maximum number of generations (here, we set it as 500). After that, for each link e_{ij}, we compute its state Z_{ij}^* based on the estimated $\Theta^{(n+1)}$ as follows:

$$\mathrm{Z}_{ij}^* \leftarrow \underset{\mathrm{Z}_{ij}\in\{0,1\}}{\arg\max}\, p(\mathrm{Z}_{ij}|\mathrm{A}_{ij};S_i,S_j,\Theta^{(n+1)}). \tag{6}$$

Algorithm 1 EMLIC

1: **Input**: Observations A.
2: **Output**: Estimated presence state Z^*.
3: Detect the potential community structures S of the network using the BGLL algorithm [33].
4: $n\leftarrow 0$ and $\Theta^{(n)}\leftarrow\emptyset$.
5: **Initialization**: Generate an initial Z^* based on the majority voting, and then generate an initial $\Theta^{(1)}$ based on Z^*.
6: **while** $||\Theta^{(n+1)}-\Theta^{(n)}||\geq 0.01$ and $n<500$ **do**
7: $\quad n\leftarrow n+1$.
8: $\quad$ **E-step**: Compute the expected log likelihood function $\mathbf{Q}(\Theta|\Theta^{(n)})$ of $p(\mathrm{A},\mathrm{Z};S,\Theta^{(n)})$ based on Eq. (3).
9: $\quad$ **M-step**: Maximize $\mathbf{Q}(\Theta|\Theta^{(n)})$ of $p(\mathrm{A},\mathrm{Z};S,\Theta^{(n)})$, and then get $\Theta^{(n+1)}$ based on Eqs. (4) and (5).
10: **end while**
11: Compute Z^* based on Eq. (6).

In EM algorithms, it is important for them to generate initial parameters [31]. Here, we use the majority voting technique to initialize Z in EMLIC as its computational complexity is linear with the number of observations and it reflects the observations of the majority of individuals. Specifically, for each edge e_{ij}, its initial Z_{ij}^* value is the observation value returned by the majority of individuals. Then, the initial values of the parameters in $\Theta^{(1)}$ are generated based on Eq. (5). Here, if $\mathrm{Z}_{ij}^*=1$, $p(\mathrm{Z}_{ij}=1|\mathrm{A}_{ij};S_i,S_j,\Theta^{(0)})=1$; and $p(\mathrm{Z}_{ij}=1|\mathrm{A}_{ij};S_i,S_j,\Theta^{(0)})=0$ otherwise.

The framework of EMLIC for the link inference of network data with community structures is given in **Algorithm** 1. In **Algorithm** 1, the initialization step in line 5 needs to make a decision for each edge $e_{ij}\in E$ by directly aggregating the observations of q platforms. Therefore, it has a computational complexity $O(m\cdot q)$, where m is the number of edges in the network data. The **While** loop in lines 6-10 is executed g_{max} times, where g_{max} is the predefined maximum number of generations (here, we set it as 500). Within the **While** loop, the **E-step** in line 8 needs to compute the $\mathbf{Q}(\Theta|\Theta^{(n)})$, which has a computational complexity $O(2\cdot m\cdot q)$. The **M-step** in line 9 needs to update the unknown parameters $R^{[\alpha]}$ for each platform and p_k for each community, and each update needs m operations at most. Therefore, the **M-step** has a computational complexity $O(m\cdot q)$. In conclusion, the EMLIC has a computational complexity $O(m\cdot q\cdot g_{max})$.

Comparisons of EMLIC with classical reliable link inference algorithms. EMLIC, EMLI [31], EML [11] and the work in [24] model the reliable truth inference problem into a maximum posteriori estimation of observations. However, they are different in the following aspects:

1) In the system models, EMLIC, EML and the work in [24] develop a reliable link inference model for network data in the presence of missing and spurious links whereas EMLI presents a reliable truth inference model in crowdsourcing platforms. Compared with EML and the work in [24], EMLIC considers the heterogeneous reliability of platforms and the impacts of community structures on the link inference of the system.

2) In the system assumptions, the work in [24] is under the assumption that the links of a network can be observed only once while EML extends this assumption by allowing multiple observations. Different from EML, EMLIC assumes that the observations to the links of a network can be collected by multiple platforms (or a platform at different times) and each platform has its own reliability. When the reliability of platforms is set to the same value in advance, EMLIC has the same system assumptions as EML.

3) In the system solutions, both the work in [24] and EMLIC consider the effects of community structures in link inferences. However, their focuses are different, resulting in distinct link inference performances. Specifically, the work in [24] uses the Metropolis algorithm to sample a set of community partitions and infers the true state of links based on their reliability under the sampled community partitions. EMLIC first discovers the most possible community structures of networks, and then uses an EM algorithm to aggregate the true states of the links in different communities based on the observations.

IV. Experimental Results

In this section, we test EMLIC on the GN [43] and LFR [44] benchmark networks and 8 real networks. Moreover, we adopt three criteria to test the performances of EMLIC, and compare EMLIC with three classical algorithms. In the following, the experimental settings are first given, and then the experimental comparisons are made. Finally, the effects of the experimental settings on the performances of comparison algorithms are analyzed.

A. Experimental Settings

Experimental networks: The GN [43] and LFR [44] benchmark networks and 8 real-world networks are chosen.

• GN benchmark networks [43]: They have 4 communities and each one consists of 32 nodes and 256 links. In these networks, each node has 8 links, including the intra-community links and extra-community links, and the fraction of extra-community links is determined by a mixing parameter $\mu \in [0,\ 1]$. Actually, with the increase of μ, the number of extra-community links of nodes increases, and the communities in the networks become less clear. Here, 11 GN benchmark networks are generated by ranging the mixing parameter μ from 0 to 0.5 with the interval 0.05. The GN benchmark networks mainly examine whether comparison algorithms can handle the effects of community structures on reliable link inference.

• LFR benchmark networks: They were proposed by Lancichinetti and Fortunato in [44]. Compared with the GN benchmark networks, they consider the heterogeneity properties, especially in the distribution of community sizes and node degrees which follow power law distributions with different exponential parameters β and γ, respectively, and they can contain a large number of communities. Similar to the GN benchmark networks, the fraction of extra-community links in the LFR benchmark networks is determined by a mixing parameter $\mu \in [0,\ 1]$. Here, 11 LFR benchmark networks are generated by ranging the mixing parameter μ from 0 to 0.5 with the interval as 0.05 under $n = 1000$, $\beta = 1$ and $\gamma = 2$. Moreover, similar to the work [35], we set the minimum community size and the maximum community size as 20 and 100, respectively. Experiments on these networks are to investigate the impacts of the community heterogeneity of networks on the link inference of comparison algorithms.

TABLE I: Basic information of real-world networks, including the number of nodes n, the number of edges m, the average node degree $\bar{k}$, the number of communities c detected by BGLL, and the modularity Q [32] of community structures.

Network	n	m	$\bar{k}$	c	Q
Poolbooks	105	441	8.4	4	0.4986
Jazz	198	2742	27.70	4	0.4431
USAir	332	2126	12.71	9	0.3497
Elegans	453	2025	8.940	13	0.4322
Email	1133	5451	9.622	14	0.5412
Power	4941	6594	2.669	624	0.7756
Geom	7743	11898	3.073	2234	0.7775
Pgp	10680	20340	4.544	540	0.8604

• Real-world networks: They mainly measure whether a link inference algorithm can effectively handle the real structure properties coming from different complex systems, including the social communication system, the collaboration system, the biological system and the email system. The link inference in these systems promotes their collaborations and functional formulation. The tested networks are introduced as follows:

The Poolbooks network (Poolbooks) was collected by Newman from http://www.orgnet.com/ and it represents the purchase of books about US politics in the online Amazon.com. In this network, nodes represent the books, and links denote the frequency of books that are copurchased by the same buyers.

