Journal of Data Science 21 (3), 508–522
July 2023

DOI: 10.6339/23-JDS1109
Statistical Data Science

A Mixed-Membership Model for Social Network Clustering

Guang Ouyang[1], Dipak K. Dey[1], and Panpan Zhang[2,*]

[1]*Department of Statistics, University of Connecticut, Storrs, CT 06269, USA*
[2]*Department of Biostatistics, Vanderbilt University Medical Center, Nashville, TN 37203, USA*

Abstract

We propose a simple mixed membership model for social network clustering in this paper. A flexible function is adopted to measure affinities among a set of entities in a social network. The model not only allows each entity in the network to possess more than one membership, but also provides accurate statistical inference about network structure. We estimate the membership parameters using an MCMC algorithm. We evaluate the performance of the proposed algorithm by applying our model to two empirical social network data, the Zachary club data and the bottlenose dolphin network data. We also conduct some numerical studies based on synthetic networks for further assessing the effectiveness of our algorithm. In the end, some concluding remarks and future work are addressed briefly.

Keywords *cosine similarity; MCMC algorithm; mixed membership; social network clustering; stochastic blockmodels*

1 Introduction

Social network analysis is part of the social science which is an academic discipline studying a society and the behavior of entities therein. A social network consists of a set of entities (called actors) with certain interactions (represented by ties) among them. Statistical modeling has been a popular and powerful tool to study social networks thanks to its solid theoretical foundation. A plethora of statistical models have been established and exploited to uncover relational structure of social networks, and dyadic ties among actors. Friendship among Facebook users, business relationship across companies on the Wall street, and collaborations among researchers in a scientific field are all social network examples that have been extensively studied in the past. Social network analysis has a long history in sociology, where classical works traced back to the 1940s and 1950s (Rapoport, 1949a,b, 1950; Harary, 1953; Cartwright and Harary, 1956).

Modern research on social network analysis within mathematics, physics and other scientific disciplines focus mainly on the following three distinctive network features. The first feature is to explore how local mechanisms of network formation produce global network structure. Two representative models are the network evolution model (Newman, 2001) and the nodal attribute model (Boguñá et al., 2004). We refer the readers to Toivonen et al. (2009) for a comparison of these two models, and to the survey paper (Snijders, 2001) for a complete review of related statistical models. The second feature is to investigate topological properties of social networks and develop methods of modeling, either analytically or numerically. Two of the most popular properties of social networks are the small world phenomenon (Watts and Strogatz, 1998) and the power law of degree distribution (Barabási and Albert, 1999). A summary of some solvable

*Corresponding author. Email: panpan.zhang@vumc.org.

Received November 7, 2021; Accepted July 4, 2023

random-graph-based social network models was given in (Newman, 2006). The third feature, which is the one that we investigate in this paper, is network clustering.

Social network clustering works under the rationale that a group of actors excessively tied in a network are inclined to forming a cluster. One of the seminal works on social network clustering was Watts and Strogatz (1998), where each pair of actors in a social network was proven to be tied with a high probability if they had a mutual acquaintance, and such "tie" was parameterized by a measure called the clustering coefficient. This natural phenomenon in social networks was also discussed extensively by Newman (2001); Newman et al. (2001). The formation of a cluster requires the connections of actors within the cluster are significantly higher than those between actors from different clusters. It was posited in some literatures, e.g., (Holland et al., 1983), that a high probability of the occurrence of ties between actors within a cluster was due to some kind of homology (also called "internal homogeneity") of the actors. For instance, students from the same department of a college tend to form a community, in which almost everybody is a friend of everybody (i.e., the students in the same community are more likely to be connected); while students with different educational background are much less likely to be connected. Such internal homogeneity is mostly reflected in a background parameter (e.g., same department) and a location parameter (e.g., same college).

In this paper, we propose a simple but effective method for accurately clustering the entities in a social network into mutually exclusive communities. The proposed model was inspired and elevated from the classical stochastic blockmodel (SBM, Nowicki and Snijders, 2001). Recently, there were a variety of models extended from SBM in the literature. For instance, Sengupta and Chen (2018) introduced an SBM adjusted by node popularity, Huang et al. (2020) established an SBM for heterogeneous networks accounting for node attribute and Noroozi and Pensky (2022) suggested a nested SBM integrating standard SBM and LSM. Different from the existing literature, we specifically consider a flexible function to measure the similarities between actors in a network. Mixed membership is allowed for each actor in our model. The fit of our model is done in a Bayesian framework. The ascendancy of our model over the classic SBM will be detailed and discussed in the subsequent section. This paper not only introduces a flexible and extensible model allowing mixed memberships for network actors, but also gives the interested researchers, especially those relatively new to the field, insights into a standard approach of conducting statistical inference for social network clustering problems.

The rest of this paper is organized as follows: We review some representative model-based methods for social network clustering, with an additional concentration on the SBM, in Section 2. We propose a mixed membership model based on a simple similarity function in Section 3. Theoretical parameter estimation and an associated MCMC algorithm are presented in Section 4. Two empirical social network examples, the Zachary karate club data and the bottlenose dolphin network data, are used to evaluate the performance of our model, shown respectively in Sections 5 and 6. We then conduct some simulation study on synthetic data in Section 7. In the end, we give some concluding remarks and propose some future work in Section 8.

2 Notations for Stochastic Blockmodels

In general, methods for social network clustering can be summarized into two categories. A metric-based method, in contrast, aims at specifying an objective function which evaluates the quality of each network clustering strategy, followed by an algorithm optimizing the objective function (e.g., Ng et al., 2001; Shi and Malik, 2000; Newman et al., 2002; Ouyang et al., 2020).

A model-based method is to propose a (parametric) graphical generative model that characterizes the community structure of a social network, followed by an algorithm estimating the membership parameters conditioning on the observed data, most done in a Bayesian framework. To date, there have emerged a variety of graphical models for social network clustering, including but not limited to SBM (Nowicki and Snijders, 2001; Airoldi et al., 2008; Abbe, 2018; Gao et al., 2018), latent space models (LSM, Hoff et al., 2002; Handcock and Raftery, 2007; Sewell and Chen, 2017) random dot product graphs (RDPG, Young and Scheinerman, 2007; Marchette and Priebe, 2008; Lyzinski et al., 2017; Athreya et al., 2018), and exponential random graph models (ERGM, Snijders et al., 2006; Hunter et al., 2008; Fronczak et al., 2013), among others.

