A Network Model for Dynamic Textual Communications with Application to Government Email Corpora

Bomin Kim[1], Aaron Schein[3], Bruce Desmarais[1], and Hanna Wallach[2,3]

[1]Pennsylvania State University
[2]Microsoft Research NYC
[3]University of Massachusetts Amherst

June 27, 2017

Abstract

In this paper, we introduce the interaction-partitioned topic model (IPTM)—a probabilistic model of who communicates with whom about what, and when. Broadly speaking, the IPTM partitions time-stamped textual communications, such as emails, according to both the network dynamics that they reflect and their content. To define the IPTM, we integrate a dynamic version of the exponential random graph model—a generative model for ties that tend toward structural features such as triangles—and latent Dirichlet allocation—a generative model for topic-based content. The IPTM assigns each topic to an "interaction pattern"—a generative process for ties that is governed by a set of dynamic network features. Each communication is then modeled as a mixture of topics and their corresponding interaction patterns. We use the IPTM to analyze emails sent between department managers in two county governments in North Carolina; one of these email corpora covers the Outer Banks during the time period surrounding Hurricane Sandy. Via this application, we demonstrate that the IPTM is effective at predicting and explaining continuous-time textual communications.

1 Introduction

In recent decades, real-time digitized textual communication has developed into a ubiquitous form of social and professional interaction (see, e.g., Kanungo and Jain, 2008; Szóstek, 2011; Burgess et al., 2004; Pew, 2016). From the perspective of the computational social scientist, this has lead to a growing need for methods of modeling interactions that manifest as text exchanged in continuous time (e.g., e-mail messages). A number of models that build upon topic modeling through Latent Dirichlet Allocation (Blei et al., 2003) to incorporate link data as well as textual content have been developed recently (McCallum et al., 2005; Lim et al., 2013; Krafft et al., 2012). These models are innovative in their extensions that incorporate network tie information. However, none of the models that are currently available in the literature integrate the rich random-graph structure offered by state of the art models for network structure—in particular, the exponential random graph model (ERGM) (Robins et al., 2007; Chatterjee et al., 2013; Hunter et al., 2008). The ERGM is the canonical model for network structure, as it is flexible enough to specify a generative model that accounts for nearly any pattern of tie formation (e.g., tie reciprocation, clustering, popularity effects) (Desmarais and Cranmer, 2017). We build upon recent extensions of ERGM that model time-stamped ties (Perry and Wolfe, 2013; Butts, 2008), and develop the interaction-partitioned topic model (IPTM) to simultaneously model the network structural patterns that govern tie formation, and the content in the communications.

ERGM, and models based on ERGM, provide a framework for explaining or predicting ties between nodes using the network sub-structures in which the two nodes are embedded (e.g., an ERGM specification may predict ties between two nodes that have many shared partners). ERGM-style models have been used for many applications in which the ties between nodes are annotated with

text. The text, despite providing rich information regarding the strength, scope, and character of the ties, has been largely excluded from these analyses, due to the inability of ERGM-style models to incorporate textual attributes of ties. These application domains include, among other applicaitons, the study of legislative networks in which networks reflect legislators' co-support of bills, but exclude bill text (Bratton and Rouse, 2011; Alemán and Calvo, 2013); the study of alliance networks in which networks reflect countries' co-signing of treaties, but exclude treaty text (Camber Warren, 2010; Cranmer et al., 2012b,a; Kinne, 2016); the study of scientific co-authorship networks that exclude the text of the co-authored papers (Kronegger et al., 2011; Liang, 2015; Fahmy and Young, 2016); and the study of text-based interaction on social media (e.g., users tied via 'mentions' on twitter) (Yoon and Park, 2014; Peng et al., 2016; Lai et al., 2017).

In defining and testing the IPTM we embed three core conceptual properties, in addition to modeling both text and network structure. First, we link the content component of the model, and network component of the model such that knowing who is communicating with whom at what time (i.e., the network component) provides information about the content of communication, and vice versa. Second, we fully specify the network dynamic component of the model such that, given the content of the communication and the history of tie formation, we can draw an exact, continuous-time prediction of when, by whom, and to whom the communication will be sent. Third, we formulate the network dynamic component of the model such that the model can represent, and be used to test hypotheses regarding, canonical processes relevant to network theory such as preferential attachment—the tendency for actors to prefer interacting with actors who have been popular in the past (Barabási and Albert, 1999; Vázquez, 2003; Jeong et al., 2003), reciprocity (Hammer, 1985; Rao and Bandyopadhyay, 1987), and transitivity—the tendency for the friends of friends to become friends (Louch, 2000; Burda et al., 2004). In what follows we (1) present the generative process for the IPTM, describing how it meets our theoretical criteria, (2) derive the sampling equations for Bayesian inference with the IPTM, and (3) illustrate the IPTM through application to email corpora of internal communications by county officials in North Carolina county governments. [**What predictive comparisons should we run to other models**]?

2 IPTM: Model Definition and Derivation

To define and derive the IPTM, we begin by describing a probabilistic process by which documents are generated, where documents include a sender, recipients, contents, and timing. We provide a fully parametric definition of each component of the generative process, which enables the model to be used to simulate distributions of who communicates with whom about what, and when. We take a Bayesian approach to inference for the parameters of the IPTM. In the next section, we derive equations for sampling from the posterior distributions of the IPTM parameters conditional on data generated by the generative process that we define in the current section.

The data generated under the IPTM consists of D unique documents. A single email, indexed by $d \in \{1, ..., D\}$, is represented by the four components $(i^{(d)}, J^{(d)}, t^{(d)}, \boldsymbol{w}^{(d)})$. The first two are the sender and recipients of the email: an integer $i^{(d)} \in \{1, ..., A\}$ indicates the identity of the sender out of A actors (or nodes) and an integer vector $J^{(d)} = \{j_r^{(d)}\}_{r=1}^{|J^{(d)}|}$, which indicates the identity of the receiver (or receivers) out of $A-1$ actors, where $|J^{(d)}| \in \{1, ..., A-1\}$ denotes the total number of receivers. Next, $t^{(d)}$ is the timestamp of the email d. Lastly, $\boldsymbol{w}^{(d)} = \{w_n^{(d)}\}_{n=1}^{N^{(d)}}$ is a set of tokens, or word type instances, that comprise the text of the email, where $N^{(d)}$ denotes the total number of words in a document.

In this section, we illustrate how the words $\boldsymbol{w}^{(d)}$ are generated according to latent Dirichlet allocation (Blei et al., 2003), and then how the other components, $(i^{(d)}, J^{(d)}, t^{(d)})$, are generated conditional on the document content. For simplicity, we assume that documents are ordered by time such that $t^{(d)} < t^{(d+1)}$ for all $d = 1, ..., D$.

2.1 Content Generating Process

The content generating process follows from the generative process of Latent Dirichlet Allocation Blei et al. (2003). First we generate the global (corpus-wide) variables. Each topic k is associated with a cluster, or interaction pattern, assignment c_k, where c_k can take one of $c = \{1, 2, ..., C\}$ values. There are two main sets of global variables—those that describe the content via topics and those that describe how people interact (interaction patterns). These variables are linked by a third set of variables that associate each topic with the pattern that best describes how people interact when talking about that topic.

There are K topics. Each topic k is a discrete distribution over V word types.

1. $\boldsymbol{\phi}^{(k)} \sim \text{Dirichlet}(\beta, \mathbf{u})$ **[See Algorithm 1]**
   - A topic k is characterized by a discrete distribution over V word types with probability vector $\phi^{(k)}$. We specify a symmetric Dirichlet prior $\mathbf{u}$ with the concentration parameter β for the probability vector $\phi^{(k)}$.

There are C interaction patterns. Each interaction pattern consists of a vector of coefficients $\boldsymbol{b}^{(c)}$ in $\boldsymbol{R}^P$ and a vector of P-dimensional dynamic network statistics for directed edge (i, j) at time t $\boldsymbol{x}_t^{(c)}(i, j)$. The inner product of $\boldsymbol{b}^{(c)}$ and $\boldsymbol{x}_t^{(c)}(i, j)$ is used to generate both the recipient vector for a document and the timing of the document.

2. $\boldsymbol{b}^{(c)} \sim \text{Multivariate Normal}(\mu_{\boldsymbol{b}}, \Sigma_{\boldsymbol{b}})$ **[See Algorithm 2]**:
   - The vector of coefficients depends on the interaction pattern c. This means that there is variation across interaction patterns in the degree to which document timing and recipients depend upon the dynamic network statistics. The prior for $\boldsymbol{b}^{(c)}$ is a P-variate multivariate Normal with mean vector $\mu_{\boldsymbol{b}}$ and covariance matrix $\Sigma_{\boldsymbol{b}}$.

The topics and interaction patterns are tied together via a set of K categorical variables.

3. $c_k \sim \text{Uniform}(1, C)$ **[See Algorithm 3]**:
   - Each topic is associated with a single interaction pattern, and topics under same interaction pattern share the network properties via $\boldsymbol{b}^{(c)}$.

We have now defined all of the variables that make up the generative process of the IPTM. We assume the following generative process for each document d in a corpus D **[See Algorithm 4]**:

4-1. Choose the number of words $\bar{N}^{(d)} = \max(1, N^{(d)})$, where $N^{(d)}$ is known.

4-2. Choose document-topic distribution $\boldsymbol{\theta}^{(d)} \sim \text{Dir}(\alpha, \boldsymbol{m})$

4-3. For $n = 1$ to $\bar{N}^{(d)}$:

   (a) Choose a topic $z_n^{(d)} \sim \text{Multinomial}(\boldsymbol{\theta}^{(d)})$

   (b) if $N^{(d)} > 0$, choose a word $w_n^{(d)} \sim \text{Multinomial}(\phi^{(z_n^{(d)})})$

2.2 Stochastic Intensity

In this section, we illustrate how a set of dynamic network features and topic-interaction assignments jointly identify the stochastic intensity of a document, which plays a key role in the tie generating process in Section 2.4. Assume that each document $d \in \{1, ..., D\}$ is associated with an $A \times A$ stochastic intensity matrix $\boldsymbol{\lambda}^{(d)}(t)$, where the $(i, j)^{th}$ element $\lambda_{ij}^{(d)}(t)$ can be interpreted as the likelihood of document d being sent from node i to node j at time t.

First, content of a document is reflected to the stochastic intensity via the distribution of interaction patterns, $\{p_c^{(d)}\}_{c=1}^C$. To calculate the distribution of interaction patterns within a document, we estimate the proportion of words in document d which are assigned the topics corresponding to the

interaction pattern c from Section 2.1:

$$p_c^{(d)} = \frac{\sum_{k:c_k=c} N^{(k|d)}}{N^{(d)}}, \tag{1}$$

where $N^{(k|d)}$ is the number of times topic k appears in the document d and $N^{(d)}$ is the total number of words, as defined earlier. By definition, $\sum_{c=1}^{C} p_c^{(d)} = 1$.

Now, we define the $(i,j)^{th}$ element of the stochastic intensity matrix $\boldsymbol{\lambda}^{(d)}(t)$ in the form of continuous-time ERGM:

$$\lambda_{ij}^{(d)}(t) = \sum_{c=1}^{C} p_c^{(d)} \cdot \exp\Big\{\lambda_0^{(c)} + \boldsymbol{b}^{(c)T}\boldsymbol{x}_t^{(c)}(i,j)\Big\}, \tag{2}$$

where $p_c^{(d)}$ is as defined in Equation (1), $\lambda_0^{(c)}$ is the baseline intensity for the interaction pattern c, $\boldsymbol{b}^{(c)}$ is an unknown vector of coefficients in $\boldsymbol{R}^p$ corresponding to the interaction pattern c, and $\boldsymbol{x}_t^{(c)}(i,j)$ is a vector of the p-dimensional dynamic network statistics for directed edge (i,j) at time t corresponding to the interaction pattern c. Detailed specifications of the dynamic network statsitics are demonstrated in Section 2.3.

2.3 Dynamic Network Statistics

For the network statistics $\boldsymbol{x}_t^{(c)}(i,j)$ of Equation (2), we use 8 different effects as the components of $\boldsymbol{x}_t^{(c)}(i,j)$, (intercept, outdegree, indegree, send, receive, 2-send, 2-receive, sibling, and cosibling) to capture common network properties such as popularity, centrality, reciprocity, and transitivity. Each network statsitic is calculated for each interaction pattern $c = 1, ..., C$, therefore we each interaction pattern can be characterized by its unique set of network statsitics. Below are the specifications of degree, dyadic, and triadic network statistics we use in this paper.

