Journal of the American Statistical Association

ISSN: (Print) (Online) Journal homepage: www.tandfonline.com/journals/uasa20

Survival Mixed Membership Blockmodel

Fangda Song, Jing Chu, Shuangge Ma & Yingying Wei

To cite this article: Fangda Song, Jing Chu, Shuangge Ma & Yingying Wei (2024) Survival Mixed Membership Blockmodel, Journal of the American Statistical Association, 119:546, 1647-1656, DOI: 10.1080/01621459.2023.2213466

To link to this article: https://doi.org/10.1080/01621459.2023.2213466

JOURNAL OF THE AMERICAN STATISTICAL ASSOCIATION
2024, VOL. 119, NO. 546, 1647–1656: Theory and Methods
https://doi.org/10.1080/01621459.2023.2213466

Survival Mixed Membership Blockmodel

Fangda Song[a], Jing Chu[b], Shuangge Ma[c], and Yingying Wei[b]

[a]School of Data Science, The Chinese University of Hong Kong (Shenzhen), Shenzhen, China; [b]Department of Statistics, The Chinese University of Hong Kong, Hong Kong, China; [c]Department of Biostatistics, Yale University, New Haven, CT, USA

ABSTRACT

Whenever we send a message via a channel such as E-mail, Facebook, WhatsApp, WeChat, or LinkedIn, we care about the response rate—the probability that our message will receive a response—and the response time—how long it will take to receive a reply. Recent studies have made considerable efforts to model the sending behaviors of messages in social networks with point processes. However, statistical research on modeling response rates and response times on social networks is still lacking. Compared with sending behaviors, which are often determined by the sender's characteristics, response rates and response times further depend on the relationship between the sender and the receiver. Here, we develop a survival mixed membership blockmodel (SMMB) that integrates semiparametric cure rate models with a mixed membership stochastic blockmodel to analyze time-to-event data observed for node pairs in a social network, and we are able to prove its model identifiability without the pure node assumption. We develop a Markov chain Monte Carlo algorithm to conduct posterior inference and select the number of social clusters in the network according to the conditional deviance information criterion. The application of the SMMB to the Enron E-mail corpus offers novel insights into the company's organization and power relations. Supplementary materials for this article are available online.

ARTICLE HISTORY
Received April 2021
Accepted April 2023

KEYWORDS
Model identifiability; Semiparametric cure rate model; Social networks; Time-to-event data

1. Introduction

With the rapid development of social media and electronic commerce, many social networks now encompass a sequence of time-to-event data for a pair of social actors. In an E-mail network, how long it takes for the receiver of an E-mail to reply to the sender is available; on Twitter and Sina Weibo, the time for a user to repost a message from another user can be recorded; in a company that uses web-based electronic document management systems such as those provided by Dropbox and ParaDM, the time an employee processes a file shared by another employee can be tracked. Compared to the number of events between a pair of actors, time-to-event data can provide alternative perspectives into interpersonal relationships. For instance, an employee may have many more E-mails with a secretary than with the CEO but replies to E-mails from the CEO much faster than to those from the secretary. However, despite the active research on time-to-event data on social networks, statistical research on response times and response rates on a social network is still lacking.

Many previous studies on time-to-event data from a social network focus on the behavior undertaken by a social actor to initiate events, either by sending E-mails or by posting tweets. Such behavior has often been modeled by point processes. Perry and Wolfe (2013) treat the time of sending E-mails between pairs of actors in a network as a multivariate point process. Fox et al. (2016) adopt self-exciting point processes to model the rate of sending E-mails, in which the intensity function comprises a baseline sending rate and triggering functions that characterize the impact of receiving E-mails on sending rates. Matias, Rebafka, and Villers (2018) assume that each actor belongs to one and only one social group, in line with the classical stochastic blockmodel (Wang and Wong 1987), and assume a separate conditional inhomogeneous Poisson process for each social group pair. They use the proposed model to analyze the times at which actor A sends E-mails to actor B. Sit, Ying, and Yu (2021) model the times that an actor A sends an E-mail to actor B as recurrent events and develop a pseudo-partial likelihood approach to capture the flexible dependence structures among pairs of actors in the network. Zhang et al. (2022) jointly model the times at which actors initiate two events: posting their own tweets and reposting tweets from others. All of these studies focus on how frequently actor A initiates events (such as by posting tweets) or how frequently actor A interacts with actor B (such as by sending an E-mail to actor B).

Nevertheless, we not only care how frequently friends or colleagues contact us, but we are also always eager to know (*a*) whether the recipient will reply to our message (response rate) and (*b*) how long it will take us to receive a reply (response time). Both the response rate and the response time encode important information. For example, even if two actors send E-mails with a similar frequency, the one with a higher response rate will be more influential in their social network. Therefore, instead of studying the sending behaviors, in this article, we will model response rates and response times on a social network.

CONTACT Yingying Wei ywei@cuhk.edu.hk Department of Statistics, The Chinese University of Hong Kong, Hong Kong, China.
Supplementary materials for this article are available online. Please go to www.tandfonline.com/r/JASA.

Noteworthy, the response behavior of actors in a social network can be highly heterogeneous. Recently, Rastelli and Fop (2020) propose the exponential stochastic blockmodel (expSBM) to examine the duration of the interaction between actors. expSBM assumes that the network structure follows the stochastic blockmodel (Wang and Wong 1987) and that the interaction lengths follow exponential distributions. To fit into their framework, we may view actors' E-mail response behaviors as interactions and the time period with no E-mail and with no E-mail reply as a noninteraction period. Because the stochastic blockmodel partitions the network into several communities and lets each actor belong to one and only one community, although expSBM allows individuals' response speed to vary, it assumes the response speed of a given social actor to be the same no matter whom he or she interacts with. However, even for the same person, his or her response speed can be very different when communicating with different people. For example, the speed of an employee's responses to the CEO, a secretary, and his or her peers can vary considerably. Moreover, expSBM cannot deal with censoring time and hence E-mails that are never replied to.

To deal with the challenges of nonresponse events, severe heterogeneity, and sparse observations, we integrate the mixed membership stochastic blockmodel (MMSB) (Airoldi et al. 2008) with the semiparametric cure rate model (SCRM) (Chen, Ibrahim, and Sinha 1999; Ibrahim, Chen, and Sinha 2001a, chap. 5) to build a survival mixed membership blockmodel (SMMB). The SMMB allows an actor to play different roles when interacting with different actors, accounts for the impact of covariates, and models nonresponse events. Globally, the SMMB can detect social clusters in a network and characterize between- and within-cluster interaction patterns. Locally, the SMMB is capable of identifying the specific role an actor plays when he or she interacts with another actor. The SMMB simultaneously learns the network structure from the data and uses the learned structure to improve actor pair-specific inference. Consequently, the SMMB pools information across the whole network.

The network structure inferred from time-to-event data can be very different from the structure learned from a binary network W that records the connectivity between actors. In this article, we use "communities" to refer to the grouping of actors according to their connectivity encoded by W and "roles" to refer to the grouping of actors according to their response times and response rates. If W follows an MMSB, then actors belonging to the same community have a high probability of communicating with each other—sending E-mails in the case of the Enron E-mail corpus. In comparison, actors sharing the same role have similar distributions of response times and response rates in the SMMB. Therefore, communities and roles reveal different aspects of a social network with time-to-event data. For example, different communities can correspond to the various departments of a company, whereas roles reflect employees' levels of seniority. Employees in the same department tend to have more E-mail contacts, and thus are grouped together by W, but have very different E-mail response speeds and rates because of their different levels of seniority.

As pointed out by Zhang, Levina, and Zhu (2020), it is nontrivial to establish identifiability for social network models that allow actors to belong to multiple social clusters such as the MMSB (Airoldi et al. 2008). Recently, Mao, Sarkar, and Chakrabarti (2021) prove that the requirement of the existence of at least one pure node, defined as a social actor who acts as a member of only one social cluster irrespective of whom he or she communicates with, for each social cluster is not only sufficient (Kaufmanna, Bonaldb, and Lelargec 2018; Zhang, Levina, and Zhu 2020) but also necessary for the MMSB for binary networks to be identifiable. Fortunately, compared with binary data, time-to-event data encode more information, and hence we are able to prove that the SMMB is identifiable up to label switching without the pure node assumption.

We conduct statistical inference under the Bayesian framework and develop a Markov chain Monte Carlo (MCMC) algorithm. The application of the SMMB to the Enron E-mail corpus reveals network structure that better represent leadership patterns than those obtained from the state-of-the-art methods.

