DSSR

Discussion Paper No. 97

Identifying Topic-based Communities by Combining Social Network Data and User Generated Content

Mirai Igarashi
Nobuhiko Terui

April, 2019

Data Science and Service Research
Discussion Paper

Center for Data Science and Service Research
Graduate School of Economic and Management
Tohoku University
27-1 Kawauchi, Aobaku
Sendai 980-8576, JAPAN

Identifying Topic-based Communities by Combining Social Network Data and User Generated Content

Mirai Igarashi* Nobuhiko Terui*

April, 2019

*Mirai Igarashi is Doctoral student, Tohoku University, Graduate School of Economics and Management, Kawauchi Aoba-ku, Sendai, 980-8576, Japan (E-mail: mirai.igarashi.s7@dc.tohoku.ac.jp). Igarashi acknowledges a grant by JSPS KAKENHI 18J20698. Nobuhiko Terui is Professor, Tohoku University, Graduate School of Economics and Management (E-mail: terui@econ.tohoku.ac.jp). Terui acknowledges a grant by JSPS KAKENHI (A) 17H01001.

Identifying Topic-based Communities by Combining Social Network Data and User Generated Content

Abstract

This study proposes a model for identifying communities by combining two types of data: social network data and user-generated-content (UGC). The existing models for detecting the community structure of a network employ only network information. However, not all people connected in a network share the same interests. For instance, even if students belong to the same community of "school," they may have various hobbies such as music, books, or sports. Hence, targeting various networks to identify communities according to their interests uncovered by their communications on social media is more realistic and beneficial for companies. In addition, people may belong to multiple communities such as family, work, and online friends. Our model explores multiple overlapping communities according to their topics identified using two types of data jointly. By way of validating the main features of the proposed model, our simulation study shows that the model correctly identifies the community structure that could not be found without considering both network data and UGC. Furthermore, an empirical analysis using Twitter data clarifies that our model can find realistic and meaningful community structures from large social networks and has a good predictive performance.

Keyword: Social network analysis, Community detection, User-generated-content, Topic modeling, Bayesian inference

1 Introduction

The product or information diffusion is affected by not only the communication between companies and consumers but also by interactions between consumers such as word-of-mouth on social media or product reviews on e-commerce sites; the impact of the latter is stronger in the modern social media development. Companies are required to implement various marketing activities considering such relationships between consumers. A significant first step towards learning about the relationship between customers is to grasp the community structure of their social networks. If nodes of a network can be divided into some (potentially overlapping) groups such that nodes are densely connected internally, the network is said to have a community structure. Furthermore, researchers know that social network structures with closely connected nodes, or consumers, can bring some benefits to companies such as sharing contents (Peng et al, 2018), achieving long-term popularity (Ansari et al, 2018), and accelerating product innovation (Peres, 2014). Therefore, grasping the community structure of the customers' social networks may prove to be useful for companies when planning their marketing activities.

A lot of attention has been paid to identifying community structures for a long time, and many methods have been proposed (e.g., Newman, 2006; Ng, Jordan, & Weiss, 2002; Nowicki & Snijders, 2001; Handcock, Raftery, & Tantrum, 2007). In addition to social network analysis, these methods are used in many other fields, including analysis of protein-protein interaction networks (Jeong et al, 2001), terrorists networks (Krebs, 2002), and co-author networks (Liu et al, 2005).

However, these methods focus only on network information, while more meaningful communities could be identified if other social network information was considered. For example, students belonging to the same community of "school" are thought to be connected to each other via some types of relationships to form social networks. Such networks are regarded as one community when considering only network information. At the same time, the students may be involved in various hobbies such as music, books, or sports. More meaningful segmentation can be achieved if researchers regard networks whose members have different properties (or interests) as multiple communities rather than one. To do so, text information on social

media representing user-generated-content (UGC) and uncovering members' interests can be used. Employing both the network and text information can allow researchers to recognize the community structure from two viewpoints: the density of the network connections and similarity of interests derived from text information.

To understand the community structure of a social network, the problem of people belonging to multiple communities such as family, work, and online friends should be considered in addition to network data and UGC. This problem is called community overlapping. The estimated network structure in this case can have a large deviation from that of the real network when applying methods such as hard clustering, where each node in considered to belong to a single community. The mixed membership stochastic block model (MMSB) proposed by Airoldi, Fienberg, & Xing (2008) is one of the most popular statistical generative models investing the community overlapping problem. In this study, we extend MMSB and propose a model, called the mixed membership stochastic topic block model (MMSTB), for identifying communities by combining social network data and UGC.

The rest of this paper is organized as follows: related work is discussed in Section 2. The proposed model, MMSTB, and its inference algorithm are introduced in Section 3. Section 4 examines the simulation studies conducted to validate the main features of MMSTB and choose numbers of communities and topics. Section 5 presents an application of the proposed model to a real-world network, namely, Twitter. Finally, Section 6 provides some concluding remarks.

2 Literature review

2.1 Social networks and consumer behavior

Many studies have focused on the impact of the structure of a social network on consumer behavior (see Muller & Peres, 2018, for a comprehensive review). As a remarkable achievement of these studies, researchers know that consumers belonging to the same community positively influence each other's consumption behavior. For example, Peng et al (2018) used the content shared on Twitter and found that a receiver is more likely to share content from a sender with whom they have more common followers, common followees, and common mutual followers.

Peres (2014) demonstrated that the higher the average degree a network has, the faster the product innovation progresses; that is, consumption behaviors are more active in the network. People belonging to the same community tend to have many common acquaintances because of the high density of connections; hence, they are likely to be more intimate. Based on these results, it is sensible for companies to grasp the groups (or communities) of customers who have positive influences on each other's consumption behavior.

2.2 Identifying communities using network information

A number of models have been proposed in the literature to identify the community structure of a network. They can be divided into approaches using deterministic algorithm and statistical models. One of the approaches using a deterministic algorithm is based on the modularity score introduced by Newman (2006), where modularity is a measure of the strength of connections within a network divided into modules; a network with high modularity forms dense connections between the nodes within modules but sparse connections between nodes in different modules. The algorithm proposed by Newman (2006) detects communities by maximizing modularity, and this algorithm is one of the most widely used methods due to its simplicity. Another approach using a deterministic algorithm is spectral clustering (Ng, Jordan, & Weiss, 2002), which is based on the eigenvalue decomposition of the graph Laplacian. The graph Laplacian is a matrix obtained by transforming the adjacency matrix, and the community structure can be clarified by applying some clustering methods such as k-means for the eigenvectors of the graph Laplacian.

The community detection methods using statistical models have been well developed in past decades, with the representative one being the stochastic block model (SBM) proposed by Wang & Wong (1987) and formulated by Snijders & Nowicki (1997) and Nowicki & Snijders (2001). The SBM assumes that when the cluster membership of each node is given, the relationship between nodes is generated according to some probability distribution such as the Bernoulli distribution. Many researchers have studied the statistical properties of the SBM (e.g., Sussman et al, 2012; Abbe, 2018) and proposed extended models (e.g., Karrer & Newman, 2011). Another statistical model for detecting community is the latent position cluster model (LPCM, Handcock, Raftery, & Tantrum, 2007) that extends the latent space model

(LSM, Hoff, Raftery, & Handcock, 2002). While the LSM models the edge probability using a parameter representing the latent position of a node, the LPCM introduces the hierarchical structure of the spherical Gaussian mixture model for the latent position parameter. The spherical Gaussian distribution corresponds to the latent position of the community to which nodes belong on their social network, and hence, the LPCM allows identifying the community structures.

Recently, some models have been proposed to accommodate community overlapping (Gormley & Murphy, 2010; Latouche, Birmele, & Ambroisé, 2011; McDaid & Hurley, 2010). In particular, the MMSB (Airoldi et al, 2008) allows each node to belong to multiple clusters by extending the SBM, which imposes the constraint that nodes can belong to only one cluster.

2.3 Simultaneous modeling of network Data and user-generated-content

The models introduced in Section 2.2 consider only network information (i.e., the connections between nodes). Conversely, simultaneously modeling network data and UGC for a deep understanding of modern social networks such as Twitter and Facebook would be useful, because the pieces of information allows researchers to recognize a social network structure that is more valuable for companies by accommodating the detection of heterogeneous relationships and interests across a specific community that are hidden in social network data. For instance, it is possible that there is a group of music lovers in a community of school.

Several studies on the community identification considering network data and UGC (particularly text) have been developed recently. In an earlier study, Zhou et al (2006) proposed the community user topic model (CUT). The CUT extends the latent Dirichlet allocation (LDA, Blei, Ng, & Jordan, 2003) model for natural language processing to accommodate the phenomenon that users with the same interests, that is, those who create texts with the same topic, tend to belong to the same community. However, people who belong to the same community do not necessarily create texts with the same topic, and the CUT can only clarify a limited community structure of a social network.

Pathak et al (2006) proposed the community author recipient topic (CART) model that incorporates both network and text information to extract well-connected and topically meaningful communities. Furthermore, CART allows the nodes to belong to multiple communities.

Conversely, the CART assumes textual edges, where text information appertains to edges, which is the case in e-mail networks and co-authorship networks of papers and is different from the focus of this research. In addition, unlike the CART designed only for directed graphs, our model can handle both directed and undirected graphs.

Liu et al (2009) proposed the topic-link LDA (TL-LDA) method that detects the community structure by considering information in a situation with textual nodes, which is similar to our research; however, this method assumes that each node has a single community membership. In addition, the probability of creating an edge between nodes is defined by the similarity of the community and topic proportion of the nodes. Hence, the probability is constant regardless of the direction of the edge and can be applied to undirected graphs only.

In a recent study, Bouveyron, Latouche, & Zreik (2018) proposed the stochastic block topic model (STBM) that extends the SBM by incorporating text information into the model and is suitable for both undirected and directed graphs. If a node belongs to community A and another node belongs to community B, the SBM handles any graph regardless of whether it is directed or not by estimating the probability separately for the cases of generating edges from A to B and from B to A. While our proposed method can handle the two types of graphs similar to the STBM, our method also overcomes the limitation of the STBM, where nodes can have only a single community membership.

Finally, we clarify the characteristics of our model. Table 1 summarizes the discussed models compared by five characteristics. When comparing to the models that consider either network or text information only (such as LDA, SBM, and MMSB), our model has an advantage of being able to extract well-connected topically meaningful communities by taking both types of information into account. When comparing to the models that consider both types of information, our model can be distinguished from the existing models according to the following three properties: nodes can have multiple community memberships; graphs can be both directed and undirected; text information appertains to nodes, which is the situation, where people send their own tweets to all users on their Twitter timeline. Considering these features, we call our model the MMSTB.

Table 1: Comparison between the proposed model and existing models

	Network	Text	Mixed membership	Direction of graph	Situation of text
LDA (Blei, Ng, & Jordan 2003)	×	◯	◯	-	-
SBM (Nowichi & Snijders, 2001)	◯	×	×	Both	-
MMSB (Airoldi et al., 2008)	◯	×	◯	Both	-
CUT (Zhou et al., 2006)	◯	◯	◯	Only directed	Textual edges
CART (Pathak et al., 2006)	◯	◯	◯	Only directed	Textual edges
TL-LDA (Liu et al., 2009)	◯	◯	×	Only undirected	Textual nodes
STBM (Bouveyron et al., 2018)	◯	◯	×	Both	Textual edges
MMSTB (This study)	◯	◯	◯	Both	Textual nodes

3 Model

This section describes the proposed model MMSTB for community identification. Our observed data consist of the adjacency matrix A as a network information and bag-of-words collection W as a text information. In the following, we explain the process of generating these data and inference procedure employed in MMSTB.

3.1 Model specification

First, we consider a directed network with D nodes. $D \times D$ adjacency matrix A represents the relationships between the nodes with their elements being $a_{ij} = 0$ (not connected) or 1 (connected). We assume that the network has no self-loops and therefore $a_{ii} = 0, \ \forall i$. For the relationship from node i to node j, we consider that sender i belongs to latent community $s_{ij} \in \{1, \ldots, K\}$ (K is the number of communities), while recipient j belongs to latent community $r_{ji} \in \{1, \ldots, K\}$. $D \times D$ matrix representations of latent communities are denoted as $S = (s_{ij})$ and $R = (r_{ji})$, respectively. These sender and recipient communities are assumed to follow

a categorical distribution, $s_{ij}|\eta_i \sim Categorical(\eta_i)$, $r_{ji}|\eta_j \sim Categorical(\eta_j)$, where $\eta_i = (\eta_{i1}, \ldots, \eta_{iK})^T$ is a community distribution which represents node i's community proportion, and $\sum_{k=1}^{K} \eta_{ik} = 1$, $\forall i$. The matrix representation of community proportions are denoted as $H = (\eta_1, \ldots, \eta_D)$. The prior distribution of H is assumed to follow a Dirichlet distribution, $\eta_i|\gamma \sim Dirichlet(\gamma)$ $(i = 1, \ldots, D)$, where $\gamma = (\gamma_1, \ldots, \gamma_K)$ is a hyperparameter.

We assume that the connection variable a_{ij} between node i to j, when s_{ij} and r_{ji} are given, follows the Bernoulli distribution that depends on the community of the nodes. That is, $a_{ij}|s_{ij}, r_{ji}, \Psi \sim Bernoulli\left(\psi_{s_{ij}, r_{ji}}\right)$, where $\psi_{kk'}$ is a probability that an edge is generated when a sender node belongs to community k and a recipient node belongs to community k'. Let $K \times K$ matrix, $\Psi = (\psi_{kk'})$ be the matrix representation of edge probabilities. Each edge probability is assumed to follow a Beta distribution, $\psi_{kk'}|\delta_{kk'}, \epsilon_{kk'} \sim Beta(\delta_{kk'}, \epsilon_{kk'})$ $(k, k' = 1, \ldots, K)$, where δ, ϵ are hyperparameters of the $K \times K$ matrix.