Following Perry and Wolfe (2013), we introduce the covariates that measure higher-order time dependence with the following form. We partition the interval $[-\infty, t)$ into $L = 4$ sub-intervals with equal length in the log-scale, by setting $\Delta_l = (6 \text{ hours}) \times 4^l$ for $l = 1, ..., L-1$ such that Δ_l takes the values 24 hours (=1 day), 96 hours (=4 days), 384 hours (=16 days):

$$\begin{aligned}
[-\infty, t) &= [-\infty, t-\Delta_3) \cup [t-\Delta_3, t-\Delta_2) \cup [t-\Delta_2, t-\Delta_1) \cup [t-\Delta_1, t-\Delta_0) \\
&= [-\infty, t-384h) \cup [t-384h, t-96h) \cup [t-96h, t-24h) \cup [t-24h, t-0) \\
&= I_t^{(4)} \cup I_t^{(3)} \cup I_t^{(2)} \cup I_t^{(1)},
\end{aligned}$$

where $\Delta_0 = 0$ and $I_t^{(l)}$ is the half-open interval $[t-\Delta_l, t-\Delta_{l-1})$.

Based on the preliminary results, we do not include the last interval $I_t^{(4)}$, history before 16 days ago, considering the strong recency effect of document exchange behaviors (e.g. email). Although the specification of these dynamic network covariates could be reformulated based on the objectives of each study, in this paper, we define the degree and dyadic efffects for each $l = 1, ..., L-1$ and $c = 1, ..., C$ as

1. $\textbf{outdegree}_{t,l}^{(c)}(i) = \sum_{d:t^{(d)} \in I_t^{(l)}} p_c^{(d)} \cdot I\{i \to \forall j\}$
2. $\textbf{indegree}_{t,l}^{(c)}(j) = \sum_{d:t^{(d)} \in I_t^{(l)}} p_c^{(d)} \cdot I\{\forall i \to j\}$
3. $\textbf{send}_{t,l}^{(c)}(i,j) = \sum_{d:t^{(d)} \in I_t^{(l)}} p_c^{(d)} \cdot I\{i \to j\}$
4. $\textbf{receive}_{t,l}^{(c)}(i,j) = \sum_{d:t^{(d)} \in I_t^{(l)}} p_c^{(d)} \cdot I\{j \to i\}$

Next, we define four triadic statistics involving pairs of messages, which are analogous to 2-path statistics commonly used in the network science literature. While Perry and Wolfe (2013) adapted full sets of triadic statistics for each combination of time intervals (e.g. $3 \times 3 = 9$), we maintain 3 intervals per each statistic, by defining 3×3 time windows and sum the combination-specific statistics based on the interval where the triads are closed. (Refer to Figure 1.) As a result, our interval-adjusted definition of triadic effects become

5. $\textbf{2-send}^{(c)}_{t,l}(i,j) = \sum_{(l_1=l \text{ or } l_2=l)} \sum_{h \neq i,j} \Big(\sum_{d:t^{(d)} \in I_t^{(l_1)}} p_c^{(d)} \cdot I\{i \to h\} \Big) \cdot \Big(\sum_{d':t^{(d')} \in I_t^{(l_2)}} p_c^{(d')} \cdot I\{h \to j\} \Big)$

6. $\textbf{2-receive}^{(c)}_{t,l}(i,j) = \sum_{(l_1=l \text{ or } l_2=l)} \sum_{h \neq i,j} \Big(\sum_{d:t^{(d)} \in I_t^{(l_1)}} p_c^{(d)} \cdot I\{h \to i\} \Big) \cdot \Big(\sum_{d':t^{(d')} \in I_t^{(l_2)}} p_c^{(d')} \cdot I\{j \to h\} \Big)$

7. $\textbf{sibling}^{(c)}_{t,l}(i,j) = \sum_{(l_1=l \text{ or } l_2=l)} \sum_{h \neq i,j} \Big(\sum_{d:t^{(d)} \in I_t^{(l_1)}} p_c^{(d)} \cdot I\{h \to i\} \Big) \cdot \Big(\sum_{d':t^{(d')} \in I_t^{(l_2)}} p_c^{(d')} \cdot I\{h \to j\} \Big)$

8. $\textbf{cosibling}^{(c)}_{t,l}(i,j) = \sum_{(l_1=l \text{ or } l_2=l)} \sum_{h \neq i,j} \Big(\sum_{d:t^{(d)} \in I_t^{(l_1)}} p_c^{(d)} \cdot I\{i \to h\} \Big) \cdot \Big(\sum_{d':t^{(d')} \in I_t^{(l_2)}} p_c^{(d')} \cdot I\{j \to h\} \Big),$

where $l_1 \in \{1, ..., 3\}$ and $l_2 \in \{1, ..., 3\}$.

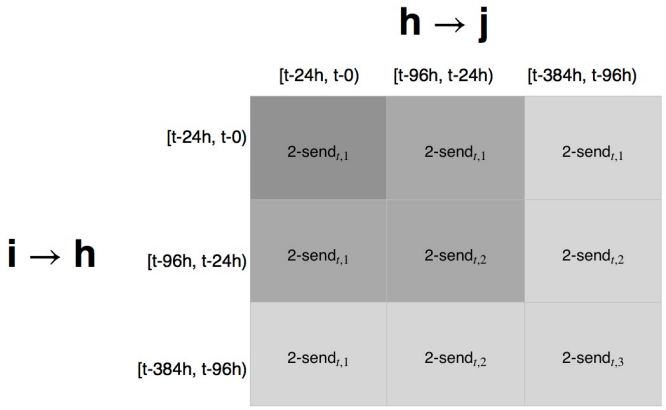

Figure 1: Example of 2-send statistic defined for each interval $l = 1, ..., 3$. Cells with same shades sum up to one statistic, based on when the triads are "closed".

2.4 Tie Generating Process

Given the contents and stochastic intensity, we then move to tie generating process which determines the sender, recipients, and time $(i^{(d)}, J^{(d)}, t^{(d)})$ of the document. We assume the following generative process for each document d in a corpus D:

1. Assume anyone can be the sender of document d, and the receiver/receivers are dependent on the sender, based on "who might be the recipient $J^{(d)}$ if the sender of document d was i". Under this assumption, we use data augmentation scheme and first generate latent ties or sender-recipient pairs.

   For each sender $i \in \{1, ..., A\}$, we create binary receiver vector of length $A-1$, $J_i^{(d)}$, by applying the non-empty Gibbs measure (Fellows and Handcock, 2017) to every $j \in \mathcal{A}_{\setminus i}$, since we exclude self-loop.

$$\mathrm{P}(J_i^{(d)}) = \frac{1}{Z(\delta, \log(\lambda_i^{(d)}))} \exp\Big\{ \log\big(\mathrm{I}(\sum_{j \in \mathcal{A}_{\setminus i}} J_{ij}^{(d)} > 0)\big) + \sum_{j \in \mathcal{A}_{\setminus i}} (\delta + \log(\lambda_{ij}^{(d)})) J_{ij}^{(d)} \Big\}, \qquad (3)$$

where δ is real-valued intercept that controls the overall recipient size or the length of $J_i^{(d)}$, with its prior distribution specified as Normal$(\mu_\delta, \sigma_\delta^2)$. As defined in Section 2.2, $\lambda_{ij}^{(d)}$ is a positive dyad-specific stochastic inensity included in the model, and we use $\lambda_i^{(d)} = \{\lambda_{ij}^{(d)}\}_{j \in \mathcal{A} \setminus i}$ to denote the vector of dyadic weights in which i is the sender. Note that we omitted the notation (t) from Equation (2) and used $\lambda_{ij}^{(d)}$ instead, since the stochastic intensity $\lambda_{ij}^{(d)}$ is always evaluated at time $t_+^{(d-1)}$, implying that λ_{ij} for d^{th} document is obtained using the history of interactions up to and including the time when the previous document was sent, $t^{(d-1)}$.

To assure that the probabilities sum to unity, we use the normalizing constant $Z(\delta, \log(\lambda_i^{(d)}))$, which is the sum of $P(J_i^{(d)})$ over the entire support, and it can be simplified as:

$$Z(\delta, \log(\lambda_i^{(d)})) = \Big(\prod_{j \in \mathcal{A} \setminus i} \Big(\exp\{\delta + \log(\lambda_{ij}^{(d)})\} + 1 \Big) \Big) - 1. \tag{4}$$

Details on how the normalizing constant ends up with this functional form are shown in APPENDIX A.

2. For every sender $i \in \mathcal{A}$, generate the time increments given the latent ties from previous step:

$$\Delta T_{iJ_i} \sim \text{Exponential}(\lambda_{iJ_i}^{(d)}), \tag{5}$$

where the mean parameter $\lambda_{iJ_i}^{(d)}$ is computed by taking the average of network effect terms $\boldsymbol{b}^{(c)T} \boldsymbol{x}_t^{(c)}(i,j)$ across the chosen receivers $J_i^{(d)}$:

$$\lambda_{iJ_i}^{(d)}(t) = \sum_{c=1}^{C} p_c^{(d)} \cdot \exp\Big\{ \lambda_0^{(c)} + \frac{1}{|J_i^{(d)}|} \sum_{j \in J_i} \boldsymbol{b}^{(c)T} \boldsymbol{x}_t^{(c)}(i,j) \Big\}. \tag{6}$$

Note that Equation (6) reduces to the stochastic intensity $\lambda_{ij}^{(d)}$ in Equation (2) in case of single receiver documents (i.e. $|J_i^{(d)}| = 1$), so we can interpret this mean parameter as weighted stochastic intensity across the chosen recievers. When there are multiple chosen receivers (i.e. $|J_i^{(d)}| > 1$), we call it as multicast interactions—those involving a single sender but multiple receivers.

3. Set the observed sender, recipient, and time of the document simultaneously by choosing the sender who generated the minimum time in step 2 and the corresponding receipient and time increment (NOTE: $t^{(0)} = 0$):

$$\begin{aligned} i^{(d)} &= i_{\min(\Delta T_{iJ_i})}, \\ J^{(d)} &= J_{i^{(d)}}, \\ t^{(d)} &= t^{(d-1)} + \min(\Delta T_{iJ_i}). \end{aligned} \tag{7}$$

The intuition behind this choice is that all possible senders $i \in \mathcal{A}$ are competing against each other to send the document to their chosen receivers $\{J_i^{(d)}\}_{i=1}^{A}$, and the one with highest urgency (or highest importance) becomes the observed sender, jointly determining the observed recipient and timestamp of d^{th} document.

2.5 Joint Generative Process of Document

Below are the joint generative process for each document in a corpus D, integrating Section 2.1, Section 2.2, Section 2.3, and Section 2.4.

Algorithm 1 Topic Word Distributions

for *k=1 to K* **do**
  draw $\boldsymbol{\phi}^{(k)} \sim \text{Dirichlet}(\beta, \mathbf{u})$
end

Algorithm 2 Interaction Pattern Parameters

for *c=1 to C* **do**
  draw $\boldsymbol{b}^{(c)} \sim$ Multivariate Normal$(\mu_{\boldsymbol{b}}, \Sigma_{\boldsymbol{b}})$
end

Algorithm 3 Topic Interaction Pattern Assginments

for *k=1 to K* **do**
  draw $C_k \sim$ Uniform(1, C)
end

Algorithm 4 Recipient Size Parameter

draw $\delta \sim$ Normal$(\mu_\delta, \sigma_\delta^2)$

Algorithm 5 Document Generating Process

for *d=1 to D* **do**
  set $\bar{N}^{(d)} = \max(1, N^{(d)})$
  draw $\boldsymbol{\theta}^{(d)} \sim$ Dirichlet$(\alpha, \boldsymbol{m})$
  for *n=1 to* $\bar{N}^{(d)}$ **do**
    draw $z_n^{(d)} \sim$ Multinomial$(\boldsymbol{\theta}^{(d)})$
    if $N^{(d)} > 0$ **then**
      draw $w_n^{(d)} \sim$ Multinomial$(\boldsymbol{\phi}^{(z_n^{(d)})})$
    end
  end
  for *c=1 to C* **do**
    set $p_c^{(d)} = \frac{\sum_{k:c_k=c} N^{(k|d)}}{N^{(d)}}$
  end
  for *i=1 to A* **do**
    for *j=1 to A* **do**
      if $j \neq i$ **then**
        calculate $\boldsymbol{x}^{(c)}_{t_+^{(d-1)}}(i,j)$
        set $\lambda_{ij}^{(d)} = \sum_{c=1}^{C} p_c^{(d)} \cdot \exp\Big\{\lambda_0^{(c)} + \boldsymbol{b}^{(c)T} \boldsymbol{x}^{(c)}_{t_+^{(d-1)}}(i,j)\Big\}$
      end
    end
    draw $J_i^{(d)} \sim$ Gibbs measure$(\{\lambda_{ij}^{(d)}\}_{j=1}^{A}, \delta)$
    draw $\Delta T_{iJ_i} \sim$ Exponential$(\lambda_{iJ_i}^{(d)})$
  end
  set $i^{(d)} = i_{\min(\Delta T_{iJ_i})}$, $J^{(d)} = J_{i^{(d)}}$, and $t^{(d)} = t^{(d-1)} + \min(\Delta T_{iJ_i})$
end