The Jazz musicians network (Jazz) [45] was extracted from http://www.redhotjazz.com. It shows the collaborations of Jazz musicians. Specifically, the Jazz musicians and the musicians' collaborations on playing the same Jazz are represented by the nodes and edges of the network, respectively.

The USAir network (USAir) [46] illustrates the network of US air transportation system. In USAir, each node is an airport while each edge denotes an airline across two airports.

The C.elegans metabolic network (Elegans) [47] denotes the metabolic relationships of the Caenorhabditis elegans. In this network, nodes correspond to the Caenorhabditis elegans while edges represent their metabolic relationships.

The Email network (Email) [48] records the interchanges of emails between individuals in the origination of the University of Rovira i Virgili. Specifically, the e-mail addresses and the corresponding e-mail communications are denoted by the nodes and edges of the network, respectively.

The Power Grid network (Power) [49] was revealed by Watts and Strogatz, and it represents the topology of the Power Grid in the Western States of United States. In the Power

network, nodes and edges correspond to the Power Grid and their topological communications, respectively.

The Geom network (Geom) [50] expresses the authors' collaboration in computational geometry. In Geom, nodes represent the authors and edges indicate the collaboration of authors who wrote at least one common paper or book.

The Pretty Good Privacy network (Pgp) [51] was generated by the Pretty-Good-Privacy algorithm [52] for privacy information communication. In Pgp, nodes represent the peers in the Internet who send private information while edges correspond to the private information flows between the peers.

The Poolbooks and Power networks can be downloaded from http://www-personal.umich.edu/~mejn/netdata/. The USAir and Geom networks are collected from http://vlado.fmf.uni-lj.si/pub/networks/data/default.htm and http://vlado.fmf.uni-lj.si/pub/networks/data/collab/geom.htm, respectively, while the Jazz, Elegans, Email and Pgp networks are revealed from http://deim.urv.cat/~alexandre.arenas/data/welcome.htm.

Table I shows the basic information of tested real networks. As shown in Table I, the tested networks are ranging from small-scale networks with hundreds of nodes to large-scale networks with ten thousands of nodes. Moreover, they are heterogeneous in the node degrees and communities. Specifically, the Power, Geom and Pgp networks have $\bar{k} < 5$ while the Poolbooks, Jazz, USAir, Elegans and Email networks have $\bar{k} > 8$, where $\bar{k}$ is the average node degree. The modularity Q of the tested networks is ranging from 0.4322 to 0.8644. Generally, the larger the Q value is, the more clear community structures the network has. Experiments on those networks are to demonstrate that EMLIC can be effectively applied to real complex systems with heterogeneous structure information.

For each real network, we randomly incorporate a ratio p_k of intra-community spurious links into the link structures of community k, and insert a ratio p_0 of inter-community spurious links. Moreover, for each platform α, its observation for each link e_{ij} in the tested spurious network is simulated based on its reliability $R^{[\alpha]}$. We test comparison algorithms on the tested spurious networks with different parameter settings. Specifically, we take $q = 30$, $p_k \in [0.7,\ 1]$, $p_0 \in [0.3,\ 1]$, $R_{1k}^{[\alpha]} \in [0.2,\ 0.9]$, $R_{0k}^{[\alpha]} \in [0.3,\ 0.9]$, $k = 0, 1, 2, \ldots, c$ and $\alpha = 1, 2, \ldots, q$. For each simulation setting, 30 datasets are generated randomly, and the averaged results over 30 independent trials are recorded. Moreover, influences of some parameter settings are analyzed by varying one of the parameters while keeping the other parameters unchanged.

Comparison algorithms: The majority voting algorithm MV, the link inference algorithm EML in [11] and the simplified version of EMLIC (recorded as EMLI) [31] are adopted.

• MV: It is one of the most classical link inference methods, which reflects the observations of the majority of platforms. Here, for each edge e_{ij}, its state Z_{ij}^{*} is estimated by MV as follows:

$$\mathrm{Z}_{ij}^{*} \leftarrow \underset{d\in\{0,1\}}{\arg\max} \sum_{\alpha=1}^{q} \mathbb{I}(\mathrm{A}_{ij}^{[\alpha]} = d).$$

where $\mathbb{I}(\cdot)$ is an indicator function. If the predicate is true, $\mathbb{I}(\cdot) = 1$, and $\mathbb{I}(\cdot) = 0$, otherwise. The comparison between MV and EMLIC aims to demonstrate the superior performances of the EM based inference method with proper assumptions on the link inference.

• EML [11]: It assumes that some edges in the tested networks can be observed by multiple times and the observations for each link may conflict with each other. It considers two rates r_1 and r_2 corresponding to the probability of observing a real edge and a spurious edge, respectively, and adopts a heuristic EM algorithm to estimate both the true state of edges and the two unknown rates. The comparison between EML and EMLIC is made to demonstrate that EMLIC outperforms the existing reliable link inference algorithms with homogeneous reliability of platforms' observations.

• EMLI [31]: In crowdsourcings, EMLI first adopts a confusion matrix to represent the reliability of individuals, and then aggregates the conflicting observations of individuals based on the confusion matrix. Here, EMLI is considered as a simplified version of EMLIC, which does not consider the effects of community structures on the link inference. For the network (e.g., the Erdő-Rényi network) with no community structures, EMLI has the same performances on the link inference as EMLIC. This is because in this case, EMLIC only needs to estimate the reliability of a platform' observations to inter-community links. The reason for choosing EMLI as the comparison algorithm is to demonstrate that by considering the impacts of the communities of networks on the link inference, EMLIC has superior performances than EMLI.

Criteria: The true error rate r_t, the spurious error rate r_s and the normalized mutual information ρ [42] are adopted. r_t and r_s are used to test the performances of comparison algorithms on the link inference while ρ is adopted to evaluate the performance of comparison algorithms on preserving the community property of networks.

• *True error rate* r_t: It measures the detection error rate on the real links, i.e., the fraction of the edges that are not detected by comparison algorithms but they exist in reality. r_t is computed as follows:

$$r_t = \frac{m_t - m_t^a}{m_t},$$

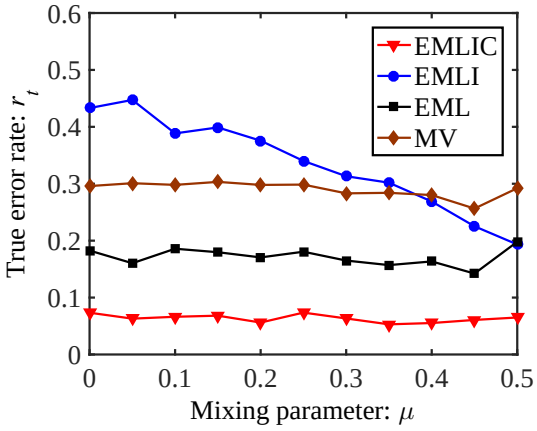
where m_t is the number of real links and m_t^a denotes the number of real links that are detected. The r_t value is in the range of [0, 1]. A comparison algorithm has good performances on the detection of real links when it has a low r_t value.

• *Spurious error rate* r_s: It evaluates the detection error rate on the spurious links, i.e., the fraction of the edges that are observed by comparison algorithms but they are absent in reality. r_s is computed as follows:

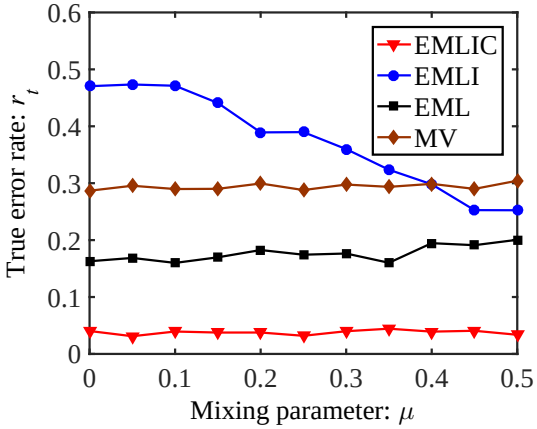
$$r_s = \frac{m_s^a}{m_s},$$

where m_s denotes the number of spurious links while m_s^a records the number of spurious links that are observed. The r_s value is in the range of [0, 1]. A small value of r_s corresponds to a good performance on the detection of spurious links.