Then, the conditional joint likelihood of the network information for parameters and latent variables, when the community distribution, H, is given, is

$$
\begin{aligned}
p(A, S, R, \Psi|H) = p(A|S, R, \Psi)p(S|H)p(R|H)p(\Psi|\delta, \epsilon) \\
\prod_{i=1}^{D} \left\{ \prod_{j=1, j\neq i}^{D} \{p(a_{ij}|s_{ij}, r_{ji}, \Psi)p(s_{ij}|\eta_i)p(r_{ji}|\eta_j)\} \right\} \times \\
\prod_{k=1}^{K} \prod_{k'=1}^{K} p(\psi_{kk'}|\delta_{kk'}, \epsilon_{kk'}),
\end{aligned} \tag{1}
$$

where δ and ϵ are hyperparameters for Ψ and fixed to some values.

For UGC, this study considers text data such as contents created by people on social media or blogs. However, other types of UGC (e.g., images, movies, or music) can also be handled by MMSTB after appropriate modeling.

Next, we consider modeling text content. Node i creates some texts that are vectorized as M_i words ignoring the order, i.e., "bag-of-words". Node i's mth word w_{im} $(m = 1, \ldots, M_i)$ is assumed to have latent community $x_{im} \in \{1, \ldots, K\}$ and latent topic $z_{im} \in \{1, \ldots, L\}$ (L is the number of topics), as in the case of the conventional LDA model. The array representations of words, word communities, and word topics are denoted as W, X, and Z,

respectively, and each component of the arrays is a M_i-dimensional vector. We assume that word community x_{im} follows a categorical distribution, $x_{im}|\eta_i \sim Categorical(\eta_i)$. We note that η_i is a parameter for generating not only word community x_{im} but also node communities s_{ij} and r_{ij} as mentioned before, that is, η_i is a common parameter for modeling networks and texts that connects the two types of information.

A word topic z_{im} is assumed to follow a categorical distribution, $z_{im}|x_{im}, \Theta \sim Categorical(\theta_{x_{im}})$, where $\theta_k = (\theta_{k1}, \ldots, \theta_{kL})^T$ is the topic distribution representing community k's topic proportion, and $\sum_{l=1}^{L} \theta_{kl} = 1, \forall k$. The matrix representations of topic proportions are denoted as $\Theta = (\theta_1, \ldots, \theta_K)$. Each topic distribution is assumed to follow a Dirichlet distribution, $\theta_k|\alpha \sim Dirichlet(\alpha)$ $(k = 1, \ldots, K)$, where $\alpha = (\alpha_1, \ldots, \alpha_L)$ is a hyperparameter.

When a word topic z_{im} is given, the corresponding word $w_{im} \in \{1, \ldots, V\}$ is assumed to follow a categorical distribution that depends on word topic, i.e., $w_{im}|z_{im}, \Phi \sim Categorical(\phi_{z_{im}})$, where $\phi_l = (\phi_{l1}, \ldots, \phi_{lV})^T$ (V is the number of unique words in the corpus) is the word distribution representing the word generation probability, and $\sum_{v=1}^{V} \phi_{lv} = 1, \forall l$. The matrix representation of word distributions is denoted as $\Phi = (\phi_1, \ldots, \phi_L)$. Each word distribution is assumed to follow a Dirichlet distribution, $\phi_l|\beta \sim Dirichlet(\beta)$ $(l = 1, \ldots, L)$, where β is a hyperparameter.

Then, the conditional joint likelihood of text information, when H is given, is

$$\begin{aligned} p(W, X, Z, \Theta, \Phi|H) = {} & p(W|Z, \Phi)p(Z|X, \Theta)p(X|H)p(\Theta|\alpha)p(\Phi|\beta) \\ & \prod_{i=1}^{D} \left\{ \prod_{m=1}^{M_i} \{p(w_{im}|z_{im}, \Phi)p(z_{im}|x_{im}, \Theta)p(x_{im}|\eta_i)\} \right\} \times \\ & \prod_{k=1}^{K} p(\theta_k|\alpha) \prod_{l=1}^{L} p(\phi_l|\beta), \end{aligned} \tag{2}$$

where α and β are hyperparameters for Θ and Φ, respectively, and these are fixed to some values. Under the assumption of conditional independence of Equations (1) and (2), when nodes' community distribution, H, is given, the full joint likelihood of MMSTB is obtained

by the product of Equations (1) and (2) multiplied by the density of H, $p(H|\gamma)$,

$$
\begin{aligned}
&p(A, W, S, R, X, Z, H, \Psi, \Theta, \Phi) \\
&= \prod_{i=1}^{D} \left\{ \prod_{j=1, j\neq i}^{D} \{p(a_{ij}|s_{ij}, r_{ji}, \Psi) p(s_{ij}|\eta_i) p(r_{ji}|\eta_j)\} \times \right. \\
&\qquad \left. \prod_{m=1}^{M_i} \{p(w_{im}|z_{im}, \Phi) P(z_{im}|x_{im}, \Theta) p(x_{im}|\eta_i)\} \right\} \times \\
&\prod_{i=1}^{D} p(\eta_i|\gamma) \prod_{k=1}^{K} \left\{ p(\theta_k|\alpha) \prod_{k'=1}^{K} p(\psi_{kk'}|\delta_{kk'}, \epsilon_{kk'}) \right\} \times \prod_{l=1}^{L} p(\phi_l|\beta). \qquad (3)
\end{aligned}
$$

The corresponding graphical model is provided in the left panel of Figure 1, while the right panel shows the model overview when $K = 3$ and $L = 3$. This illustrates the generative process of the relationship between node i and j, a_{ij}, and the mth word of the text created by node i, w_{im}. A sender community s_{ij} (orange) and a recipient community r_{ji} (blue) are generated according to their community distributions with parameters η_i and η_j, respectively. Then the link probability, $p(a_{ij} = 1|s_{ij}, r_{ji}, \Psi)$, is denoted as $\psi_{s_{ij}, r_{ji}}$, which is an (s_{ij}, r_{ji}) element of the edge probability matrix Ψ. Conversely, a word community x_{im} (orange) is also generated according to the same distribution with parameter η_i. Then, a word topic z_{im} (yellow) is generated according to topic distribution with θ_2 corresponding to x_{im}. Finally, a word w_{im} is generated according to word distribution ϕ_1 because z_{im} equals to 1 (yellow).

3.2 Conditional posterior distributions and parameter estimation

Many methods for estimating topic models have been proposed (e.g., the variational Bayesian method and sequential learning method). Among them, the most widely used method is the collapsed Gibbs sampler (CGS) proposed by Griffiths & Steyvers (2004), which samples only latent variables by integrating out parameters. CGS can estimate topic models more efficiently compared to the Gibbs sampler that directly samples all parameters. This study uses CGS for estimating MMSTB's parameters.

MMSTB has four types of model parameters: namely, community distributions H, edge probabilities Ψ, topic distributions Θ, and word distributions Φ. First, we derive the conditional posterior distributions from the full joint distribution according to Equation (3) by

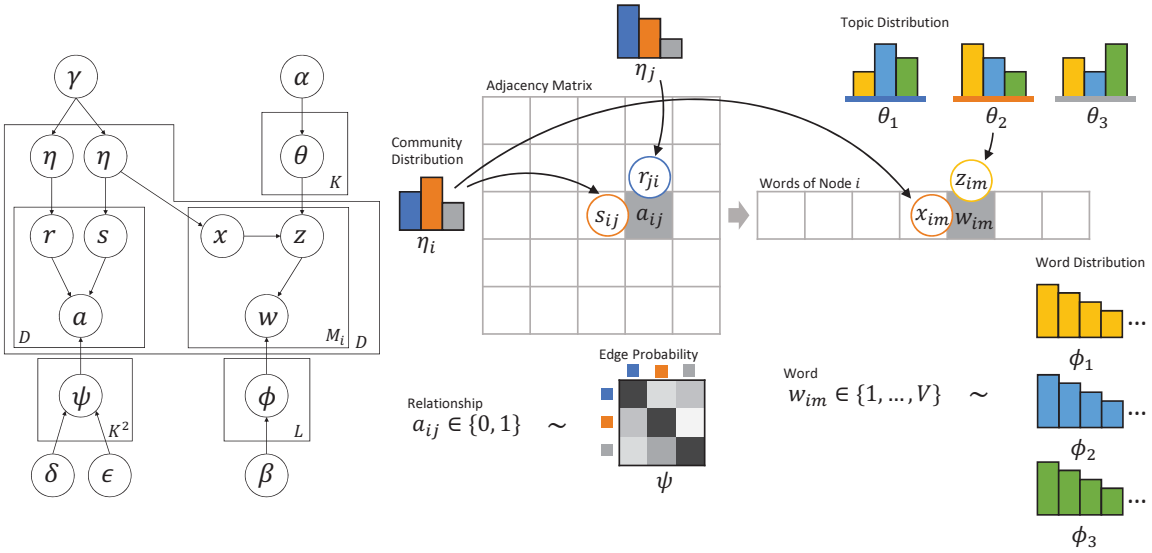


Figure 1: Graphical representation of the mixed membership stochastic block model (left) and the model overview when the number of community (K) is 3 and the number of topics (L) is 3 (right)

using the conjugacy, Dirichlet-Categorical relationship for H, Θ, and Φ and Beta-Bernoulli relationship for Ψ. Then, the full conditional posterior distributions, when data (A, W), the state of latent variables (S, R, X, Z), and hyperparameters are given, can be easily derived as follows:

$$P(\eta_i|S,R,X,\gamma) = \frac{\Gamma\left(\sum_k N_{ik} + M_{ik} + \gamma_k\right)}{\prod_k \Gamma(N_{ik} + M_{ik} + \gamma_k)} \prod_{k=1}^{K} \eta_{ik}^{N_{ik}+M_{ik}+\gamma_k} \tag{4}$$

$$P(\psi_{kk'}|A,S,R,\delta,\epsilon) = \frac{\Gamma(n_{kk'}^{(+)} + n_{kk'}^{(-)} + \delta_{kk'} + \epsilon_{kk'})}{\Gamma(n_{kk'}^{(+)} + \delta_{kk'})\Gamma(n_{kk'}^{(-)} + \epsilon_{kk'})} \psi_{kk'}^{\mathbb{I}(a_{ij}=1)}(1-\psi_{kk'})^{\mathbb{I}(a_{ij}=0)} \tag{5}$$

$$P(\theta_k|X,Z,\alpha) = \frac{\Gamma\left(\sum_l M_{kl} + \alpha_l\right)}{\Pi_l \Gamma(M_{kl} + \alpha_l)} \prod_{l=1}^{L} \theta_{kl}^{M_{kl}+\alpha_l} \tag{6}$$

$$P(\phi_l|W,Z,\beta) = \frac{\Gamma\left(\sum_v M_{lv} + \beta_v\right)}{\Pi_v \Gamma(M_{lv} + \beta_v)} \prod_{v=1}^{V} \phi_{lv}^{M_{lv}+\beta_v}, \tag{7}$$

where N_{ik} is the count number of when node i is assigned community k on the edges from node i to other nodes and from other nodes to node i. M_{ik} is the count number of when words in node i's document are assigned to community k. $n_{kk'}^{(+)}$ ($n_{kk'}^{(-)}$) is the number of links

(non-links) from nodes in community k to nodes in community k'. M_{kl} is the count number of when words are assigned to community k and topic l. M_{lv} is the count number of when word v is assigned to topic l. Γ is the gamma function, and $\mathbb{I}$ is the indicator function that returns 1, if the condition is satisfied, and 0 otherwise.

MMSTB has four types of latent variables: two latent variables for a relationship between node i and j, s_{ij} (sender community) and r_{ji} (recipient community), and two latent variables for a mth word of node i, x_{im} (word community) and z_{im} (word topic). The conditional posterior distributions of these four latent variables are derived by integrating out parameters (H, Ψ, Θ, Φ) as follows:

$$
\begin{aligned}
&P(s_{ij}=k, r_{ji}=k'|a_{ij}, A_{\backslash ij}, S_{\backslash ij}, R_{\backslash ji}, X, \gamma, \delta, \epsilon) \\
&\propto \int_{\eta_i}\int_{\eta_j} P(s_{ij}=k|\eta_i)P(r_{ji}=k'|\eta_j)P(x_i|\eta_i)P(x_j|\eta_j) \\
&\qquad P(\eta_i|S_{\backslash ij}, R_{\backslash ji}, X, \gamma)P(\eta_j|S_{\backslash ij}, R_{\backslash ji}, X, \gamma)d\eta_i d\eta_j \times \\
&\quad \int_{\psi_{kk'}} P(a_{ij}|\psi_{kk'})P(\psi_{kk'}|A_{\backslash ij}, S_{\backslash ij}, R_{\backslash ji}, \delta, \epsilon)d\psi_{kk'} \\
&= \frac{N_{ik\backslash ij}+M_{ik}+\gamma_k}{\sum_t\left(N_{it\backslash ij}+M_{it}+\gamma_t\right)} \times \frac{N_{jk'\backslash ji}+M_{jk'}+\gamma_{k'}}{\sum_t\left(N_{jt\backslash ji}+M_{jt}+\gamma_t\right)} \times \\
&\quad \frac{\left(n^{(+)}_{kk'\backslash ij}+\delta_{kk'}\right)^{\mathbb{I}(a_{ij}=1)}\left(n^{(-)}_{kk'\backslash ij}+\epsilon_{kk'}\right)^{\mathbb{I}(a_{ij}=0)}}{n^{(+)}_{kk'\backslash ij}+n^{(-)}_{kk'\backslash ij}+\delta_{kk'}+\epsilon_{kk'}}
\end{aligned} \tag{8}
$$

$$
\begin{aligned}
&P(x_{im}=k, z_{im}=l|w_{im}=v, W_{\backslash im}, S, R, X_{\backslash im}, Z_{\backslash im}, \alpha, \beta, \gamma) \\
&\propto \int_{\eta_i} P(s_i|\eta_i)P(r_i|\eta_i)P(x_{im}=k|\eta_i)P(\eta_i|S, R, X_{\backslash im}, \gamma)d\eta_i \times \\
&\quad \int_{\theta_k} P(z_{im}=l|\theta_k)P(\theta_k|X_{\backslash im}, Z_{\backslash,im}, \alpha)d\theta_k \times \\
&\quad \int_{\phi_l} P(w_{im}=v|\phi_l)P(\phi_l|W_{\backslash im}, Z_{\backslash im}, \beta)d\phi_l \\
&= \frac{M_{lv\backslash im}+\beta_v}{\sum_u\left(M_{lu\backslash im}+\beta_u\right)} \times \frac{M_{kl\backslash im}+\alpha_l}{\sum_q\left(M_{kq\backslash im}+\alpha_q\right)} \times \frac{N_{ik}+M_{ik\backslash im}+\gamma_k}{\sum_t\left(N_{it}+M_{it\backslash im}+\gamma_t\right)},
\end{aligned} \tag{9}
$$

where the symbol $\backslash$ represents the exclusion of an edge or a word from the count number.