3 Inference

For real-world data, we make inference on the latent variables to find out the values that most likely would have generated the data we observe, under the generative process in Section 2.5. In this section, we derive the joint distribution over the variables $\Phi = \{\boldsymbol{\phi}^{(k)}\}_{k=1}^{K}, \Theta = \{\boldsymbol{\theta}^{(d)}\}_{d=1}^{D}, \mathcal{Z} =$

$\{\boldsymbol{z}^{(d)}\}_{d=1}^{D}, \mathcal{C} = \{c_k\}_{k=1}^{K}, \mathcal{B} = \{\boldsymbol{b}^{(c)}\}_{c=1}^{C}, \delta, \mathcal{J}_{\mathrm{a}} = \{\{J_i^{(d)}\}_{i \neq i_o^{(d)}}\}_{d=1}^{D}$, and $\mathcal{T}_{\mathrm{a}} = \{\{t_{iJ_i}^{(d)}\}_{i \neq i_o^{(d)}}\}_{d=1}^{D}$, and $\mathcal{P} = \{(i, J, t)^{(d)}\}_{d=1}^{D}$ given the observed four components $\mathcal{W} = \{\boldsymbol{w}^{(d)}\}_{d=1}^{D}$, $\mathcal{I}_{\mathrm{o}} = \{i_o^{(d)}\}_{d=1}^{D}$, $\mathcal{J}_{\mathrm{o}} = \{J_o^{(d)}\}_{d=1}^{D}$, and $\mathcal{T}_{\mathrm{o}} = \{t^{(d)}\}_{d=1}^{D}$, and the hyperparameters $(\beta, \boldsymbol{u}, \alpha, \boldsymbol{m}, \mu_{\boldsymbol{b}}, \Sigma_{\boldsymbol{b}}, \mu_\delta, \sigma_\delta^2)$.

After integreting out Φ and Θ using Dirichlet-multinomial conjugacy (Griffiths and Steyvers, 2004) we sample the remaining unobserved variables from their joint posterior distribution using Markov chain Monte Carlo methods. Additionally, we integrate out the latent time-increments $\mathcal{T}_{\mathrm{a}}$ using the property of the minimum of Exponential random variables, as shown in B.1. Now, our inference goal is to draw samples from the posterior distribution

$$\begin{aligned} &P(\mathcal{Z}, \mathcal{C}, \mathcal{B}, \delta, \mathcal{J}_{\mathrm{a}} | \mathcal{W}, \mathcal{I}_{\mathrm{o}}, \mathcal{J}_{\mathrm{o}}, \mathcal{T}_{\mathrm{o}}, \beta, \boldsymbol{u}, \alpha, \boldsymbol{m}, \mu_{\boldsymbol{b}}, \Sigma_{\boldsymbol{b}}, \mu_\delta, \sigma_\delta^2) \\ &\propto P(\mathcal{Z}, \mathcal{C}, \mathcal{B}, \delta, \mathcal{W}, \mathcal{J}_{\mathrm{a}}, \mathcal{I}_{\mathrm{o}}, \mathcal{J}_{\mathrm{o}}, \mathcal{T}_{\mathrm{o}} | \beta, \boldsymbol{u}, \alpha, \boldsymbol{m}, \mu_{\boldsymbol{b}}, \Sigma_{\boldsymbol{b}}, \mu_\delta, \sigma_\delta^2) \\ &= P(\mathcal{Z}|\alpha, \boldsymbol{m}) P(\mathcal{C}) P(\mathcal{B}|\mathcal{C}, \mu_{\boldsymbol{b}}, \Sigma_{\boldsymbol{b}}) P(\delta|\mu_\delta, \sigma_\delta^2) P(\mathcal{W}|\mathcal{Z}, \beta, \boldsymbol{u}) P(\mathcal{J}_{\mathrm{a}}, \mathcal{I}_{\mathrm{o}}, \mathcal{J}_{\mathrm{o}}, \mathcal{T}_{\mathrm{o}} | \mathcal{Z}, \mathcal{C}, \mathcal{B}, \delta), \end{aligned} \tag{8}$$

where all the detailed derivation of sampling equations can be found in B.

To implement the inference procedure outlined above, we provide a pseudocode for Markov Chain Monte Carlo (MCMC) sampling. For better performance, we implement n_1 iteration of the hyperparameter optimization technique called "new fixed-point iterations using the Digamma recurrence relation" in Wallach (2008) based on Minka's fixed-point iteration (Minka, 2000), for every outer iteration o. Also, while we update the categorical variables $\mathcal{Z}$ and $\mathcal{C}$ once per outer iteration, we specify larger number of inner iterations (n_2 and n_3) for the continuous variables $\mathcal{B}$ and δ, which converge slowly than the discrete variables. When summarizing model results, we only use the samples from the last O^{th} outer loop.

Algorithm 6 MCMC

set initial values $\mathcal{Z}^{(0)}, \mathcal{C}^{(0)}$, and $(\mathcal{B}^{(0)}, \delta^{(0)})$
for *o=1 to O* **do**
  for *n=1 to* n_1 **do**
    optimize α and $\boldsymbol{m}$ using hyperparameter optimization in Wallach (2008)
  end
  for *d=1 to D* **do**
    for $i \in \mathcal{A}_{\setminus i_o^{(d)}}$ **do**
      sample the augmented data $J_i^{(d)}$ following Section B.2
    end
    for *n=1 to* $N^{(d)}$ **do**
      draw of $z_n^{(d)} \sim$ Multinomial$(p^{\mathcal{Z}})$ following Section B.3
    end
  end
  for *k=1 to K* **do**
    draw $C_k \sim$ Multinomial$(p^{\mathcal{C}})$ following Section B.4
  end
  for *n=1 to* n_2 **do**
    sample $\mathcal{B}$ using Metropolis-Hastings following Section B.5
  end
  for *n=1 to* n_3 **do**
    sample δ using Metropolis-Hastings following Section B.6
  end
end
Summarize the results with:
last sample of $\mathcal{C}$, last sample of $\mathcal{Z}$, last n_2 length chain of $\mathcal{B}$, last n_3 length chain of δ

4 Getting It Right (GiR)

4.1 Collapsed-time Tie Generating Process

Considering that we integrated out latent time $\mathcal{T}_{\text{a}}$ in the inference, we develop the new generative process for (sender, recipients, timestamp) parts with latent time integrated out. Note that this is built upon the property of the minimum of independent Exponential random variables (the probability ΔT_{iJ_i} being the minimum is $\frac{\lambda_{iJ_i}^{(d)}}{\sum_{i=1}^{A}\lambda_{iJ_i}^{(d)}}$). Details are illustrated in Algorithm 7.

Algorithm 7 Collapsed-time Tie Generating Process

for *d=1 to D* **do**
  for *i=1 to A* **do**
    for *j=1 to A* **do**
      if $j \neq i$ **then**
        calculate $\boldsymbol{x}^{(c)}_{t_+^{(d-1)}}(i,j)$, the network statisitcs evaluated at time $t_+^{(d-1)}$
        set $\lambda_{ij}^{(d)} = \sum_{c=1}^{C} p_c^{(d)} \cdot \exp\Big\{\lambda_0^{(c)} + \boldsymbol{b}^{(c)T}\boldsymbol{x}^{(c)}_{t_+^{(d-1)}}(i,j)\Big\} \cdot 1\{j \in \mathcal{A}_{\setminus i}\}$
      end
    end
    draw $J_i^{(d)} \sim$ Gibbs measure$(\{\lambda_{ij}^{(d)}\}_{j=1}^{A}, \delta)$
    calculate $\lambda_{iJ_i}^{(d)}$
  end
  choose $i^{(d)} \sim$ Multinomial$(\{\frac{\lambda_{iJ_i}^{(d)}}{\sum_{i\in\mathcal{A}}\lambda_{iJ_i}^{(d)}}\}_{i=1}^{A})$
  set $J^{(d)} = J_{i^{(d)}}$
  draw $\Delta T_{i^{(d)}J^{(d)}} \sim$ Exponential$(\sum_{i\in\mathcal{A}}\lambda_{iJ_i}^{(d)})$ and set $t^{(d)} = t^{(d-1)} + \Delta T_{i^{(d)}J^{(d)}}$
end

With this generative process, the joint likelihood (comparable to Equation (12)) becomes:

$$
\begin{aligned}
&P(\mathcal{J}_{\text{a}}^{(d)}, \mathcal{T}_{\text{a}}^{(d)}, i_{\text{o}}^{(d)}, J_{\text{o}}^{(d)}, t_{\text{o}}^{(d)} | \mathcal{I}_{\text{o}}^{(<d)}, \mathcal{J}_{\text{o}}^{(<d)}, \mathcal{T}_{\text{o}}^{(<d)}, \mathcal{Z}, \mathcal{C}, \mathcal{B}, \delta) \\
&= P\Big(\text{latent receivers generation}\Big) \times P\Big(\text{choose the sender}\Big) \times P\Big(\text{observed minimum time generation}\Big) \\
&= \prod_{i\in\mathcal{A}} \Big(J_i^{(d)} \sim \text{Gibbs measure}(\{\lambda_{ij}^{(d)}\}_{j=1}^{A}, \delta)\Big) \times \Big(i_o^{(d)} \sim \text{Multinom}(\{\frac{\lambda_{iJ_i}^{(d)}}{\sum_{i\in\mathcal{A}}\lambda_{iJ_i}^{(d)}}\}_{i=1}^{A})\Big) \times \Big(\Delta T^{(d)}_{i^{(d)}J^{(d)}} \sim \text{Exp}(\sum_{i\in\mathcal{A}}\lambda^{(d)}_{i_o^{(d)}J_o^{(d)}})\Big) \\
&= \Big(\prod_{i\in\mathcal{A}} \frac{1}{Z(\delta, \log(\lambda_i^{(d)})} \exp\Big\{\sum_{j\in\mathcal{A}_{\setminus i}} (\delta + \log(\lambda_{ij}^{(d)}))J_{ij}^{(d)}\Big\}\Big) \times \Big(\frac{\lambda^{(d)}_{i_o^{(d)}J_o^{(d)}}}{\sum_{i\in\mathcal{A}}\lambda_{iJ_i}^{(d)}}\Big) \times \Big((\sum_{i\in\mathcal{A}}\lambda_{iJ_i}^{(d)}) e^{-\Delta T^{(d)}_{i_o^{(d)}J_o^{(d)}} \sum_{i\in\mathcal{A}}\lambda_{iJ_i}^{(d)}}\Big) \\
&= \Big(\prod_{i\in\mathcal{A}} \frac{1}{Z(\delta, \log(\lambda_i^{(d)})} \exp\Big\{\sum_{j\in\mathcal{A}_{\setminus i}} (\delta + \log(\lambda_{ij}^{(d)}))J_{ij}^{(d)}\Big\}\Big) \times \Big(\frac{\lambda^{(d)}_{i_o^{(d)}J_o^{(d)}}}{\sum_{i\in\mathcal{A}}\lambda_{iJ_i}^{(d)}}\Big) \times \Big((\sum_{i\in\mathcal{A}}\lambda_{iJ_i}^{(d)}) e^{-\Delta T^{(d)}_{i_o^{(d)}J_o^{(d)}} \sum_{i\in\mathcal{A}}\lambda_{iJ_i}^{(d)}}\Big) \\
&\propto \Big(\prod_{i\in\mathcal{A}} \frac{1}{\Big(\prod_{j\in\mathcal{A}_{\setminus i}}\Big(\exp\{\delta + \log(\lambda_{ij}^{(d)})\} + 1\Big)\Big) - 1} \exp\Big\{\sum_{j\in\mathcal{A}_{\setminus i}} (\delta + \log(\lambda_{ij}^{(d)}))J_{ij}^{(d)}\Big\}\Big) \\
&\qquad\qquad \times \Big(\lambda^{(d)}_{i_o^{(d)}J_o^{(d)}}\Big) \times \Big(e^{-\Delta T^{(d)}_{i_o^{(d)}J_o^{(d)}} \sum_{i\in\mathcal{A}}\lambda_{iJ_i}^{(d)}}\Big),
\end{aligned}
\tag{9}
$$

which is exactly the same as Equation (14), thus we will use this collpased-time generative process as a forward/backward generative process in Geweke's "Getting it Right" test (Geweke, 2004).