• *Normalized mutual information* ρ [42]: It measures the similarity of communities between the real network and the inferred network. Let S and $\widehat{S}$ be the communities of the real network and the inferred network, respectively, and ρ is computed as follows:

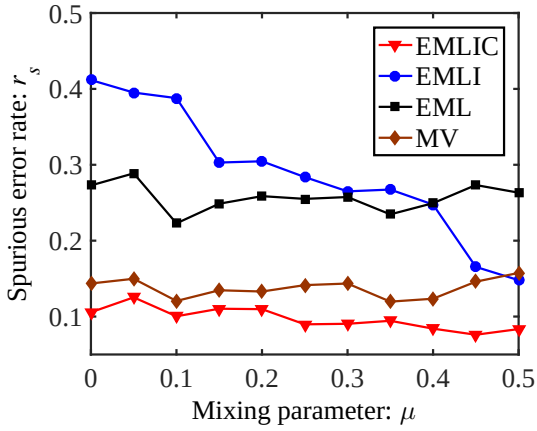


(a)

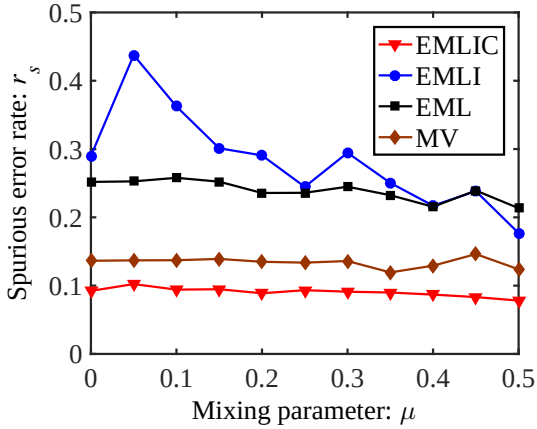


(b)

Fig. 4: (Color online) True error rate r_t VS. Mixing parameter μ on the benchmark networks. (a) the GN benchmark networks and (b) the LFR benchmark networks.



(a)



(b)

Fig. 5: (Color online) Spurious error rate r_s VS. Mixing parameter μ on the benchmark networks. (a) the GN benchmark networks and (b) the LFR benchmark networks.

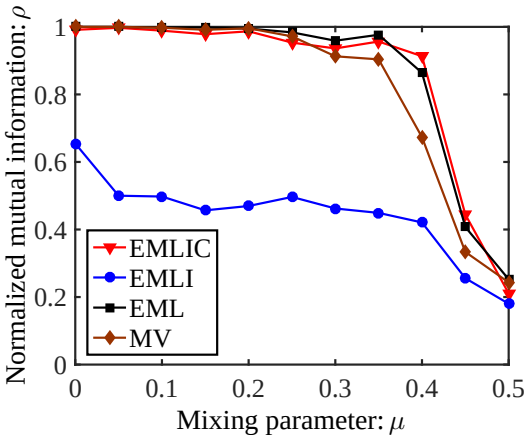
$$\rho(S,\widehat{S}) = \frac{-2\sum_{i=1}^{k_S}\sum_{j=1}^{k_{\widehat{S}}} \mathrm{F}_{ij}\log\left(\mathrm{F}_{ij}\cdot n/(||\mathrm{F}_{i.}||\cdot||\mathrm{F}_{.j}||)\right)}{\sum_{i=1}^{k_S}||\mathrm{F}_{i.}||\log(||\mathrm{F}_{i.}||/n)+\sum_{j=1}^{k_{\widehat{S}}}||\mathrm{F}_{.j}||\log(||\mathrm{F}_{.j}||/n)}$$

where k_S is the number of communities in S and F represents a confusion matrix with element F_{ij} denoting the number of common nodes in the community i of S and the community j of $\widehat{S}$. $\mathrm{F}_{i.}$ and $\mathrm{F}_{.j}$ are the sets of elements of the i-th row and j-th column of F, respectively. The value of $\rho(S,\widehat{S})$ is in the range of [0, 1]. The higher the ρ value is, the more similar community structures the partitions S and $\widehat{S}$ have.

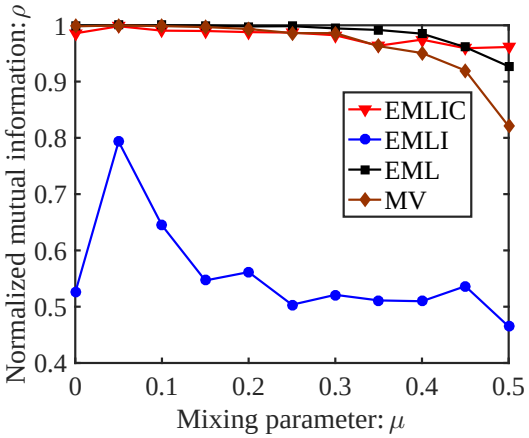
B. Experimental Results on Tested Networks

Firstly, all comparison algorithms are tested on the GN and LFR benchmark networks with different mixing parameters μ, and the corresponding true error rates r_t are recorded in Fig. 4. The results illustrate that EMLIC performs significantly better than EMLI, EML and MV on all networks, in terms of r_t. Specifically, the averaged r_t value of EMLIC over the 22 networks is about 5.060% lower than that of EMLI (35.49%), EML (17.39%) and MV (29.20%). On average, EMLIC has reduced 85.74%, 70.90% and 82.67% of the r_t value obtained by EMLI, EML and MV, respectively. The high performance of EMLIC may be attributed to its heterogeneous reliability setting of platforms' observations to the communities of networks, which is further demonstrated by the comparisons among EMLIC, EMLI and EML.

The comparison between EML and EMLI in Fig. 4 shows that the setting of the heterogeneous reliability of platforms would degrade the performances of EML on the detection of real links. This is to be expected as the maximization of the likelihood function in EMs is nonconvex and EMs cannot guarantee convergence to a maximum likelihood estimator. Therefore, for some practical applications with multimodal distributions, the EM algorithms may converge to a local optimal solution. Generally, the number of local optimal solutions exponentially increases with that of unknown parameters. Here, EMLI has more unknown reliability parameters than EML, which makes EMLI easier to trap into a local optimal solution. Moreover, the parameters of EM algorithms are



(a)



(b)

Fig. 6: (Color online) Normalized mutual information ρ VS. Mixing parameter μ on the benchmark networks. (a) the GN benchmark networks and (b) the LFR benchmark networks.

estimated from the observations under a given prior distribution, and hence their performances rather strongly depend on the adopted stochastic blockmodel. When the estimated parameters are far from the underlying distributions of parameters, EM algorithms may have poor performances. Note that, although it has more unknown parameters than EML and EMLI, EMLIC has lower r_t values than EMLI and EML. This is because the incorporation of structure information into the statistic blockmodel can improve the performances of EMs [28], [29]. Here, EMLIC incorporates the underlying community structures of networks into its statistic blockmodel, thus generating good performances for network data with community structures.

Fig. 4 also shows the impacts of the μ setting on the link inference of comparison algorithms. Generally, with the increase of μ, the corresponding benchmark network has less clear community structures, and the impacts of the community structures of the network on the link inference become smaller. This is consistent with the reported performances of EMLI which does not consider the impacts of community structures on the link inference in Fig. 4. As shown in Fig. 4, the r_t

TABLE II: Results on different real-world networks. All results are averaged over 30 independent trials, and the best result is marked in boldface for each network.