The algorithm of CGS for MMSTB is provided in the Appendix 1. In CGS, according to

Equations (8) and (9), the latent community and topic for each edge and word are sampled G times. Finally, using the samples of the latent variables excluding the burn-in samples before b, model parameters are estimated as follows;

$$\hat{\eta}_{ik} = \frac{1}{G}\sum_{g=b+1}^{G} \frac{N_{ik}^{(g)} + M_{ik}^{(g)} + \gamma_k}{\sum_t \left(N_{it}^{(g)} + M_{it}^{(g)} + \gamma_t\right)} \tag{10}$$

$$\hat{\psi}_{kk'} = \frac{1}{G}\sum_{g=b+1}^{G} \frac{n_{kk'}^{(+,g)} + \delta_{kk'}}{n_{kk'}^{(+,g)} + n_{kk'}^{(-,g)} + \delta_{kk'} + \epsilon_{kk'}} \tag{11}$$

$$\hat{\theta}_{kl} = \frac{1}{G}\sum_{g=b+1}^{G} \frac{M_{kl}^{(g)} + \alpha_l}{\sum_q \left(M_{kl}^{(g)} + \alpha_q\right)} \tag{12}$$

$$\hat{\phi}_{lv} = \frac{1}{G}\sum_{g=b+1}^{G} \frac{M_{lv}^{(g)} + \beta_v}{\sum_u \left(M_{lu}^{(g)} + \beta_u\right)}. \tag{13}$$

4 Numerical experiments

This section described the numerical experiments we conducted to highlight the main features of the proposed approach and provide the validity of our inference algorithm.

4.1 Experimental settings

The main features of our modeling are the mixed membership of nodes and simultaneous modeling of network data and text content. The characteristic of a mixed membership captures the situation of people belonging to multiple communities on a social network and building relationships with other members of these communities. Furthermore, extracting more meaningful segments from social networks by considering both network data and text content is possible.

To highlight these two properties of MMSTB, we have designed three different scenarios. Table 2 provides the settings of each scenario, while Figure 2 depicts an example of the generated adjacency matrix, where black (white) cells mean the presence (absence) of a relationship between two nodes. We set some values for the community distribution, edge probability, and topic distribution but did not set any values for the word distribution. Instead, for all scenarios, 150 words are sampled per node according to their word topics from

the BBC news document dataset (Greene & Cunningham 2006) as virtual text contents; this dataset contains three topics: namely, business, entertainment, and sports.

Scenario A

The network and text content are composed of $K = 3$ communities and $L = 2$ topics. Each node belongs to only one community (Node 1-20, 41-60, 81-100) or two communities (Node 21-40, 61-80); that is, these communities are overlapping. But the edge probabilities across the communities are lower ($\psi_{kk'} = 0.1$) than within the communities ($\psi_{kk} = 0.5$), and each community has a unique topic proportion ($\theta_1 \neq \theta_2 \neq \theta_3$). Therefore, both MMSTB and other models using only one source of information such as LDA and MMSB can be expected to detect these communities accurately.

Scenario B

Similar to scenario A, each node belongs to one or two communities, and $K = 4$ communities are overlapping. Unlike scenario A, the community 1 and 4 have the same topic proportions ($\theta_1 = \theta_4$). Therefore, the models using only text content information cannot distinguish between the nodes that belong to only community 1 (Node 1-20) or community 4 (Node 91-100). Conversely, the edge probabilities across the communities are low; hence, both MMSTB and models using only network information should be able to distinguish all communities.

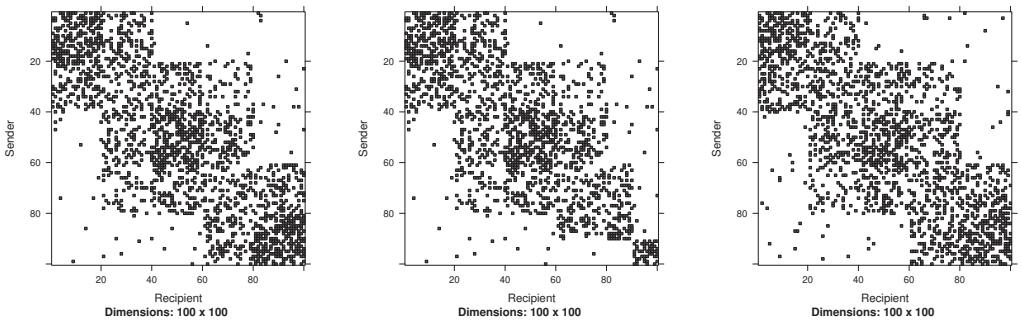
Scenario C

The community 1 and 4 have the same topic proportion, and the text content-based models cannot distinguish between these two communities. Furthermore, the edge probabilities between communities 3 and 4 (ψ_{34}, ψ_{43}) are high; that is, people in these communities are well-connected even if they have different interests (topics). Therefore, the network-based models cannot identify these two communities. Only MMSTB can detect all communities and recover the community structure properly.

We note that nodes are divided into some clusters where they belong to the same community (communities) with the same proportion and generate virtual texts of the same topic(s). Each row of H in Table 2 corresponds to each cluster, and, for example, in scenario A, nodes 1-20 are classified into the same cluster. Whether models can recover these cluster structures

Table 2: The settings of three simulation scenarios (see text for details)

Scenario	A	B	C
D (nodes)	100	100	100
K (communities)	3	4	4
L (topics)	2	2	3
Community dist. H	$\{\eta_1, \ldots, \eta_{20}\}: (1,0,0)$ $\{\eta_{21}, \ldots, \eta_{40}\}: (.5,.5,0)$ $\{\eta_{41}, \ldots, \eta_{60}\}: (0,1,0)$ $\{\eta_{61}, \ldots, \eta_{80}\}: (0,.5,.5)$ $\{\eta_{81}, \ldots, \eta_{100}\}: (0,0,1)$	$\{\eta_1, \ldots, \eta_{20}\}: (1,0,0,0)$ $\{\eta_{21}, \ldots, \eta_{40}\}: (.5,.5,0,0)$ $\{\eta_{41}, \ldots, \eta_{60}\}: (0,1,0,0)$ $\{\eta_{61}, \ldots, \eta_{80}\}: (0,.5,.5,0)$ $\{\eta_{81}, \ldots, \eta_{90}\}: (0,0,1,0)$ $\{\eta_{91}, \ldots, \eta_{100}\}: (0,0,0,1)$	$\{\eta_1, \ldots, \eta_{20}\}: (1,0,0,0)$ $\{\eta_{21}, \ldots, \eta_{40}\}: (.5,.5,0,0)$ $\{\eta_{41}, \ldots, \eta_{60}\}: (0,1,0,0)$ $\{\eta_{61}, \ldots, \eta_{80}\}: (0,.5,.5,0)$ $\{\eta_{81}, \ldots, \eta_{90}\}: (0,0,1,0)$ $\{\eta_{91}, \ldots, \eta_{100}\}: (0,0,0,1)$
Topic dist. Θ	$\theta_1 = (1,0)$ $\theta_2 = (.5,.5)$ $\theta_3 = (0,1)$	$\theta_1 = (1,0)$ $\theta_2 = (.5,.5)$ $\theta_3 = (0,1)$ $\theta_4 = (1,0)$	$\theta_1 = (.5,0,.5)$ $\theta_2 = (.5,.5,0)$ $\theta_3 = (0,1,0)$ $\theta_4 = (.5,0,.5)$
Edge prob. Ψ	$\psi_{11}, \psi_{22}, \psi_{33} = .5$ *otherwise* .1	$\psi_{11}, \psi_{22}, \psi_{33}, \psi_{44} = .5$ *otherwise* .1	$\psi_{11}, \psi_{22}, \psi_{33}, \psi_{34}, \psi_{43}, \psi_{44} = .5$ *otherwise* .1



Scenario A Scenario B Scenario C

Figure 2: Adjacency matrix for each scenario; the black cells represent the link presence, whereas the white cells represent the link absence

depends on the situation of each scenario as described above. In the next section, we validate whether our model and the models that are popular in the literature, namely, LDA as a text-based model and MMSB as a network-based model, are able to correctly estimate parameters and identify true cluster structures.

4.2 Reproducibility of parameters and recovery of cluster structures

This section presents the experiments we conducted to verify whether the considered models (LDA, MMSB, and MMSTB) can reproduce parameters and recover cluster structures as described in the previous section. The modeling assumptions for LDA and MMSB are

taken from the original papers (Blei, Ng, & Jordan, 2003 and Airoldi et al, 2008), while the generative process of these models is outlined in the Appendix 2. As LDA is a model for text content and MMSB is a model for network data, we provide only text data of the simulated dataset for LDA, only network data for MMSB, and the entire dataset for MMSTB. Similar to MMSTB, we use CGS to estimate parameters of LDA and MMSB. The number of iterations is set to 5,000, and the first 2,000 samples are excluded as burn-in samples. The values of the hyperparameters for the respective prior distributions are listed in Table 3.

First, we carry out an experiment to verify the reproducibility of parameters. Figure 3 and Table 4 show the results of scenario C estimating MMSTB. The three panels in Figure 3 show the estimated parameters, community distribution (left), edge probability (top-right), and topic distribution (right-bottom). The results show that MMSTB reproduces the values provided in Table 2 with high accuracy. Table 4 lists the top 10 words for each topic in descending order of the estimated word distribution values. From left to right, words related to business, entertainment, and sports are lined up, which implies that MMSTB extracts all topics correctly. Therefore, MMSTB behaves appropriately when detecting more meaningful communities by allowing nodes to have mixed memberships and considering network data and text content simultaneously. The results of the other scenarios and models are provided in the Appendix 2.

Next, we conducted an experiment to demonstrate the recovery of cluster structures from the simulated dataset. These cluster structures can be found using the estimated node-specific parameters. In particular, MMSB and MMSTB have a node-specific community distribution, whereas LDA has a node-specific topic distribution. For example, MMSTB's community distribution affects the generation of both the network and text data as explained in Section 3, while the nodes having similar values for the node-specific parameter (e.g., nodes 1-20 in scenario A have the same value for community distribution) should generate similar network and text data. Therefore, it is natural that these nodes are classified into the same cluster. In this experiment, we apply a clustering method, spectral clustering (Ng, Jordan, & Weiss, 2002), to the estimated node-specific parameter of each model and compare the clustering results with the true labels listed in Table 2.

The process of the experiment is as follows. First, we simulate datasets for each scenario

according to Table 2. Second, we estimate the model parameters while providing text data for LDA, network data for MMSB, and both datasets for MMSTB. The number of iterations and hyperparameter values are the same as described above. Next, we classify nodes according to the estimated node-specific parameters using spectral clustering (Ng, Jordan, & Weiss, 2002) [1]. Then, we calculate the adjusted Rand index (ARI, Hubert & Arabie, 1985) between the cluster and true labels, with higher ARIs representing higher similarity between these labels (when the labels perfectly match, ARI is 1). Because spectral clustering employs the k-means method, its result depends on the initial value of the number of clusters. Hence, we independently calculate ARIs for 20 different initial values of the the number of clusters and set the latter according to the obtained maximum ARI value. Finally, we repeat this process 50 times with the different seed value when generating datasets but the same parameter settings (Table 2).

Table 5 lists the medians of 50 ARIs calculated for the three models and three scenarios. According to the result of scenario A (first column), all medians of ARIs are 1.0; that is, all models can recover the true clusters. This result can be explained by the fact that the links within (between) communities are dense (sparse) while the topics of texts within a community are distinct from that of other communities. Even if only network data or text information are employed, differences between clusters can be identified.

According to the result of scenario B (second column), ARIs of MMSB and MMSTB are 1.0, whereas LDA's ARI is lower than before. In scenario B, community 1, to which nodes 1-20 (cluster 1) belong, and community 4, to which nodes 91-100 (cluster 6) belong, have the same value of the topic proportion; therefore, these clusters cannot be distinguished when looking at text data only. Conversely, non-diagonal elements of the edge probability are low; that is, the difference between these clusters is clear when considering network data. This is the reason why MMSB and MMSTB are able to recover the true clusters.

Finally, according to the result of scenario C, ARI is 1.0 only for MMSTB, whereas the ARI values of LDA and MMSB are far less than 1.0; that is, the latter models are unable to correctly cluster nodes. The reason for this result is that the text data in scenario C have the same topic structures as that of scenario B (topic distributions of communities 1 and 4

[1]The algorithm for spectral clustering we use is implemented as a function "specc" in R-package "'kernlab".