The core idea of model-based methods is to theoretically uncover the probabilistic and statistical properties of the proposed models. To begin with, we introduce some notations that will be used all through the paper. In general, a network is modeled by a mathematical undirected (or directed) graph consisting of a set of nodes which represent actors (e.g., Facebook users) in the network, and a set of undirected (or directed) edges which represent the relational ties between each pair of nodes (e.g., friendship connections between Facebook users). Let n be the number of nodes in an undirected social network. The observation of the network can be mathematically represented by an $n \times n$ dyadic adjacency matrix $\boldsymbol{A} = (A_{ij})_{n\times n}$, where A_{ij} equals 1 if nodes i and j are connected; 0, otherwise. For undirected networks, adjacency matrices are symmetric. If a network is directed, $A_{ij} = 1$ refers to a directed relation from i (initiator) to j (receiver), and the associated adjacency matrix $\boldsymbol{A}$ may be asymmetric.

More specifically, we consider the SBM first proposed by Snijders and Nowicki (1997). Although directed networks were considered in Snijders and Nowicki (1997), we simplify the problem to undirected networks for the sake of explanation. The model and the related methods can be extended to directed networks effortlessly. Our goal is to cluster a network of order n into h distinct communities. For each node $i = 1, 2, \ldots, n$, let $c(i)$ denote the community membership function for i. Assuming that $c(1), c(2), \ldots, c(n)$ are independently and identically distributed (i.i.d.) multinomial random variables with a hyperparameter vector $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_h)$, one defines $\boldsymbol{B}$ as an $h \times h$ symmetric probability matrix indicating linkages across different communities. Conditioning on $c(1), c(2), \ldots, c(n)$, the distribution of A_{ij} for each pair of nodes i and j is Bernoulli with probability $B_{c(i)c(j)}$.

As $\boldsymbol{A}$ is observed, our goal turns to estimate hyperparameters in $\boldsymbol{\theta}$ and the probability matrix $\boldsymbol{B}$, and ultimately to uncover the network structure by inferring $\boldsymbol{c} = (c(1), c(2), \ldots, c(n))$. The estimation can be performed in a Bayesian framework:

1. Establish the likelihood function, $\mathbf{Pr}(\boldsymbol{A}, \boldsymbol{c}; \boldsymbol{\theta}, \boldsymbol{B})$;
2. Estimate $\boldsymbol{\theta}$ and $\boldsymbol{B}$ jointly, which usually can be done by some Bayesian methods, such as Markov Chain Monte Carlo (MCMC) algorithms;
3. Determine the posterior distribution of $\boldsymbol{c}$ given $\boldsymbol{A}$, which is given by

$$\mathbf{Pr}(\boldsymbol{c} \mid \boldsymbol{A}) \propto \int \mathbf{Pr}(\boldsymbol{A}, \boldsymbol{c} \mid \boldsymbol{\theta}, \boldsymbol{B}) \pi(\boldsymbol{\theta}, \boldsymbol{B})\, d\boldsymbol{\theta} d\boldsymbol{B},$$

where $\pi(\boldsymbol{\theta}, \boldsymbol{B})$ denotes a joint prior distribution of $\boldsymbol{\theta}$ and $\boldsymbol{B}$.

The community prediction of node i is the index of the membership with the largest posterior probability.

There are several shortcomings of the classical SBM. One is that each actor in the network only can be assigned to one community, which may not be the case for many real social networks. A mixed membership SBM, inspired from the latent Dirichlet allocation (LDA, Blei et al., 2003), was proposed by Airoldi et al. (2008) to break this limitation. The model in Airoldi et al. (2008)

allows each actor in the network to possess multiple community memberships. In addition, it seems natural to define a function to quantitatively measure the similarities (or dissimilarities) between the actors in a social network space. Such functions are viewed as an indispensable part in clustering analysis, but are not considered in the classic SBM. In Section 3, we propose a mixed membership probabilistic model based on a simple and well-defined similarity function.

3 A Mixed Membership Model

In this section, we propose a simple generative model which admits multiple membership (of actors) for social network clustering. The development of the model is based on a probabilistic relationship between the observed adjacency matrix $\boldsymbol{A}$ and a similarity function—more specifically, the cosine similarity.

We start by introducing some additional notations and preliminaries. Consider a social network consisting of n nodes to be clustered into h distinct communities with $h \leqslant n$. For each node $i = 1, 2, \ldots, n$, let $Z_i = (Z_{i1}, Z_{i2}, \ldots, Z_{ih})^\top$, where $\sum_{k=1}^{h} Z_{ik} = 1$, be an $h \times 1$ vector that represents the mixed membership of node i across h communities. To be specific, for $1 \leqslant k \leqslant h$, Z_{ik} refers to the probability that node i belongs to community k. The very special case $Z_{ik} = 1$ for some k indicates that node i is assigned to community k with probability 1 without uncertainty, though it is very rare in practice.

For any two s-dimensional vectors $\boldsymbol{x}$ and $\boldsymbol{y}$, the cosine similarity between $\boldsymbol{x}$ and $\boldsymbol{y}$ is

$$\cos(\boldsymbol{x}, \boldsymbol{y}) = \frac{\boldsymbol{x}^\top \boldsymbol{y}}{\|\boldsymbol{x}\|_2 \|\boldsymbol{y}\|_2} = \frac{\sum_{r=1}^{s} x_r y_r}{\sqrt{\sum_{r=1}^{s} x_r^2} \sqrt{\sum_{r=1}^{s} y_r^2}}, \tag{1}$$

where $\|\cdot\|_2$ refers to the standard ℓ_2 norm. Thus, the corresponding dissimilarity function is (1 − cosine similarity).

We choose the cosine similarity as the measure of similarity in our model for three major reasons:

1. Cosine similarity is a simple measure, and it can be easily applied to high-dimensional data.
2. Cosine similarity has a standard statistical interpretation, as it is equivalent to the Pearson correlation coefficient for the data that are centered by mean.
3. Cosine similarity is defined on [0, 1], so it is ready for modeling link density.