2. Model Formulation

Suppose that a social network consists of N actors and has sequences of time-to-event data, instead of binary outcomes W_{ij}s, recorded for actor pairs. For a pair of actors (i, j), when actor i sends a sequence of E-mails to actor j, the response (failure) time and the censoring time for the gth E-mail by actor j are denoted as T_{ijg} and C_{ijg}, respectively. The associated p-dimensional covariates are denoted as $\boldsymbol{X}_{ijg} = (X_{ijg1}, X_{ijg2}, \ldots, X_{ijgp})^T$, where $X_{ijg1} = 1$ corresponds to the intercept term. If actor i sends n_{ij} E-mails to actor j, then we observe a sequence of survival times $\boldsymbol{y}_{ij} = (y_{ij1}, \ldots, y_{ijn_{ij}})^T$ and censoring indicators $\boldsymbol{v}_{ij} = (v_{ij1}, \ldots, v_{ijn_{ij}})^T$ satisfying $y_{ijg} = \min(t_{ijg}, c_{ijg})$ and $v_{ijg} = \mathbf{1}(t_{ijg} < c_{ijg})$, and the corresponding covariates $\boldsymbol{x}_{ij} = (\boldsymbol{x}_{ij1}, \ldots, \boldsymbol{x}_{ijn_{ij}})^T$. Here, $\mathcal{N} = \{1, \ldots, N\}$ denotes the set of actors and $\mathcal{E} = \{(i, j) | n_{ij} > 0, i, j \in \mathcal{N}\}$ represents the set of actor pairs with at least one failure or censoring observation.

Following the MMSB, assume that there exist K roles with Dirichlet parameters $\boldsymbol{\xi} = (\xi_1, \ldots, \xi_K)^T \in (0, +\infty)^K$. Here, $\frac{\xi_k}{\sum_{l=1}^K \xi_l}$ represents the abundance of the kth role in the network. Each pair of roles has a unique SCRM.

For the SCRM, we adopt the formulation $S(t) = \exp(-\theta + \theta S_0(t)) = \tau^{1-S_0(t)}$ (Chen, Ibrahim, and Sinha 1999; Ibrahim, Chen, and Sinha 2001a, chap. 5), which enjoys a population level proportional hazards structure. Here, $S_0(t)$ is a proper survival function for the baseline, and its hazard function $h_0(t)$ is piecewise constant: $h_0(t) = \lambda_m$ if $t \in (s_{m-1}, s_m]$, with $0 < s_1 < s_2 < \cdots < s_M < \infty$ being a partition of the time axis and $h_0(t) = 0$ for $t > s_M$. Meanwhile, $\tau = \exp(-\theta)$ is the cure proportion. The survival time of this model can be viewed as the earliest event time of $\tilde{N}$ iid competing risks with survival probability $S_0(t)$ and $\tilde{N}$ following a Poisson distribution Pois(θ), so θ can be interpreted as the average number of competing risks (Chen, Ibrahim, and Sinha 1999). Similarly, we may suppose that there exist $\tilde{N}$ latent intentions to reply to an E-mail and view θ as the average number of latent response intentions. Furthermore, we can model covariate effects $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^T$ via $\theta = \exp(\boldsymbol{x}^T\boldsymbol{\beta})$, which corresponds to a log-log link for τ: $\log(-\log(\tau)) = \boldsymbol{x}^T\boldsymbol{\beta}$. Consequently, the hazard function of the SCRM is $h(t|\boldsymbol{x}) = \exp(\boldsymbol{x}^T\boldsymbol{\beta}) f_0(t)$, where $f_0(t)$ denotes the probability density function for $S_0(t)$. In the SMMB, we let both the baseline hazards and covariate effects to be role-pair-specific.

When role k replies to the E-mails sent by role l, the piecewise constant baseline hazard function in interval $t \in (s_{m-1}, s_m]$ is $h_0^{lk}(t) = \lambda_m^{lk}$. We collect $\boldsymbol{\lambda}^{lk} = (\lambda_1^{lk}, \ldots, \lambda_M^{lk})$ and denote the role-pair-specific covariate effects as $\boldsymbol{\beta}^{lk} = (\beta_1^{lk}, \ldots, \beta_p^{lk})^T$. Notably, the time-to-event data from role l to role $k \neq l$ do not necessarily have the same distribution as those from role k to role l. In other words, $f^{lk}(y, v|\boldsymbol{x}, \boldsymbol{\beta}^{lk}, \boldsymbol{\lambda}^{lk}) = f^{kl}(y, v|\boldsymbol{x}, \boldsymbol{\beta}^{kl}, \boldsymbol{\lambda}^{kl})$ is not required. Instead, the interactions are allowed to be asymmetric.

When actor j replies to an E-mail sent by actor i, actor i and actor j pick up their latent social role Q_{ij} and R_{ij} according to their latent role proportions $\boldsymbol{\pi}_i = (\pi_{i1}, \ldots, \pi_{iK})^T$ and $\boldsymbol{\pi}_j$, respectively. Given the latent roles $Q_{ij} = l$ and $R_{ij} = k$, a sequence of failure times $T_{ijg}, g = 1, \ldots, n_{ij}$ are iid generated according to a SCRM with the piecewise constant baseline hazards $\boldsymbol{\lambda}^{lk}$ and the coefficients $\boldsymbol{\beta}^{lk}$.

We denote $\boldsymbol{\xi} = (\xi_1, \ldots, \xi_K)^T$, $\boldsymbol{\beta} = \{\boldsymbol{\beta}^{lk}\}_{k=1,\ldots,K}^{l=1,\ldots,K}$ and $\boldsymbol{\Lambda} = \{\boldsymbol{\lambda}^{lk}\}_{k=1,\ldots,K}^{l=1,\ldots,K}$. The proposed SMMB can then be described by the following data generation mechanism:

$$\begin{aligned}
\boldsymbol{\pi}_i &\sim \text{Dirichlet}(\boldsymbol{\xi}) && i = 1, \ldots, N;\\
Q_{ij}|\boldsymbol{\pi}_i &\sim \text{Categorical}(\boldsymbol{\pi}_i) && (i,j) \in \mathcal{E};\\
R_{ij}|\boldsymbol{\pi}_j &\sim \text{Categorical}(\boldsymbol{\pi}_j) && (i,j) \in \mathcal{E};\\
Y_{ijg}, v_{ijg}|Q_{ij} = l, R_{ij} = k &\sim f(y_{ijg}, v_{ijg}|Q_{ij} = l, R_{ij} = k, \boldsymbol{x}_{ijg}, \boldsymbol{\beta}, \boldsymbol{\Lambda})\\
&= f(y_{ijg}, v_{ijg}|\boldsymbol{x}_{ijg}, \boldsymbol{\beta}^{lk}, \boldsymbol{\lambda}^{lk}) \quad g = 1, \ldots, n_{ij},
\end{aligned} \tag{1}$$

where $\boldsymbol{\pi}_i, i = 1, \ldots, N$, are iid given $\boldsymbol{\xi}$; Q_{ij} (R_{ij}) for $(i,j) \in \mathcal{E}$ are independent conditional on $\boldsymbol{\pi}_i$ ($\boldsymbol{\pi}_j$); conditional on Q_{ij}, R_{ij} and given the parameters $\boldsymbol{\beta}, \boldsymbol{\Lambda}$ and covariates $\boldsymbol{x}_{ijg}$, failure time T_{ijg} and censoring time C_{ijg} are independent, and for $g = 1, \ldots, n_{ij}$, the corresponding observed survival times and censoring indicators (Y_{ijg}, v_{ijg})s are independent. It is also interesting to note the link between the SMMB and the MMSB. When y_{ij} is a binary random variable and there is exactly one observation y_{ij} on each edge with no censoring ($v_{ij} = 1$), the SMMB reduces to the MMSB.

Following the inference for the SCRM model (Chen, Ibrahim, and Sinha 1999), we can augment a latent variable U_{ijg} for each E-mail by each pair of actors to facilitate the computation. If actor i plays role $Q_{ij} = l$ and actor j plays role $R_{ij} = k$ when actor j replies to actor i's E-mails, then U_{ijg} follows $\text{Pois}(\exp(\boldsymbol{x}_{ijg}^T\boldsymbol{\beta}^{lk})) + v_{ijg}$. Consequently, in the SMMB, $\boldsymbol{\Theta} = (\boldsymbol{\xi}, \boldsymbol{\beta}, \boldsymbol{\Lambda})$ are the parameters of interest; $\boldsymbol{y} = \{y_{ijg}\}_{(i,j)\in\mathcal{E}}^{g=1,\ldots,n_{ij}}$, $\boldsymbol{v} = \{v_{ijg}\}_{(i,j)\in\mathcal{E}}^{g=1,\ldots,n_{ij}}$ and $\boldsymbol{X} = \{x_{ijgr}\}_{(i,j)\in\mathcal{E}}^{g=1,\ldots,n_{ij};r=1,\ldots,p}$ are the observed data; and $\boldsymbol{Q} = \{Q_{ij}\}_{(i,j)\in\mathcal{E}}$, $\boldsymbol{R} = \{R_{ij}\}_{(i,j)\in\mathcal{E}}$, $\boldsymbol{\Pi} = (\boldsymbol{\pi}_1, \boldsymbol{\pi}_2, \ldots, \boldsymbol{\pi}_N)$ and $\boldsymbol{U} = \{U_{ijg}\}_{(i,j)\in\mathcal{E}}^{g=1,\ldots,n_{ij}}$ are latent variables. Let $\delta_{ijgm} = 1$ if $y_{ijg} \in (s_{m-1}, s_m]$ and zero otherwise. Then, the complete data likelihood function becomes:

$$\begin{aligned}
L_c(\boldsymbol{\Theta}|\boldsymbol{y}, \boldsymbol{v}, \boldsymbol{X}, \boldsymbol{\Pi}, \boldsymbol{Q}, \boldsymbol{R}, \boldsymbol{U}) = \prod_{i\in\mathcal{N}} p(\boldsymbol{\pi}_i|\boldsymbol{\xi}) \prod_{(i,j)\in\mathcal{E}} \prod_{l=1}^{K}\prod_{k=1}^{K} [\pi_{il}\pi_{jk} \prod_{g=1}^{n_{ij}} [S_0(y_{ijg}|\boldsymbol{\lambda}^{lk})^{U_{ijg}-v_{ijg}} (U_{ijg} f_0(y_{ijg}|\boldsymbol{\lambda}^{lk}))^{v_{ijg}}]\\
\cdot \exp\{\sum_{g=1}^{n_{ij}} [U_{ijg}\boldsymbol{x}_{ijg}^T\boldsymbol{\beta}^{lk} - \log(U_{ijg}!) - \exp(\boldsymbol{x}_{ijg}^T\boldsymbol{\beta}^{lk})]\}]^{\mathbf{1}(Q_{ij}=l)\mathbf{1}(R_{ij}=k)},
\end{aligned} \tag{2}$$

where $p(\boldsymbol{\pi}_i|\boldsymbol{\xi}) = \frac{\Gamma(\sum_{l=1}^K \xi_l)}{\prod_{k=1}^K \Gamma(\xi_k)} \prod_{k=1}^K \pi_{ik}^{\xi_k - 1}$, $S_0(y_{ijg}|\boldsymbol{\lambda}^{lk}) = \prod_{m=1}^M \exp\{-\delta_{ijgm}[\lambda_m^{lk}(y_{ijg} - s_{m-1}) + \sum_{q=1}^{m-1} \lambda_q^{lk}(s_q - s_{q-1})]\}$,, and $f_0(y_{ijg}|\boldsymbol{\lambda}^{lk}) = \prod_{m=1}^M (\lambda_m^{lk})^{\delta_{ijgm}} \exp\{-\delta_{ijgm}[\lambda_m^{lk}(y_{ijg} - s_{m-1}) + \sum_{q=1}^{m-1} \lambda_q^{lk}(s_q - s_{q-1})]\}$.
After marginalizing over the latent variables, the observed data likelihood is as follows:

$$L_o(\boldsymbol{\Theta}|\boldsymbol{y}, \boldsymbol{v}, \boldsymbol{X}) = \int_{\boldsymbol{\Pi}} \prod_{(i,j)\in\mathcal{E}} \sum_{l=1}^{K}\sum_{k=1}^{K} \pi_{il}\pi_{jk} \prod_{g=1}^{n_{ij}} \exp\{v_{ijg}[\boldsymbol{x}_{ijg}^T\boldsymbol{\beta}^{lk} + \log(f_0(y_{ijg}|\boldsymbol{\lambda}^{lk}))] - \exp(\boldsymbol{x}_{ijg}^T\boldsymbol{\beta}^{lk})(1 - S_0(y_{ijg}|\boldsymbol{\lambda}^{lk}))\} \prod_{i\in\mathcal{N}} [p(\boldsymbol{\pi}_i|\boldsymbol{\xi}) d\boldsymbol{\pi}_i]. \tag{3}$$

For reference, we have summarized all the notations in Table S1.

3. Model Identifiability

There is emerging research on the identifiability of the MMSB (Mao, Sarkar, and Chakrabarti 2017; Zhang, Levina, and Zhu 2020; Mao, Sarkar, and Chakrabarti 2021). As mixture models, both the regular stochastic blockmodel (Wang and Wong 1987) and the MMSB are only identifiable up to the label switching of communities. Theorem 1 shows that the SMMB is also identifiable up to the label switching of roles given easily met regularity conditions.

Theorem 1. In a network $\mathcal{G} = (\mathcal{N}, \mathcal{E})$, $\mathcal{N}$ denotes a set of actors and $\mathcal{E} = \{(i,j)|n_{ij} > 0, i, j \in \mathcal{N}\}$ represents the set of actor pairs with at least one observation. Suppose that we observe n_{ij} time-to-event data for each actor pair (i,j). Let $K \geq 2$ denote the number of roles and V be an open set that consist of all possible covariate values. Given that

(C1) There exist three actors (i, j_1, j_2) such that $n_{ij_1} > 0$ and $n_{ij_2} > 0$;

(C2) There exists a covariate set $V^* \subset V$ satisfying

$$V^* = \{(x_1, x_2, \ldots, x_p)^T \in \mathbb{R}^p | x_1 = 1 \text{ and } x_r \in \{x_{r1}^*, x_{r2}^*\}, r = 2, 3, \ldots, p\},$$

where $(x_{r1}^*, x_{r2}^*), r = 1, \ldots, p$, are two distinct real numbers such that for two distinct role pairs (l, k) and $(\tilde{l}, \tilde{k})$, and $x^* \in V^*$,

$$\exp((x^*)^T \beta^{lk}) \exp(-\sum_{q=1}^{m} \lambda_q^{lk}(s_q - s_{q-1})) \neq \exp((x^*)^T \beta^{\tilde{l}\tilde{k}}) \exp(-\sum_{q=1}^{m} \lambda_q^{\tilde{l}\tilde{k}}(s_q - s_{q-1}))$$

at knot $m, m = 1, \ldots, M$;

then the SMMB is identifiable (up to label switching) in the sense that if two sets of parameters $\Theta = \{\xi, \lambda^{lk}, \beta^{lk}\}_{l,k=1,\cdots,K}$ and $\widetilde{\Theta} = \{\delta, \mu^{rs}, \gamma^{rs}\}_{r,s=1,\ldots,\widetilde{K}}$ give rise to the same likelihood function as (3) for any y_{ijg} and $x_{ijg} = (1, x_{ijg2}, x_{ijg3}, \ldots, x_{ijg,p})^T \in V \subset \mathbb{R}^p$, then $K = \widetilde{K}, \xi_l = \delta_{\rho(l)}, \lambda^{lk} = \mu^{\rho(l)\rho(k)}$ and $\beta^{lk} = \gamma^{\rho(l)\rho(k)}$, where ρ is a permutation of $\{1, \ldots, K\}$.

Condition 1 requires the existence of an actor who has time-to-event data with at least two actors, which is easily met for real social networks. Condition 2 requires each covariate to take at least two values so that the survival functions of different role pairs are distinct at all of the knots; this is similar to the non-singular block matrix assumption that ensures the identifiability of the MMSB for binary networks (Mao, Sarkar, and Chakrabarti 2021). For the identifiability of the MMSB, Mao, Sarkar, and Chakrabarti (2021) require the existence of pure nodes who each play one and only one role. Fortunately, because a sequence of survival outcomes y_{ijg} rather than a single binary or continuous weight is observed for each pair of actors, much more information is available for the SMMB to distinguish between different role pairs. Consequently, the assumption of the existence of pure nodes is released in our theorem.

In the proof, we need to show that if two sets of parameters $\Theta = \{\xi, \lambda^{lk}, \beta^{lk}\}_{l,k=1,\ldots,K}$ and $\widetilde{\Theta} = \{\delta, \mu^{rs}, \gamma^{rs}\}_{r,s=1,\ldots,\widetilde{K}}$ give rise to the same observed data likelihood function, then $\widetilde{\Theta}$ must be a permutation of Θ with $K = \widetilde{K}$, $\xi_l = \delta_{\rho(l)}$, $\lambda^{lk} = \mu^{\rho(l)\rho(k)}$ and $\beta^{lk} = \gamma^{\rho(l)\rho(k)}$. It is difficult to show the existence of such a permutation $\rho(\cdot)$ directly. Instead, by first viewing the model as a mixture of cure rate models with K^2 components, we are able to find a permutation $\tilde{\rho}(\cdot, \cdot)$ of all K^2 role pairs such that these two sets of parameters satisfy $\lambda^{lk} = \mu^{\tilde{\rho}(l,k)}$ and $\beta^{lk} = \gamma^{\tilde{\rho}(l,k)}$. Then, we show that the permutation $\tilde{\rho}(\cdot, \cdot)$ can be decomposed with the single permutation $\rho(\cdot)$ so that $\tilde{\rho}(l, k) = (\rho(l), \rho(k))$. Finally, these two sets of parameters can be equal up to the label switching $\rho(\cdot)$, that is, $\xi_l = \delta_{\rho(l)}, \lambda^{lk} = \mu^{\rho(l),\rho(k)}$, and $\beta^{lk} = \gamma^{\rho(l),\rho(k)}$. The proof of Theorem 1 is outlined in the Appendix and provided in details in Supplementary Sections S2–S4.