Table 3: The setting of hyperparameters for the simulation experiments

Hyperparameters	γ	δ	ϵ	α	β
Prior distributions	$\eta_i \sim Dir(\gamma)$	$\psi_{kk'} \sim Beta(\delta_{kk'}, \epsilon_{kk'})$		$\theta_k \sim Dir(\alpha)$	$\phi_l \sim Dir(\beta)$
Value	$\gamma_k = 1.0$, $\forall k$	$\delta_{kk'} = 0.1$, $\forall k, k'$	$\epsilon_{kk'} = 0.1$, $\forall k, k'$	$\alpha_l = 0.1$, $\forall l$	$\beta_v = 0.1$, $\forall v$

Table 4: Top 10 words in descending order of the word distribution for each topic

Topic 1 (Business)	Topic 2 (Entertainment)	Topic 3 (Sports)
bank	film	champion
growth	award	cup
oil	actor	coach
profit	album	rugbi
euro	nomin	ireland
stock	band	season
yuko	song	injuri
investor	oscar	olymp
award	chart	championship
deficit	actress	goal

are the same). Furthermore, the edge probabilities between communities 3 and 4 are equal to the probabilities within these communities; that is, both communities completely overlap in the network. Therefore, these communities cannot be identified when considering network data only. On the other hand, MMSTB takes both network and text data into account and hence is able to recover the true cluster structures. This numerical experiment reveals that our proposed model can correctly identify structures of communities and topics even if these structures overlap, which is one of the most notable features of our model.

Table 5: Medians of the adjusted Rand indices for the clustering result using the estimated node-specific parameters of the three models (LDA, MMSB, and proposed model) in the three scenarios

	Scenario A	Scenario B	Scenario C
LDA	1.0	0.85	0.85
MMSB	1.0	1.0	0.93
MMSTB	1.0	1.0	1.0

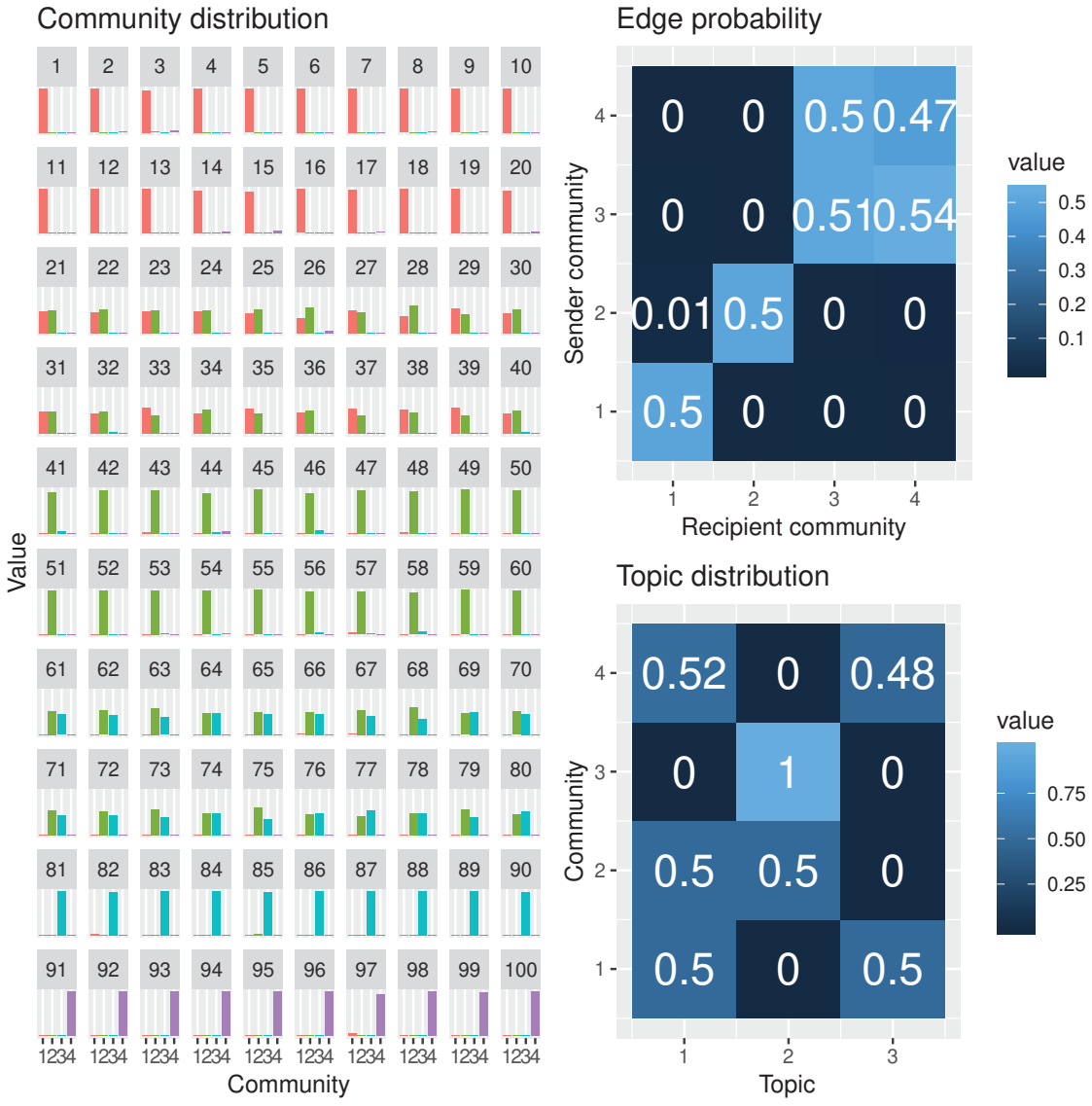


Figure 3: The estimation results of scenario C (see text for details)

4.3 Choosing the number of communities and the number of topics

The numbers of communities K and topics L need to be fixed before applying SBM or one of its extended models (DCSB, MMSB, MMSTB, etc.). A variety of approaches has been proposed in the literature for choosing these numbers, including information criteria such as BIC (Handcock, Raftery, & Tantrum, 2007; Saldana, Yu, & Feng, 2017), integrated completed likelihood (Daubin, Picard, & Robin, 2008; Bouveyron, Latouche, & Zreik, 2018), cross-validation (Chen & Lei, 2018), and Bayesian inference (Latouche, Birmele, & Ambroise, 2012; McDaid et al., 2013).

In this study, the numbers of communities and topics are determined using an information criteria based on its solid theoretical ground and convenience of calculating from the outputs of CGS. However, the topic models (including MMSTB), which have latent variables, are known as singular models, and information criteria for regular models such as AIC and BIC are not appropriate. Therefore, we employ the widely applicable information criterion (WAIC, Watanabe 2010) because it can be applied to both regular and singular models. WAIC estimates the expected pointwise predictive density for a new dataset. It is defined as $-2(lppd - p_{waic})$, where $lppd$ denotes the log pointwise predictive density representing the predictive accuracy of the fitted model to data, and p_{waic} denotes a term to correct for bias due to overfitting[2]. The definition of WAIC for MMSTB is provided in the Appendix 3.

In addition to the reproducibility of MMSTB parameters of described above, we confirm that the numbers of communities and topics can be correctly estimated by the model selection using WAIC. The procedure of the model selection simulation is as follows. For each scenario, we generate simulation data according to the values listed in Table 2. We estimate the models within the range of numbers of communities and topics from 2 to 6, and the model with the smallest WAIC is selected. The results of repeating these procedures 50 times are shown in Table 6. In all three scenarios, the model selection using WAIC succeeds in identifying the correct combination of the numbers of communities and topics. These experiments allow us to validate WAIC as a model selection criterion for MMSTB.

[2]In this study, we use the Gelman et al (2013)'s scale with $-2n$ times Watanabe's original definition (n is the number of data). This scale enables us to compare with other information criterion such as AIC and DIC

Table 6: The number of times WAIC selects each MMSTB model (K, L) in 50 simulations of each of the three scenarios

Scenario A ($K = 3$, $L = 2$)							Scenario B ($K = 4$, $L = 2$)							Scenario C ($K = 4$, $L = 3$)						
		Topics (L)							Topics (L)							Topics (L)				
		2	3	4	5	6			2	3	4	5	6			2	3	4	5	6
Communities (K)	2	0	0	0	0	0	Communities (K)	2	0	0	0	0	0	Communities (K)	2	0	0	0	0	0
	3	46	2	0	0	0		3	0	0	0	0	0		3	0	0	0	0	0
	4	2	0	0	0	0		4	38	1	0	0	0		4	0	46	1	1	0
	5	0	0	0	0	0		5	11	0	0	0	0		5	0	2	0	0	0
	6	0	0	0	0	0		6	0	0	0	0	0		6	0	0	0	0	0

5 Empirical analysis

5.1 Dataset

In this section, we apply our model to empirical data to demonstrate the usefulness of MMSTB for actual social networks. In particular, we employ the Twitter platform and user-generated text data collected by the authors. We focus on a Twitter ego network centered on the official account (@NintendoAmerica) operated by a subsidiary company of Nintendo Co., Ltd. in U.S., Nintendo of America Inc. We created a dataset for analysis according to the following procedure.

First, users were randomly sampled from the users who follow the official account of Nintendo of America based on the following-followed relationship on May 1, 2018. Next, additional users were randomly sampled from the users who follow the users following the Nintendo account. The users whose average of the numbers of followers and followees is less than 3 in this network were excluded as outliers (note that the numbers of followers and followees are the numbers in the dataset and not the actual numbers). As a result, the number of selected users is 3,500, the number of total link edges are 68,949 (i.e., each user has 19.7 edges on average), and their directed relationships are used as network information.

Next, we collected the tweets posted by the selected users on their timelines from September 1, 2017 to February 28, 2018 [3]. These tweet data were preprocessed as follows: decomposing into word sets for each user, changing to lowercase letters, excluding numbers, symbols,

[3]We confirmed that the majority of users posted about the presentation of a new game software, called Nintendo Direct, in March 2018. Hence, in this study, to avoid the effect of such text information commonly posted by many users, we decided to limit the period of data to be until February 28, 2018.

Table 7: WAIC of each model of MMSTB estimated for the Twitter dataset (bold represents the smallest value)

		Topics (L)					
		5	6	7	8	9	10
Communities (K)	5	4,601,682	4,591,215	4,547,102	4,651,380	4,651,888	4,521,875
	6	4,633,828	4,580,391	4,564,193	4,568,752	4,629,114	5,563,824
	7	4,607,615	4,588,504	4,627,986	4,564,135	4,596,299	4,553,339
	8	4,613,074	4,637,185	4,623,877	4,517,891	4,564,046	4,537,160
	9	4,612,382	4,626,961	4,557,745	4,540,766	**4,500,094**	4,571,307
	10	4,598,036	4,580,622	4,580,856	4,544,071	4,534,666	6,629,801

and some popular stop-words (a, the, I, etc.) and reducing inflected words to their word stem. Among the preprocessed words, we excluded those with low frequencies (words having the number of occurrences in the corpus less than 20 or used by less than 20 users) or high frequencies (words used by more than 50 users) because these words may adversely affect the topic extraction. Then, the users whose number of words is less than five are also excluded. As a result, the number of unique words in the corpus is 9,001, and the average number of words per node is 98.2 (the average unique word number is 59.3). Next, we applied MMSTB to this Twitter dataset. The model selected by WAIC was $(K, L) = (9, 9)$ as shown in Table 7.

5.2 Empirical results

Because the considered dataset is a large-scale network, unlike the simulation network described in the previous section, understanding the results of the model estimation is difficult even if we showed the community distributions of all nodes and the entire network image. Therefore, we focus only on a certain node when discussing the results of this experiment. Figure 5 illustrates a sub-network consisting of a specific node (node 95), its neighbors, and the partial results of the estimated parameters related to them. Furthermore, Figure 4 shows the top 10 words for each topic.

First, interpreting the meaning of each topic from the top 10 words is necessary to understand what kind of interest people in the community display. The meaning of topics and their related words are as follows:

Topic 1

animation (e.g., blackclover, hunter×hunter, and jojos_bizarre_adventure are the titles of animations);

Topic 2

game (e.g., steinsgate, xenovers, and acnl, Animal Crossing: New Leaf, are the titles of game software);

Topic 3

e-sports (e.g., hori and mkleosaga are words related to fighting-games, while wnf and mdva are e-sports specific words);

Topic 4

music (e.g., vevo, spinrilla, and wshh are websites for music);

Topic 5

everyday life (e.g., people post texts and images of their everyday life with the hashtags of dogsoftwitter and momlife);

Topic 6 and 7

business (e.g., digitalmarket, socialmediamarket, and contentmarket are the hashtags which are sometimes used in a business-related tweet);

Topic 8

streaming and broadcasting (e.g., teamemmmmsi, twitchkitten, roku, and wizebot are words related to streaming or broadcasting);

Topic 9

sports (e.g., orton and sdlive, oiler, horford, and herewego are wrestling, ice hockey, basketball, and american football specific words, respectively).

Based on the interpretation of the topics provided above, obtaining some interesting insights from the sub-network of node 95 is possible. In Figure 5, the pie-charts show the values of the node's community distribution, η_i; bar plots are the values of the community's topic distribution, θ_k; and arrows represent that there is a following relationship between the nodes,

where the start node of the arrow is a sender of the following, the end node of the arrow is a recipient of the following, while the bidirected arrow means the mutual-following relationship.

The interpretation of a huge network, such as these Twitter data, is hardly achievable even if we looked at the entire network image. However, the local sub-network and the estimated parameters corresponding to them provide useful insights on not only the relationship between nodes but also overlapping communities, their proportions of belonging communities, and characteristic topics within each community. For instance, nodes 95 and 336 belong to community 5, in which people often post sports-related tweets (Topic 9). Node 95 belongs to not only community 5 but also community 1 related to music (Topic 4) together with other nodes (804, 2241, 3476). Thus, the communities detected by our model represent a subset of nodes with not only dense links on the network but also similar topics in their texts. In the field of consumer behavior analysis, researchers know that product or information diffusion tends to become faster among people located in a well-connected area of their social network (i.e., they belong to the same community) as discussed in Muller & Peres (2018). In addition to the network effect, people in the community identified by our model share their interests owing to the same topic of texts posted on Twitter. Therefore, our model can help companies to detect some useful community structures that positively affect the consumption behaviors. By analyzing the relationship between a company's followers and its created text content using our model, marketing managers can understand community structures and the interests of the customers connected through these communities. The managers can then use the obtained knowledge to update their marketing strategies accordingly.