Recall the adjacency matrix $\boldsymbol{A} = (A_{ij})$. Assuming that A_{ij}'s are mutually independent, we incorporate a Bernoulli model into the link distribution of A_{ij} for nodes i and j, given their mixed community membership Z_i and Z_j; that is,

$$\mathbf{Pr}(A_{ij} = a_{ij}) = p(Z_i, Z_j)^{a_{ij}} \left(1 - p(Z_i, Z_j)\right)^{1-a_{ij}},$$

where

$$p(Z_i, Z_j) = \cos(Z_i, Z_j) = \left(\frac{Z_i^\top Z_j}{\|Z_i\|_2 \|Z_j\|_2}\right)$$

and

$$a_{ij} = \begin{cases} 1, & \text{if nodes } i \text{ and } j \text{ are connected;} \\ 0, & \text{otherwise.} \end{cases}$$

By the assumption of conditional independence, we obtain the likelihood function of the adjacency matrix $\boldsymbol{A}$,

$$\mathbf{Pr}(\boldsymbol{A} \mid \boldsymbol{Z}) = \prod_{1 \leqslant i < j \leqslant n} p(Z_i, Z_j)^{a_{ij}} (1 - p(Z_i, Z_j))^{1-a_{ij}}, \tag{2}$$

where $\boldsymbol{Z} = (Z_1, Z_2, \ldots, Z_n)$ is an $h \times n$ matrix which represents community memberships of all nodes. It is worth mentioning that $\boldsymbol{Z}$ directly reflects node memberships, so should not be interpreted as latent positions for LSM (Hoff et al., 2002). Membership parameter and latent position are conceptually nonequivalent, though the latter usually has impact on network connectivity and is implicitly related to node membership. Our goal is to predict $\boldsymbol{Z}$ given the observation of $\boldsymbol{A}$, which can be done via an algorithm shown in Section 4.

4 Parameter Estimation

In this section, we estimate the parameters in our mixed membership model via a standard Bayesian method. At first, we posit a prior distribution for $\boldsymbol{Z}$. Notice that each component in $\boldsymbol{Z}$, Z_i, consists of h elements representing probabilities adding up to 1. Dirichlet distribution appears a reasonable and widely-accepted choice for its prior. For $1 \leqslant i \leqslant n$, let Z_i's be random variables such that

$$Z_i \overset{i.i.d.}{\sim} \text{Dirichlet}(\boldsymbol{\alpha}),$$

where $\boldsymbol{\alpha}$ is an h-dimensional hyperparameter vector. The initial selection of $\boldsymbol{\alpha}$ is flexible unless related information is available. In practice, one may choose each element in $\boldsymbol{\alpha}$ to be equal to $1/h$. For each Z_i in $\boldsymbol{Z}$, our goal is to approximate the posterior distribution of Z_i given $\boldsymbol{A}$. We exploit the Gibbs sampling algorithm proposed by Gelfand and Smith (1990).

The Gibbs sampling is a well-developed MCMC algorithm, which is popular for its simplicity and versatility. The Gibbs sampling was first appeared in Geman and Genman (1984), and the theoretical properties of the algorithm were discussed extensively by Casella and George (1992); Gelfand and Smith (1990). It was proven in Geman and Genman (1984) that the distribution of simulated samples converges to the posterior distribution of true parameters given the observations, regardless of the starting state (i.e., the prior distribution). The key of Gibbs sampling is to simulate the next generation of unknown parameters based on the estimates at the current state. Let $\boldsymbol{Z}^{(m)} = (Z_1^{(m)}, Z_2^{(m)}, \ldots, Z_n^{(m)})$ be the estimate in the current iteration. We simulate $\boldsymbol{Z}^{(m+1)}$ in the following way:

1. Simulate $Z_1^{(m+1)}$ from the posterior distribution of Z_1 given $Z_2^{(m)}, \ldots, Z_n^{(m)}$, $\boldsymbol{\alpha}$ and $\boldsymbol{A}$.
2. For $i = 2, 3, \ldots, n-1$, simulate $Z_i^{(m+1)}$ from the posterior distribution of Z_i given $Z_1^{(m+1)}, \ldots, Z_{i-1}^{(m+1)}, Z_{i+1}^{(m)}, \ldots, Z_n^{(m)}$, $\boldsymbol{\alpha}$ and $\boldsymbol{A}$.
3. Simulate $Z_n^{(m+1)}$ from the posterior distribution of Z_1 given $Z_1^{(m+1)}, \ldots, Z_{n-1}^{(m+1)}$, $\boldsymbol{\alpha}$ and $\boldsymbol{A}$.

In order to implement the algorithm, we derive the posterior distribution of Z_i, given $Z_1, \ldots, Z_{i-1}, Z_{i+1}, \ldots, Z_n$, $\boldsymbol{\alpha}$ and $\boldsymbol{A}$. For brevity, denote $Z_{-i} = (Z_1, \ldots, Z_{i-1}, Z_{i+1}, \ldots, Z_n)$. According to the definition of conditional probability, we have

$$\begin{aligned}
\mathbf{Pr}(Z_i \mid Z_{-i}, \boldsymbol{\alpha}, \boldsymbol{A}) &= \frac{\mathbf{Pr}(\boldsymbol{Z}, \boldsymbol{A} \mid \boldsymbol{\alpha})}{\mathbf{Pr}(Z_{-i}, \boldsymbol{A} \mid \boldsymbol{\alpha})} \\
&\propto \mathbf{Pr}(\boldsymbol{A} \mid \boldsymbol{Z}, \boldsymbol{\alpha}) \mathbf{Pr}(\boldsymbol{Z} \mid \boldsymbol{\alpha}) \\
&\propto \prod_{1 \leqslant i < j \leqslant n} p(Z_i, Z_j)^{a_{ij}} \left(1 - p(Z_i, Z_j)\right)^{1 - a_{ij}} \prod_{k=1}^{h} Z_{ik}^{\alpha_k - 1}. \qquad (3)
\end{aligned}$$

Since the density function expressed in Equation (3) is not from any well-known distribution, we use another well-studied MCMC algorithm—the Metropolis Hastings sampling—to simulate the density function at each Gibbs iteration. We present the Gibbs sampling procedures in

Algorithm 1. Notice that *burninNum* in the input of Algorithm 1 refers to a burn-in number—a threshold of the Gibbs iterations, after which the distribution of our simulated samples converges to the posterior distribution of the target parameters. We thus only keep the simulated estimates after the burn-in number (as reflected in Line 11 in Algorithm 1). We usually choose a large burn-in number such that with a high probability, the MCMC iterations have converged to the true posterior distribution.

Algorithm 1: The Gibbs sampling algorithm for the proposed mixed membership model.