Network	Index	EMLIC	EMLI	EML	MV
Poolbooks	r_t	**0.0919**	0.4215	0.1373	0.2840
	r_s	**0.1092**	0.2935	0.3000	0.1377
	ρ	**0.8112**	0.3992	0.7794	0.6718
Jazz	r_t	**0.0396**	0.3145	0.1814	0.2849
	r_s	**0.1028**	0.2946	0.2356	0.1372
	ρ	**0.8520**	0.4786	**0.8520**	0.8079
USAir	r_t	**0.0681**	0.3215	0.1831	0.3012
	r_s	**0.0998**	0.2655	0.2584	0.1363
	ρ	**0.7090**	0.5132	0.6496	0.6067
Elegans	r_t	**0.0732**	0.3049	0.1954	0.3050
	r_s	**0.0904**	0.2284	0.2267	0.1294
	ρ	0.5837	0.4400	**0.5990**	0.5469
Email	r_t	**0.0500**	0.3375	0.2004	0.2979
	r_s	**0.0793**	0.2854	0.2157	0.1289
	ρ	**0.6868**	0.4573	0.5892	0.5398
Power	r_t	0.2108	0.3990	**0.1477**	0.2840
	r_s	**0.0956**	0.1839	0.2632	0.1347
	ρ	0.8872	0.8609	**0.9129**	0.8911
Geom	r_t	0.1454	0.5176	**0.1394**	0.2708
	r_s	**0.0904**	0.1974	0.2633	0.1295
	ρ	0.9405	0.8711	**0.9411**	0.9283
Pgp	r_t	**0.0923**	0.4655	0.1314	0.2845
	r_s	**0.0974**	0.2810	0.3047	0.1493
	ρ	**0.9034**	0.7966	0.8975	0.8569

values of EMLI decrease with the increase of μ. Note that, compared with EML and MV, EMLI is more sensitive to the parameter μ. This is because EMLI needs to estimate more reliability parameters which may be influenced by the community structures of networks.

Fig. 5 shows the spurious error rate r_s of comparison algorithms on the GN and LFR benchmark networks with different mixing parameters μ. From Fig. 5, we can obtain similar observations from the r_t comparison results. Specifically, for the GN networks, the averaged r_s of EMLIC is about 9.720% lower than that of EMLI (28.90%), EML (25.68%) and MV (13.75%), while for the LFR networks, EMLIC obtains an averaged $r_s = 9.040\%$ lower than the 28.22%, 23.92% and 13.39% of EMLI, EML, and MV, respectively. It is also shown that the r_s values of EMLIC on the LFR benchmark networks are smaller than those on the GN benchmark networks and they decrease with increasing network heterogeneity in community sizes and degrees. In addition, Fig. 5 demonstrates that the r_s values of EMLIC and EMLI are robust and sensitive to μ, respectively. Those phenomena validate the superior performances of EMLIC on the detection of spurious links.

The detected networks by comparison algorithms include both real links and spurious links, which may change the community properties of original networks. To demonstrate the consistency of community properties, we record the ρ values of the community structures between the detected networks by all link inference algorithms and the original ones in Fig. 6. Fig. 6 illustrates that EMLIC achieves very competitive ρ values on most cases, and outperforms or compares well with the other comparison algorithms. It also indicates that the ρ values of all comparison algorithms decrease with increasing μ. This is to be expected that with the increase of μ, the potential community structures of networks become more unclear and the missing of intra-community links and the adding of spuri-

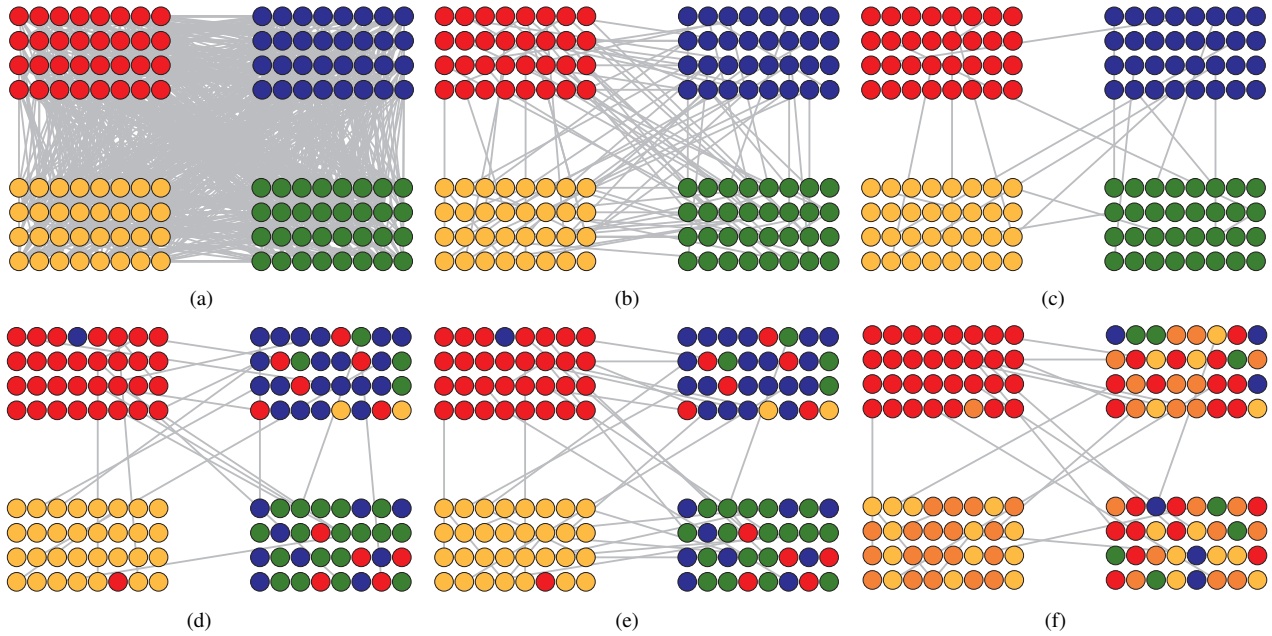


Fig. 7: (Color online) Community preservation and undetected spurious links of comparison algorithms on the GN benchmark network with $\mu = 0.4$. (a) the real GN network, (b) the GN network only with spurious links and the GN network detected by (c) EMLIC, (d) EMLI, (e) EML and (f) MV. Nodes plotted by different colors are in different communities.

TABLE III: Performance of comparison algorithms VS. Reliability of platforms R_{1k} & R_{0k} on the Email network.

R_{1k} & R_{0k}	Index	EMLIC	EMLI	EML	MV
[0.2, 0.6] & [0.3, 0.9]	r_t	0.5996	**0.5268**	0.9966	0.8643
	r_s	0.4056	0.3902	**0.0047**	0.1349
	ρ	0.3790	0.4222	**0.4797**	0.4438
[0.2, 0.6] & [0.2, 0.7]	r_t	0.9576	**0.6993**	0.9473	0.8634
	r_s	0.7771	0.6584	**0.5128**	0.7122
	ρ	0.4008	**0.4890**	0.4192	0.4042
[0.2, 0.9] & [0.3, 0.9]	r_t	**0.0500**	0.3375	0.2004	0.2979
	r_s	**0.0793**	0.2854	0.2157	0.1289
	ρ	**0.6868**	0.4573	0.5892	0.5398
[0.2, 0.9] & [0.2, 0.7]	r_t	0.4599	0.4734	**0.0560**	0.2863
	r_s	**0.5178**	0.5624	0.9496	0.7133
	ρ	0.5159	0.5138	**0.6753**	0.5589
[0.4, 0.9] & [0.3, 0.9]	r_t	**0.0105**	0.1436	0.0275	0.0469
	r_s	**0.0607**	0.1177	0.1895	0.1353
	ρ	**0.7661**	0.6275	0.6807	0.6661
[0.4, 0.9] & [0.2, 0.7]	r_t	**0.0436**	0.3692	0.0069	0.0483
	r_s	**0.2385**	0.2479	0.8930	0.6955
	ρ	**0.7040**	0.3935	0.6623	0.6534

ous links would result in the change of community structures. Those phenomena demonstrate the superior performance of EMLIC on preserving the community structures of networks.