5.3 Predicting on holdout samples

In this section, we demonstrate the predictive performance of our model on some test data generated by holding out a part of the dataset described in Section 5.1. Unlike the experiment outlined in the previous section, where the entire dataset was used for the model estimation, 90% of edges of each node with $D-1$ edges are selected randomly in this experiment and used as training data, while the remaining 10% of edges are used as test data. For the text data, all words of each node are selected as training data. The settings of the hyperparameters are the same as listed in Table 3, and the numbers of communities and topics are fixed at

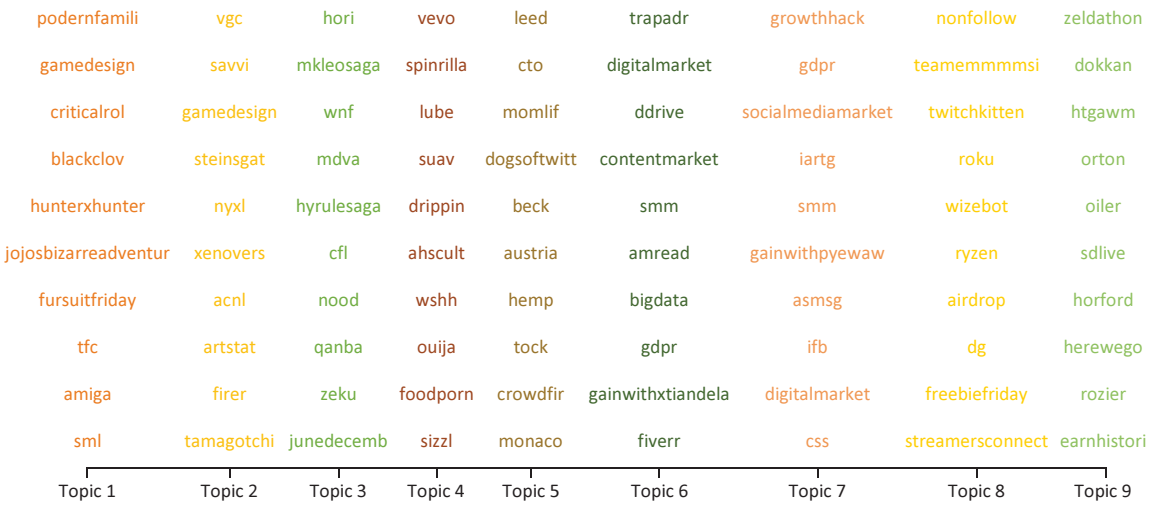


Figure 4: Top 10 words in descending order of the word distribution for each topic of Twitter data

the same values as memtioned in the previous section, $(K, L) = (9, 9)$. Using these settings, we estimate the model parameters using the training dataset. Let the estimated community distributions and edge probabilities be $\hat{H}$ and $\hat{\Psi}$. Then, the predicted probabilities for the test data $a_{ij} \in A^{test}$ can be calculated as follows:

$$P(a_{ij} = 1) = \sum_{k=1}^{K} \sum_{k'=1}^{K} \hat{\eta}_{ik} \hat{\eta}_{jk'} \hat{\psi}_{kk'}. \tag{14}$$

First, to see the predictive performance of our model, we use the receiver operating characteristic (ROC) curve shown in Table 6. The ROC curve connects plots of the true positive rate as the y-axis against the false positive rate as the x-axis when the cutoff value varies. If the ROC curve overlaps the line at 45 degrees, the model randomly predicts the test data. When the ROC curve is above this line, the model splits two groups (link or non-link) more clearly. Therefore, the area under the curve (AUC) shows the predictive performance. The AUC of Figure 6 is 0.93, which means that MMSTB has a good predictive performance.

Next, we search the best cutoff value based on the predictive performance of our model. By determining some cutoff value, we can predict labels of test edges (link or non-link). Our network data are imbalanced having many non-link edges (about 99.4% of edges are non-link).

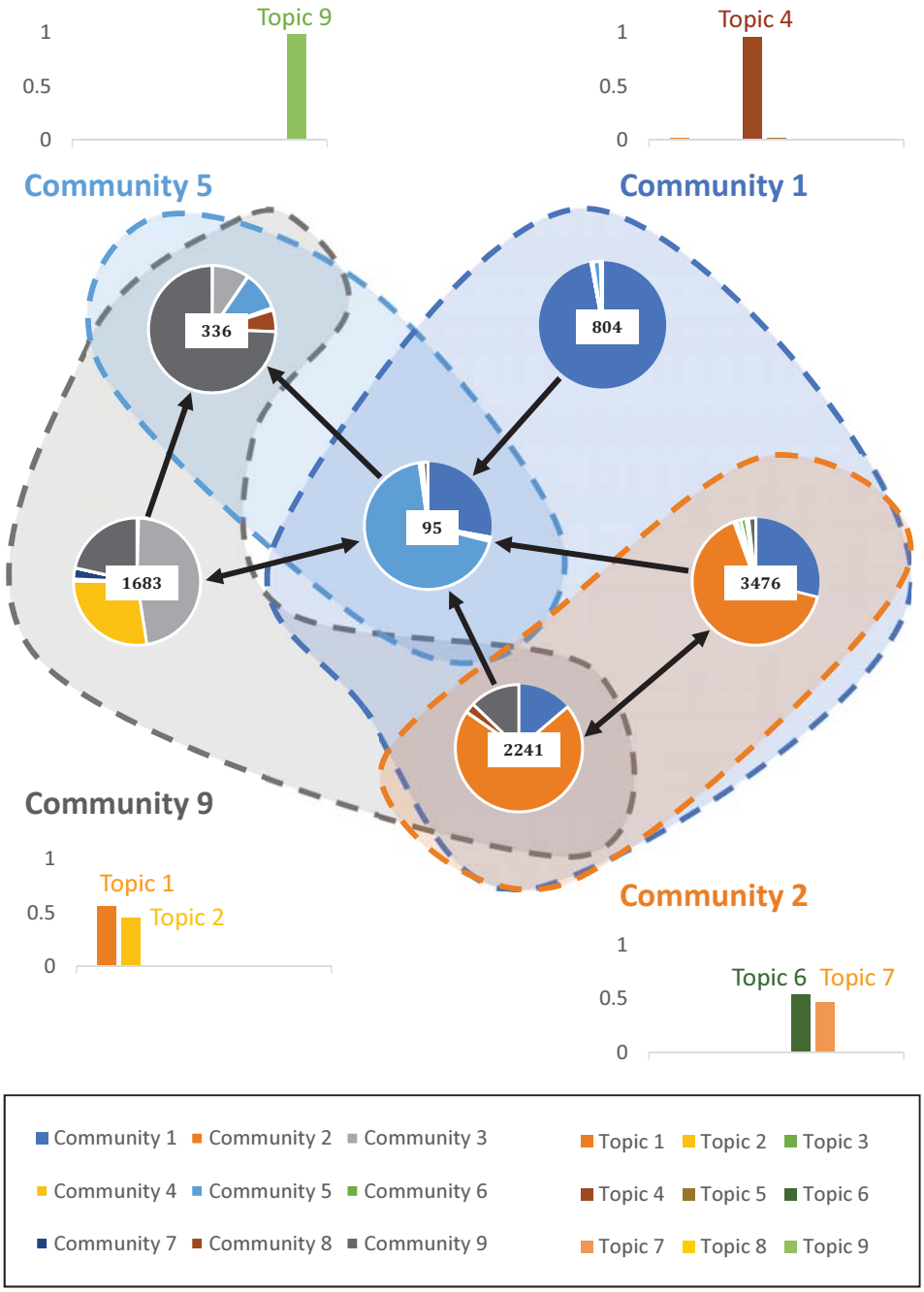


Figure 5: Sub-network consisting of a specific node (node 95) and its neighbors, and the results estimated by MMSTB. Circles represent nodes (numbers in circles are node indicators); arrows represent that there is a following relationship between the nodes; pie-charts represent the estimated community distributions and surrounding bar-graphs represent the estimateed topic distributions.

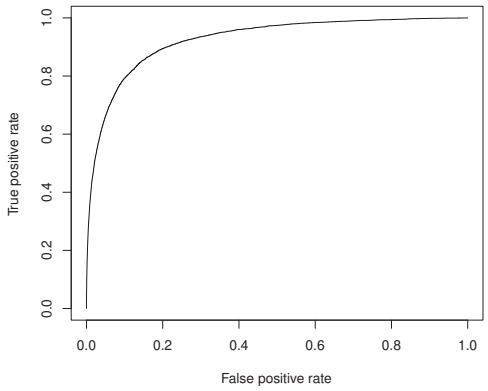


Figure 6: Receiver operating characteristic curve for test edge samples

Table 8: Confusion matrix when cutoff is set to 0.088, where the Matthews correlation coefficient is the highest

		Prediction	
		Link	Non-link
Data	Link	2,041	4,786
	Non-link	7,079	1,211,094

Therefore, we employ the Matthews correlation coefficient (MCC) to imply the correlation coefficient between the true class labels and predicted labels. Using the obtained predicted probabilities as described above, the cutoff value with the highest MCC of 0.254 is determined to be 0.088. Table 8 represents the confusion matrix for the test data computed using this cutoff value. According to this confusion matrix, our model predicts not only non-link edges with high accuracy but also link edges with more than 40% accuracy despite the difficulty of predicting correctly due to the imbalanced nature of the data (the class size of the link data is much smaller than that of the non-link data). This result demonstrates that our model performs well as a predictive model.

6 Conclusion

This study proposed a model, called MMSTB, for identifying realistic and beneficial communities based on not only the relationships within a social network but also interests of its members reflected by UGC. The main features of MMSTB are (1) extracting communities and topics considering network information, which represents the relationships between network nodes, and text information posted on social media, which uncovers people interests, (2) allowing each node to belong to multiple communities, which is called mixed membership, and (3) being applicable to both directed and undirected graphs.

The CGS was used for the model inference, and the numbers of communities and topics were chosen according to WAIC. Numerical experiments using simulated data confirmed the above-mentioned model features and showed that the procedures of the model inference and model selection work properly. As an empirical application, we analyzed Twitter data and found realistic and beneficial community structures that could not be obtained unless both network and text information were considered. Furthermore, the proposed model demonstrated a good predictive performance on holdout samples.

Further work may include the problem of node heterogeneity. SBM and other extended models assume that all nodes belonging to the same community are homogeneous and edges within or between communities are generated according to the corresponding community's edge probability. However, in general social networks, a few hub nodes tend to have many edges, while many other nodes tend to have a few edges, even if they belong to the same community. This property is called scale-free. Ignoring this node heterogeneity may lead to a deviation of model results from the real network structure. Krivitsky et al (2009) introduced a parameter representing node heterogeneity for the edge generating part of the LSM. Karrer & Newman (2011) proposed an extended SBM that corrects the probability of generating an edge between a pair of nodes considering the node degrees to address the problem of node heterogeneity. This problem cannot be avoided in social network analysis; hence, our model also needs to be extended to solve the problem.

Appendix

Appendix 1: Algorithm of the collapsed Gibbs sampler (CGS) for the proposed model (MMSTB)

In Section 3.2, we derived the posterior distributions of latent variables. CGS repeats sampling according to Equations (8) and (9) and updates the allocation counters of latent communities and topics. The pseudo algorithm of CGS for MMSTB is provided in algorithm 1.

Algorithm 1 collapsed Gibbs sampler for MMSTB

Assign randomly communities and topics to S, R, X, Z
for $g = 1, \ldots, G$ **do**
  for $i = 1, \ldots, D$ **do**
    for $j = 1, \ldots, D$ **do**
      Set $N_{ik\backslash ij}, N_{jk'\backslash ji}, n^{(+)}_{kk'\backslash ij}, n^{(-)}_{kk'\backslash ij}$
      Sample edge communities, $s^{(g)}_{ij}, r^{(g)}_{ji}$, from (8)
      Update $N_{ik}, N_{jk'}, n^{(+)}_{kk'}, n^{(-)}_{kk'}$
    end for
    for $m = 1, \ldots, M_i$ **do**
      Set $M_{ik\backslash im}, M_{kl\backslash im}, M_{lv\backslash im}$
      Sample word community and word topic, $x^{(g)}_{im}, z^{(g)}_{im}$, from (9)
      Update M_{ik}, M_{kl}, M_{lv}
    end for
  end for
end for

Appendix 2: Supplementary results of numerical experiments

In this Appendix, we show the remaining results of the numerical experiments that could not be explained in the Section 4.2 due to the limited space. Before listing the results, we briefly introduce two statistical models widely used in the literature: namely, latent Dirichlet allocation (LDA) and mixed membership stochastic block model (MMSB).