Input: *burninNum* = 5000, *size* = 10000, empty set *posteriorSample*

1 Initialization $Z_{ik} \leftarrow \frac{1}{h}$ for all $i = 1, \ldots, n$ and $k = 1, \ldots, h$;

2 Initialization *iterNum* $\leftarrow 1$;

3 **repeat**

4 **for** $i = 1$ to n **do**

5 Simulate $T_i \sim \text{Dirichlet}(\boldsymbol{\alpha})$;

6 Simulate $U \sim \text{Uniform}(0, 1)$;

7 **if** $U < \frac{\prod_{1 \leqslant i \neq j \leqslant n} p(T_i, Z_j)^{a_{ij}} \left(1 - p(T_i, Z_j)\right)^{1 - a_{ij}} \prod_{k=1}^{h} T_{ik}^{\alpha_k - 1}}{\prod_{1 \leqslant i \neq j \leqslant n} p(Z_i, Z_j)^{a_{ij}} \left(1 - p(Z_i, Z_j)\right)^{1 - a_{ij}} \prod_{k=1}^{h} Z_{ik}^{\alpha_k - 1}}$ **then**

8 Set $Z_i \leftarrow T_i$

9 **if** *iterNum* > *burninNum* **then**

10 Add $\boldsymbol{Z} = (Z_1, Z_2, \ldots, Z_n)$ to *posterioSample*

11 Set *iterNum* $\leftarrow$ *iterNum*+1 ;

12 **until** *iterNum* > *size* + *burninNum*;

Output: *posteriorSample*

After obtaining the *posteriorSample* of $\boldsymbol{Z}$, we compute the sample mean $\bar{Z}_i$ as the Bayes estimate for Z_i, for each $i = 1, 2, \ldots, n$. For hard clustering, i.e., each of the nodes in the network only belongs to one community, so we assign every node to the community with the associated probability dominating the estimated membership parameter, i.e., $\text{argmax}_k(\bar{Z}_{ik})$.

5 Example: Zachary Karate Club Data

In this section and the next, we evaluate the performance of our mixed membership model by applying it to two empirical social network data. The first that we consider is the Zachary karate club data, which was collected and used to study conflict and fission in small groups by Zachary (1977). The data was from a university-based karate club of 34 members, who were tentatively divided into two groups due to an incipient conflict between the president of the club and the opposing faction. Consider the club as a social network consisting 34 nodes that represent club members. Each pair of the nodes are formalized by adding an edge in between if they are observed to interact outside normal activities, interpreted as "extra" friendship in Zachary (1977). A total of 78 (undirected) edges are observed; see Zachary (1977, Figure 1). The corresponding adjacency matrix was presented in Zachary (1977, Figure 2).

We apply the mixed membership model proposed in Section 3 to split the karate club members into two factions, and compare our clustering result with the ground truth released by Zachary (1977). Based on the feature of the karate club network data and the background

Table 1: Mixed membership result for the Zachary karate club data.

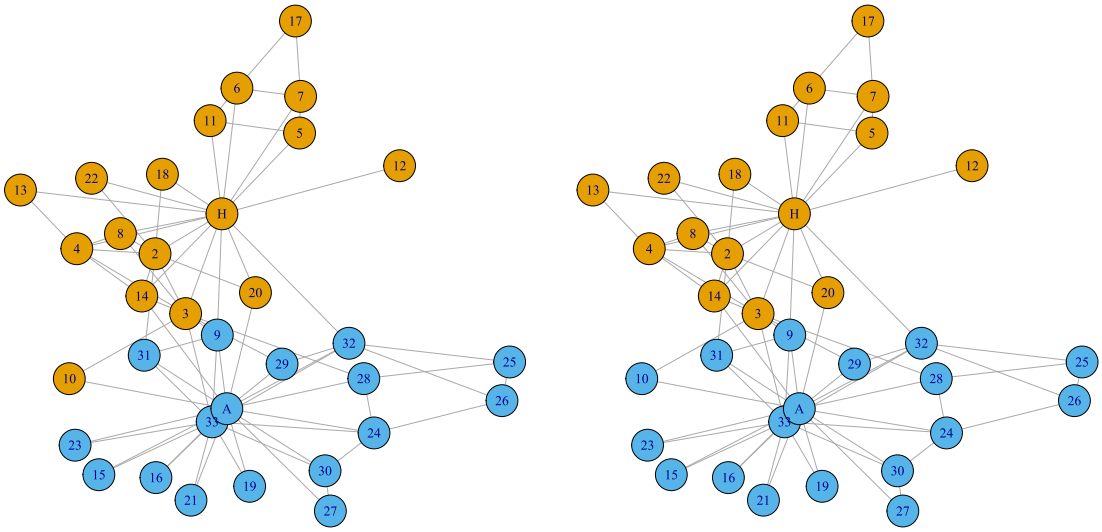
Node	Z_{i1}	Z_{i2}	Cluster	Node	Z_{i1}	Z_{i2}	Cluster
H	0.8490	0.1510	1	18	0.7764	0.2236	1
2	0.7190	0.2810	1	19	0.0683	0.9317	2
3	0.6144	0.3856	1	20	0.6932	0.3068	1
4	0.7034	0.2966	1	21	0.1063	0.8937	2
5	0.9466	0.0534	1	22	0.7694	0.2306	1
6	0.9741	0.0259	1	23	0.1181	0.8819	2
7	0.9791	0.0209	1	24	0.2443	0.7557	2
8	0.6973	0.3027	1	25	0.3895	0.6105	2
9	0.1943	0.8057	2	26	0.3777	0.6223	2
10	0.5001	0.4999	1	27	0.1871	0.8129	2
11	0.9399	0.0601	1	28	0.3620	0.6380	2
12	0.7631	0.2369	1	29	0.4419	0.5581	2
13	0.7624	0.2376	1	30	0.1815	0.8185	2
14	0.6757	0.3243	1	31	0.1878	0.8122	2
15	0.0765	0.9235	2	32	0.3934	0.6066	2
16	0.0855	0.9145	2	33	0.0602	0.9398	2
17	0.9775	0.0225	1	*A*	0.0880	0.9120	2

story, we set the number of communities $h = 2$. We implement Algorithm 1, for which the *burninNum* and *size* respectively take values 5,000 and 10,000. The posterior mean $\bar{Z}_{ik}$, for $i = 1, 2, \ldots, n$ and $k = 1, 2$, is used as the Bayes estimate for the mixed membership parameter Z_{ik}. The result is presented in Table 1. If a hard clustering framework is considered, we present a graphic summary in Figure 1a, where the nodes in different clusters are distinguished by different colors: orange for community 1 and blue for community 2.