4.2 Backward Generating Process

We need to define a "backward" generative process in order to perform Geweke's "Getting it Right" (GiR) test because when we are generating our "backwards" samples, we only want to resample the token word types given the token-topic assignments (using Collapsed Gibbs sampling) and (sender, recipients, timestamp) pairs, and not any of our latent variables. This means we take the latent variables we got by running our inference procedure (latent edges from data augmentation, token topic assignments, topic interaction pattern assignments, interaction pattern parameters, and receiver size parameter δ) as inputs, and simply condition on these to draw new data.

For backward sampling, we let NKV be a $V \times K$ dimensional matrix where each entry will record the count of the number of tokens of word-type v that are currently assigned to topic k. Also let NK be a K dimensional vector recording the total count of tokens currently assigned to topic k. In addition, here we do not generate latent edges, instead, we use inferred ones from the inference as an input. Only the receivers for previously observed sender ($i_o^{(d)}$), which is not inferred, will be sampled according to the generative process, same as forward sampling (with inferred parameter values). This "backward" version of the generative process is detailed below in Algorithm 8.

Algorithm 8 Generate data with backward sampling

Input:
1) latent edges $\{\{iJ_i^{(d)}\}_{i\neq i_o^{(d)}}\}_{d=1}^D$ (where $\{i_o^{(d)}\}_{d=1}^D$ from previous backward sample),
2) token topic assignments $\{\{z_n^{(d)}\}_{n=1}^{N^{(d)}}\}_{d=1}^D$,
3) topic interaction pattern assignments, $\{C_k\}_{k=1}^K$,
4) interaction pattern parameters $\{\boldsymbol{b}^{(c)}\}_{c=1}^C$,
5) receiver size parameter δ.

for *d=1 to D* **do**
  set $NKV = 0$ and $NK = 0$
  for *n=1 to* $\bar{N}^{(d)}$ **do**
    for *v=1 to V* **do**
      token-word-type-distribution$_n^{(d)}[v] = \frac{NKV_{v,z_n^{(d)}}+\beta \boldsymbol{u}_v}{NK_{z_n^{(d)}}+\beta}$
    end
    draw $w_n^{(d)} \sim$ (token-word-type-distribution$_n^{(d)}$)
    $NKV_{w_n^{(d)},z_n^{(d)}} += 1$
    $NK_{z_n^{(d)}} += 1$
  end
  only for the previously observed sender $i_o^{(d)}$,
  for $j \neq i_o^{(d)}$ **do**
    calculate $\boldsymbol{x}^{(c)}_{t_+^{(d-1)}}(i_o^{(d)}, j)$, the network statisitcs evaluated at time $t_+^{(d-1)}$
    set $\lambda^{(d)}_{i_o^{(d)}j} = \sum_{c=1}^{C} p_c^{(d)} \cdot \exp\Big\{\lambda_0^{(c)} + \boldsymbol{b}^{(c)T}\boldsymbol{x}^{(c)}_{t_+^{(d-1)}}(i_o^{(d)}, j)\Big\} \cdot 1\{j \in \mathcal{A}_{\setminus i_o^{(d)}}\}$
  end
  draw $J^{(d)}_{i_o^{(d)}} \sim$ Gibbs measure$(\{\lambda^{(d)}_{i_o^{(d)}j}\}_{j=1}^A, \delta)$
  choose $i^{(d)} \sim$ Multinomial$(\{\frac{\lambda^{(d)}_{iJ_i}}{\sum_{i\in\mathcal{A}}\lambda^{(d)}_{iJ_i}}\}_{i=1}^A)$
  set $J^{(d)} = J_{i^{(d)}}$
  draw $\Delta T_{i^{(d)}J^{(d)}} \sim$ Exponential$(\sum_{i\in\mathcal{A}} \lambda^{(d)}_{iJ_i})$ and set $t^{(d)} = t^{(d-1)} + \Delta T_{i^{(d)}J^{(d)}}$
end

4.3 Intitialization of History $\boldsymbol{x}_t^{(c)}$

Considering that our network statistics $\boldsymbol{x}_t^{(c)}$ are generated by the network itself, it is necessary to use the same initial value of $\boldsymbol{x}_t^{(c)}$ across the forward and backward samples. If not, when we generate fixed number of documents (e.g. nDocs = 10), we cannot guarantee the same number of documents used for the inference, since only the documents with its timestamp greater than 384 hours (=16 days) are used in the inference. In the extreme cases, we may end up with two types of failure:

1. Zero document generated after 384 hours (i.e. $t^{(10)} < 384$), making no documents to be used for inference,
2. Zero document generated before 384 hours (i.e. $t^{(1)} > 384$), making the estimate of $\mathcal{B}$ totally biased since $\forall \boldsymbol{x}_t^{(c)}(i,j) = 0$.

Therefore, we fix the initial state of $\boldsymbol{x}_t^{(c)}$ over the entire GiR process. Specifically, we fix 30 (for example) baseline documents where the timestamps are all smaller than 384 and use as an input for forward sampling, backward sampling, and the inference. Then, in the forward and backward generative process, we set the starting point of the timestamp as $t^{(0)} = 384$ and generate nDocs =

10 documents given the initial $\boldsymbol{x}^{(c)}_{t^{(0)}=384}$ so that we can acheive consistency in the generated number of documents with $t^{(d)} > 384$.

4.4 GiR Implementation Details

While we tried a number of different parameter combinations in the course of testing, we outline our standard setup. We selected the following parameter values:

- nDocs (number of documents) = 5
- nwords (tokens per document) = 4
- node (number of actors)= 4
- W (unique word types) = 5
- nIP (number of interaction patterns) = 2
- K (number of topics) = 4
- α (Dirichlet concentration prior) = 2
- $\boldsymbol{m}$ (Dirichlet base prior) = $\boldsymbol{u}$
- β (Dirichlet concentration prior)= 2
- $\boldsymbol{n}$ (Dirichlet base prior) = $\boldsymbol{u}$
- netstat = "intercept" and "dyadic"
- prior for $\boldsymbol{b}^{(c)}$: $\mu_{\boldsymbol{b}^{(c)}} = (-3, \mathbf{0}_6), \Sigma_{\boldsymbol{b}^{(c)}} = 0.005 \times I_7$
- prior for δ: $\mu_\delta = 0$, $\sigma^2_\delta = 0.1$
- I (outer iteration) = 3
- n_1 (iteration for hyperparamter optimization) = 0
- n_2 (M-H sampling iteration of $\mathcal{B}$) = 330
- burn (M-H sampling burn-in of $\mathcal{B}$)= 30
- thin (M-H sampling thinning of $\mathcal{B}$)= 3
- σ^2_{Q1} (proposal variance for $\mathcal{B}$) = 0.04
- n_3 (M-H sampling iteration of δ) = 10
- σ^2_{Q2} (proposal variance for δ) = 2

Next, we list the selection of statistics we save for each forward and backward sample. Note that these statistics are not sensitive to the label swithches across the updates. Therefore, at each iteration, we calculate and save the statistics below:

1. Mean of interaction pattern parameters $(\boldsymbol{b}^{(1)}_p, ..., \boldsymbol{b}^{(C)}_p)$ for every $p = 1, ..., P$,
2. Three istory statistic 'send' calculated for the last document
3. δ value used to generate the samples
4. Mean number of recipients,
5. Mean of time-increments $t^{(d)} - t^{(d-1)}$ for every $d = 2, ..., \text{nDocs}$,
6. Mean topic-interaction pattern assignment (for interaction patterns indexed by 1 and 2),
7. Number of tokens in topics assigned to each interaction pattern $c = 1, ..., C$,
8. Number of tokens assigned to each topic $k = 1, ..., K$,
9. Number of tokens assigned to each unique word type $w = 1, ..., W$.

4.5 GiR Results

Having saved a set of samples, we generated PP (Probability-Probability) plots for each of the 25 statistics we saved. We calculated 1,000 quantiles for each of the interaction pattern statistics (1.), and 50 quantiles for the rest of the statistics. Exactly follwoing CPME, we also calculated a t-test p-value for the equivalence of statistic means between forward and backward samples, and a Mann-Whitney test p-value for the equivalence of statistic distributions between forward and backward samples. Before we calculated these statistics, we first thinned our sample of statistics by taking every 90th sample starting at the 100,000th sample for a resulting sample size of 10,000, to reduce the autocorrection in the Markov chain. In each case, if we observe a large p-value, this gives us evidence that the statistics have the same mean and distribution respectively. We included a diagonal line in each plot that we expect these PP dots to line up on if we are passing GiR. The PP-plots are depicted in Figure below.

5 Appliction to North Carolina email data

To see the applicability of the model, we used the North Carolina email data using two counties, Vance county and Dare county, which are the two counties whose email corpus cover the date of Hurricane Sandy (October 22, 2012 – November 2, 2012). Especially, Dare county geographically covers the Outer Banks, so we would like to see how the communication pattern changes during the time period surrounding Hurricane Sandy. Here we apply IPTM to both data and demonstrate the effectiveness of the model at predicting and explaining continuous-time textual communications.

5.1 Vance county email data

Vance county data contains $D = 185$ emails sent between $A = 18$ actors, including $W = 620$ vocabulary in total. We used $K = 5$ topics and $C = 2$ interaction patterns. MCMC sampling was implemented based on the order and scheme illustrated in Section 3. We set the outer iteration number as $I = 500$, the inner iteration numbers as $n_1 = 1, n_2 = 1$, and $n_3 = 3300$. First 50 outer iterations and first 300 iterations of third inner iteration were used as a burn-in, and every 20^{th} sample was taken as a thinning process of third inner iteration. In addition, after some experimentation, σ_Q^2 was set as 0.2, to ensure sufficient acceptance rate. MCMC diagnostic plots are attached in APPENDIX D, as well as the geweke test statistics.

Below are the summary of IP-topic-word assignments. Each interaction pattern is paired with (a) posterior estimates of dynamic network effects $\boldsymbol{b}^{(c)}$ corresponding to the interaction pattern, and (b) the top 10 most likely words to be generated conditioned on the topic and their corresponding interaction pattern. By examining the estimates in Figure 2 and their corresponding interpretation, it

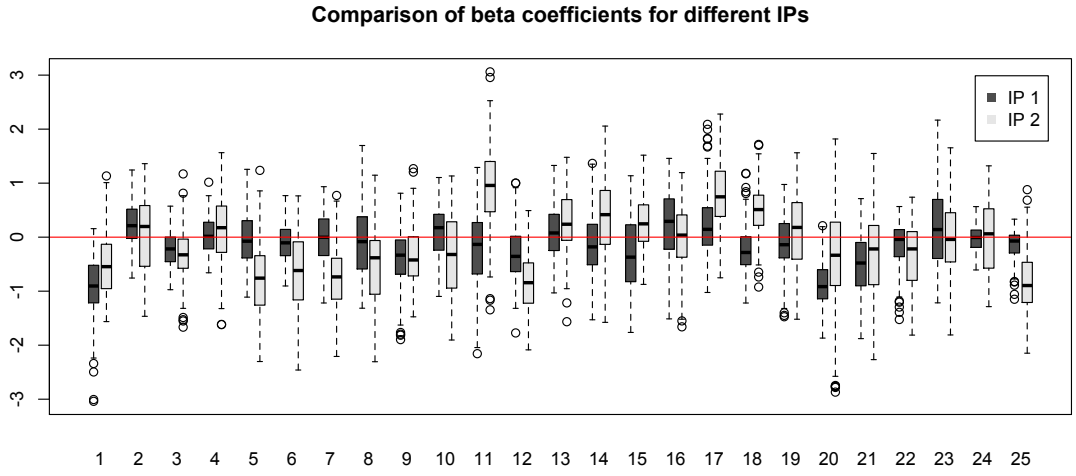

Figure 2: Posterior distribution of $\boldsymbol{b}^{(c)}$ for Vance county emails

covariate	1	2 - 4	5 - 7	8 - 10	11 - 13	14 - 16	17 - 19	20 - 22	23 - 25
name	intercept	outdegree	indegree	send	receive	2-send	2-receive	sibling	cosibling

Table 1: Network statistics

seems that there exist strong effects of dynamic network covariates. That is, whether the sender and receiver previously had dyadic or triangle interaction strongly increase the rate of their interactions. What are the findings here?