4. Statistical Inference

We conduct the statistical inference under the Bayesian framework. We first reparameterize ξ as $\alpha = \sum_{k=1}^{K} \xi_k$ and $\eta_k = \frac{\xi_k}{\alpha}, k = 1, \ldots, K$. For the MMSB, Airoldi et al. (2008) let $\frac{\alpha}{K}$ vary between 0.05 and 0.25, so here following the same philosophy, we impose a Beta(a, a) prior on α and a non-informative Dirichlet$(1, 1, \ldots, 1)$ prior on η independently. We follow the common practice in choosing independent Gaussian priors for $\beta_r^{lk} \sim N(b_r, \sigma^2)$ with $b_r = 0$, for $r = 2, \ldots, p$, and independent Gamma priors for the baseline hazard $\lambda_m^{lk} \sim$ Gamma$(\frac{\omega_{0m}}{\kappa}, \frac{1}{\kappa}), m = 1, \ldots, M$. By default, (a, σ^2, κ) is set as $(1, 5, 50)$, and b_1 and ω_{0m}s rely on their empirical estimators. In particular, we first use the Kaplan-Meier estimator (Kaplan and Meier 1958) to estimate an overall survival function for all of the observations. As a result, an empirical estimation b_1^{emp} is $\log(-\log(S_{\text{KM}}(s_M)))$, where $S_{KM}(t)$ represents the Kaplan-Meier estimator of the survival probability at time t. Given $b_1 = b_1^{\text{emp}}$ and $S_{\text{KM}}(t)$, we then sample latent variables U_{ijg}^{emp} from Pois$(S_{\text{KM}}(y_{ijg}) \exp(b_1^{\text{emp}})) + \nu_{ijg}$ and set ω_{0m} as the estimated baseline hazard (Robbins 1955):

$$\omega_{0m} = \frac{\sum_{(i,j)\in\mathcal{E}} \sum_{g=1}^{n_{ij}} \delta_{ijgm} \nu_{ijg}}{\sum_{(i,j)\in\mathcal{E}} \sum_{g=1}^{n_{ij}} \{U_{ijg}^{\text{emp}} [\delta_{ijgm}(y_{ijg} - s_{m-1}) + \sum_{q=m+1}^{M} \delta_{ijgq}(s_q - s_{q-1})]\}}.$$

Let us now discuss the asymptotic behavior of the posterior distribution of the SMMB. Without loss of generality, let $n_{ij} = n$ for all actor pairs $(i, j) \in \mathcal{E}$. Let us denote $E = |\mathcal{E}|$ as the number of active edges. When two actor pairs share an actor, for example, (i, j_1) and (i, j_2), their survival outcomes are dependent. Inspired by the Bayesian consistency results for non-iid data (Ghosal and Van Der Vaart 2007; Ghosal and Van der Vaart 2017), we provide sufficient conditions under which the posterior of the SMMB concentrates on the true parameters as E goes to infinity.

Theorem 2. Assume that the conditions in Theorem 1 hold so that the SMMB is identifiable and that X_{ijg} are generated iid from a bounded distribution $\tilde{f}_X(x)$ such that $f_X(x) = \prod_{(i,j)\in\mathcal{E}} \prod_{g=1}^{n} \tilde{f}_X(x_{ijg})$. Let $P_{\Theta}^{(E,n)}$ denote the joint distribution of E observations $\{Y_{ij}, \nu_{ij}, X_{ij}, (i, j) \in \mathcal{E}\}$ given Θ that admits the density $f(y, \nu, x|\Theta) = f(y, \nu|\Theta, x) f_X(x)$, where $f(y, \nu|\Theta, x)$ follows the SMMB. Let $P_{\Theta_0}^{(E,n)}$ indicate the joint distribution under the true parameters Θ_0 and $\Pi(\Theta)$ represent the prior distribution of Θ. If for $\varepsilon > 0$, there exists an ε-neighborhood $\mathcal{N}_\varepsilon(\Theta_0) \subset \Omega$ such that $\Pi(\mathcal{N}_\varepsilon(\Theta_0)) > 0$ and it satisfies

(C1) For some constant $b > 0$ and $B > 0$, there exist test functions $\Phi_E(Y, \nu, X)$ for testing $H_0 : \Theta = \Theta_0$ versus $H_1 : \Theta \notin \mathcal{N}_\varepsilon(\Theta_0)$ such that

$$P_{\Theta_0}^{(E,n)}(\Phi_E) := \int \Phi_E dP_{\Theta_0}^{(E,n)} \leq Be^{-bE}$$

$$\sup_{\Theta\in\mathcal{N}_\varepsilon^c(\Theta_0)} P_{\Theta}^{(E,n)}(1 - \Phi_E) := \sup_{\Theta\in\mathcal{N}_\varepsilon^c(\Theta_0)} \int (1 - \Phi_E) dP_{\Theta}^{(E,n)} \leq Be^{-bE}$$

(C2) There exists some positive $b_0 < b$ such that for any $\Theta \in \mathcal{N}_\varepsilon(\Theta_0)$

$$\frac{1}{E} \log\left(\frac{f(y, \nu|\Theta_0, x)}{f(y, \nu|\Theta, x)}\right) \leq \frac{b_0}{2}, P_{\Theta_0}^{(E,n)}\text{-a.s.},$$

then as E goes to infinity, the posterior distribution $\tilde{\Pi}(\Theta \in \mathcal{N}_\varepsilon^c(\Theta_0)|y, \nu, x) \to 0, P_{\Theta_0}^{(E,n)}$-a.s.

The proof of Theorem 2 is provided in Supplementary Section S5.

We develop an MCMC algorithm to draw samples from the posterior distribution. Here, we use superscript $[t]$ to denote the corresponding posterior samples of parameters or latent variables at the tth iteration of the MCMC algorithm. At the tth iteration:

1. To update the latent variable $U_{ijg}^{[t]}$ associated with the gth E-mail sent from actor i to actor j, we first sample from $\text{Pois}(S_0(y_{ijg}|\boldsymbol{\lambda}^{lk[t-1]})\exp(\boldsymbol{x}_{ijg}^T\boldsymbol{\beta}^{lk[t-1]}))$ and then add v_{ijg}.
2. To update the roles $Q_{ij}^{[t]}$ and $R_{ij}^{[t]}$ of actor i and actor j when j replies to an E-mail sent from i, we first collapse down the latent variable $\boldsymbol{U}_{ij} = (U_{ij1}, U_{ij2}, \ldots, U_{ijn_{ij}})^T$ and then sample these roles from a categorical distribution with probabilities (Liu 1994)

$$\begin{aligned}\Pr(Q_{ij}^{[t]} = l, R_{ij}^{[t]} = k|\boldsymbol{\lambda}^{lk[t-1]}, \boldsymbol{\beta}^{lk[t-1]}, \boldsymbol{\pi}_i^{[t-1]}, \boldsymbol{\pi}_j^{[t-1]}) \\ \propto \pi_{il}^{[t-1]}\pi_{jk}^{[t-1]} \cdot \prod_{g=1}^{n_{ij}}\{[f_0(y_{ijg}|\boldsymbol{\lambda}^{lk[t-1]})\exp(\boldsymbol{x}_{ijg}^T\boldsymbol{\beta}^{lk[t-1]})]^{v_{ijg}} \\ \cdot \exp[-\exp(\boldsymbol{x}_{ijg}^T\boldsymbol{\beta}^{lk[t-1]})(1 - S_0(y_{ijg}|\boldsymbol{\lambda}^{lk[t-1]}))]\}\end{aligned}$$

3. To update $\beta_r^{lk[t]}$, we apply a Metropolis-Hasting (MH) step with a symmetric Gaussian proposal distribution $g(\beta_r^{lk*}|\beta_r^{lk[t-1]}) \sim \text{N}(\beta_r^{lk[t-1]}, 0.1^2)$.
4. To update $\lambda_m^{lk[m]}$, we sample from its full conditional, which is a Gamma distribution with a shape parameter equal to

$$\sum_{(i,j)\in\mathcal{E}} \mathbf{1}(Q_{ij}^{[t]} = l, R_{ij}^{[t]} = k)\sum_{g=1}^{n_{ij}} v_{ijg}\delta_{ijgm} + \frac{\omega_{0m}}{\kappa}$$

and a rate parameter equal to

$$\begin{aligned}\sum_{(i,j)\in\mathcal{E}} \mathbf{1}(Q_{ij}^{[t]} = l, R_{ij}^{[t]} = k)\sum_{g=1}^{n_{ij}} U_{ijg}^{[t]}(\delta_{ijgm}(y_{ijg} - s_{m-1}) \\ + \sum_{q=m+1}^{M}\delta_{ijgq}(s_q - s_{q-1})) + \frac{1}{\kappa}.\end{aligned}$$

5. To update $\boldsymbol{\pi}_i^{[t]}$, we sample from its full conditional, a Dirichlet distribution with the following parameters

$$\begin{aligned}\Bigg(\sum_{j:(i,j)\in\mathcal{E}} \mathbf{1}(Q_{ij}^{[t]} = 1) + \sum_{j:(j,i)\in\mathcal{E}} \mathbf{1}(R_{ji}^{[t]} = 1) + \xi_1^{[t-1]}, \ldots, \\ \sum_{j:(i,j)\in\mathcal{E}} \mathbf{1}(Q_{ij}^{[t]} = K) + \sum_{j:(j,i)\in\mathcal{E}} \mathbf{1}(R_{ji}^{[t]} = K) + \xi_K^{[t-1]}\Bigg).\end{aligned}$$

6. To update $\alpha^{[t]}$, we apply an MH step with the proposal distribution $g(\frac{\alpha^*}{K}|\alpha^{[t-1]}) \sim \text{Beta}(\alpha^{[t-1]}, K - \alpha^{[t-1]})$.
7. To update $\boldsymbol{\eta}^{[t]}$, we adopt an MH step with the proposal distribution $g(\boldsymbol{\eta}^*|\boldsymbol{\eta}^{[t-1]}) \sim \text{Dirichlet}(10\boldsymbol{\eta}^{[t-1]})$.