The generative model for LDA is as follows:

$$\begin{aligned}
&\text{word topic} \quad z_{im}|\theta_i \sim Categorical(\theta_i), \quad i = 1, \ldots, D, \quad m = 1, \ldots, M_i \\
&\text{word} \quad w_{im}|z_{im}, \Phi \sim Categorical(\phi_{z_{im}}), \quad i = 1, \ldots, D, \quad m = 1, \ldots, M_i \\
&\text{topic distribution} \quad \theta_i|\alpha \sim Dirichlet(\alpha), \quad i = 1, \ldots, D \\
&\text{word distribution} \quad \phi_l|\beta \sim Dirichlet(\beta), \quad l = 1, \ldots, L.
\end{aligned}$$

The generative model for MMSB is as follows:

$$\begin{aligned}
\text{sender community} \quad & s_{ij}|\eta_i \sim Categorical(\eta_i), \quad i, j = 1, \ldots, D \\
\text{recipient community} \quad & r_{ji}|\eta_j \sim Categorical(\eta_j), \quad i, j = 1, \ldots, D \\
\text{edge} \quad & a_{ij}|s_{ij}, r_{ji}, \Psi \sim Bernoulli(\psi_{s_{ij}r_{ji}}), \quad i, j = 1, \ldots, D \\
\text{community distribution} \quad & \eta_i|\gamma \sim Dirichlet(\gamma), \quad i = 1, \ldots, D \\
\text{edge probability} \quad & \psi_{kk'}|\delta, \epsilon \sim Beta(\delta_{kk'}, \epsilon kk'), \quad k, k' = 1, \ldots, K.
\end{aligned}$$

The parameters estimated by LDA, MMSB, and MMSTB for the dataset of the considered three scenarios are shown in Figures 8-14 (the result of MMSTB for scenario C is described in Section 4.2).