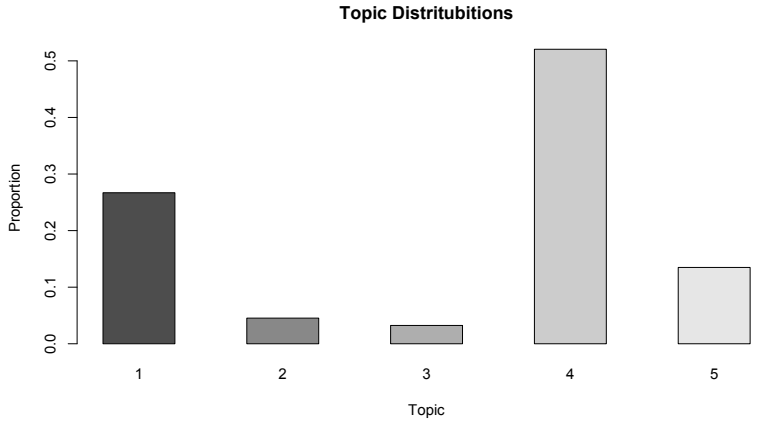

Figure 3: Posterior distribution of $\mathcal{Z}$ for Vance county emails

Next, we scrutinize the topic distributions in Figure 3. There is some distinctive differences in the topic distributions $\mathcal{Z}$, given the assignment of interaction patterns to the documents $\mathcal{C}$. Specifically, each interaction pattern has different topics as the topic with highest probability.
Furthermore, we look at the distribution of words given the topics, which corresponds to Algorithm 4 in the generative process. Since the topic-word distribution ϕ does not depend on the interaction patterns as previous cases, Table 3 lists top 10 topics with top 10 words that have the highest probability conditioned on the topic. In addition, this time we try to check the interaction pattern-word distribution by listing top 10 words that have the highest probability conditioned on the interaction pattern. It seems that the words are not significantly different, having several words like 'director', 'phones', 'department', 'description', or 'henderson' (county seat of Vance county) appeared repetitively across the most of the topics or interaction patterns. The word 'will' was ranked the top in most of the lists, probably because it was not deleted during the text mining process while other similar type of words like 'am', 'is', 'are', or 'can' are all removed.

IP1	**IP2**
K=5, K=2	K=3, K=4, K=1
operations, description	emergency, electronic, will
emergency, phase	operations, message, meeting
henderson, planning	fax, request, director
director, development	henderson, review, phone
street, director	street, response, october
church, henderson	center, records, extension
suite, fax	director, manager, phones
office, church	church, pursuant, latest
center, phone	office, ncgs, directory
communications, email	communications, chapter, attached

Table 2: Summary of top 10 words that have the highest probability conditioned on the topic

5.2 Dare county email data

Dare county data contains $D = 2247$ emails between $A = 27$ actors, including $W = 2907$ vocabulary in total. Again, we used $K = 10$ topics and $C = 3$ interaction patterns. MCMC sampling was implemented based on the order and scheme illustrated earlier. We set the outer iteration number as $I = 1000$, and inner iteration numbers as $n_1 = 3, n_2 = 3$, and $n_3 = 3300$. In addition, after some experimentation, σ_Q^2 was set as 0.02, to ensure sufficient acceptance rate. In our case, the average

acceptance rate for $\boldsymbol{b}$ was 0.277. As demonstrated in Algorithm 5, the last value of $\mathcal{C}$, the last value of $\mathcal{Z}$, and the last n_3 length chain of $\mathcal{B}$ were taken as the final posterior samples. Among the $\mathcal{B}$ samples, 300 were discarded as a burn-in and every 10^{th} samples were taken. After these post-processing, MCMC diagnostic plots are attached in APPENDIX D, as well as geweke test statistics.

6 Posterior predictive experiments

We use a set of posterior predictive experiments to evaluate the performance of the IPTM as compared to alternative modeling approaches, and with respect to alternative parameterizations of the IPTM. For documents $d = \{M, M+1, \ldots, D-1\}$, we fit the IPTM to the first d documents, then use the inferred posterior distributions to generate a distribution of predicted tie data $(i^{(d+1)}, J^{(d+1)}, t^{(d+1)})$ for document $d+1$ conditional on the content in document $d+1$, $(\boldsymbol{w}^{(d+1)})$. A reasonable choice for M would be $D/2$, to assure a sufficient size training set. The variables that need to be sampled are the token topic assignments, $\mathcal{Z}^{d+1}$, and the tie data $(i^{(d+1)}, J^{(d+1)}, t^{(d+1)})$.

Algorithm 9 Predicting tie data for the next document

Input
1. O, number of outer iterations of inference from which to generate predictions
2. d, the last document to use in inference
3. R, the number of iterations to sample predicted data within each outer iteration

Run burnin iterations
for *o=1 to O* **do**
  run an outer iteration of inference on documents 1 through d
  initialize values for $i^{(d+1)}$, $J^{(d+1)}$, $t^{(d+1)}$, and $\mathcal{Z}^{d+1}$
  for *r=1 to R* **do**
    sample $i^{(d+1)}$, $J^{(d+1)}$, and $t^{(d+1)}$ conditional on $\mathcal{Z}^{d+1}$, via the generative process
    sample $\mathcal{Z}^{d+1}$ via Equation 24
  end
  store $i^{(d+1)}$, $J^{(d+1)}$, $t^{(d+1)}$, and $\mathcal{Z}^{d+1}$
end

References

Alemán, E. and Calvo, E. (2013). Explaining policy ties in presidential congresses: A network analysis of bill initiation data. *Political Studies*, 61(2):356–377.

Barabási, A.-L. and Albert, R. (1999). Emergence of scaling in random networks. *science*, 286(5439):509–512.

Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022.

Bratton, K. A. and Rouse, S. M. (2011). Networks in the legislative arena: How group dynamics affect cosponsorship. *Legislative Studies Quarterly*, 36(3):423–460.

Burda, Z., Jurkiewicz, J., and Krzywicki, A. (2004). Network transitivity and matrix models. *Physical Review E*, 69(2):026106.

Burgess, A., Jackson, T., and Edwards, J. (2004). Email overload: Tolerance levels of employees within the workplace. In *Innovations Through Information Technology: 2004 Information Resources Management Association International Conference, New Orleans, Louisiana, USA, May 23-26, 2004*, volume 1, page 205. IGI Global.

Butts, C. T. (2008). A relational event framework for social action. *Sociological Methodology*, 38(1):155–200.

Camber Warren, T. (2010). The geometry of security: Modeling interstate alliances as evolving networks. *Journal of Peace Research*, 47(6):697–709.

Chatterjee, S., Diaconis, P., et al. (2013). Estimating and understanding exponential random graph models. *The Annals of Statistics*, 41(5):2428–2461.

Cranmer, S. J., Desmarais, B. A., and Kirkland, J. H. (2012a). Toward a network theory of alliance formation. *International Interactions*, 38(3):295–324.

Cranmer, S. J., Desmarais, B. A., and Menninga, E. J. (2012b). Complex dependencies in the alliance network. *Conflict Management and Peace Science*, 29(3):279–313.

Desmarais, B. A. and Cranmer, S. J. (2017). Statistical inference in political networks research. In Victor, J. N., Montgomery, A. H., and Lubell, M., editors, *The Oxford Handbook of Political Networks*. Oxford University Press.

Fahmy, C. and Young, J. T. (2016). Gender inequality and knowledge production in criminology and criminal justice. *Journal of Criminal Justice Education*, pages 1–21.

Fellows, I. and Handcock, M. (2017). Removing phase transitions from gibbs measures. In *Artificial Intelligence and Statistics*, pages 289–297.

Geweke, J. (2004). Getting it right: Joint distribution tests of posterior simulators. *Journal of the American Statistical Association*, 99(467):799–804.

Griffiths, T. L. and Steyvers, M. (2004). Finding scientific topics. *Proceedings of the National academy of Sciences*, 101(suppl 1):5228–5235.

Hammer, M. (1985). Implications of behavioral and cognitive reciprocity in social network data. *Social Networks*, 7(2):189–201.

Hunter, D. R., Handcock, M. S., Butts, C. T., Goodreau, S. M., and Morris, M. (2008). ergm: A package to fit, simulate and diagnose exponential-family models for networks. *Journal of statistical software*, 24(3):nihpa54860.

Jeong, H., Néda, Z., and Barabási, A.-L. (2003). Measuring preferential attachment in evolving networks. *EPL (Europhysics Letters)*, 61(4):567.

Kanungo, S. and Jain, V. (2008). Modeling email use: a case of email system transition. *System Dynamics Review*, 24(3):299–319.

Kinne, B. J. (2016). Agreeing to arm: Bilateral weapons agreements and the global arms trade. *Journal of Peace Research*, 53(3):359–377.

Krafft, P., Moore, J., Desmarais, B., and Wallach, H. M. (2012). Topic-partitioned multinetwork embeddings. In Pereira, F., Burges, C., Bottou, L., and Weinberger, K., editors, *Advances in Neural Information Processing Systems 25*, pages 2807–2815. Curran Associates, Inc.

Kronegger, L., Mali, F., Ferligoj, A., and Doreian, P. (2011). Collaboration structures in slovenian scientific communities. *Scientometrics*, 90(2):631–647.

Lai, C.-H., She, B., and Tao, C.-C. (2017). Connecting the dots: A longitudinal observation of relief organizations' representational networks on social media. *Computers in Human Behavior*, 74:224–234.

Liang, X. (2015). The changing impact of geographic distance: A preliminary analysis on the co-author networks in scientometrics (1983-2013). In *System Sciences (HICSS), 2015 48th Hawaii International Conference on*, pages 722–731. IEEE.

Lim, K. W., Chen, C., and Buntine, W. (2013). Twitter-network topic model: A full bayesian treatment for social network and text modeling. In *NIPS2013 Topic Model workshop*, pages 1–5.

Louch, H. (2000). Personal network integration: transitivity and homophily in strong-tie relations. *Social networks*, 22(1):45–64.

McCallum, A., Corrada-Emmanuel, A., and Wang, X. (2005). The author-recipient-topic model for topic and role discovery in social networks, with application to enron and academic email. In *Workshop on Link Analysis, Counterterrorism and Security*, page 33.

Minka, T. (2000). Estimating a dirichlet distribution.

Peng, T.-Q., Liu, M., Wu, Y., and Liu, S. (2016). Follower-followee network, communication networks, and vote agreement of the us members of congress. *Communication Research*, 43(7):996–1024.

Perry, P. O. and Wolfe, P. J. (2013). Point process modelling for directed interaction networks. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 75(5):821–849.

Pew, R. C. (2016). Social media fact sheet. *Accessed on 03/07/17*.

Rao, A. R. and Bandyopadhyay, S. (1987). Measures of reciprocity in a social network. *Sankhyā: The Indian Journal of Statistics, Series A*, pages 141–188.

Robins, G., Pattison, P., Kalish, Y., and Lusher, D. (2007). An introduction to exponential random graph (p*) models for social networks. *Social networks*, 29(2):173–191.

Szóstek, A. M. (2011). ?dealing with my emails?: Latent user needs in email management. *Computers in Human Behavior*, 27(2):723–729.

Vázquez, A. (2003). Growing network with local rules: Preferential attachment, clustering hierarchy, and degree correlations. *Physical Review E*, 67(5):056104.

Wallach, H. M. (2008). *Structured topic models for language*. PhD thesis, University of Cambridge.

Yoon, H. Y. and Park, H. W. (2014). Strategies affecting twitter-based networking pattern of south korean politicians: social network analysis and exponential random graph model. *Quality & Quantity*, pages 1–15.