Next, all comparison algorithms are tested on eight real-world networks coming from different areas, and the corresponding r_t, r_s and ρ are recorded in Table II. From Table II, we can see that i) for the Poolbooks, Jazz, USAir, Elegans, Email and Pgp networks which have a high node degree or a clear community partition, EMLIC has the lowest r_t value among comparison algorithms; ii) for the low-degree Power and Geom networks, EMLIC only gives a slightly larger r_t value than EML; and iii) for all real-world networks, EMLI achieves a high true error rate r_t as it needs to estimate a large amount of reliability parameters and neglects the impacts of community structures on the link inference.

Table II also shows that for all real networks, EMLIC has the lowest r_s values among comparison algorithms. Moreover, all comparison algorithms achieve high ρ values for the Power, Geom and Pgp networks which have a low node degree, whereas they obtain low ρ values for the Poolbooks, Jazz and Email networks which have a high node degree. This may be explained by the fact that the numbers of undetected spurious and real links are relevant to the networks' node degree, resulting in changing the underlying community structures of original networks.

To further demonstrate the superior performances of EMLIC on preserving community structures and identifying spurious links, we plot the community divisions of the small-scale GN benchmark and USAir networks and their spurious versions detected by the comparison algorithms in Figs. 7 and 8, respectively. Moreover, the undetected spurious links of networks by comparison algorithms are also recorded in Figs. 7 and 8. Fig. 7 illustrates that only the GN benchmark network detected by EMLIC has the same community division as the original and spurious networks, which has 4 communities. The GN benchmark networks detected by the other comparison algorithms tend to divide a large community of the original network into two or more small communities. Fig. 7 also shows that the GN benchmark network detected by EMLIC has less spurious links than those by the other algorithms, especially for the spurious inter-community links, which demonstrates the effectiveness of the heterogeneous reliability strategy. From Fig. 8, we can see that the USAir network detected by the EMLIC algorithm has the most similar community division as the real one, and its spurious links are far less than those of the USAir networks detected by the other comparison algorithms.

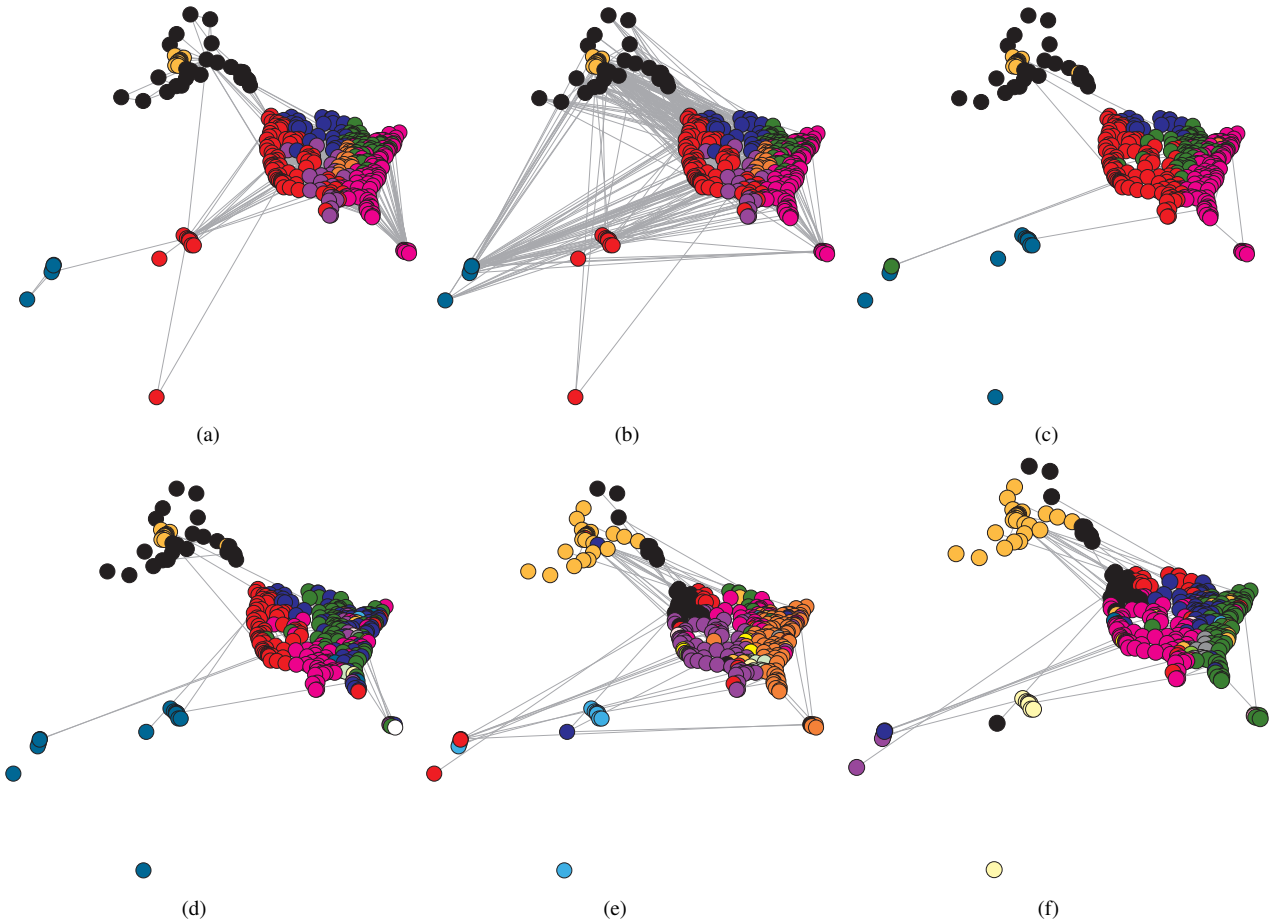


Fig. 8: (Color online) Community preservation and undetected spurious links of comparison algorithms on the USAir network. (a) the real USAir network, (b) the USAir network only with spurious links and the USAir network detected by (c) EMLIC, (d) EMLI, (e) EML and (f) MV. Nodes plotted by different colors are in different communities.

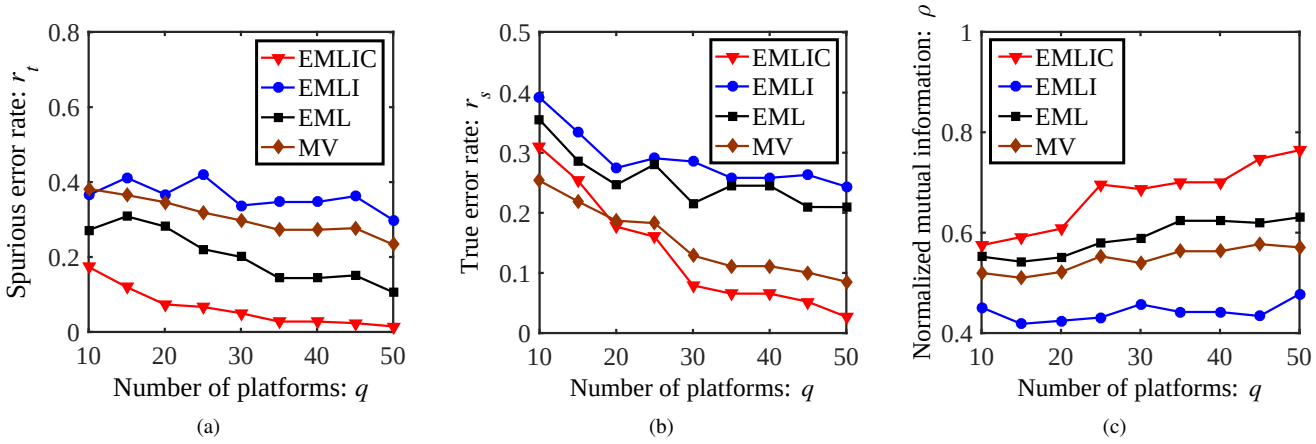


Fig. 9: (Color online) Performance of comparison algorithms VS. Number of platforms q on the Email network. (a) the true error rate r_t. (b) the spurious error rate r_s. (c) the normalized mutual information ρ.