For the purpose of comparison, the ground truth corresponding to Zachary (1977, Figure 1) and Zachary (1977, Table 1) is portrayed in Figure 1b. We observe that the entity labeled with 10 is the only misclassified node according to our model. We cluster node 10 into community 1, but in reality node 10 joins community 2. The occurrence of misclassification of node 10 is probably because the node is connected with one node (node 3) from community 1, and is also connected with one node (node *A*) from community 2. However, node *A* is the center of community 2, hence more influential in the network. Additionally, Table 1 shows that the membership parameter estimate for node 10 is 0.5001 for community 1 versus 0.4999 for community 2, so the difference is minimal.

6 Example: Bottlenose Dolphin Network Data

In this section, we analyze the bottlenose dolphin network data from Lusseau et al. (2003). A study of identifying the roles that bottlenose dolphins played in their social network was conducted by Lusseau and Newman (2004). The network data was collected for 62 bottlenose dolphins living in Doubtful Sound, New Zealand, over a period of seven years from 1994 to 2001. The bottlenose dolphins are represented by nodes in the network, and ties between nodes are interpreted as associations between dolphin pairs occurring more often (due to some sort of



(a) Clustering result of the Zachary karate club data based on the proposed mixed membership model.

(b) Ground truth of the clustering of the Zachary karate club data.

Figure 1: Comparison between the clustering result and the ground truth of the Zachary karate club data.

homophily) than expected by chance. There is a total of 318 edges observed in the network.

A natural division of the bottlenose dolphin network was discussed in Lusseau and Newman (2004), and it was done via an accurate and sensitive clustering algorithm proposed by Girvan and Newman (2002). The algorithm therein was based on a newly-defined "betweenness" measure generalized from the one defined in Freeman (1977). Two communities were detected for the bottlenose dolphin network, shown in Lusseau and Newman (2004, Figure 1(a)), as well as in Figure 2b, for the purpose of comparison.

We set $h = 2$ in our mixed membership model based on the conclusion from Lusseau and Newman (2004). Both of the *burninNum* and *size* take value 50000. Executing Algorithm 1 with the new *burninNum* and *size*, we obtain the mixed membership probabilities for the bottlenose dolphins, organized in Table 2. We also depict the hard clustering result in Figure 2 for a better visualization. The nodes in community 1 are colored with orange, while the nodes in community 2 are colored with blue.

Comparing the clustering result of our mixed membership model and that of the betweenness-based model in Lusseau and Newman (2004), we realize that the community classification matches for most of the nodes in the network, except for "Beak", "Bumper", "Fish", "Oscar", "PL", "SN89", "SN96" and "TR77" on the boundary. Lusseau and Newman (2004), in fact, assigned these dolphins to a sub-community of Community 1 using their algorithm.

7 Simulations

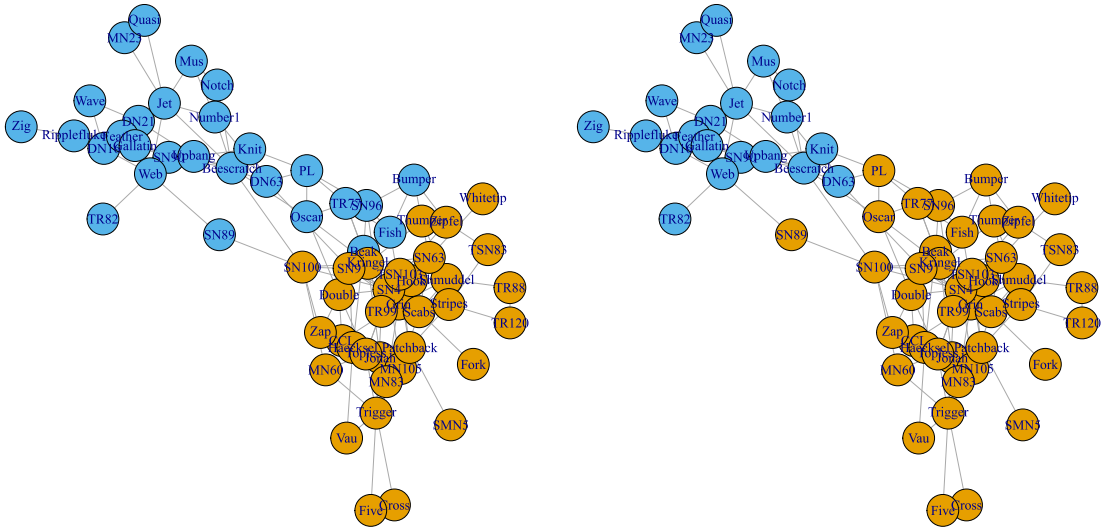
We show the identifiability and reliability of our model as well as the proposed algorithm through two empirical social network examples in Sections 5 and 6. However, both of those networks only contain a relatively small number of nodes which are only divided into two communities. In this

Table 2: Mixed membership result for the bottlenose dolphin network data.

Node	Z_{i1}	Z_{i2}	Cluster	Node	Z_{i1}	Z_{i2}	Cluster
1	0.0126	0.9874	2	32	0.3661	0.6339	2
2	0.1125	0.8875	2	33	0.3610	0.6390	2
3	0.0029	0.9971	2	34	0.7588	0.2412	1
4	0.9005	0.0995	1	35	0.6634	0.3366	1
5	0.5303	0.4697	1	36	0.5429	0.4571	1
6	0.2785	0.7215	2	37	0.8460	0.1540	1
7	0.2313	0.7687	2	38	0.8132	0.1868	1
8	0.0721	0.9279	2	39	0.7190	0.2810	1
9	0.8991	0.1009	1	40	0.4481	0.5519	2
10	0.2386	0.7614	2	41	0.8124	0.1876	1
11	0.0082	0.9918	2	42	0.1883	0.8117	2
12	0.5440	0.4560	1	43	0.0080	0.9920	2
13	0.5295	0.4705	1	44	0.7278	0.2722	1
14	0.2291	0.7709	2	45	0.6549	0.3451	1
15	0.7747	0.2253	1	46	0.9868	0.0132	1
16	0.9554	0.0446	1	47	0.5691	0.4309	1
17	0.7469	0.2531	1	48	0.0175	0.9825	2
18	0.2039	0.7961	2	49	0.3791	0.6209	2
19	0.9943	0.0057	1	50	0.5713	0.4287	1
20	0.0717	0.9283	2	51	0.7497	0.2503	1
21	0.7326	0.2674	1	52	0.9920	0.0080	1
22	0.9909	0.0091	1	53	0.7884	0.2116	1
23	0.3629	0.6371	2	54	0.5116	0.4884	1
24	0.9187	0.0813	1	55	0.1423	0.8577	2
25	0.9905	0.0095	1	56	0.9420	0.0580	1
26	0.1431	0.8569	2	57	0.3022	0.6978	2
27	0.1348	0.8652	2	58	0.2325	0.7675	2
28	0.1317	0.8683	2	59	0.5491	0.4509	1
29	0.0403	0.9597	2	60	0.9095	0.0905	1
30	0.9912	0.0088	1	61	0.3816	0.6184	2
31	0.0400	0.9600	2	62	0.5052	0.4948	1