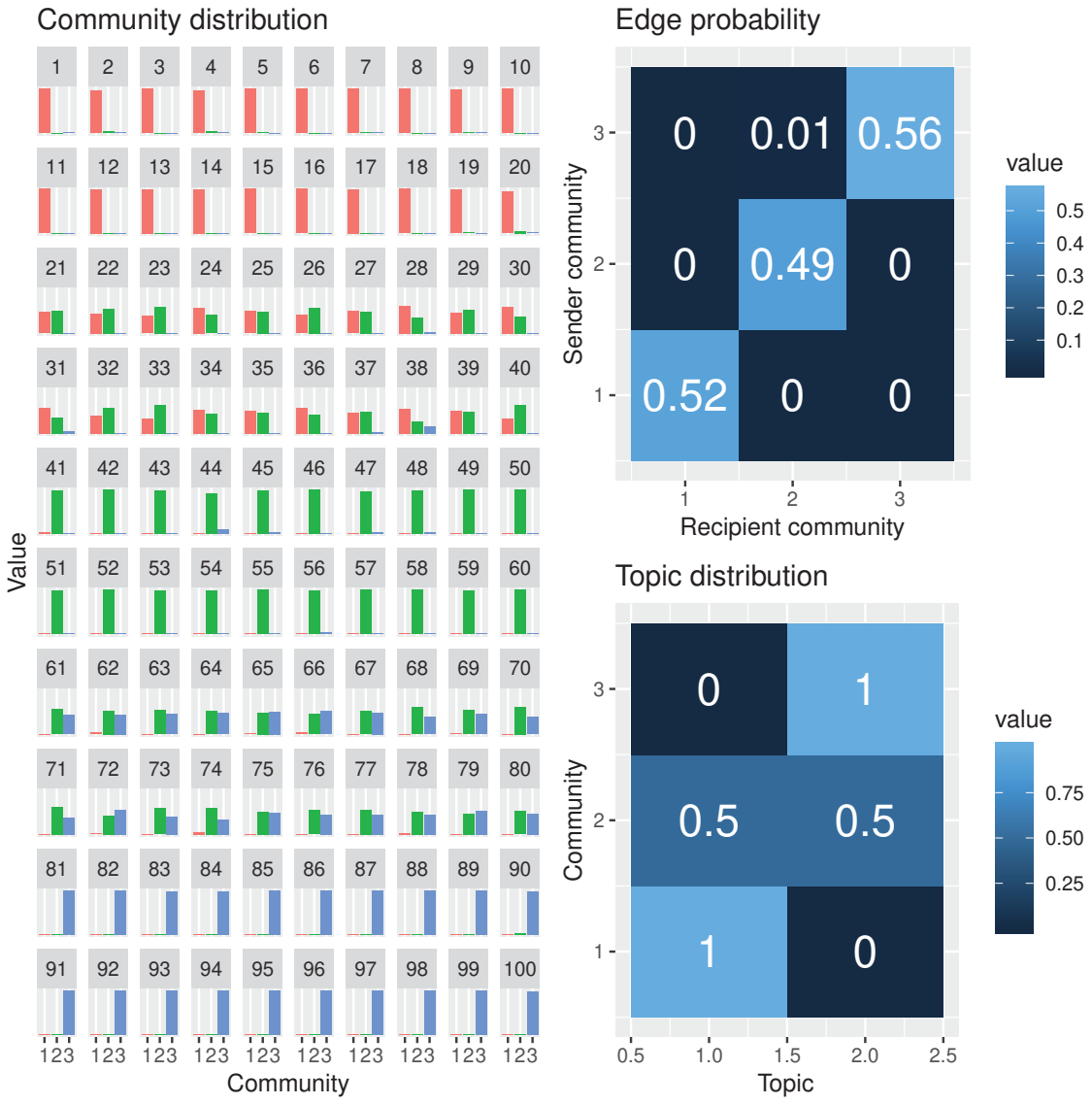


Figure 7: Estimation results of scenario A by MMSTB

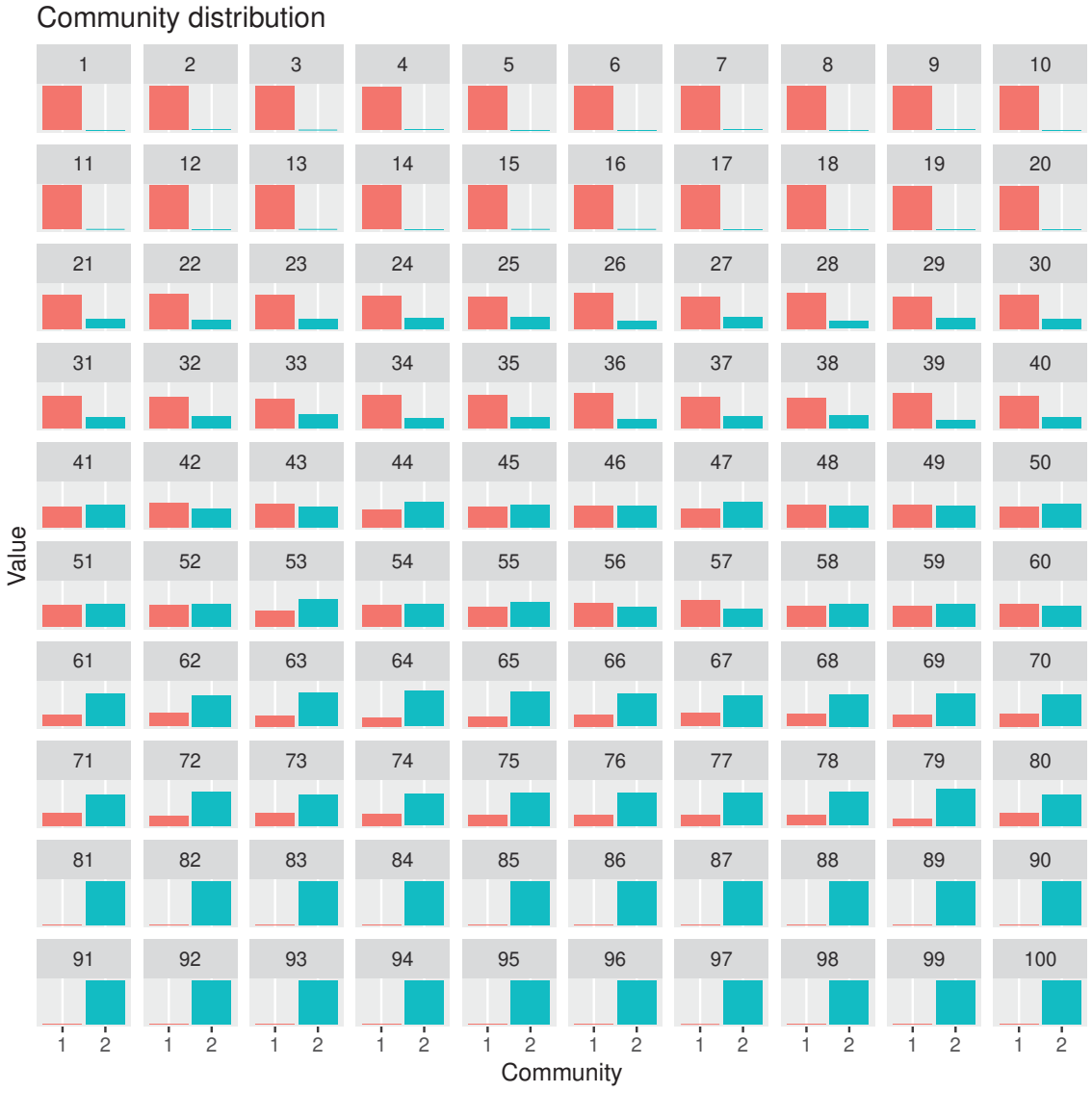


Figure 8: Estimation results of scenario A by LDA

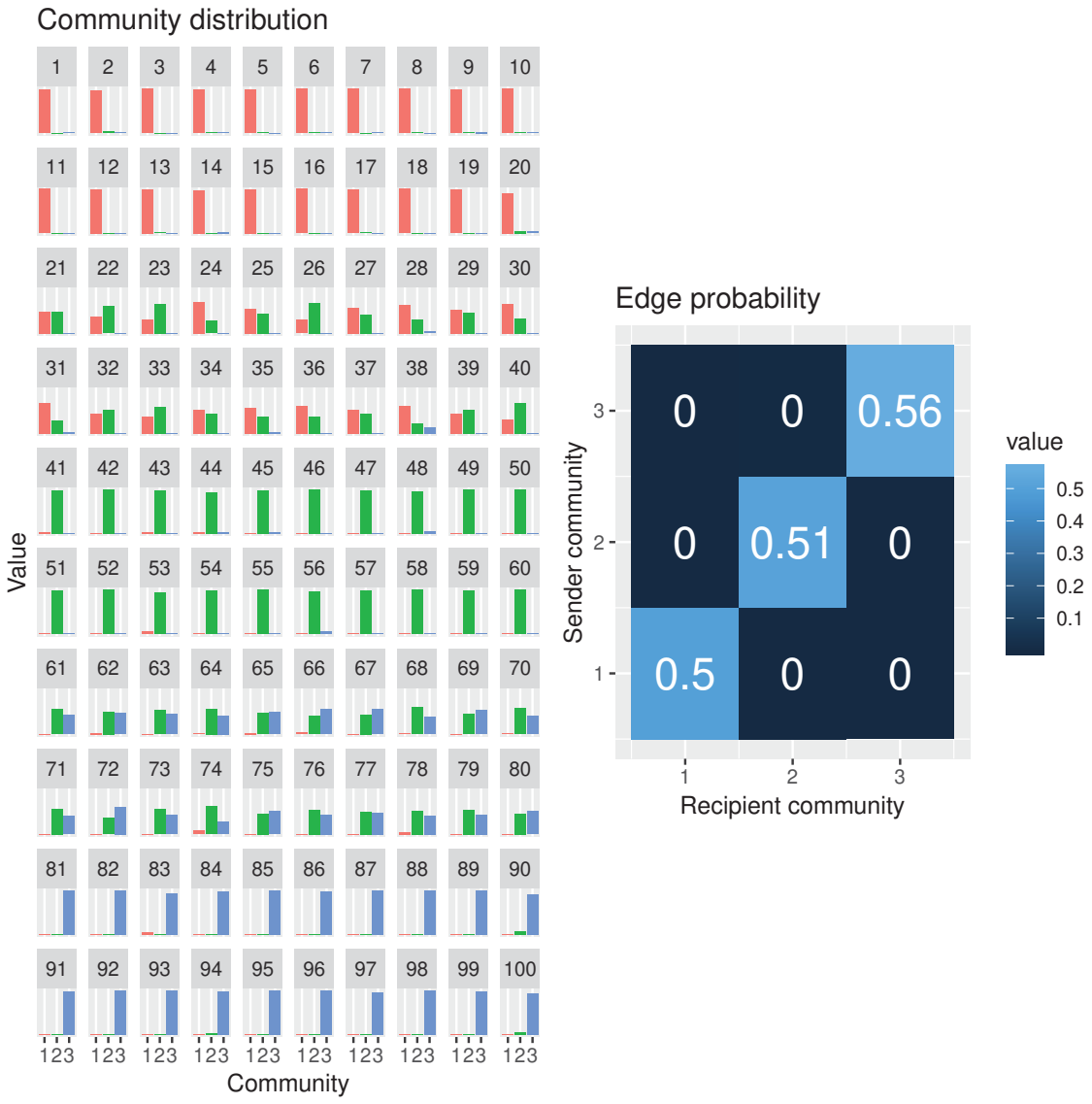


Figure 9: Estimation results of scenario A by MMSB

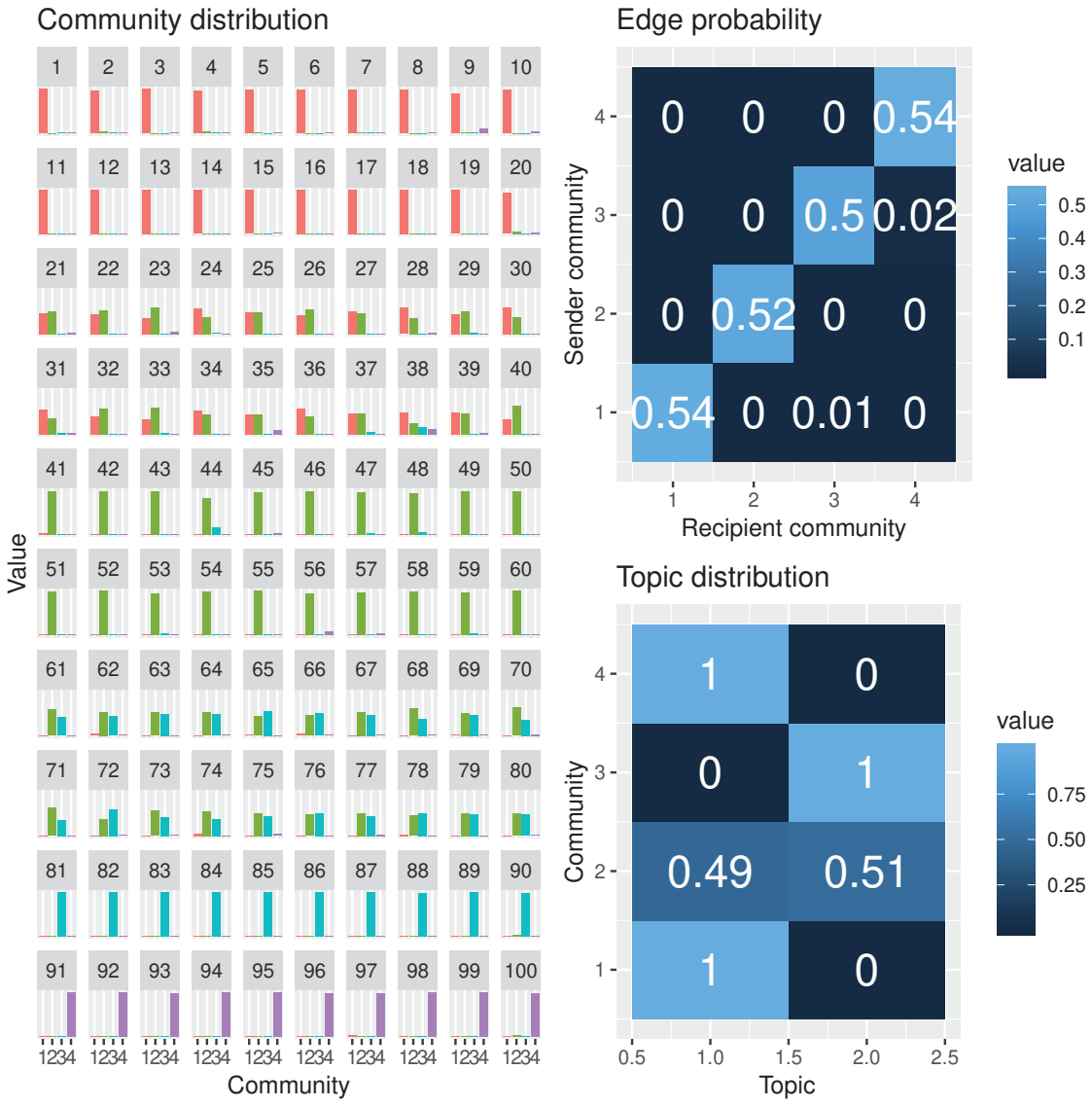


Figure 10: Estimation results of scenario B by MMSTB

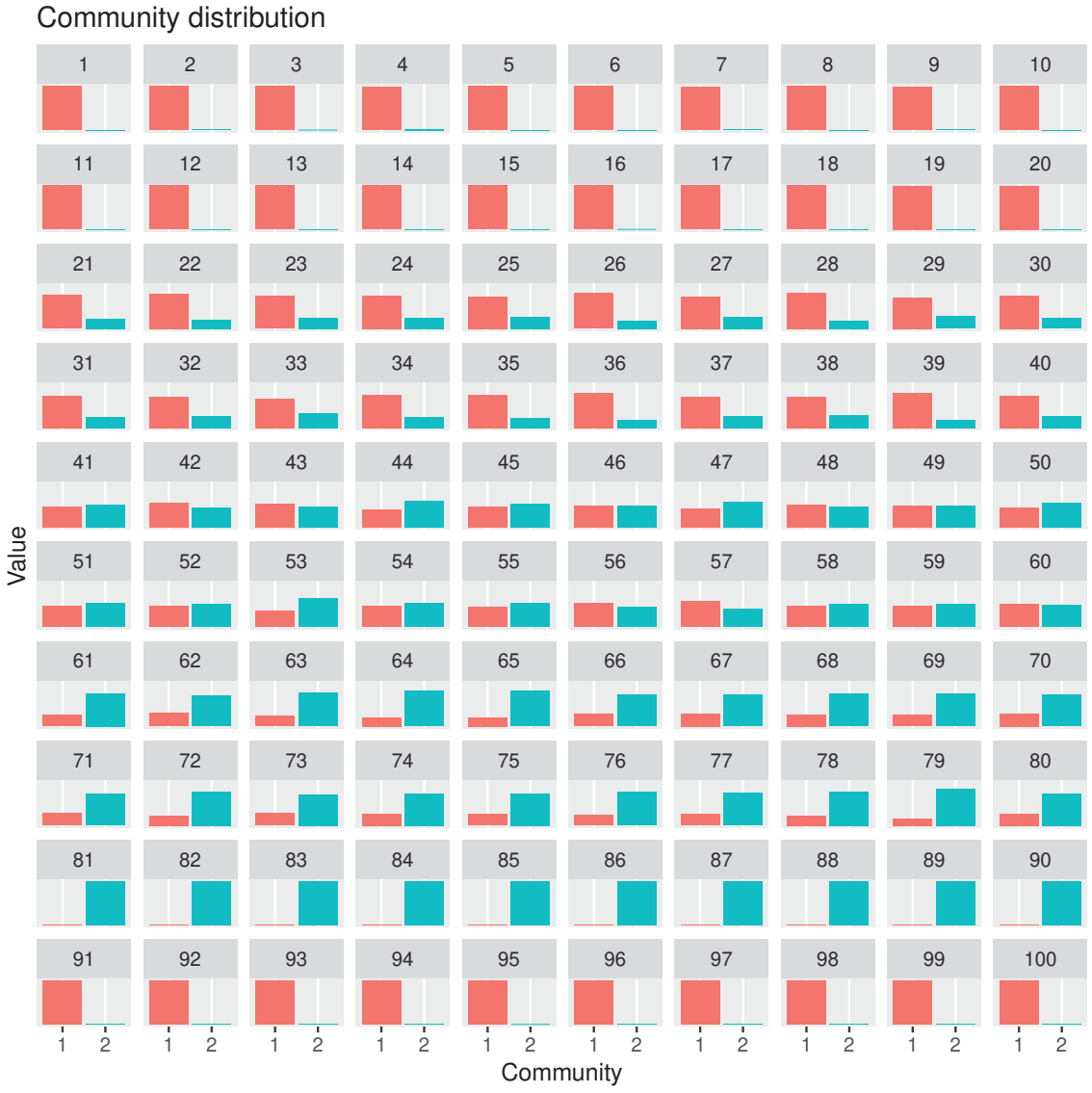


Figure 11: Estimation results of scenario B by LDA

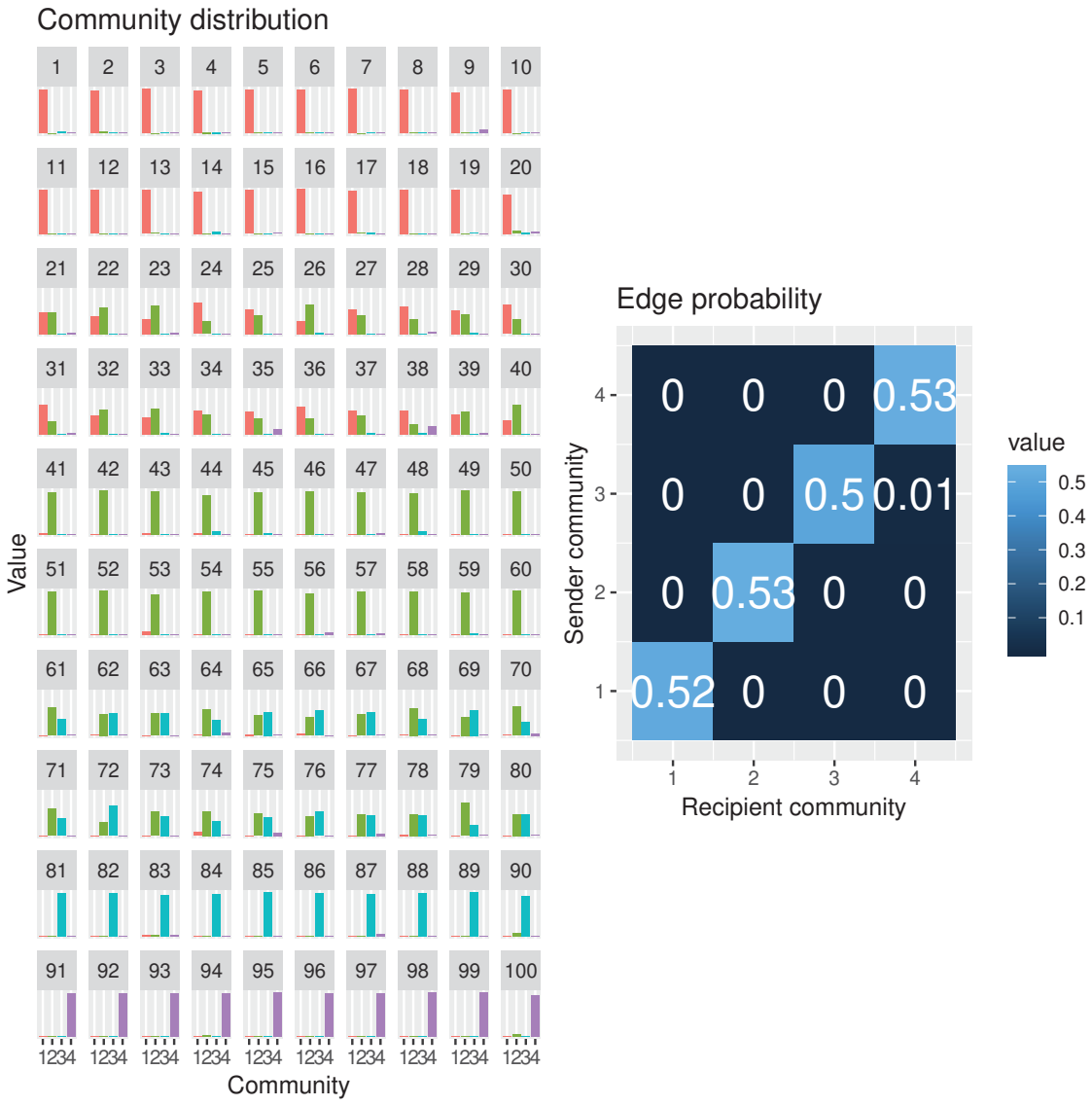


Figure 12: The estimation results of scenario B by MMSB

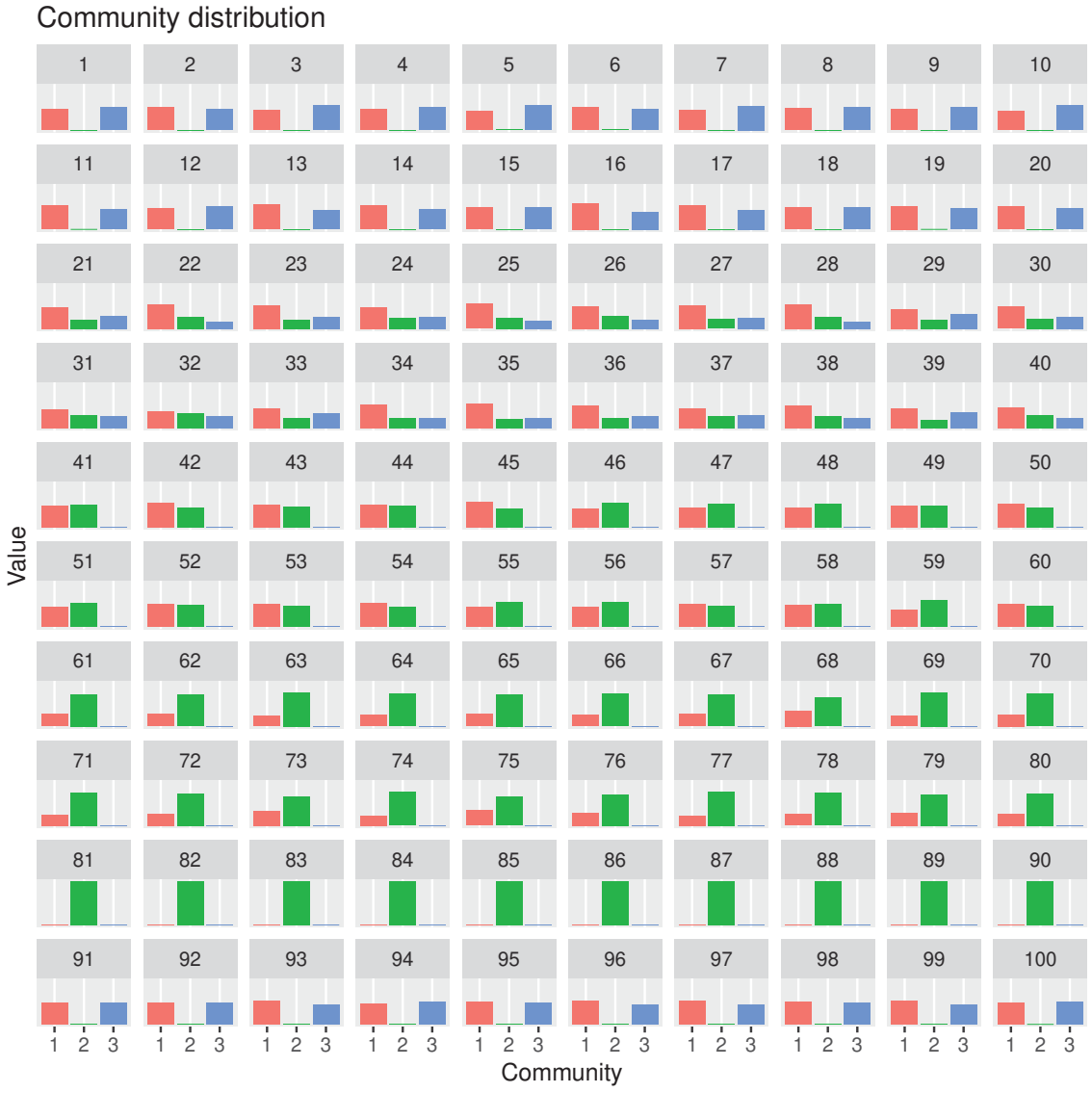


Figure 13: Estimation results of scenario C by LDA

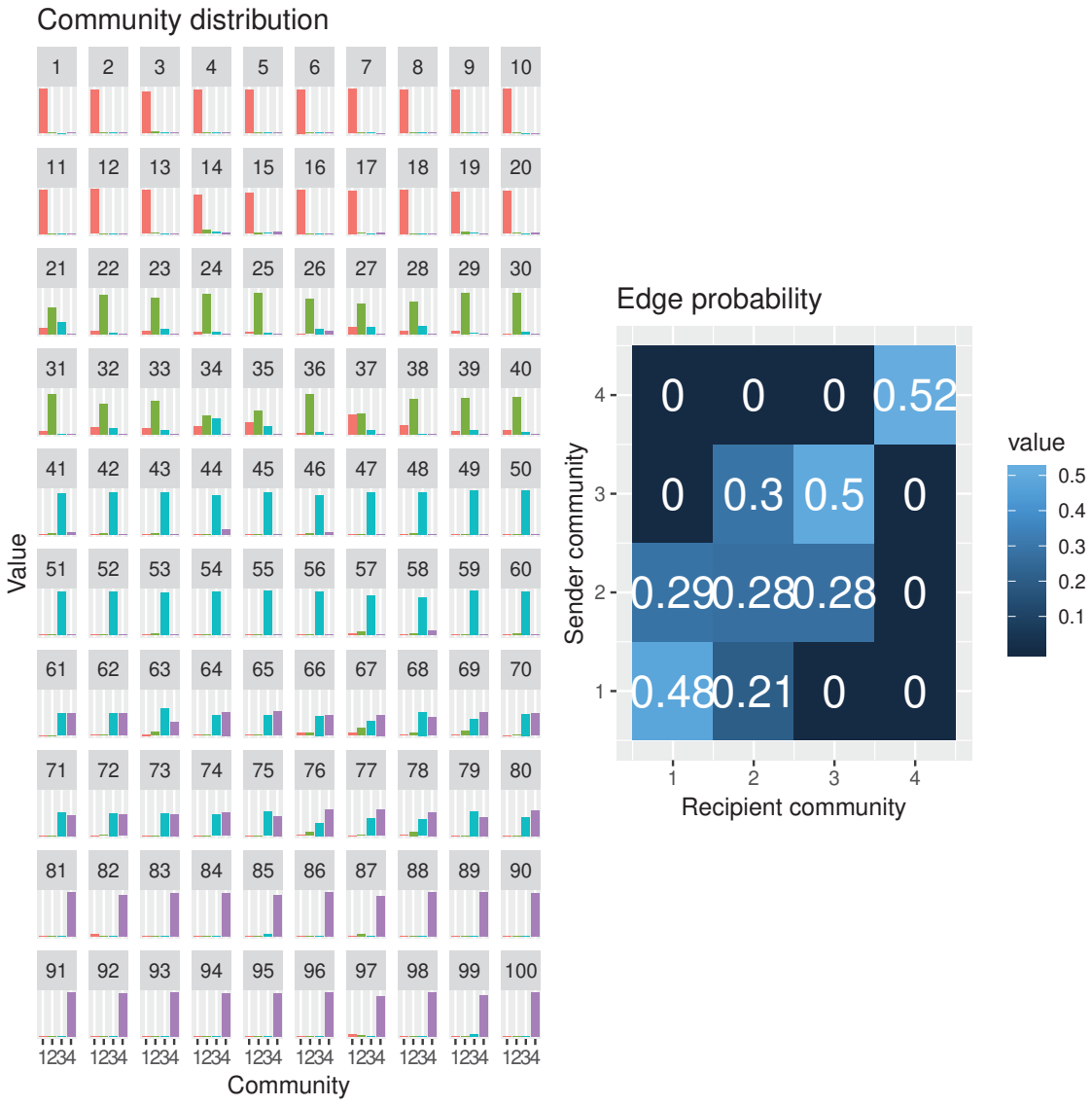


Figure 14: Estimation results of scenario C by MMSB

Appendix 3: Definition of WAIC for MMSTB

The definition of WAIC for MMSTB is as follows:

$$lpd = \sum_{i=1}^{D} \left(\log \left(\frac{1}{G} \sum_{g=b+1}^{G} \prod_{j=1}^{D} P\left(a_{ij} | H^{(g)}, \Psi^{(g)}\right) \prod_{m=1}^{M_i} P\left(w_{im} | H^{(g)}, \Theta^{(g)}, \Phi^{(g)}\right) \right) \right) \tag{15}$$

$$\begin{aligned} P_{waic} = \sum_{i=1}^{D} \Bigg(\frac{G}{G-1} \Bigg(\frac{1}{G} \sum_{g=b+1}^{G} \Bigg(\sum_{j=1}^{D} \log P\left(a_{ij} | H^{(g)}, \Psi^{(g)}\right)^2 + \sum_{m=1}^{M_i} \log P\left(w_{im} | H^{(g)}, \Theta^{(g)}, \Phi^{(g)}\right)^2 \Bigg) \\ - \left(\frac{1}{G} \sum_{g=b+1}^{G} \left(\sum_{j=1}^{D} \log P\left(a_{ij} | H^{(g)}, \Psi^{(g)}\right) + \sum_{m=1}^{M_i} \log P\left(w_{im} | H^{(g)}, \Theta^{(g)}, \Phi^{(g)}\right) \right) \right)^2 \Bigg) \Bigg) \end{aligned} \tag{16}$$

$$WAIC = -2\left(lpd - P_{waic}\right), \tag{17}$$

where $P\left(a_{ij} | H^{(g)}, \Psi^{(g)}\right)$ and $P\left(w_{im} | H^{(g)}, \Theta^{(g)}, \Phi^{(g)}\right)$ are the model likelihood conditioned with the parameters estimated using samples at sth iteration

$$P\left(a_{ij} | H^{(g)}, \Psi^{(g)}\right) = \sum_{k=1}^{K} \sum_{k'=1}^{K} \eta_{ik} \cdot \eta_{jk'}^{(g)} \cdot \psi_{kk'}^{(g)\mathbb{I}(a_{ij}=1)} \cdot \left(1 - \psi_{kk'}\right)^{(g)\mathbb{I}(a_{ij}=0)} \tag{18}$$

$$P\left(w_{im} | H^{(g)}, \Theta^{(g)}, \Phi^{(g)}\right) = \sum_{k=1}^{K} \sum_{l=1}^{L} \eta_{ik}^{(g)} \cdot \theta_{kl}^{(g)} \cdot \phi_{lw_{im}}^{(g)}. \tag{19}$$

References

[1] Abbe, E. (2017), "Community detection and stochastic block models: recent developments," *The Journal of Machine Learning Research*, 18(1), 6446-6531.

[2] Airoldi, E. M., Blei, D. M., Fienberg, S. E., & Xing, E. P. (2008), "Mixed membership stochastic blockmodels," *Journal of Machine Learning Research*, 9(Sep), 1981-2014.

[3] Ansari, A., Stahl, F., Heitmann, M., & Bremer, L. (2018), "Building a social network for success," *Journal of Marketing Research*, 55(3), 321-338.

[4] Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003), "Latent dirichlet allocation," *Journal of Machine Learning Research*, 3(Jan), 993-1022.

[5] Bouveyron, C., Latouche, P., & Zreik, R. (2018), "The stochastic topic block model for the clustering of vertices in networks with textual edges," *Statistics and Computing*, 28(1), 11-31.

[6] Chen, K., & Lei, J. (2018), "Network cross-validation for determining the number of communities in network data," *Journal of the American Statistical Association*, 113(521), 241-251.

[7] Daudin, J. J., Picard, F., & Robin, S. (2008), "A mixture model for random graphs," *Statistics and Computing*, 18(2), 173-183.

[8] Gelman, A., Stern, H. S., Carlin, J. B., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2013), "*Bayesian data Analysis*," Chapman and Hall/CRC.

[9] Gormley, I. C., & Murphy, T. B. (2010), "A mixture of experts latent position cluster model for social network data," *Statistical Methodology*, 7(3), 385-405.

[10] Greene, D., & Cunningham, P. (2006, June), "Practical solutions to the problem of diagonal dominance in kernel document clustering," *In Proceedings of the 23rd International Conference on Machine learning*, 377-384.

[11] Griffiths, T. L., & Steyvers, M. (2004), "Finding scientific topics," *Proceedings of the National Academy of Sciences*, 101(suppl 1), 5228-5235.

[12] Handcock, M. S., Raftery, A. E., & Tantrum, J. M. (2007), "Model-based clustering for social networks," *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 170(2), 301-354.

[13] Hoff, P. D., Raftery, A. E., & Handcock, M. S. (2002), "Latent space approaches to social network analysis," *Journal of the American Statistical Association*, 97(460), 1090-1098.

[14] Hubert, L., & Arabie, P. (1985), "Comparing partitions," *Journal of Classification*, 2(1), 193-218.