C. Effects of Parameter Settings on Reliable Link Inference

In this part, the effects of the parameters, including the number of platforms q, the reliability of platforms R and the ratio of spurious links $p_k\&p_0$, on the link inference are investigated in the real Email network.

Fig. 9 shows the performances of comparison algorithms in the Email network with different settings of q. It can be observed that the performances of all comparison algorithms improve with the increase of q, in terms of r_t, r_s and ρ. It also illustrates EMLIC has the lowest r_t and r_s values

TABLE IV: Performance of comparison algorithms VS Ratios p_k&p_0 of spurious links on the real Email network.

p_k & p_0	Index	EMLIC	EMLI	EML	MV
[0, 1] & [0, 1]	r_t	**0.0264**	0.3616	0.2393	0.2953
	r_s	**0.0551**	0.2062	0.1812	0.1281
	ρ	**0.6847**	0.4141	0.5846	0.5509
[0, 1] & [0.3, 1]	r_t	**0.0293**	0.3628	0.2461	0.3022
	r_s	**0.0586**	0.2182	0.1843	0.1434
	ρ	**0.6909**	0.4469	0.6162	0.5919
[0.5, 1] & [0, 1]	r_t	**0.0264**	0.3994	0.1983	0.2838
	r_s	**0.0695**	0.1892	0.2164	0.1450
	ρ	**0.7116**	0.3970	0.5845	0.5553
[0.5, 1] & [0.3, 1]	r_t	**0.0408**	0.3697	0.1948	0.2893
	r_s	**0.0724**	0.1614	0.2111	0.1281
	ρ	**0.6847**	0.4141	0.5846	0.5509
[0.7, 1] & [0, 1]	r_t	**0.0413**	0.3925	0.2034	0.3008
	r_s	**0.0789**	0.1650	0.2206	0.1335
	ρ	**0.6949**	0.4008	0.5897	0.5692
[0.7, 1] & [0.3, 1]	r_t	**0.0500**	0.3375	0.2004	0.2979
	r_s	**0.0793**	0.2854	0.2157	0.1289
	ρ	**0.6868**	0.4573	0.5892	0.5398

TABLE V: Performance of comparison algorithms VS. Ratio of platforms r_p with low reliability on the real Email network.

r_p	Index	EMLIC	EMLI	EML	MV
0%	r_t	**0.0500**	0.3375	0.2004	0.2979
	r_s	**0.0793**	0.2854	0.2157	0.1289
	ρ	**0.6868**	0.4573	0.5892	0.5398
10%	r_t	**0.0446**	0.4392	0.1811	0.3012
	r_s	**0.0757**	0.1164	0.2507	0.1372
	ρ	**0.6807**	0.3607	0.5774	0.5323
20%	r_t	**0.0654**	0.4500	0.3699	0.4079
	r_s	**0.1009**	0.2042	0.1484	0.1255
	ρ	**0.6976**	0.3610	0.5983	0.5746
30%	r_t	**0.0350**	0.3046	0.3708	0.4262
	r_s	**0.0740**	0.3133	0.1418	0.1053
	ρ	**0.6823**	0.4748	0.5487	0.5290
40%	r_t	**0.0594**	0.3234	0.4070	0.4291
	r_s	**0.0526**	0.2422	0.0844	0.0765
	ρ	**0.6245**	0.4316	0.4925	0.4963
50%	r_t	**0.0635**	0.3455	0.5566	0.5091
	r_s	**0.0524**	0.2260	0.0770	0.0975
	ρ	**0.6667**	0.4618	0.4874	0.4889

and the highest ρ value when $q > 20$ (in this case, the link inference systems have enough information about the network). Note that, when $q < 20$, the EM based link inference algorithms, including EMLIC, EMLI and EML, have higher r_s values than MV. Those phenomena suggest that the amount of observations may have a large effect on the performances of link inference algorithms, especially for the EM based link inference algorithms. They also demonstrate the superior performances of EMLIC on the link inference when the systems have enough information about the network.

Table III records the impacts of the reliability of platforms $R^{[\alpha]}$ on the link inference of comparison algorithms in the Email network. It can be seen that i) when R_{0k} (R_{1k}) keeps unchanged, the performances of all comparison algorithms improve with the increase of R_{1k} (R_{0k}), generally; ii) when all platforms have a relatively low reliability (i.e., $R_{1k} \in [0.2,\ 0.6]$ and $R_{0k} \in [0.2,\ 0.7]$), all algorithms cannot infer the true links and the spurious links. When all platforms have a relatively high reliability (i.e., $R_{1k} \in [0.4,\ 0.9]$ and $R_{0k} \in [0.3,\ 0.9]$), all algorithms have good performances in terms of r_t, r_s and ρ; and iii) when the reliability of platforms is in an acceptable range, among comparison algorithms EMLIC has the best performances in terms of r_t, r_s and ρ.

Table IV records the performances of comparison algorithms in the Email network varying with different ratios p_k and p_0 of spurious links. The results illustrate that EMLIC is the most influenced one among comparison algorithms by the settings of p_k and p_0. Specifically, the performances of EMLIC degrade with the increase of p_k and p_0, whereas those of other algorithms have no obvious relation with p_k and p_0. This is to be expected as the settings of p_k and p_0 may have a large impact on the community divisions of networks and only EMLIC would be influenced by the change of community divisions. The results in Table IV also demonstrate that EMLIC has the best performances for all settings of p_k and p_0, in terms of r_t, r_s and ρ.

In order to demonstrate whether the platforms' observations with relatively low reliability positively influence the link inference when they are coupled with those with high reliability, we incorporate a ratio r_p of platforms with low reliability (i.e., $R_{1k} \in [0.3,\ 0.5], k = 1, 2, \ldots, c$) into the simulations, and record the performances of comparison algorithms with different settings of r_p in Table V. It can be seen that the performances of EMLIC and EMLI are robust to the parameter r_p whereas those of EML and MV decrease with the increase of r_p values. This is because EMLIC and EMLI consider the heterogeneous reliability of platforms whereas EML and MV take into account the homogeneous reliability of platforms.

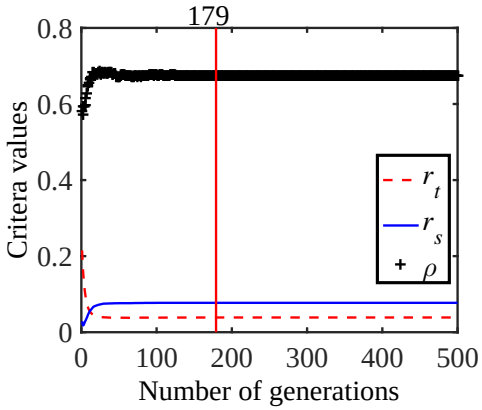


Fig. 10: Convergence performance of EMLIC on the Email network.

Fig. 10 presents the convergence of EMLIC on the real Email network. It records the r_t, r_s and ρ values averaged over 30 independent trials versus the number of generations. The results illustrate that for all 30 independent trials, EMLIC can converge within 500 generations. Specifically, the r_t, r_s and ρ obtained by EMLIC are not changed in the 179 generation.