section, we run a few more simulations to further evaluate the performance of our algorithm. We simulate several SBMs with different predetermined community structure. Each block in the simulated SBMs is generated by implementing an algorithm for the Erdös-Rényi graph (Gilbert, 1959). There are three key parameters for simulated SBMs: class size, within-cluster link density and cross-cluster link density. Noticing that the community structure of the simulated networks is known, we can use this information as the ground truth for assessment.

Two well-defined metrics, the Normalized Mutual Information (NMI) (Meilă, 2007) and the Adjusted Rand Index (ARI) (Rand, 1971), are adopted to examine the closeness between the clustering results of our algorithm and the ground truths. In addition, we implement another commonly-used method for social network clustering—the modularity maximization algo-



(a) Clustering result of the bottlenose dolphin network data based on the proposed model.

(b) Ground truth of the clustering of the bottlenose dolphin network data.

Figure 2: Comparison between the clustering result and the ground truth of the bottlenose dolphin network data.

Table 3: Link density summaries for the simulated SBMs.

	SBM1		SBM2		SBM3		SBM4			SBM5			
	C_1	C_2	C_1	C_2	C_1	C_2	C_1	C_2	C_3	C_1	C_2	C_3	C_4
size	80	20	80	20	400	100	100	10	10	40	30	20	20
C_1	0.8	0.02	0.8	0.2	0.8	0.02	0.8	0.02	0.02	0.8	0.02	0.02	0.02
C_2	0.02	0.8	0.2	0.8	0.02	0.8	0.02	0.8	0.02	0.02	0.8	0.02	0.02
C_3							0.02	0.02	0.8	0.02	0.02	0.8	0.02
C_4										0.02	0.02	0.02	0.8

rithm (Newman, 2006)—to the simulated SBMs for further comparison.

We simulate a total of five SBMs with known community structure as summarized in Table 3. SBM1 and SBM2 are both in moderate size (i.e., of order 100), containing two communities (of sizes 80 and 20, respectively). The within-cluster link densities are significantly high (0.8) for both SBMs, whereas the cross-cluster link density of SBM1 is much smaller than that of SBM2 (0.02 vs. 0.2). Next, we consider a larger network. We simulate SBM3 of order 500, of which 400 nodes form one community, and the rest 100 nodes form the other. Although one community is much lager than the other, both of the community sizes are generally large in SBM3. Then, we consider networks containing more than two communities. There are three communities in SBM4 and four communities in SBM5, respectively. In SBM4, the size of one community (100) is significantly larger than those of the other two (10 for each). Empirically, extremely small-size communities are likely to cause problems for network clustering. In SBM5, all the communities are quite close in size.

For each SBM, we set *burninNum* at 1000 and *size* at 2000, respectively. The proposed algorithm is run for 30 times, and for each result, both ARI and NMI are computed. The averages of all 30 ARI's (i.e., $\widehat{\text{ARI}}$) and NMI's (i.e., $\widehat{\text{NMI}}$) are used as estimates for evaluating the performance of the algorithm. In addition, we implement the modularity maximization algorithm to all five simulated SBMs, and compute the corresponding ARI and NMI. These results are presented in Table 4.

Table 4: Evaluation of clustering results.

	Algorithm 1		Mod. max.	
	$\widehat{\text{ARI}}$	$\widehat{\text{NMI}}$	ARI	NMI
SBM1	0.8215	0.7392	0.2661	0.4102
SBM2	0.7246	0.6466	0.2025	0.2555
SBM3	0.8863	0.8112	1.0000	1.0000
SBM4	0.8772	0.7633	0.2194	0.3218
SBM5	1.0000	1.0000	1.0000	1.0000

We observe that the proposed algorithm performs well in general for all simulated SBMs. On the other hand, it seems that the modularity maximization algorithm undergoes several severe clustering problems. The first problem that we notice is over-clustering. In theory, there is no cluster structure in the Erdös-Rényi graph. However, the modularity maximization algorithm divides predetermined communities (i.e., the Erdös-Rényi graphs) to reach a higher modularity index for small networks, reflected in the clustering results for SBM1 and SBM2. Both of the clustering results indicate that four communities are needed for these two simulated networks so as to attain the global maximum of the modularity index. Second, the modularity maximization algorithm also has under-clustering problem sometimes, especially when communities are extremely small. In SBM4, the modularity index reaches the global maximum when the two smaller communities merge together. The inconsistency of the modularity maximization algorithm was discussed extensively by Bickle and Chen (2009). However, it seems that the modularity maximization algorithm overperforms when all the communities are large in size, for instance, SBM3. Besides, our algorithm and the modularity maximization algorithm both perform perfectly well when the sizes of communities are similar in the network. Nevertheless, we conclude that the proposed algorithm is more robust for social network clustering.