To accelerate the convergence of the MCMC algorithm, inspired by the shift-mode Metropolis step proposed by Liu (1994), we incorporate another Metropolis step into the above algorithm to allow the global swapping of two randomly selected role pairs. Because this step involves extra sampling and computation, we insert it into the above algorithm every few iterations, such as every 10 iterations. We call the new step a *global swapping Metropolis step*. Specifically, a pair of role pairs (l_1, k_1) and (l_2, k_2) are randomly selected and their role-pair-specific parameters $\boldsymbol{\lambda}^{lk}$ and $\boldsymbol{\beta}^{lk}$ are swapped to obtain the proposed values $(\boldsymbol{\Lambda}^*, \boldsymbol{\beta}^*)$. We correspondingly swap the role labels Q_{ij}'s and R_{ij}'s as (Q_{ij}^*, R_{ij}^*). To propose $(\boldsymbol{\Pi}^*, \boldsymbol{\xi}^*)$, we incorporate an MH step to draw a sequence of $\boldsymbol{\pi}_i$ and $\boldsymbol{\xi}$ from their full conditional distributions given (Q_{ij}^*, R_{ij}^*). To reach convergence, we run the MH step for 50 iterations and take the samples at the 50th iteration as the proposed values $\boldsymbol{\Pi}^*$ and $\boldsymbol{\xi}^*$. If the swapping is rejected, we keep all of the parameters and latent variables unchanged; otherwise, we swap the role pair and update the parameters and latent variables with the proposed values. The detailed derivations of the full conditional distributions and the acceptance ratios are listed in Supplementary Section S6.

The Dirichlet parameters $\boldsymbol{\xi}$, the role-pair-specific piecewise constant hazards $\boldsymbol{\lambda}^{lk}$'s, the covariate coefficients $\boldsymbol{\beta}^{lk}$'s, and the user-specific role proportions $\boldsymbol{\pi}_i$'s are estimated by their posterior means, and the latent role pairs (Q_{ij}, R_{ij}) of each pair $(i,j) \in \mathcal{E}$ are estimated by their posterior modes.

Denoting $L := \sum_{(i,j)\in\mathcal{E}} n_{ij}$ as the total number of observed survival times, the time complexity of each step of the MCMC algorithm is shown in Table 1. In general, the number of covariates p is fixed, and the number of knots M is not very large. Therefore, the scalability of the proposed MCMC algorithm is determined by the number of roles K and the total number of observed survival times L. The total time complexity, which is the summation of all of the terms in Table 1, is linear in L and quadratic in K.

We determine the number of roles K present in the network according to the conditional deviance information criterion (DIC) (Celeux et al. 2006; Lu and Song 2012) and select K as the one that attains the minimum (see Supplementary Section S7 for details).

5. Simulation

5.1. A Simulation Dataset Mimicking the Enron E-mail Corpus

To evaluate the performance of the SMMB, we first generate a simulation dataset that mimics the Enron E-mail corpus. We generate a binary network with $N = 150$ actors using a

Table 1. Time complexity of MCMC steps.

Variables	$\boldsymbol{U}$	$(\boldsymbol{Q}, \boldsymbol{R})$	$\boldsymbol{\Pi}$	$\boldsymbol{\Lambda}$	$\boldsymbol{\beta}$	$\boldsymbol{\xi}$
Time complexity	$O(L(p+M))$	$O(LK^2(p+M))$	$O(E+NK)$	$O(LM)$	$O(Lp(p+M))$	$O(NK)$

NOTE: Here, E denotes the number of active edges in the network, and $L = \sum_{(i,j)\in\mathcal{E}} n_{ij}$ represents the total number of observed survival times.

stochastic block model (Wang and Wong 1987). Specifically, we assume the existence of four communities, which contain 25, 30, 40, and 55 actors, respectively (Figure S1). For each active edge, we generate the number of E-mails n_{ij} from a shifted negative binomial distribution. We assume that when replying to E-mails, actors can play $K = 3$ roles, incorporate two covariates, one binary and the other continuous, and divide the time axis by $M = 5$ knots. Consequently, the true underlying response times T_{ijg}'s are generated sequentially according to the SMMB. Finally, we generate the censoring time $C_{ijg} \sim \text{Unif}(0, 100)$ (see Supplementary Section S8 and Table S2 for more details).

We run 50,000 MCMC iterations with the first 25,000 as burnins. The estimated potential scale reduction (EPSR) criterion (Gelman et al. 2013) shows that the Markov chain has reached convergence after 25,000 iterations (Supplementary Section S9 and Table S7). To identify the number of roles, we vary the number of roles K from 1 to 5, and the conditional DIC correctly selects the number of roles as the true value $K = 3$. The estimated $\hat{\boldsymbol{\xi}} = (0.215, 0.368, 0.597)^T$ is consistent with the true $\boldsymbol{\xi}$. Figure S2 shows that the posterior means of piecewise constant hazards $\hat{\boldsymbol{\lambda}}^{lk}$ and intercept terms $\hat{\beta}_1^{lk}$ recover the true values of the baseline survival function of each role pair well (see also Tables S4–S6 and Figures S3–S4). The credible intervals of all the covariate coefficients β_r^{lk}s, $r = 2, 3, \ldots, p$, cover the true values (Table S2).

The contingency table of the estimated versus the true role pairs shows that the roles of the actors in each pair are inferred accurately (Table S3). Extensive sensitivity analyses (Supplementary Section S10) show that SMMB is robust to the choice of hyperparameters (Table S8) and the selection of knots (Table S9).

We perform posterior predictive check and use the L measure (Ibrahim, Chen, and Sinha 2001b) to compare the SMMB with (a) fitting a single SCRM (overall SCRM) to all the data and (b) fitting a separate SCRM to each actor pair (pairwise SCRM) (Supplementary Section S11). The SMMB achieves the smallest L measure and hence outperforms the two benchmark methods in goodness of fit (Table S10).

5.2. Simulation Datasets with Different Heterogeneity and Sparsity Levels

We further test the performance of the SMMB using datasets generated with different heterogeneity levels, different numbers of active edges and different numbers of observed survival times per active edge. In order to vary the sparsity of the network, here we generate the binary networks W with $N = 150$ using the Erdős-Rényi model (Erdős and Rényi 1960) with the connectivity probability p_c, which denotes the probability that a directed edge is drawn between two arbitrary actors. We generate the baseline probabilities, the censoring times C_{ijg}, and the underlying response times T_{ijg} similarly as in Section 5.1. Details are presented in Supplementary Section S12.

First, we fix the number of roles at $K = 3$ and vary the connectivity probability p_c over $(0.2, 0.3, 0.4, 0.5)$ to investigate how the network sparsity affects the recovery performance. Next, we fix the number of roles at $K = 3$ and the connectivity probability $p_c = 0.3$ but vary the average number of observations on each edge. In particular, we sample the number of E-mails n_{ij} from $\text{Pois}(\mu) + 5$ and vary μ over $(10, 15, 20, 25, 30, 35)$. Finally, we are also interested in the impact of pattern heterogeneity on parameter estimation. Thus, we fix the connectivity probability $p_c = 0.3$ and generate the number of E-mails per active edge from $\text{Pois}(25) + 5$ while varying the number of roles K from 2 to 5. We generate 100 replicated datasets for each setting. In total, there are 1,200 simulation datasets.

For each simulation dataset, we run 50,000 MCMC iterations with the first 25,000 as burnins. Almost all of the parameters enjoy small biases, standard deviation (SD), and root mean square errors (RMSE), and their coverage probabilities (CP) of the 95% credible interval are high (Figures 1 and Supplement Tables S12–S47).

Moreover, these simulation studies allow us to verify the theoretical calculation of time complexity of our MCMC algorithm. Figure 2 confirms that the time complexity of the algorithm is linear in the total number of observed survival times L and quadratic in the number of roles K.

We also examine the performance of the conditional DIC in determining the number of roles in the network. For the 100 sets of simulated data with $K = 3$ roles, $p_c = 0.3$ and $n_{ij} \sim \text{Pois}(25) + 5$, we run the MCMC algorithm with the number of roles K varying from 2 to 4 for each dataset. It turns out that the conditional DIC correctly identifies the optimal K as three for all of the replicates. Moreover, the conditional DIC also performs well in selecting the optimal number of knots and outperforms the Bayesian information criterion (BIC) (Airoldi

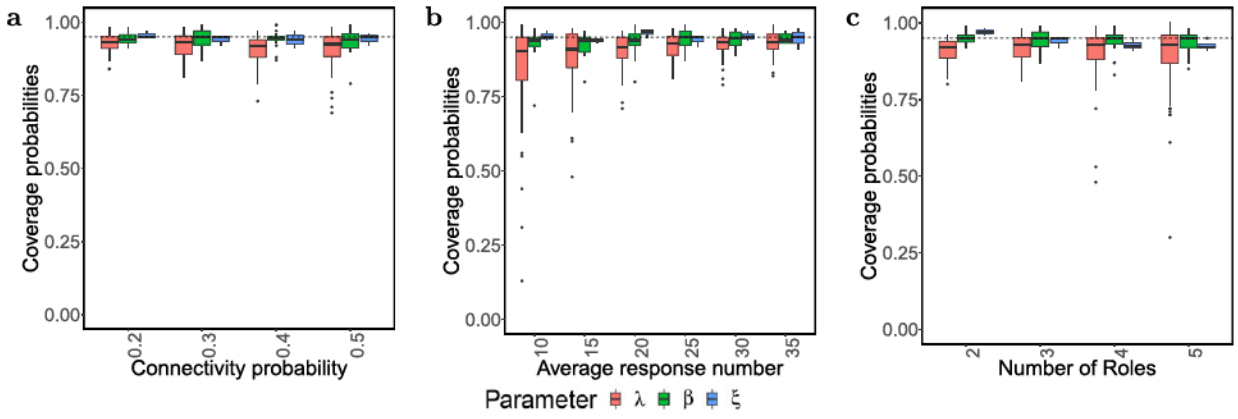

Figure 1. Boxplot of coverage probabilities among 100 synthetic replicated datasets (a) when $K = 3$, $n_{ij} \sim \text{Pois}(25) + 5$ and the connectivity probabilities p_c varies over $(0.2, 0.3, 0.4, 0.5)$; (b) when $K = 3$, $p_c = 0.3$, $n_{ij} \sim \text{Pois}(\mu) + 5$ and μ varies over $(10, 15, 20, 25, 30, 35)$; (c) when $p_c = 0.3$, $n_{ij} \sim \text{Pois}(25) + 5$ and K varies over $(2, 3, 4, 5)$.