V. Conclusions

In this paper, we presented a link inference problem, in which the links in a network cannot be directly revealed but they can be observed by multiple platforms or a platform at different times. We derived a link inference model which considers the heterogeneous reliability of platforms' observations and enables a comprehensive understanding of

the impacts of community structures on the link inference of networks. To solve this link inference problem, we proposed an efficient expectation maximization algorithm EMLIC with a low computational complexity $O(m \cdot q \cdot g_{max})$ for inferring the true state of links, the reliability of platforms' observations to links in different communities and the link errors in different communities. Simulations on the GN and LFR benchmark networks and eight real-world networks demonstrated that the EMLIC algorithm has lower error rates than the EMLI, EML and MV algorithms, including the true error rate and the spurious error rate. The results also illustrated that the spurious networks detected by the EMLIC algorithm can effectively preserve the community structures of real networks.

In this work, all links of a network can be observed directly by the system. In future work, we will consider the partial observations of platforms to the network's links. Moreover, the presented system may not work well for the networks with many spurious links due to the differences between the community structures detected by the BGLL algorithm on the spurious network and the real one. In this case, inspired by these works [28]–[30], we will construct a general inference model which jointly infers the community structure and the links of networks with a maximum posterior probability of observations. In addition, we will study the link inference problem in network data with temporal community structures and dynamic link information, in which the links are observed by multiple platforms at a series of times and the network data in different platforms have their own community structures. Finally, we will consider multiple classes of platforms (e.g., expert platforms with high reliability and malicious platforms with low reliability), and jointly infer the true state of the links of networks and the classes of platforms.

REFERENCES

[1] S. Boccaletti, G. Bianconi, R. Criado, C. I. Del Genio, J. Gómez-Gardenes, M. Romance, I. Sendina-Nadal, Z. Wang, and M. Zanin, "The structure and dynamics of multilayer networks," *Physics Reports*, vol. 544, no. 1, pp. 1–122, 2014.

[2] Q. Xuan, Z. Y. Zhang, C. Fu, H. X. Hu, and V. Filkov, "Social synchrony on complex networks." *IEEE Transactions on Cybernetics*, vol. 48, no. 5, pp. 1420–1431, 2018.

[3] C. Liu, J. Liu, and Z. Jiang, "A multiobjective evolutionary algorithm based on similarity for community detection from signed social networks," *IEEE Transactions on Cybernetics*, vol. 44, no. 12, pp. 2274–2287, 2014.

[4] C. Shi, Y. Li, J. Zhang, Y. Sun, and S. Y. Philip, "A survey of heterogeneous information network analysis," *IEEE Transactions on Knowledge and Data Engineering*, vol. 29, no. 1, pp. 17–37, 2017.

[5] Y. Zhu, D. Li, W. Xu, W. Wu, L. Fan, and J. Willson, "Mutual-relationship-based community partitioning for social networks," *IEEE Transactions on Emerging Topics in Computing*, vol. 2, no. 4, pp. 436–447, 2014.

[6] Q. Xu, Z. Su, K. Zhang, P. Ren, and X. S. Shen, "Epidemic information dissemination in mobile social networks with opportunistic links," *IEEE Transactions on Emerging Topics in Computing*, vol. 3, no. 3, pp. 399–409, 2015.

[7] W. Kolch, M. Halasz, M. Granovskaya, and B. N. Kholodenko, "The dynamic control of signal transduction networks in cancer cells," *Nature reviews. Cancer*, vol. 15, no. 9, p. 515, 2015.

[8] T. Martin, B. Ball, and M. Newman, "Structural inference for uncertain networks," *Physical Review E*, vol. 93, no. 1, p. 012306, 2016.

[9] Y. Q. Zhang, X. Li, and A. V. Vasilakos, "Spectral analysis of epidemic thresholds of temporal networks," *IEEE Transactions on Cybernetics*, vol. PP, no. 99, pp. 1–13, 2017.

[10] M. Newman and A. Clauset, "Structure and inference in annotated networks," *Nature Communications*, vol. 7, 2016.

[11] M. Newman, "Measurement errors in network data," *Nature Physics*, vol. In press, pp. doi:10.1038/s41 567–018–0076–1, 2018.

[12] V. Lyzinski, M. Tang, A. Athreya, Y. Park, and C. E. Priebe, "Community detection and classification in hierarchical stochastic blockmodels," *IEEE Transactions on Network Science and Engineering*, vol. 4, no. 1, pp. 13–26, 2017.

[13] D. Hric, T. P. Peixoto, and S. Fortunato, "Network structure, metadata, and the prediction of missing nodes and annotations," *Physical Review X*, vol. 6, no. 3, p. 031038, 2016.

[14] L. Peel, D. B. Larremore, and A. Clauset, "The ground truth about metadata and community detection in networks," *Science Advances*, vol. 3, no. 5, p. e1602548, 2017.

[15] P. Killworth and H. Bernard, "Informant accuracy in social network data," *Human Organization*, vol. 35, no. 3, pp. 269–286, 1976.

[16] A. Clauset and C. Moore, "Accuracy and scaling phenomena in internet mapping," *Physical Review Letters*, vol. 94, no. 1, p. 018701, 2005.

[17] Z. Xu, F. Yan, and Y. Qi, "Bayesian nonparametric models for multiway data analysis," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 37, no. 2, pp. 475–487, 2015.

[18] S. Sun, C. Zhang, and G. Yu, "A bayesian network approach to traffic flow forecasting," *IEEE Transactions on Intelligent Transportation Systems*, vol. 7, no. 1, pp. 124–132, 2006.

[19] L. Zhu, D. Guo, J. Yin, G. Ver Steeg, and A. Galstyan, "Scalable temporal latent space inference for link prediction in dynamic social networks," *IEEE Transactions on Knowledge and Data Engineering*, vol. 28, no. 10, pp. 2765–2777, 2016.

[20] L. Duan, S. Ma, C. Aggarwal, T. Ma, and J. Huai, "An ensemble approach to link prediction," *IEEE Transactions on Knowledge and Data Engineering*, vol. 29, no. 11, pp. 2402–2416, 2017.

[21] X. Li and X. Li, "Reconstruction of stochastic temporal networks through diffusive arrival times," *Nature Communications*, vol. 8, 2017.

[22] L. Lü and T. Zhou, "Link prediction in complex networks: A survey," *Physica A*, vol. 390, no. 6, pp. 1150–1170, 2011.

[23] L. Lü, M. Medo, C. H. Yeung, Y.-C. Zhang, Z.-K. Zhang, and T. Zhou, "Recommender systems," *Physics Reports*, vol. 519, no. 1, pp. 1–49, 2012.

[24] R. Guimerà and M. Sales-Pardo, "Missing and spurious interactions and the reconstruction of complex networks," *Proceedings of the National Academy of Sciences*, vol. 106, no. 52, pp. 22 073–22 078, 2009.

[25] C. De Bacco, E. A. Power, D. B. Larremore, and C. Moore, "Community detection, link prediction, and layer interdependence in multilayer networks," *Physical Review E*, vol. 95, no. 4, p. 042317, 2017.

[26] M. Newman and G. Reinert, "Estimating the number of communities in a network," *Physical Review Letters*, vol. 117, no. 7, p. 078301, 2016.

[27] 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.

[28] B. Ball and M. Newman, "An efficient and principled method for detecting communities in networks," *Physical ReviewE*, vol. 84, p. 036103, 2011.

[29] D. He, Z. Feng, D. Jin, X. Wang, and W. Zhang, "Joint identification of network communities and semantics via integrative modeling of network topologies and node contents," in *Thirty-First AAAI Conference on Artificial Intelligence*, San Francisco, California USA, Feb. 2017.

[30] D. Jin, X. Wang, R. He, D. He, J. Dang, and W. Zhang, "Robust detection of link communities in large social networks by exploiting link semantics," in *Thirty-Second AAAI Conference on Artificial Intelligence*, New Orleans Riverside, New Orleans, Louisiana, USA, Feb. 2018.

[31] Q. Liu, J. Peng, and A. T. Ihler, "Variational inference for crowdsourcing," in *Advances in Neural Information Processing Systems*, Lake Tahoe, Nevada, USA, Dec. 2012.