[15] Jeong, H., Mason, S. P., Barabási, A. L., & Oltvai, Z. N. (2001), "Lethality and centrality in protein networks," *Nature*, 411(6833), 41.

[16] Karrer, B., & Newman, M. E. (2011), "Stochastic blockmodels and community structure in networks," *Physical Review E*, 83(1), 016107.

[17] Krebs, V. E. (2002), "Mapping networks of terrorist cells," *Connections*, 24(3), 43-52.

[18] Krivitsky, P. N., Handcock, M. S., Raftery, A. E., & Hoff, P. D. (2009), "Representing degree distributions, clustering, and homophily in social networks with latent cluster random effects models," *Social Networks*, 31(3), 204-213.

[19] Latouche, P., Birmelé, E., & Ambroise, C. (2011), "Overlapping stochastic block models with application to the french political blogosphere," *The Annals of Applied Statistics*, 5(1), 309-336.

[20] Latouche, P., Birmelé, E., & Ambroise, C. (2012), "Variational Bayesian inference and complexity control for stochastic block models," *Statistical Modelling*, 12(1), 93-115.

[21] Liu, X., Bollen, J., Nelson, M. L., & Van de Sompel, H. (2005), "Co-authorship networks in the digital library research community," *Information Processing & Management*, 41(6), 1462-1480.

[22] Liu, Y., Niculescu-Mizil, A., & Gryc, W. (2009, June), "Topic-link LDA: joint models of topic and author community," *In Proceedings of the 26th Annual International Conference on Machine Learning*, 665-672.

[23] McDaid, A., & Hurley, N. (2010, August), "Detecting highly overlapping communities with model-based overlapping seed expansion," *In 2010 International Conference on Advances in Social Networks Analysis and Mining*, 112-119.

[24] McDaid, A. F., Murphy, T. B., Friel, N., & Hurley, N. J. (2013), "Improved Bayesian inference for the stochastic block model with application to large networks," *Computational Statistics and Data Analysis*, 60, 12-31.

[25] Muller, E., & Peres, R. (2018 online), "The effect of social networks structure on innovation performance: A review and directions for research," *International Journal of Research in Marketing*.

[26] Newman, M. E. (2006), "Modularity and community structure in networks," *Proceedings of the National Academy of Sciences*, 103(23), 8577-8582.

[27] Ng, A. Y., Jordan, M. I., & Weiss, Y. (2002), "On spectral clustering: Analysis and an algorithm." *In Advances in Neural Information Processing Systems*, 849-856.

[28] Nowicki, K., & Snijders, T. A. B. (2001), "Estimation and prediction for stochastic blockstructures," *Journal of the American Statistical Association*, 96(455), 1077-1087.

[29] Pathak, N., DeLong, C., Banerjee, A., & Erickson, K. (2008, August), "Social topic models for community extraction," *In The 2nd SNA-KDD Workshop*, 8, 2008.

[30] Peng, J., Agarwal, A., Hosanagar, K., & Iyengar, R. (2018), "Network overlap and content sharing on social media platforms," *Journal of Marketing Research*, 55(4), 571-585.

[31] Peres, R. (2014), "The impact of network characteristics on the diffusion of innovations," *Physica A: Statistical Mechanics and its Applications*, 402, 330-343.

[32] Saldana, D. F., Yu, Y., & Feng, Y. (2017), "How many communities are there?," *Journal of Computational and Graphical Statistics*, 26(1), 171-181.

[33] Schwarz, G. (1978), "Estimating the dimension of a model," *The Annals of statistics*, 6(2), 461-464.

[34] Snijders, T. A., & Nowicki, K. (1997), "Estimation and prediction for stochastic blockmodels for graphs with latent block structure," *Journal of Classification*, 14(1), 75-100.

[35] Sussman, D. L., Tang, M., Fishkind, D. E., & Priebe, C. E. (2012), "A consistent adjacency spectral embedding for stochastic blockmodel graphs," *Journal of the American Statistical Association*, 107(499), 1119-1128.

[36] Wang, Y. J., & Wong, G. Y. (1987), "Stochastic blockmodels for directed graphs," *Journal of the American Statistical Association*, 82(397), 8-19.

[37] Watanabe, S. (2010), "Asymptotic equivalence of Bayes cross validation and widely applicable information criterion in singular learning theory," *Journal of Machine Learning Research*, 11(Dec), 3571-3594.

[38] Xing, E. P., Fu, W., & Song, L. (2010), "A state-space mixed membership blockmodel for dynamic network tomography," *The Annals of Applied Statistics*, 4(2), 535-566.

[39] Zhou, D., Manavoglu, E., Li, J., Giles, C. L., & Zha, H. (2006, May), "Probabilistic models for discovering e-communities," *In Proceedings of the 15th International Conference on World Wide Web*, 173-182.