APPENDIX

A Normalizing constant of non-empty Gibbs measure

In Section 2.4, we define the non-empty Gibbs measure such that the probability of sender i selecting the binary receiver vector of length $(A-1)$, $J_i^{(d)}$ is given by

$$\mathrm{P}(J_i^{(d)}) = \frac{1}{Z(\delta, \log(\lambda_i^{(d)}))} \exp\Big\{\log\big(\mathrm{I}(\sum_{j\in\mathcal{A}_{\setminus i}} J_{ij}^{(d)} > 0)\big) + \sum_{j\in\mathcal{A}_{\setminus i}} (\delta + \log(\lambda_{ij}^{(d)}))J_{ij}^{(d)}\Big\}.$$

To use this distribution efficiently, we need to derive a closed-form expression for $Z(\delta, \log(\lambda_i^{(d)}))$ that does not require brute-force summation over the support of $J_i^{(d)}$. We begin by recognizing that if $J_i^{(d)}$ were drawn via independent Bernoulli distributions in which $\mathrm{P}(J_{ij}^{(d)}{=}1)$ was given by $\mathrm{logit}\Big(\delta + \lambda_{ij}^{(d)}\Big)$, then

$$\mathrm{P}(J_i^{(d)}) \propto \exp\Big\{\sum_{j\in\mathcal{A}_{\setminus i}} (\delta + \log(\lambda_{ij}^{(d)}))J_{ij}^{(d)}\Big\}.$$

This is straightforward to verify by looking at

$$\begin{aligned}\mathrm{P}(J_{ij}^{(d)} = 1 | J_{i,-j}) &= \frac{\exp\left(\delta + \log(\lambda_{ij}^{(d)})\right)\exp\Big\{\sum_{h\neq i,j}(\delta + \log(\lambda_{ih}^{(d)}))J_{ih}^{(d)}\Big\}}{\exp\left(\delta + \log(\lambda_{ij}^{(d)})\right)\exp\Big\{\sum_{h\neq i,j}(\delta + \log(\lambda_{ih}^{(d)}))J_{ih}^{(d)}\Big\} + \exp(0)\exp\Big\{\sum_{h\neq i,j}(\delta + \log(\lambda_{ih}^{(d)}))J_{ih}^{(d)}\Big\}}, \\ &= \frac{\exp\left(\delta + \log(\lambda_{ij}^{(d)})\right)}{\exp\left(\delta + \log(\lambda_{ij}^{(d)})\right) + 1}.\end{aligned}$$

We denote the logistic-Bernoulli normalizing constant as $Z^l(\delta, \lambda_i^{(d)})$, which is defined as

$$Z^l(\delta, \log(\lambda_i^{(d)})) = \sum_{J_i\in[0,1]^{(A-1)}} \exp\Big\{\sum_{j\neq i}(\delta + \log(\lambda_{ij}^{(d)}))J_{ij}^{(d)}\Big\}.$$

Now, since

$$\exp\Big\{\log\big(\mathrm{I}(\sum_{j\in\mathcal{A}_{\setminus i}} J_{ij} > 0)\big) + \sum_{j\in\mathcal{A}_{\setminus i}} (\delta + \log(\lambda_{ij}^{(d)}))J_{ij}^{(d)}\Big\} = \exp\Big\{\sum_{j\in\mathcal{A}_{\setminus i}} (\delta + \log(\lambda_{ij}^{(d)}))J_{ij}^{(d)}\Big\},$$

except when $\sum_{j\in\mathcal{A}_{\setminus i}} J_{ij}^{(d)} = 0$, in which case the left-hand side

$$\exp\Big\{\log\big(\mathrm{I}(\sum_{j\in\mathcal{A}_{\setminus i}} J_{ij} > 0)\big) + \sum_{j\in\mathcal{A}_{\setminus i}} (\delta + \log(\lambda_{ij}^{(d)}))J_{ij}^{(d)}\Big\} = 0.$$

As such, we note that

$$\begin{aligned}Z(\delta, \log(\lambda_i^{(d)})) &= Z^l(\delta, \log(\lambda_i^{(d)})) - \exp\Big\{\sum_{j\in\mathcal{A}_{\setminus i}, J_{ij}^{(d)}=0} (\delta + \log(\lambda_{ij}^{(d)}))J_{ij}^{(d)}\Big\} \\ &= Z^l(\delta, \log(\lambda_i^{(d)})) - 1.\end{aligned}$$

We can therefore derive a closed form expression for $Z(\delta, \log(\lambda_i^{(d)}))$ via a closed form expression for $Z^l(\delta, \log(\lambda_i^{(d)}))$. This can be done by looking at the probability of the zero vector under the

logistic-Bernoulli model:

$$\frac{\exp\left\{\sum_{j\neq i, J_{ij}^{(d)}=0}(\delta+\log(\lambda_{ij}^{(d)}))J_{ij}^{(d)}\right\}}{Z^l(\delta,\log(\lambda_{ij}^{(d)}))} = \prod_{j\in\mathcal{A}_{\setminus i}}\frac{\exp\{(-(\delta+\log(\lambda_{ij}^{(d)})))\}}{\exp\{(-(\delta+\log(\lambda_{ij}^{(d)})))\}+1},$$

$$\frac{1}{Z^l(\delta,\log(\lambda_{ij}^{(d)}))} = \prod_{j\in\mathcal{A}_{\setminus i}}\frac{\exp\left(-(\delta+\log(\lambda_{ij}^{(d)}))\right)}{\exp\left(-(\delta+\log(\lambda_{ij}^{(d)}))\right)+1},$$

$$Z^l(\delta,\log(\lambda_{ij}^{(d)})) = \frac{1}{\prod_{j\in\mathcal{A}_{\setminus i}}\frac{\exp\left(-(\delta+\log(\lambda_{ij}^{(d)}))\right)}{\exp\left(-(\delta+\log(\lambda_{ij}^{(d)}))\right)+1}}.$$

The closed form expression for the normalizing constant under the non-empty Gibbs measure is therefore

$$Z(\delta,\lambda_i^{(d)}) = \Big(\prod_{j\in\mathcal{A}_{\setminus i}}\Big(\exp\{\delta+\log(\lambda_{ij}^{(d)})\}+1\Big)\Big)-1.$$

B Sampling Equations

Note that since $p_c^{(d)}$ is a deterministic function of $(\mathcal{Z},\mathcal{C})$, $\boldsymbol{x}^{(c)}_{t_+^{(d-1)}}$ is a deterministic function of $(p_c^{(d)},\mathcal{P}')$ and $\boldsymbol{\lambda}^{(d)}$ is a deterministic function of $(p_c^{(d)},\boldsymbol{x}^{(c)}_{t_+^{(d-1)}},\mathcal{B})$, we do not include them as variables in the joint distribution. Given that $\boldsymbol{\lambda}^{(d)}$ is a function of the three latent variables $\mathcal{Z},\mathcal{C}$, and $\mathcal{B}$, we use any parts of joint distribution that involves the term $\boldsymbol{\lambda}^{(d)}$ to make inference on $(\mathcal{Z},\mathcal{C},\mathcal{B})$.

B.1 Joint distribution of Tie variables

As mentioned earlier in Section 2.4, we use data augmentation in the tie generating process. Since we should include both the observed and augmented data to make inferences on the related latent variables, the derivation steps for the contribution of tie is not as smiple as other variables. Therefore, here we provide the detailed derivation steps for the last term of conditional probability:

$$\begin{aligned}
&P(\mathcal{J}_{\mathrm{a}},\mathcal{T}_{\mathrm{a}},\mathcal{I}_{\mathrm{o}},\mathcal{J}_{\mathrm{o}},\mathcal{T}_{\mathrm{o}}|\mathcal{Z},\mathcal{C},\mathcal{B},\delta)\\
&=\prod_{d=1}^{D}P(\mathcal{J}_{\mathrm{a}}^{(d)},\mathcal{T}_{\mathrm{a}}^{(d)},i_{\mathrm{o}}^{(d)},J_{\mathrm{o}}^{(d)},t_{\mathrm{o}}^{(d)}|\mathcal{I}_{\mathrm{o}}^{(-d)},\mathcal{J}_{\mathrm{o}}^{(-d)},\mathcal{T}_{\mathrm{o}}^{(-d)},\mathcal{Z},\mathcal{C},\mathcal{B},\delta)\\
&=\prod_{d=1}^{D}P(\mathcal{J}_{\mathrm{a}}^{(d)},\mathcal{T}_{\mathrm{a}}^{(d)},i_{\mathrm{o}}^{(d)},J_{\mathrm{o}}^{(d)},t_{\mathrm{o}}^{(d)}|\mathcal{I}_{\mathrm{o}}^{(<d)},\mathcal{J}_{\mathrm{o}}^{(<d)},\mathcal{T}_{\mathrm{o}}^{(<d)},\mathcal{Z},\mathcal{C},\mathcal{B},\delta).
\end{aligned}\tag{10}$$

Note that the conditional probability only depends on the past documents $(\mathcal{I}_{\mathrm{o}}^{(<d)},\mathcal{J}_{\mathrm{o}}^{(<d)},\mathcal{T}_{\mathrm{o}}^{(<d)})$, but not on the future ones $(\mathcal{I}_{\mathrm{o}}^{(>d)},\mathcal{J}_{\mathrm{o}}^{(>d)},\mathcal{T}_{\mathrm{o}}^{(>d)})$, since the network covariates $\boldsymbol{x}_t^{(c)}$ is calculated only based on the past interaction history.

Now we tackle the problem by deriving $P(\mathcal{J}_{\mathrm{a}}^{(d)},\mathcal{T}_{\mathrm{a}}^{(d)},i_{\mathrm{o}}^{(d)},J_{\mathrm{o}}^{(d)},t_{\mathrm{o}}^{(d)}|\mathcal{I}_{\mathrm{o}}^{(<d)},\mathcal{J}_{\mathrm{o}}^{(<d)},\mathcal{T}_{\mathrm{o}}^{(<d)},\mathcal{Z},\mathcal{C},\mathcal{B},\delta)$ for d^{th} document. There are three steps involved. First is the generation of the latent receivers J_i for each i, which corresponds to the Bernoulli part of tie generation, Equation (4); second is the generation of the observed time increment $\Delta T^{(d)} = t^{(d)} - t^{(d-1)}$ from the observed sender-receiver pairs $(i_{\mathrm{o}}^{(d)},J_{\mathrm{o}}^{(d)})$, which corresponds to the Exponential tie generation in Equation (6); and the last part is the simultaneous selection process of the observed sender, receivers, and timestamp in Equation (7), implying that the latent time increments generated from the latent sender-receiver pairs were greater than the observed time increment. Reflecting the three steps, the joint distribution is:

$$
\begin{aligned}
&P(\mathcal{J}_{\mathrm{a}}^{(d)}, \mathcal{T}_{\mathrm{a}}^{(d)}, i_{\mathrm{o}}^{(d)}, J_{\mathrm{o}}^{(d)}, t_{\mathrm{o}}^{(d)} | \mathcal{I}_{\mathrm{o}}^{(<d)}, \mathcal{J}_{\mathrm{o}}^{(<d)}, \mathcal{T}_{\mathrm{o}}^{(<d)}, \mathcal{Z}, \mathcal{C}, \mathcal{B}, \delta) \\
&= P\Big(\text{latent receivers generation}\Big) \times P\Big(\text{latent time generation}\Big) \times P\Big(\text{choose the observed}\Big) \\
&= \prod_{i \in \mathcal{A}} \Big(J_i^{(d)} \sim \text{Gibbs measure}(\{\lambda_{ij}^{(d)}\}_{j=1}^{A}, \delta)\Big) \times \prod_{i \in \mathcal{A}} \Big(\Delta T_{iJ_i}^{(d)} \sim \text{Exp}(\lambda_{iJ_i}^{(d)})\Big) \times \prod_{i \in \mathcal{A}_{\setminus i_o^{(d)}}} P\Big(\Delta T_{iJ_i}^{(d)} > \Delta T_{i_o^{(d)} J_o^{(d)}}\Big) \\
&= \Big(\prod_{i \in \mathcal{A}} \frac{1}{Z(\delta, \log(\lambda_i^{(d)}))} \exp\Big\{\log\big(\mathrm{I}(\sum_{j \in \mathcal{A}_{\setminus i}} J_{ij}^{(d)} > 0)\big) + \sum_{j \in \mathcal{A}_{\setminus i}} (\delta + \log(\lambda_{ij}^{(d)})) J_{ij}^{(d)}\Big\}\Big) \\
&\qquad\qquad \times \Big(\prod_{i \in \mathcal{A}} \lambda_{iJ_i}^{(d)} e^{-\Delta T_{iJ_i}^{(d)} \lambda_{iJ_i}^{(d)}}\Big) \times \Big(\prod_{i \in \mathcal{A}_{\setminus i_o^{(d)}}} e^{-\Delta T_{i_o^{(d)} J_o^{(d)}}^{(d)} \lambda_{iJ_i^{(d)}}^{(d)}}\Big) \\
&\propto \Big(\prod_{i \in \mathcal{A}} \frac{1}{\Big(\prod_{j \in \mathcal{A}_{\setminus i}} \Big(\exp\{\delta + \log(\lambda_{ij}^{(d)})\} + 1\Big)\Big) - 1} \exp\Big\{\sum_{j \in \mathcal{A}_{\setminus i}} (\delta + \log(\lambda_{ij}^{(d)})) J_{ij}^{(d)}\Big\}\Big) \\
&\qquad\qquad \times \Big(\lambda_{i_o^{(d)} J_o^{(d)}}^{(d)} e^{-\Delta T_{i_o^{(d)} J_o^{(d)}}^{(d)} \lambda_{i_o^{(d)} J_o^{(d)}}^{(d)}}\Big) \times \Big(\prod_{i \in \mathcal{A}_{\setminus i_o^{(d)}}} \lambda_{iJ_i}^{(d)} e^{-(\Delta T_{iJ_i}^{(d)} + \Delta T_{i_o^{(d)} J_o^{(d)}}^{(d)}) \lambda_{iJ_i}^{(d)}}\Big),
\end{aligned}
\tag{11}
$$