8 Concluding Remarks

In this paper, we develop a simple but novel model-based method for social network clustering. We adopt the cosine function to measure similarities between nodes. In addition, we propose an algorithm based on the Gibbs sampling to simulate posterior samples for mixed community membership for entities in the network. Our model is not only flexible for fuzzy clustering, but also amenable for hard clustering. We would like to point out that our model is reliable due to solid theoretical foundation of Bayesian approach and MCMC algorithms. We evaluate the performance of our model through two empirical social network data and simulations. Based on comparisons with ground truth, we conclude that our model provides accurate clustering for social network data

At last, we discuss several limitations of our model, and propose some future studies. First, it is known that MCMC algorithms are slow to achieve stationary distribution. The complexity of the proposed algorithm in this paper is $O(n^2)$ for each Gibbs iteration. In addition, a large number of *burninNum* is usually needed for ensuring convergence. Admittedly, the algorithm is not efficient especially when the number of parameters or the size of network data or both are large. There is an urge of developing faster algorithms for our mixed membership model. One alternative is the Hamiltonian Monte Carlo (HMC) algorithm, which can accelerate convergence to the target distribution by simulating Hamiltonian dynamics. We refer the interested readers to Neal (2011) for a detailed explanation of HMC, and to Betancourt (2017) for an exposition of the intuition behind HMC. Another possible approach is to use variational Bayesian methods to convert simulation procedures to optimization problems, and then implement some appropriate approximation algorithms.

Second, our current model itself can be improved. (1) The proposed model measures node relationship based on cosine similarity, which is analogous to Pearson correlation, so it may fail to preserve membership homophily for the nodes close or on the cluster boundary. These nodes usually have similar entries in their corresponding membership variables, but the proposed model favors connections among these nodes regardless of their actual membership information. For instance, suppose $Z_i = (0.51, 0.49)$ and $Z_j = (0.49, 0.51)$, we then have $p(Z_i, Z_j) = 0.999$ albeit i and j belonging to different communities in our setting; (2) The proposed model does not focus on sparse networks particularly. One may consider tweaking cosine similarity as $p(Z_i, Z_j) = \rho_n \cos(Z_i, Z_j)$ with a scaling factor $\rho_n \to 0$ as n increases to incorporate network sparsity. the proposed model yet accounts for the information possibly contained in the nodes. We would like to consider a more complete model which utilizes those auxiliary variables so as to further improve clustering accuracy.

Third, the communities considered in our model are distinct. One of our future work is to look into a possibility to extend our model to overlapping communities like in Xie et al. (2013).

Lastly, our model, as well as most other graphical generative models, requires a prior knowledge about the number of communities to which nodes are assigned. However, this number is usually unavailable. Estimating the number of communities and membership parameters simultaneously could be a challenging task. A recent research paper (Geng et al., 2019) provides us some guidance about future study in this direction.

Supplementary Material

The codes for Algorithm 1 and the implementations can be found on the journal website. The results of empirical data applications are saved in RDS files.

Acknowledgments

We would like to thank the handling AE and two anonymous reviewers for their valuable suggestions that help improve the paper quality.

References

Abbe E (2018). Community detection and stochastic block models: Recent developments. *Journal of Machine Learning Research*, 18(177): 1–86.

Airoldi EM, Blei DM, Fienberg SE, Xing EP (2008). Mixed membership stochastic blockmodels. *Journal of Machine Learning Research*, 9(65): 888–896.

Athreya A, Fishkind DE, Tang M, Priebe CE, Park Y, Vogelstein JT, et al. (2018). Statistical inference on random dot product graphs: A survey. *Journal of Machine Learning Research*, 18(226): 1–92.

Barabási AL, Albert R (1999). Emergence of scaling in random networks. *Nature*, 286(5439): 509–512.

Betancourt M (2017). A conceptual introduction to Hamiltonian Monte Carlo. arXiv preprint: https://arxiv.org/abs/1701.02434

Bickle PJ, Chen A (2009). A nonparametric view of network models and Newman-Girvan and other modularities. *Proceedings of the National Academy of Sciences of the United States of America*, 160(50): 21068–21073. https://doi.org/10.1073/pnas.0907096106

Blei DM, Ng AY, Jordan MI (2003). Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3: 993–1022.

Boguñá M, Pastor-Satorras R, Díaz-Guilera A, Arenas A (2004). Models of social networks based on social distance attachment. *Physical Review E*, 70(5): 056122. https://doi.org/10.1103/PhysRevE.70.056122

Cartwright D, Harary F (1956). Structure balance: A generalization of Heider's theory. *Psychological Review*, 63(5): 277–293. https://doi.org/10.1037/h0046049

Casella G, George EI (1992). Explaining the Gibbs sampler. *The American Statistician*, 46(3): 167–174. https://doi.org/10.1080/00031305.1992.10475878

Freeman LC (1977). A set of measures of centrality based on betweenness. *Sociometry*, 40(1): 35–41. https://doi.org/10.2307/3033543

Fronczak P, Fronczak A, Bujok M (2013). Exponential random graph models for networks with community structure. *Physical Review E*, 88(3): 032810. https://doi.org/10.1103/PhysRevE.88.032810

Gao C, Ma Z, Zhang AY, Zhou HH (2018). Community detection in degree-corrected block models. *The Annals of Statistics*, 46(5): 2153–2185.

Gelfand AE, Smith AFE (1990). Sampling-based approaches to calculating marginal densities. *Journal of the American Statistical Association*, 85(410): 398–409. https://doi.org/10.1080/01621459.1990.10476213

Geman S, Genman D (1984). Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 6(6): 721–741. https://doi.org/10.1109/TPAMI.1984.4767596

Geng J, Bhattacharya A, Pati D (2019). Probabilistic community detection with unknown number of communities. *Journal of the American Statistical Association*, 114(526): 893–905. https://doi.org/10.1080/01621459.2018.1458618

Gilbert EN (1959). Random graphs. *The Annals of Mathematical Statistics*, 30(4): 1141–1144. https://doi.org/10.1214/aoms/1177706098

Girvan M, Newman MEJ (2002). Community structure in social and biological networks. *Proceedings of the National Academy of Sciences of the United States of America*, 99(12): 7821–7826. https://doi.org/10.1073/pnas.122653799

Handcock MS, Raftery AE (2007). Model-based clustering for social networks. *Journal of the Royal Statistical Society. Series A. Statistics in Society*, 170: 301–354. https://doi.org/10.1111/j.1467-985X.2007.00471.x

Harary F (1953). On the notion of balance of a signed graph. *The Michigan Mathematical*

Journal, 2(2): 143–146. https://doi.org/10.1307/mmj/1028989917

Hoff PD, Raftery AE, Handcock MS (2002). Latent space approaches to social network analysis. *Journal of the American Statistical Association*, 97(460): 1090–1098. https://doi.org/10.1198/016214502388618906

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

Huang W, Liu Y, Chen Y (2020). Mixed membership stochastic blockmodels for heterogeneous networks. *Bayesian Analysis*, 15(3): 711–736. https://doi.org/10.1214/19-BA1163

Hunter DR, Handcock MS, Butts CT, Goodreau Morris M SM (2008). ergm: A package to fit, simulate and diagnose exponential-family models for networks. *Journal of Statistical Software*, 24(3): 1–29. https://doi.org/10.18637/jss.v024.i03

Lusseau D, Newman MEJ (2004). Identifying the role that animals play in their social networks. *Proceedings of the Royal Society B*, 271(supp(6)): 477–481.