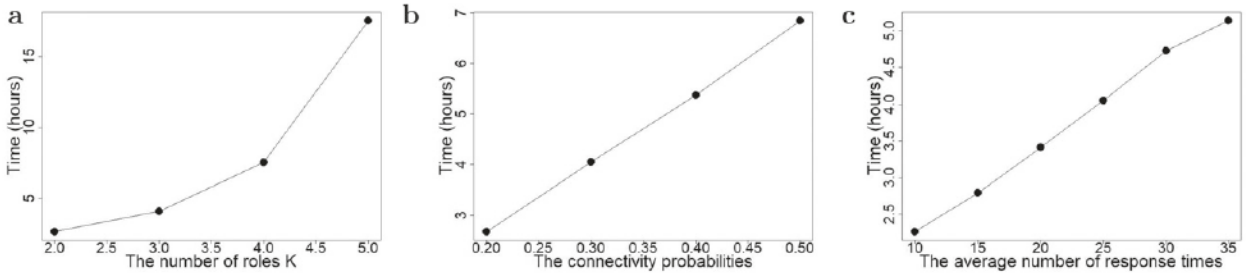

Figure 2. Running time of 100,000 iterations of the MCMC algorithm on one core of an Intel Xeon Gold 6226R Processor. The number of observations per edge $n_{ij} \sim \text{Pois}(\mu) + 5$. (a) We fix the connectivity probability $p_c = 0.3$ and $\mu = 25$ while varying the number of roles K from 2 to 5. (b) We fix $K = 3$ and $\mu = 25$ but vary p_c from 0.2 to 0.5. (c) We fix $K = 3$ and $p_c = 0.3$ but vary μ from 10 to 35.

et al. 2008) and the Watanabe-Akaike information criterion (WAIC) (Watanabe and Opper 2010) (Supplementary Section S13 and Figures S5–S6).

6. The Enron E-mail Corpus

The Enron E-mail corpus (Klimt and Yang 2004) is the largest publicly available E-mail dataset to date and was released by the Federal Energy Regulatory Commission during its investigation of Enron's bankruptcy. The dataset contains E-mails generated by 158 Enron employees between November 13, 1998 and June 21, 2002. Because "It's always about the people. Enron is no different," (Diesner, Frantz, and Carley 2005) the Enron E-mail corpus provides a unique opportunity to study the communication patterns, company organization, and power spread inside a company. The corpus contains the user information and timestamp of each E-mail. Following the *enrondata* GitHub repository, we focus on the E-mail folders of 148 Enron users whose positions in the company are known.

We apply the SMMB to the preprocessed dataset (see Supplementary Section S14 for preprocessing details and Table S48). The dataset contains the response times for 25,629 E-mails, 1886 of which are observed and 23,743 are censored. The Kaplan-Meier estimate of the survival probability after three weeks is 92.36%; thus, a cure rate model is necessary to fit the survival function of the response time. We set the number of knots as five and let $(s_0, s_1, s_2, s_3, s_4, s_5) = (0, 0.152, 0.717, 2.83, 19.82, 504)$ hours such that each interval $(s_{m-1}, s_m], m = 1, 2, 3, 4, 5$ contains the same number of failure times.

6.1. Analysis Without Confidential Information

We first consider the following four covariates: whether the E-mail was sent over the weekend, whether the E-mail was forwarded, the number of recipients of the E-mail, and the number of words in the E-mail. We log-transform the latter two covariates to reduce the impact of skewness. If a series of E-mails had the same subject, we denote the first E-mail as the original E-mail and record the fourth covariate as the number of words in the contents of the original E-mail. The hyperparameters (a, κ, σ^2) for (ξ, Λ, β) are set as $(1, 50, 5)$. We vary the number of roles K from one to five and run 100,000 MCMC iterations with the first 50,000 iterations as burnins for each value of K. The EPSR values shows that the Markov chain has reached convergence after 50,000 iterations (Supplementary Section S9 and Table S7).

The conditional DIC attains its minimum at $K = 2$ (Figure S7a), and the SMMB fits the data well (Supplementary Table S11). For $K = 2$, when we vary the number of knots, the conditional DIC also selects the optimal number of knots as five, thus, $M = 5$ is a reasonable choice for the data analysis (Figure S7b). In Figure 3(a), we plot the probability π_{i1} of belonging to the first role as opposed to the second role. Warmer colors (red and deep orange) represent higher positions (CEO and president). The warmer colored points are concentrated on the left-hand side, thus, preferring to belong to the second role, especially the four CEOs shown as red (Table 2). Thus, we regard the second role as the senior group and the first role as the junior group.

The estimated baseline survival function of each role pair reveals communication patterns between employees (Figure 3(c)). Let us recall that the role pair (l, k) denotes role k replying to the E-mails from role l. Comparing the survival function of the role pair $(1, 2)$ with that of the other three role pairs, the longest response time was that of a senior employee replying to an E-mail from a junior employee. Meanwhile, the response times between junior employees, which correspond to the role pair $(1, 1)$, were the shortest among all of the role pairs. This may be because junior employees often need to communicate via E-mails about routine daily work. As expected, when a senior employee communicated with a junior employee, which corresponds to the role pairs $(1, 2)$ and $(2, 1)$, junior employees replied to the E-mails from senior employees much faster than the reverse situation. Therefore, the level of seniority affected the speed of E-mail response.

An employee can play different roles when communicating with different colleagues. For example, John Zufferli (User 148) held the position of vice president. When he replied to the E-mails from John Lavorato (User 62), a CEO, the estimated role pair $(\hat{Q}_{62,148}, \hat{R}_{62,148}) = (2, 1)$. Meanwhile, when John Lavorato replied to the E-mails from John Zufferli, the estimated role pair became $(\hat{Q}_{148,62}, \hat{R}_{148,62}) = (1, 2)$. Thus, John Zufferli occupied a junior position in the communications with the CEO. However, when John Zufferli communicated with Chris Dorland (User 28), a manager, the estimated role pair $(\hat{Q}_{148,28}, \hat{R}_{148,28})$ and $(\hat{Q}_{28,148}, \hat{R}_{28,148})$ became $(2, 1)$ and $(1, 2)$, respectively. In other words, compared with Chris Dorland, John Zufferli held a higher position. Thus, although the positions of senders and receivers are not incorporated as covariates, the SMMB is able to reveal the seniority of employees within a company according to E-mail response times.