We can simplify this further by integreting out the latent time $\mathcal{T}_{\mathrm{a}}^{(d)} = \{\Delta T_{iJ_i}^{(d)}\}_{i \in \mathcal{A}_{\setminus i_o^{(d)}}}$ in the last term:

$$
\begin{aligned}
&\int_0^\infty \cdots \int_0^\infty \Big(\prod_{i \in \mathcal{A}_{\setminus i_o^{(d)}}} \lambda_{iJ_i}^{(d)} e^{-(\Delta T_{iJ_i}^{(d)} + \Delta T_{i_o^{(d)} J_o^{(d)}}^{(d)}) \lambda_{iJ_i}^{(d)}}\Big) d\Delta T_{1J_1}^{(d)} \cdots d\Delta T_{AJ_A}^{(d)} \\
&= \prod_{i \in \mathcal{A}_{\setminus i_o^{(d)}}} e^{-\Delta T_{i_o^{(d)} J_o^{(d)}}^{(d)} \lambda_{iJ_i}^{(d)}} \Big(\int_0^\infty \lambda_{iJ_i}^{(d)} e^{-\Delta T_{iJ_i}^{(d)} \lambda_{iJ_i}^{(d)}} d\Delta T_{iJ_i}^{(d)}\Big) \\
&= \prod_{i \in \mathcal{A}_{\setminus i_o^{(d)}}} e^{-\Delta T_{i_o^{(d)} J_o^{(d)}}^{(d)} \lambda_{iJ_i}^{(d)}} \Big(\Big[-e^{-\Delta T_{iJ_i}^{(d)} \lambda_{iJ_i}^{(d)}}\Big]_{\Delta T_{iJ_i}^{(d)} = 0}^{\infty}\Big) \\
&= e^{-\Delta T_{i_o^{(d)} J_o^{(d)}}^{(d)} \sum_{i \in \mathcal{A}_{\setminus i_o^{(d)}}} \lambda_{iJ_i}^{(d)}},
\end{aligned}
\tag{12}
$$

where $\Delta T_{i_o^{(d)} J_o^{(d)}}^{(d)}$ is the observed time difference between d^{th} and $(d-1)^{th}$ document. Therefore, we can simplify Equation (12) as below:

$$
\begin{aligned}
&P(\mathcal{J}_{\mathrm{a}}^{(d)}, i_{\mathrm{o}}^{(d)}, J_{\mathrm{o}}^{(d)}, t_{\mathrm{o}}^{(d)} | \mathcal{I}_{\mathrm{o}}^{(<d)}, \mathcal{J}_{\mathrm{o}}^{(<d)}, \mathcal{T}_{\mathrm{o}}^{(<d)}, \mathcal{Z}, \mathcal{C}, \mathcal{B}, \delta) \\
&\propto \Big(\prod_{i \in \mathcal{A}} \frac{1}{\Big(\prod_{j \in \mathcal{A}_{\setminus i}} \Big(\exp\{\delta + \log(\lambda_{ij}^{(d)})\} + 1\Big)\Big) - 1} \exp\Big\{\sum_{j \in \mathcal{A}_{\setminus i}} (\delta + \log(\lambda_{ij}^{(d)})) J_{ij}^{(d)}\Big\}\Big) \\
&\qquad\qquad \times \Big(\lambda_{i_o^{(d)} J_o^{(d)}}^{(d)}\Big) \times \Big(e^{-\Delta T_{i_o^{(d)} J_o^{(d)}}^{(d)} \sum_{i \in \mathcal{A}} \lambda_{iJ_i^{(d)}}^{(d)}}\Big),
\end{aligned}
\tag{13}
$$

where this joint distribution can be interpreted as 'probability of latent and observed edges from non-empty Gibbs measure $\times$ probability of the observed time comes from Exponential distribution $\times$ probability of all latent time greater than the observed time, given that the latent time also come from Exponential distribution.' Finally for implementation, we need to compute these equations in log space to prevent underflow:

$$
\begin{aligned}
&\log\Big(P(\mathcal{J}_{\mathrm{a}}^{(d)}, i_{\mathrm{o}}^{(d)}, J_{\mathrm{o}}^{(d)}, t_{\mathrm{o}}^{(d)} | \mathcal{I}_{\mathrm{o}}^{(<d)}, \mathcal{J}_{\mathrm{o}}^{(<d)}, \mathcal{T}_{\mathrm{o}}^{(<d)}, \mathcal{Z}, \mathcal{C}, \mathcal{B}, \delta)\Big) \\
&\propto \Big(\sum_{i \in \mathcal{A}} \Big(-\log\Big(\Big(\prod_{j \in \mathcal{A}_{\setminus i}} \Big(\exp\{\delta + \log(\lambda_{ij}^{(d)})\} + 1\Big)\Big) - 1\Big) + \sum_{j \in \mathcal{A}_{\setminus i}} (\delta + \log(\lambda_{ij}^{(d)})) J_{ij}^{(d)}\Big)\Big) \\
&\qquad\qquad + \Big(\log(\lambda_{i_o^{(d)} J_o^{(d)}}^{(d)})\Big) - \Big(\Delta T_{i_o^{(d)} J_o^{(d)}}^{(d)} \sum_{i \in \mathcal{A}} \lambda_{iJ_i^{(d)}}^{(d)}\Big),
\end{aligned}
\tag{14}
$$

and sequentially resample the values of each of our latent variables from their posterior distribution, conditional on all of our other variables.

B.2 Resampling $\mathcal{J}_{\mathrm{a}}$

First of all, for each document d, we update the latent sender-receiver(s) pairs. That is, given the observed sender of the document $i_o^{(d)}$, we sample the latent receivers for each sender $i \in \mathcal{A}_{\setminus i_o^{(d)}}$. Here we illustrate how each sender-receiver pair in the document d is updated.

Define $\mathcal{J}_i^{(d)}$ be the $(A-1)$ length random vector of indicators with its realization being $J_i^{(d)}$, representing the latent receivers corresponding to the sender i in the document d. For each latnet sender i, we are going to resample $J_{ij}^{(d)}$, which is the j^{th} element of the receiver vector $J_i^{(d)}$, one at a time with random order. The full conditional probability of $J_{ij}^{(d)}$ is:

$$P(\mathcal{J}_{ij}^{(d)} = J_{ij}^{(d)} | \mathcal{J}_{i\setminus j}^{(d)}, \mathcal{Z}, \mathcal{C}, \mathcal{B}, \delta, \mathcal{W}, \mathcal{J}_{\mathrm{a},-i}, \mathcal{I}_{\mathrm{O}}, \mathcal{J}_{\mathrm{O}}, \mathcal{T}_{\mathrm{O}}, \beta, \boldsymbol{u}, \alpha, \boldsymbol{m}, \mu_{\boldsymbol{b}}, \Sigma_{\boldsymbol{b}}, \mu_\delta, \sigma_\delta^2), \tag{15}$$

which we can drop some independent terms and move to

$$\begin{aligned}
&P(\mathcal{J}_{ij}^{(d)} = J_{ij}^{(d)} | \mathcal{J}_{i\setminus j}^{(d)}, i_{\mathrm{O}}^{(d)}, J_{\mathrm{O}}^{(d)}, t_{\mathrm{O}}^{(d)}, \mathcal{I}_{\mathrm{O}}^{(<d)}, \mathcal{J}_{\mathrm{O}}^{(<d)}, \mathcal{T}_{\mathrm{O}}^{(<d)}, \mathcal{Z}, \mathcal{C}, \mathcal{B}, \delta) \\
&\propto P(\mathcal{J}_{ij}^{(d)} = J_{ij}^{(d)}, \mathcal{J}_{i\setminus j}^{(d)}, i_{\mathrm{O}}^{(d)}, J_{\mathrm{O}}^{(d)}, t_{\mathrm{O}}^{(d)} | \mathcal{I}_{\mathrm{O}}^{(<d)}, \mathcal{J}_{\mathrm{O}}^{(<d)}, \mathcal{T}_{\mathrm{O}}^{(<d)}, \mathcal{Z}, \mathcal{C}, \mathcal{B}, \delta) \\
&\propto \Big(\frac{1}{\Big(\prod_{j \in \mathcal{A}_{\setminus i}} \Big(\exp\{\delta + \log(\lambda_{ij}^{(d)})\} + 1 \Big) \Big) - 1} \exp\Big\{ \log\big(\mathrm{I}(\sum_{j \in \mathcal{A}_{\setminus i}} J_{ij}^{(d)} > 0) \big) + \sum_{j \in \mathcal{A}_{\setminus i}} (\delta + \log(\lambda_{ij}^{(d)})) J_{ij}^{(d)} \Big\} \Big) \\
&\qquad\qquad \times \Big(\lambda_{i_o^{(d)} J_o^{(d)}}^{(d)} \Big) \times \Big(e^{-\Delta T_{i_o^{(d)} J_o^{(d)}}^{(d)} \lambda_{i J_i^{(d)}}^{(d)}} \Big) \\
&\propto \Big(\exp\Big\{ \log\big(\mathrm{I}(\sum_{j \in \mathcal{A}_{\setminus i}} J_{ij}^{(d)} > 0) \big) + \sum_{j \in \mathcal{A}_{\setminus i}} (\delta + \log(\lambda_{ij}^{(d)})) J_{ij}^{(d)} \Big\} \Big) \times \Big(e^{-\Delta T_{i_o^{(d)} J_o^{(d)}}^{(d)} \lambda_{i J_i^{(d)}}^{(d)}} \Big),
\end{aligned} \tag{16}$$

where we replace typical use of $(-d)$ to $(<d)$ on the right hand side of the conditional probability, due to the fact that $d^{(th)}$ document only depends on the past documents, not on the future ones. The last line of Equation (17) is obtained by dropping the terms that do not include $J_{ij}^{(d)}$, such as the normalizing constan of Gibbs measure.

To be more specific, since $J_{ij}^{(d)}$ could be either 1 or 0, we divide into two cases as below:

$$\begin{aligned}
&P(\mathcal{J}_{ij}^{(d)} = 1 | \mathcal{J}_{i\setminus j}^{(d)}, i_{\mathrm{O}}^{(d)}, J_{\mathrm{O}}^{(d)}, t_{\mathrm{O}}^{(d)}, \mathcal{I}_{\mathrm{O}}^{(<d)}, \mathcal{J}_{\mathrm{O}}^{(<d)}, \mathcal{T}_{\mathrm{O}}^{(<d)}, \mathcal{Z}, \mathcal{C}, \mathcal{B}, \delta) \\
&\propto \exp\Big(\log(1) + \sum_{j \in \mathcal{A}_{\setminus i}} (\delta + \log(\lambda_{ij}^{(d)})) J_{i[+j]}^{(d)} - \Delta T_{i_o^{(d)} J_o^{(d)}}^{(d)} \lambda_{i J_{i[+j]}^{(d)}}^{(d)} \Big) \\
&\propto \exp\Big(\delta + \log(\lambda_{ij}^{(d)}) - \Delta T_{i_o^{(d)} J_o^{(d)}}^{(d)} \lambda_{i J_{i[+j]}^{(d)}}^{(d)} \Big),
\end{aligned} \tag{17}$$

where $J_{i[+j]}^{(d)}$ meaning that the j^{th} element of $J_i^{(d)}$ is fixed as 1 (thus making $\log\big(\mathrm{I}(\sum_{j \in \mathcal{A}_{\setminus i}} J_{ij}^{(d)} > 0)\big) = 0$ for sure). On the other hand,

$$\begin{aligned}
&P(\mathcal{J}_{ij}^{(d)} = 0 | \mathcal{J}_{i\setminus j}^{(d)}, i_{\mathrm{O}}^{(d)}, J_{\mathrm{O}}^{(d)}, t_{\mathrm{O}}^{(d)}, \mathcal{I}_{\mathrm{O}}^{(<d)}, \mathcal{J}_{\mathrm{O}}^{(<d)}, \mathcal{T}_{\mathrm{O}}^{(<d)}, \mathcal{Z}, \mathcal{C}, \mathcal{B}, \delta) \\
&\propto \exp\Big(\log\big(\mathrm{I}(\sum_{j \in \mathcal{A}_{\setminus i}} J_{i[-j]}^{(d)} > 0) \big) + \sum_{j \in \mathcal{A}_{\setminus i}} (\delta + \log(\lambda_{ij}^{(d)})) J_{i[-j]}^{(d)} - \Delta T_{i_o^{(d)} J_o^{(d)}}^{(d)} \lambda_{i J_{i[-j]}^{(d)}}^{(d)} \Big) \\
&\propto \exp\Big(\log\big(\mathrm{I}(\sum_{j \in \mathcal{A}_{\setminus i}} J_{i[-j]}^{(d)} > 0) \big) - \Delta T_{i_o^{(d)} J_o^{(d)}}^{(d)} \lambda_{i J_{i[-j]}^{(d)}}^{(d)} \Big),
\end{aligned} \tag{18}$$

where $J_{i[-j]}^{(d)}$ meaning similarly that the j^{th} element of $J_i^{(d)}$ is fixed as 0. In this case, we cannot guarantee that $\mathrm{I}(\sum_{j \in \mathcal{A}_{\setminus i}} J_{ij}^{(d)} > 0)$ is 0 or 1, so we have to leave the term. When it is zero,

$\exp\{\log\big(\mathrm{I}(\sum_{j\in\mathcal{A}_{\setminus i}} J_{ij}^{(d)} > 0)\big)\} = 0$, thus we will sample 1 with probability 1. From this property of non-empty Gibbs measure, we prevent from the instances where the sender has no recipients to send the document. Now we can use multinomial sampling using the two probabilities, Equation (18) and Equation (19), which is equivalent to Bernoulli sampling with probability $\frac{P(\mathcal{J}_{ij}^{(d)}=1)}{P(\mathcal{J}_{ij}^{(d)}=0)+P(\mathcal{J}_{ij}^{(d)}=0)}$.