[32] M. Newman and M. Girvan, "Finding and evaluating community structure in networks," *Physical Review E*, vol. 69, no. 2, p. 026113, 2004.

[33] V. D. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre, "Fast unfolding of communities in large networks," *Journal of Statistical Mechanics: Theory and Experiment*, vol. 2008, no. 10, p. P10008, 2008.

[34] P. J. Mucha, T. Richardson, K. Macon, M. A. Porter, and J.-P. Onnela, "Community structure in time-dependent, multiscale, and multiplex networks," *Science*, vol. 328, no. 5980, pp. 876–878, 2010.

[35] L. Ma, M. Gong, J. Liu, Q. Cai, and L. Jiao, "Multi-level learning based memetic algorithm for community detection," *Applied Soft Computing*, vol. 19, no. 2, pp. 121–133, 2014.

[36] P. K. Gopalan and D. M. Blei, "Efficient discovery of overlapping communities in massive networks," *Proceedings of the National Academy of Sciences*, vol. 110, no. 36, pp. 14 534–14 539, 2013.

[37] S. Fortunato and D. Hric, "Community detection in networks: A user guide," *Physics Reports*, vol. 659, pp. 1–44, 2016.

[38] M. Rosvall and C. T. Bergstrom, "Maps of random walks on complex networks reveal community structure," *Proceedings of the National Academy of Sciences*, vol. 105, no. 4, pp. 1118–1123, 2008.

[39] M. Gong, Q. Cai, X. Chen, and L. Ma, "Complex network clustering by multiobjective discrete particle swarm optimization based on decomposition," *IEEE Transactions on Evolutionary Computation*, vol. 18, no. 1, pp. 82–97, 2014.

[40] C. Gao, M. Liang, X. Li, Z. Zhang, Z. Wang, and Z. Zhou, "Network community detection based on the physarum-inspired computational framework," *IEEE/ACM transactions on computational biology and bioinformatics*, p. doi: 10.1109/TCBB.2016.2638824, 2016.

[41] L. Ma, M. Gong, Q. Cai, and L. Jiao, "Enhancing community integrity of networks against multilevel targeted attacks," *Physical Review E*, vol. 88, no. 2, p. 022810, 2013.

[42] A. Lancichinetti and S. Fortunato, "Community detection algorithms: a comparative analysis," *Physical Review E*, vol. 80, no. 5, p. 056117, 2009.

[43] M. Girvan and M. Newman, "Community structure in social and biological networks," *Proceedings of the National Academy of Sciences*, vol. 99, no. 12, pp. 7821–7826, 2002.

[44] A. Lancichinetti and S. Fortunato, "Benchmarks for testing community detection algorithms on directed and weighted graphs with overlapping communities," *Physical Review E*, vol. 80, no. 1, p. 016118, 2009.

[45] P. M. Gleiser and L. Danon, "Community structure in jazz," *Advances in Complex Systems*, vol. 6, no. 04, pp. 565–573, 2003.

[46] V. Batageli and A. Mrvar, "Pajek datasets," http://vlado.fmf.uni-lj.si/pub/networks/data/default.htm, accessed Feb, 2007.

[47] J. Duch and A. Arenas, "Community identification using extremal optimization," *Physical Review E*, vol. 72, p. 027104, 2005.

[48] R. Guimer, L. Danon, A. Dłazguilera, F. Giralt, and A. Arenas, "Self-similar community structure in a network of human interactions," *Physical Review E*, vol. 68, no. 2, p. 065103, 2003.

[49] D. J. Watts and S. H. Strogatz, "Collective dynamics of small-world networks," *Nature*, vol. 393, pp. 440–442, 1998.

[50] V. Batageli and M. Zaveršnik, "Pajek datasets: Geom," http://vlado.fmf.uni-lj.si/pub/networks/data/collab/geom.htm, accessed February, 2002.

[51] M. Boguñà, R. Pastorsatorras, A. Díazguilera, and A. Arenas, "Models of social networks based on social distance attachment," *Physical Review E*, vol. 70, no. 2, p. 056122, 2004.

[52] S. Garfinkel and D. Russell, *PGP: Pretty Good Privacy*. O'Reilly & Associates, Inc., 1996.

Lijia Ma received his Ph.D. Degree from Xidian University, Xi'an, China in 2015. From Oct. 2015 to Oct. 2016, he was a postdoc in Hong Kong Baptist University, Hong Kong, and from Nov. 2016 to Dec. 2017, he was a postdoc in Nanyang Technological University, Singapore.

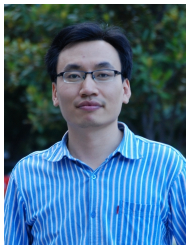
He is an assistant professor at the College of Computer and Software Engineering of Shenzhen University. His research interests mainly include evolutionary computation, machine learning and complex networks.

Jianqiang Li received his B.S and Ph.D. Degree from South China University of Technology in 2003 and 2008, respectively.

He is a professor at the College of Computer and Software Engineering of Shenzhen University. He led a project of the National Natural Science Foundation, and a project of the Natural Science Foundation of Guangdong Province, China. His major research interests include embedded systems and Internet of Things.

Qiuzhen Lin received the B.S. degree from Zhaoqing University and the M.S. degree from Shenzhen University, China, in 2007 and 2010, respectively. He received the Ph.D. degree from Department of Electronic Engineering, City University of Hong Kong, Kowloon, Hong Kong, in 2014.

He is currently a Lecturer in College of Computer Science and Software Engineering, Shenzhen University. He has published over twenty research papers since 2008. His current research interests include artificial immune system, multi-objective optimization and dynamic system.



Maoguo Gong (M'07-SM'14) received the B.S. degree and Ph.D. degree in electronic science and technology from Xidian University, Xi'an, China, in 2003 and 2009, respectively.

Since 2006, he has been a Teacher with Xidian University. In 2008 and 2010, he was promoted as an Associate Professor and as a Full Professor, respectively, both with exceptive admission. His research interests are in the area of computational intelligence with applications to optimization, learning, data mining and image understanding.

Dr. Gong received the prestigious National Program for the support of Top-Notch Young Professionals from the Central Organization Department of China, the Excellent Young Scientist Foundation from the National Natural Science Foundation of China, and the New Century Excellent Talent in University from the Ministry of Education of China. He is the Vice Chair of the IEEE Computational Intelligence Society Task Force on Memetic Computing, an Executive Committee Member of the Chinese Association for Artificial Intelligence, and a Senior Member of the Chinese Computer Federation. He is also the associate editor of *IEEE Trans. Evolutionary Computation*.

Carlos A. Coello Coello received PhD degree in computer science from Tulane University, USA, in 1996. He is currently Professor (CINVESTAV-3F Researcher) at the Computer Science Department of CINVESTAV-IPN, in Mexico City, Mxico. Dr. Coello has authored and co-authored over 450 technical papers and book chapters. He has also co-authored the book *Evolutionary Algorithms for Solving Multi-Objective Problems* (Second Edition, Springer, 2007). His publications report over 29,000 citations in Google Scholar (his h-index is 67). Currently, he is associate editor of the *IEEE Transactions on Evolutionary Computation* and serves in the editorial board of 12 other international journals. His major research interests are: evolutionary multi-objective optimization and constraint-handling techniques for evolutionary algorithms. He received the *2007 National Research Award* from the Mexican Academy of Sciences in the area of Exact Sciences, the *2013 IEEE Kiyo Tomiyasu Award* and the *2012 National Medal of Science and Arts* in the area of *Physical, Mathematical and Natural Sciences*. He is a Fellow of the IEEE, and a member of the ACM, Sigma Xi, and the Mexican Academy of Science.

Zhong Ming is a professor at the College of Computer and Software Engineering of Shenzhen University.

He is a senior member of the Chinese Computer Federation (CCF). He led three projects of the National Natural Science Foundation, and two projects of the Natural Science Foundation of Guangdong Province, China. His major research interests include Internet of Things and Cloud Computing.