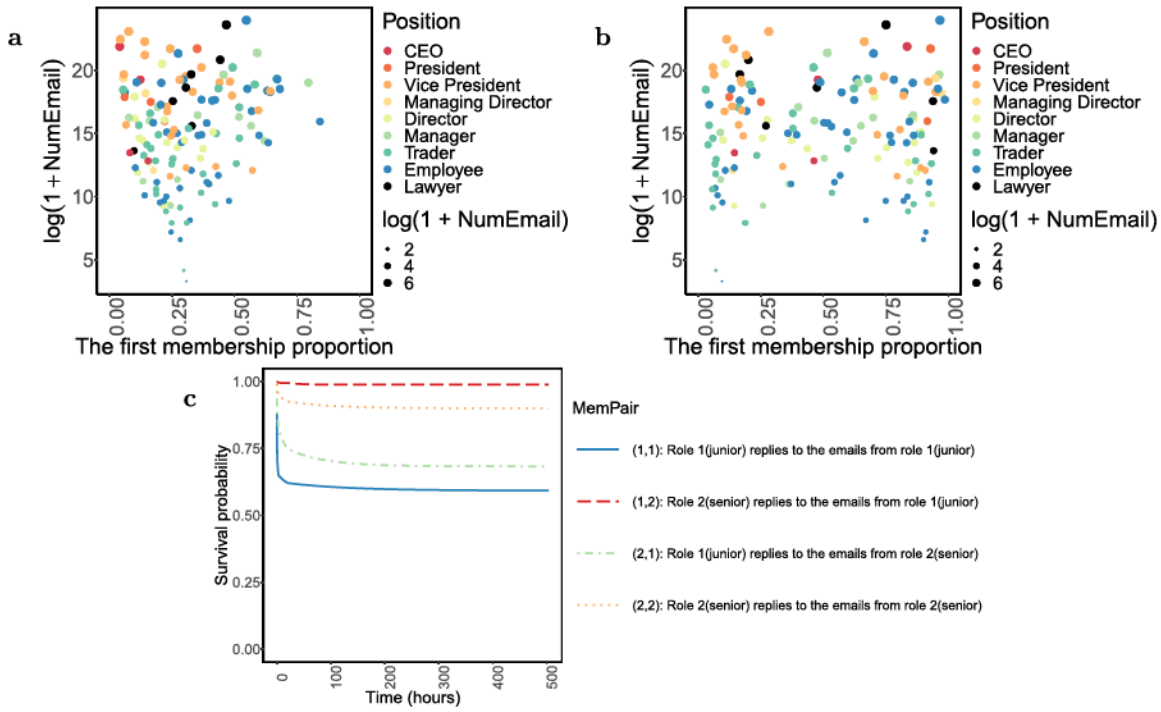

Figure 3. Patterns learned from the Enron E-mail corpus. (a), (b) The scatterplots of the employee-specific probability π_{i1} of belonging to the first role and the log-scale E-mail numbers (a) when the SMMB is applied to time-to-event data, and (b) when the MMSB is applied to relational data. Each node represents an employee, with the color indicating his or her position. The size of each node represents the number of E-mails related to the employee. (c) The estimated baseline survival curve $\exp\{-\exp(\hat{\beta}_1^{lk})[1-S_0(t|\hat{\lambda}^{lk})]\}$ of role l replying E-mails to role k when $x = (1, 0, 0, \ldots, 0)^T$.

Table 2. The estimated role probabilities π_i by the SMMB from time-to-event data and by the MMSB from binary data of the four CEOs, respectively.

User ID	User Name	SMMB		MMSB	
		$\hat{\pi}_{i1}$	$\hat{\pi}_{i2}$	$\hat{\pi}_{i1}$	$\hat{\pi}_{i2}$
23	David Delainey	0.049	0.951	0.478	0.522
62	John Lavorato	0.011	0.989	0.835	0.165
63	Kenneth Lay	0.208	0.792	0.463	0.537
117	Jeffrey Skilling	0.014	0.986	0.145	0.855

Table 3. The posterior mean, posterior standard deviation (SD), and 95% credible interval (CI) of coefficient β_r^{lk}'s for the SMMB learned from the Enron E-mail corpus without considering confidential information.

Covariates	Parameters	Posterior mean	Posterior SD	95% CI
Weekend effect	β_2^{11}	−2.6923*	1.1340	[−5.4631, −0.7872]
	β_2^{12}	−0.9623	1.6866	[−4.6645, 1.6364]
	β_2^{21}	0.1191	0.1813	[−0.2510, 0.4611]
	β_2^{22}	0.2810	0.2051	[−0.1370, 0.6667]
Forwarded effect	β_3^{11}	−1.6664*	0.2363	[−2.1566, −1.2218]
	β_3^{12}	−0.9986*	0.5689	[−2.2854, −0.0571]
	β_3^{21}	−0.6419*	0.1321	[−0.9028, −0.3852]
	β_3^{22}	−0.4578*	0.0960	[−0.6499, −0.2718]
Receiver number effect	β_4^{11}	−1.2633*	0.1600	[−1.5760, −0.9518]
	β_4^{12}	−1.4220*	0.4894	[−2.4875, −0.5576]
	β_4^{21}	−0.7037*	0.0917	[−0.8965, −0.5373]
	β_4^{22}	−0.8923*	0.0822	[−1.0588, −0.7376]
Word count effect	β_5^{11}	−0.1086*	0.0477	[−0.2033, −0.0140]
	β_5^{12}	0.1578	0.1532	[−0.1421, 0.4717]
	β_5^{21}	0.0256	0.0399	[−0.0517, 0.1062]
	β_5^{22}	−0.0113	0.0334	[−0.0781, 0.0504]

NOTE: The superscript * denotes that the corresponding 95% credible interval does not contain zero.

With regard to the weekend effect (Table 3), among all role pairs, only the 95% credible interval [−5.4631, −0.7872] for the coefficient β_2^{11} do not cover zero. The negativity of β_2^{11} indicates that at weekends, junior employees replied to the E-mails sent by other junior employees more slowly, which is unsurprising, as most junior employees are likely to rest over the weekend. The coefficients for the forwarded E-mails are negative across all of the role pairs, indicating that the response times for forwarded E-mails were often longer (Table 3). Moreover, the coefficients for the number of recipients are also negative across role pairs. Thus, an E-mail with more recipients was likely to receive longer response times or was never replied, which is consistent with our experience that people are less likely to respond to circulating E-mails. The effects of E-mail length are only negative for the role pair (1, 1). In other words, for E-mail communications between junior employees, longer E-mails were usually replied to more slowly. The most likely reason for this is that junior employees use E-mails to discuss routine daily work, and a longer E-mail indicates that the task may be more complicated and require more time to handle. Consequently, the SMMB is able to uncover heterogeneous effects across different role pairs.

For comparison, we apply the MMSB (Airoldi et al. 2008) to the binary social network W by running the C# program provided by Burnap et al. (2015). The estimated employee-specific community proportions fail to reflect the leadership among users (Figure 3(b)). For example, the four CEOs had very distinct community proportions (Table 2); therefore, compared with the relational data, the time-to-event data contain much more information about the leadership and power spread of

a company. The true position information of these employees shows that the estimated employee-specific role proportions $\pi_i, i \in \mathcal{N}$ learned from the response times are much more indicative of the leadership among the employees than those estimated from E-mail counts (Figure 3(a)).

We also apply the expSBM (Rastelli and Fop 2020) by regarding E-mail responses as the interactions in the setting of expSBM. expSBM identifies $K = 6$ clusters (Table S52). However, the distribution of employees in different groups is very uneven: 127 of 148 employees are clustered into group 2 whereas group 3 contains only one person, the CEO John Lavorato. Moreover, the patterns learned by expSBM fail to characterize the response rates (Tables S53–S54 and Supplementary Section S16).

6.2. Analysis with Confidential Information

Although we usually have no access to E-mail contents or the genders of senders and receivers because of privacy and potential legal issues, we have the unique opportunity to identify each employee and retrieve the E-mail contents for the Enron E-mail corpus. Therefore, we first conduct a sentiment analysis for the E-mail contents. The Loughran and McDonald lexicon (Loughran and McDonald 2011) categorizes words into classes of constraining, litigious, negative, positive, superfluous, and uncertainty (Table S49). For each E-mail, we count the number of words belonging to each category, log-transform them, and add them as six extra covariates to the model. To incorporate gender into the model, we take the male-male actor pair—both the sender and the recipient being male—as the baseline—and create three dummy variables: male–female, female–male, and female–female.

We then refit the SMMB model with 13 covariates. We run 500,000 iterations of the MCMC algorithm, regard the first 250,000 iterations as burnins, and set the variance of the normal prior distribution of β as 1. The role proportions (Figure S8) and the baseline survival functions of all role pairs (Figure S9) have similar patterns before and after we incorporate sentiment and gender covariates.

For the four covariates analyzed in Section 6.1, the direction and significance of their effects stay the same in the new analysis except that the 95% credible interval of the weekend effect for the role pair $(1, 2)$, β_2^{12}, no longer covers zero (Table S50). The positivity of β_2^{12} indicates that at weekends junior employees replied to the E-mails sent by the senior employees even faster. All of the sentiment covariates' credible intervals cover zero, so they have limited effects on response behavior. Meanwhile, the effects of gender are heterogeneous across different role pairs. As β_{12}^{11} and β_{13}^{11} are positive, when two junior employees communicated with each other, they tended to respond faster to E-mails sent by colleagues of the other gender than to those sent by colleagues of the same gender. Moreover, when the sender played a junior role and the receiver played a senior role, male senders were responded faster than female senders as β_{12}^{12} is significantly greater than 0 but β_{13}^{12} and β_{14}^{12} are significantly smaller than 0. In contrast, when senior employees communicated with each other, female senders were responded faster than male senders. Therefore, the analysis with the sentiment and gender covariates again confirms that by allowing different role pairs to have different survival distributions the SMMB model provides novel insights into the company organization.

6.3. Analysis Considering Sending Behavior

To explore whether recent E-mails speed up responses and increase response rates, we introduce a new covariate indicating whether actor i had received an E-mail from actor j within the previous week into the SMMB model in Section 6.1. Table S51 shows that the effects of recent communications are significant and positive for all of the role pairs. However, the patterns of role proportions and baseline survival functions become quite different from those in Section 6.1 (Figures S10–S11). The difference is because the covariate of recent communications encodes information of the frequency of communications, which implies the connectivity of the company but not the leadership relationship. Therefore, to reveal the leadership patterns within the company, we recommend the SMMB model in Section 6.1. Nevertheless, if one is interested in building a prediction model for response times and response rates, we suggest adding covariates related to sending behavior to the SMMB (see Supplementary Section S15 for details).