Lusseau D, Schneider K, Boisseau OJ, Haase P, Slooten E, Dawson SM (2003). The bottlenose dolphin community of Doubtful Sound features a large proportion of long-lasting associations. *Behavioral Ecology and Sociobiology*, 54: 396–405. https://doi.org/10.1007/s00265-003-0651-y

Lyzinski V, Tang M, Athreya A, Park Y, Priebe CE (2017). Community detection and classification in hierarchical stochastic blockmodels. *IEEE Transactions on Network Science and Engineering*, 4(1): 13–26. https://doi.org/10.1109/TNSE.2016.2634322

Marchette DJ, Priebe CE (2008). Predicting unobserved links in incompletely observed networks. *Computational Statistics & Data Analysis*, 52(3): 1373–1386. https://doi.org/10.1016/j.csda.2007.03.016

Meilă M (2007). Comparing clustering—an information based distance. *Journal of Multivariate Analysis*, 98(5): 873–895. https://doi.org/10.1016/j.jmva.2006.11.013

Neal RM (2011). MCMC using Hamiltonian dynamics. In: *Handbook of Markov Chain Monte Carlo* (S Brooks, A Gelman, G Jones, XL Meng, eds.), 113–162. Chapman & Hall/CRC, Boca Raton, FL, USA.

Newman MEJ (2001). The structure of scientific collaboration networks. *Proceedings of the National Academy of Sciences of the United States of America*, 98(2): 404–409. https://doi.org/10.1073/pnas.98.2.404

Newman MEJ (2006). Modularity and community structure in networks. *Proceedings of the National Academy of Sciences of the United States of America*, 103: 8577–8582 (2006). https://doi.org/10.1073/pnas.0601602103

Newman MEJ, Strogatz SH, Watts DJ (2001). Random graphs with arbitrary degree distributions and their applications. *Physical Review E*, 64(2): 026118. https://doi.org/10.1103/PhysRevE.64.026118

Newman MEJ, Watts DJ, Strogatz SH (2002). Random graph models of social networks. *Proceedings of the National Academy of Sciences of the United States of America*, 99(supp(1): 2566–2572. https://doi.org/10.1073/pnas.012582999

Ng AY, Jordan MI, Weiss Y (2001). On spectral clustering: Analysis and an algorithm. In: *Advances in Neural Information Processing Systems 14* (TG Dietterich, S Becker, Z Ghahramani, eds.), 849–856. MIT Press, Cambridge, MA, USA.

Noroozi M, Pensky M (2022). The hierarchy of block models. *Sankhya. Series A*, 84: 64–107. https://doi.org/10.1007/s13171-021-00247-2

Nowicki K, Snijders TAB (2001). Estimation and prediction for stochastic blockstructures. *Journal of the American Statistical Association*, 96(455): 1077–1087. https://doi.org/10.1198/

016214501753208735
Ouyang G, Dipak DK, Zhang P (2020). Clique-based method for social network clustering. *Journal of Classification*, 37: 254–274. https://doi.org/10.1007/s00357-019-9310-5
Rand WM (1971). Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical Association*, 66(336): 846–850. https://doi.org/10.1080/01621459.1971.10482356
Rapoport A (1949a). Outline of a probabilistic approach to animal sociology: I. *The Bulletin of Mathematical Biophysics*, 11(3): 183–196. https://doi.org/10.1007/BF02478364
Rapoport A (1949b). Outline of a probabilistic approach to animal sociology: II. *The Bulletin of Mathematical Biophysics*, 11(4): 273–281. https://doi.org/10.1007/BF02477980
Rapoport A (1950). Outline of a probabilistic approach to animal sociology: III. *The Bulletin of Mathematical Biophysics*, 12(1): 7–17. https://doi.org/10.1007/BF02477340
Sengupta S, Chen Y (2018). A block model for node popularity in networks with community structure. *Journal of the Royal Statistical Society, Series B, Statistical Methodology*, 80(2): 365–386. https://doi.org/10.1111/rssb.12245
Sewell DK, Chen Y (2017). Latent space approaches to community detection in dynamic networks. *Bayesian Analysis*, 12(2): 351–377. https://doi.org/10.1214/16-BA1000
Shi J, Malik J (2000). Normalized cuts and image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(8): 888–905. https://doi.org/10.1109/34.868688
Snijders TAB (2001). Statistical models for social networks. *Annual Review of Sociology*, 37: 131–153. https://doi.org/10.1146/annurev.soc.012809.102709
Snijders TAB, Nowicki K (1997). Estimation and prediction for stochastic blockmodels for graphs with latent block structure. *Journal of Classification*, 14: 75–100. https://doi.org/10.1007/s003579900004
Snijders TAB, Pattison PE, Robins GL, Handcock MS (2006). New specifications for exponential random graph models. *Sociological Methodology*, 36(1): 99–153. https://doi.org/10.1111/j.1467-9531.2006.00176.x
Toivonen R, Kovanen L, Kivelä M, Onnela JP, Saramäki J, Kaski K (2009). A comparative study of social network models: Network evolution models and nodal attribute models. *Social Networks*, 31(4): 240–254. https://doi.org/10.1016/j.socnet.2009.06.004
Watts DJ, Strogatz SH (1998). Collective dynamics of "small-world" networks. *Nature*, 393: 440–442. https://doi.org/10.1038/30918
Xie J, Kelley S, Szymański BK (2013). Overlapping community detection in networks: The state-of-the-art and comparative study. *ACM Computing Surveys*, 45(4): 43. https://doi.org/10.1145/2501654.2501657
Young SJ, Scheinerman ER (2007). Random dot product graph models for social networks. In: *WAW 2007: Algorithms and Models for the Web-Graph* (A Bonato, FRK Chung, eds.), 138–149. Springer, Berlin, Heidelberg.
Zachary WW (1977). An information flow model for conflict and fission in small groups. *Journal of Anthropological Research*, 33(4): 452–473. https://doi.org/10.1086/jar.33.4.3629752