7. Discussion

In this article, we propose the SMMB to analyze time-to-event data between actor pairs in a social network. In the SMMB, we assume that actor pairs belonging to the same role pairs share the same SCRM, whereas actor pairs belonging to different role pairs have distinct cure rate models. Thus, while keeping the heterogeneity in the response patterns between different role pairs, the SMMB enables us to borrow information across actor pairs for a given role pair. We prove the model identifiability and posterior consistency of the SMMB. We develop an efficient MCMC algorithm for statistical inference.

In our analysis of the Enron E-mail corpus, we did not consider any time-dependent covariates, as they are difficult to construct given the data we have. Nevertheless, for any external time-dependent covariates, we can directly add them into the SMMB (Kalbfleisch and Prentice 2011). In contrast, internal time-dependent covariates may influence the rate of failures, so additional modeling will be needed, which we will investigate in the future.

Thus, far, for a pair of actors, conditional on their roles $(Q_{ij}, R_{ij}) = (l, k)$, we assume that the sequence of response times is generated independently. When multiple E-mails are about the same topic, there might be dependence between these E-mails. To model dependence, we can further add multiplicative random effects into hazard functions as classic frailty models (Duchateau and Janssen 2008) do (see Supplementary Section S17 for details). As we only have 25,629 E-mails for the Enron E-mail corpus, of which 1886 are observed and the rest are censored, we do not fit these frailty models. Nevertheless, these models might be of interest to large social media and electronic document management systems companies.

With the rapid development of social media and electronic commerce, many social networks now encompass time-to-event data. Although we focus on an E-mail network, many other

social networks have similar time-to-event data. For example, the times taken for users to repost messages from other users are available for Twitter. Moreover, in Web-based electronic document management systems such as the services provided by Dropbox and ParaDM, the time that a team member spends on an assigned task can be recorded, and how soon a user responds to the action of another user is also of great interest in this context. We envision that the proposed SMMB can help to analyze these networks and provide insights into information flow, company organization, and working efficiency.

Supplementary Materials

The supplementary materials provide technical details, figures and tables referred in the main text, including the detailed proof of model identifiability and posterior consistency, the derivation of the posterior inference, and more detailed results of the simulation studies, sensitivity analysis, and real applications. We also provide the source code for generating the figures and tables.

Acknowledgments

We sincerely thank the Editor, Associate Editor, and three reviewers for their helpful comments that have greatly improved the quality of the article.

Disclosure Statement

The authors report there are no competing interests to declare.

Funding

The authors gratefully acknowledge the funding support by the Hong Kong PhD Fellowship PF15-17417, GRF 14305318, GRF 14305319, and GRF 14306020 from the Research Grants Council of Hong Kong SAR, China, NSF 2209685 from the National Science Foundation of United States, the Youth Fund No. 12201533 from the National Natural Science Foundation of China, and Shenzhen Stability Support Program from the Shenzhen Science and Technology Innovation Committee.

ORCID

Fangda Song http://orcid.org/0000-0001-6007-3517
Shuangge Ma http://orcid.org/0000-0001-9001-4999
Yingying Wei http://orcid.org/0000-0003-3826-336X

References

Airoldi, E. M., Blei, D. M., Fienberg, S. E., and Xing, E. P. (2008), "Mixed Membership Stochastic Blockmodels," *Journal of Machine Learning Research*, 9, 1981–2014. [1648,1650,1653,1654]

Burnap, A., Cruz, E., Rong, X., and Segal, B. (2015), "An Implementation of Mixed Membership Stochastic Blockmodel," available at *https://github.com/aburnap/Mixed-Membership-Stochastic-Blockmodel*. [1654]

Celeux, G., Forbes, F., Robert, C. P., and Titterington, D. M. (2006), "Deviance Information Criteria for Missing Data Models," *Bayesian Analysis*, 1, 651–673. [1651]

Chen, M.-H., Ibrahim, J. G., and Sinha, D. (1999), "A New Bayesian Model for Survival Data with a Surviving Fraction," *Journal of the American Statistical Association*, 94, 909–919. [1648,1649]

Diesner, J., Frantz, T. L., and Carley, K. M. (2005), "Communication Networks from the Enron Email Corpus: It's Always about the People. Enron is no Different," *Computational & Mathematical Organization Theory*, 11, 201–228. [1653]

Duchateau, L., and Janssen, P. (2008), *The Frailty Model*, New York: Springer. [1655]

Erdős, P., and Rényi, A. (1960), "On the Evolution of Random Graphs," *Mathematical Institute of the Hungarian Academy of Sciences*, 5, 17–60. [1652]

Fox, E. W., Short, M. B., Schoenberg, F. P., Coronges, K. D., and Bertozzi, A. L. (2016), "Modeling E-mail Networks and Inferring Leadership using Self-Exciting Point Processes," *Journal of the American Statistical Association*, 111, 564–584. [1647]

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., and Rubin, D. B. (2013), *Bayesian Data Analysis*, Boca Raton ,FL: CRC Press. [1652]

Ghosal, S., and Van Der Vaart, A. (2007), "Convergence Rates of Posterior Distributions for Noniid Observations," *The Annals of Statistics*, 35, 192–223. [1650]

Ghosal, S., and Van der Vaart, A. (2017), *Fundamentals of Nonparametric Bayesian Inference* (Vol. 44), Cambridge: Cambridge University Press. [1650]

Ibrahim, J. G., Chen, M.-H., and Sinha, D. (2001a), *Bayesian Survival Analysis* (Vol. 2), New York: Springer. [1648]

——— (2001b), "Criterion-based Methods for Bayesian Model Assessment," *Statistica Sinica*, 11, 419–443. [1652]

Kalbfleisch, J. D., and Prentice, R. L. (2011), *The Statistical Analysis of Failure Time Data*, Hoboken: Wiley. [1655]

Kaplan, E. L., and Meier, P. (1958), "Nonparametric Estimation from Incomplete Observations," *Journal of the American Statistical Association*, 53, 457–481. [1650]

Kaufmanna, E., Bonaldb, T., and Lelargec, M. (2018), "A Spectral Algorithm with Additive Clustering for the Recovery of Overlapping Communities in Networks," *Theoretical Computer Science*, 742, 3–26. [1648]

Klimt, B., and Yang, Y. (2004), "The Enron Corpus: A New Dataset for Email Classification Research," in *Machine Learning: ECML 2004*, eds. J. F. Boulicaut, F. Esposito, F. Giannotti, and D. Pedreschi, pp. 217–226, Berlin: Springer. [1653]

Liu, J. S. (1994), "The Collapsed Gibbs Sampler in Bayesian Computations with Applications to a Gene Regulation Problem," *Journal of the American Statistical Association*, 89, 958–966. [1651]

Loughran, T., and McDonald, B. (2011), "When is a Liability not a Liability? Textual Analysis, Dictionaries, and 10-ks," *The Journal of Finance*, 66, 35–65. [1655]

Lu, Z., and Song, X. (2012), "Finite Mixture Varying Coefficient Models for Analyzing Longitudinal Heterogenous Data," *Statistics in Medicine*, 31, 544–560. [1651]

Mao, X., Sarkar, P., and Chakrabarti, D. (2017), "On Mixed Memberships and Symmetric Nonnegative Matrix Factorizations," in *International Conference on Machine Learning*, pp. 2324–2333. [1649]

Mao, X., Sarkar, P., and Chakrabarti, D. (2021), "Estimating Mixed Memberships with Sharp Eigenvector Deviations," *Journal of the American Statistical Association*, 116, 1928–1940. [1648,1649,1650]

Matias, C., Rebafka, T., and Villers, F. (2018), "A Semiparametric Extension of the Stochastic Block Model for Longitudinal Networks," *Biometrika*, 105, 665–680. [1647]

Perry, P. O., and Wolfe, P. J. (2013), "Point Process Modelling for Directed Interaction Networks," *Journal of the Royal Statistical Society*, Series B, 75, 821–849. [1647]

Rastelli, R., and Fop, M. (2020), "A Stochastic Block Model for Interaction Lengths," *Advances in Data Analysis and Classification*, 14, 485–512. [1648,1655]

Robbins, H. (1955), *An Empirical Bayes Approach to Statistics*. Office of Scientific Research, US Air Force. [1650]

Sit, T., Ying, Z., and Yu, Y. (2021), "Event History Analysis of Dynamic Networks," *Biometrika*, 108, 223–230. [1647]

Wang, Y. J., and Wong, G. Y. (1987), "Stochastic Blockmodels for Directed Graphs," *Journal of the American Statistical Association*, 82, 8–19. [1647,1648,1649,1652]

Watanabe, S., and Opper, M. (2010), "Asymptotic Equivalence of Bayes Cross Validation and Widely Applicable Information Criterion in Singular Learning Theory," *Journal of Machine Learning Research*, 11, 3571–3594. [1653]

Zhang, J., Cai, B., Zhu, X., Wang, H., Xu, G., and Guan, Y. (2022), "Learning Human Activity Patterns Using Clustered Point Processes with Active and Inactive States," *Journal of Business & Economic Statistics*, 41, 388–398. [1647]

Zhang, Y., Levina, E., and Zhu, J. (2020), "Detecting Overlapping Communities in Networks Using Spectral Methods," *SIAM Journal on Mathematics of Data Science*, 2, 265–283. [1648,1649]