B.3 Resampling $\mathcal{Z}$

Second, we are going to resample the topic assignments, one words in a document at a time. The new values of $z_n^{(d)}$ are sampled using the conditional posterior probability of being topic k as we derived in APPENDIX C:

$$
\begin{aligned}
&P(z_n^{(d)} = k|\mathcal{Z}_{\setminus d,n}, \mathcal{C}, \mathcal{B}, \delta, \mathcal{W}, \mathcal{J}_{\mathrm{a}}, \mathcal{I}_{\mathrm{o}}, \mathcal{J}_{\mathrm{o}}, \mathcal{T}_{\mathrm{o}}, \beta, \boldsymbol{u}, \alpha, \boldsymbol{m}, \mu_{\boldsymbol{b}}, \Sigma_{\boldsymbol{b}}, \mu_\delta, \sigma_\delta^2)\\
&\propto P(z_n^{(d)} = k, w_n^{(d)}, \mathcal{J}_{\mathrm{a}}^{(\geq d)}, i_{\mathrm{o}}^{(\geq d)}, J_{\mathrm{o}}^{(\geq d)}, t_{\mathrm{o}}^{(\geq d)}|\mathcal{Z}_{\setminus d,n}, \mathcal{C}, \mathcal{B}, \delta, \mathcal{W}_{\setminus d,n}, \mathcal{I}_{\mathrm{o}}^{(<d)}, \mathcal{J}_{\mathrm{o}}^{(<d)}, \mathcal{T}_{\mathrm{o}}^{(<d)}, \beta, \boldsymbol{u}, \alpha, \boldsymbol{m})\\
&\propto P(z_n^{(d)} = k|\mathcal{Z}_{\setminus d,n}, \alpha, \boldsymbol{m}) P(w_n^{(d)}|z_n^{(d)} = k, \mathcal{W}_{\setminus d,n}, \mathcal{Z}_{\setminus d,n}, \beta, \boldsymbol{u}) \times \prod_{d=d}^{D} P(\mathcal{J}_{\mathrm{a}}^{(d)}, i_{\mathrm{o}}^{(d)}, J_{\mathrm{o}}^{(d)}, t_{\mathrm{o}}^{(d)}|z_n^{(d)} = k, \mathcal{Z}_{\setminus d,n}, \mathcal{C}, \mathcal{B}, \delta)
\end{aligned}
\tag{19}
$$

where the subscript "$\setminus d, n$" denotes the exclsuion of position n in d^{th} document. Note that since selecting a topic for any token influences the histories acting on all documents from d on, we use the product from d through D for the tie contribution part. We know that:

$$
P(z_n^{(d)} = k|\mathcal{Z}_{\setminus d,n}, \alpha, \boldsymbol{m}) = \frac{N_{\setminus d,n}^{(k|d)} + \alpha \boldsymbol{m}_k}{N^{(d)} - 1 + \alpha} \tag{20}
$$

which is the well-known form of collapsed Gibbs sampling equation for LDA. We also know that

$$
P(w_n^{(d)}|z_n^{(d)} = k, \mathcal{W}_{\setminus d,n}, \mathcal{Z}_{\setminus d,n}, \beta, \boldsymbol{u}) = \frac{N_{\setminus d,n}^{(w_n^{(d)}|k)} + \frac{\beta}{W}}{N_{\setminus d,n}^{(k)} + \beta}, \tag{21}
$$

where $N^{(w_n^{(d)}|k)}$ is the number of tokens assigned to topic k whose type is the same as that of $w_n^{(d)}$, excluding $w_n^{(d)}$ itself, and $N_{\setminus d,n}^{(k)} = \sum_{w=1}^{W} N_{\setminus d,n}^{(w_n^{(d)}|k)}$. Finally, we already have shown that

$$
\begin{aligned}
&P(\mathcal{J}_{\mathrm{a}}^{(d)}, i_{\mathrm{o}}^{(d)}, J_{\mathrm{o}}^{(d)}, t_{\mathrm{o}}^{(d)}|z_n^{(d)} = k, \mathcal{Z}_{\setminus d,n}, \mathcal{C}, \mathcal{B}, \delta)\\
&= \Big(\prod_{i\in\mathcal{A}} \frac{1}{\Big(\prod_{j\in\mathcal{A}_{\setminus i}} \Big(\exp\{\delta + \log(\lambda_{ij}^{(d)})\} + 1\Big)\Big) - 1} \exp\Big\{\sum_{j\in\mathcal{A}_{\setminus i}} (\delta + \log(\lambda_{ij}^{(d)})) J_{ij}^{(d)}\Big\}\Big) \times \Big(\lambda_{i_o^{(d)} J_o^{(d)}}^{(d)}\Big) \times \Big(e^{-\Delta T_{i_o^{(d)} J_o^{(d)}}^{(d)} \lambda_{i J_i^{(d)}}^{(d)}}\Big),
\end{aligned}
\tag{22}
$$

where every part includes $\lambda_{ij}^{(d)}$ such that we cannot simplify any further.

Therefore, if $N^{(d)} > 0$, then the conditional probability of n^{th} word in document d being topic k is:

$$
\begin{aligned}
&P(z_n^{(d)} = k|\mathcal{Z}_{\setminus d,n}, \mathcal{C}, \mathcal{B}, \delta, \mathcal{W}, \mathcal{J}_{\mathrm{a}}, \mathcal{I}_{\mathrm{o}}, \mathcal{J}_{\mathrm{o}}, \mathcal{T}_{\mathrm{o}}, \beta, \boldsymbol{u}, \alpha, \boldsymbol{m}, \mu_{\boldsymbol{b}}, \Sigma_{\boldsymbol{b}}, \mu_\delta, \sigma_\delta^2)\\
&\propto (N_{\setminus d,n}^{(k|d)} + \alpha \boldsymbol{m}_k) \times \frac{N_{\setminus d,n}^{(w_n^{(d)}|k)} + \frac{\beta}{W}}{N_{\setminus d,n}^{(k)} + \beta} \times\\
&\prod_{d=d}^{D} \Bigg(\Big(\prod_{i\in\mathcal{A}} \frac{1}{\Big(\prod_{j\in\mathcal{A}_{\setminus i}} \Big(\exp\{\delta + \log(\lambda_{ij}^{(d)})\} + 1\Big)\Big) - 1} \exp\Big\{\sum_{j\in\mathcal{A}_{\setminus i}} (\delta + \log(\lambda_{ij}^{(d)})) J_{ij}^{(d)}\Big\}\Big) \times \Big(\lambda_{i_o^{(d)} J_o^{(d)}}^{(d)} e^{-\Delta T_{i_o^{(d)} J_o^{(d)}}^{(d)} \lambda_{i J_i^{(d)}}^{(d)}}\Big)\Bigg),
\end{aligned}
\tag{23}
$$

and if $N^{(d)} = 0$, then the first term becomes $\alpha \boldsymbol{m_k}$ and disappears because it is a constant. The second term disappears since there are no tokens, thus we just have the term remaining as below.

$$P(z_1^{(d)} = k | \mathcal{Z}_{\setminus d,1} = \emptyset, \mathcal{C}, \mathcal{B}, \delta, \mathcal{W}, \mathcal{J}_{\mathrm{a}}, \mathcal{I}_{\mathrm{o}}, \mathcal{J}_{\mathrm{o}}, \mathcal{T}_{\mathrm{o}}, \beta, \boldsymbol{u}, \alpha, \boldsymbol{m}, \mu_{\boldsymbol{b}}, \Sigma_{\boldsymbol{b}}, \mu_\delta, \sigma_\delta^2)$$
$$\propto \prod_{d=d}^{D} \Bigg(\Big(\prod_{i \in \mathcal{A}} \frac{1}{\Big(\prod_{j \in \mathcal{A}_{\setminus i}} \Big(\exp\{\delta + \log(\lambda_{ij}^{(d)})\} + 1 \Big) \Big) - 1} \exp\Big\{ \sum_{j \in \mathcal{A}_{\setminus i}} (\delta + \log(\lambda_{ij}^{(d)})) J_{ij}^{(d)} \Big\} \Big) \times \Big(\lambda^{(d)}_{i_o^{(d)} J_o^{(d)}} e^{-\Delta T^{(d)}_{i_o^{(d)} J_o^{(d)}} \lambda^{(d)}_{i J_i^{(d)}}} \Big) \Bigg). \tag{24}$$

B.4 Resampling $\mathcal{C}$

The next variable to resample is the topic-interaction pattern assignments, one topic at a time. We derive the posterior conditional probability for the interaction pattern $\mathcal{C}$ for k^{th} topic as below:

$$\begin{aligned} &P(c_k = c | \mathcal{Z}, \mathcal{C}_{\setminus k}, \mathcal{B}, \delta, \mathcal{W}, \mathcal{J}_{\mathrm{a}}, \mathcal{I}_{\mathrm{o}}, \mathcal{J}_{\mathrm{o}}, \mathcal{T}_{\mathrm{o}}, \beta, \boldsymbol{u}, \alpha, \boldsymbol{m}, \mu_{\boldsymbol{b}}, \Sigma_{\boldsymbol{b}}, \mu_\delta, \sigma_\delta^2) \\ &\propto P(c_k = c, \mathcal{J}_{\mathrm{a}}, \mathcal{I}_{\mathrm{o}}, \mathcal{J}_{\mathrm{o}}, \mathcal{T}_{\mathrm{o}} | \mathcal{Z}, \mathcal{C}_{\setminus k}, \mathcal{B}, \delta) \\ &\propto P(c_k = c) P(\mathcal{J}_{\mathrm{a}}, \mathcal{I}_{\mathrm{o}}, \mathcal{J}_{\mathrm{o}}, \mathcal{T}_{\mathrm{o}} | \mathcal{Z}, c_k = c, \mathcal{C}_{\setminus k}, \mathcal{B}, \delta) \end{aligned} \tag{25}$$

where $P(c_k = c) = \frac{1}{C}$ so this term disappears. Therefore,

$$\begin{aligned} &P(c_k = c | \mathcal{Z}, \mathcal{C}_{\setminus k}, \mathcal{B}, \delta, \mathcal{W}, \mathcal{I}_{\mathrm{a}}, \mathcal{J}_{\mathrm{a}}, \mathcal{I}_{\mathrm{o}}, \mathcal{J}_{\mathrm{o}}, \mathcal{T}_{\mathrm{o}}, \beta, \boldsymbol{u}, \alpha, \boldsymbol{m}, \mu_{\boldsymbol{b}}, \Sigma_{\boldsymbol{b}}, \mu_\delta, \sigma_\delta^2) \\ &\propto P(\mathcal{J}_{\mathrm{a}}, \mathcal{I}_{\mathrm{o}}, \mathcal{J}_{\mathrm{o}}, \mathcal{T}_{\mathrm{o}} | \mathcal{Z}, c_k = c, \mathcal{C}_{\setminus k}, \mathcal{B}, \delta) \\ &= \prod_{d=1}^{D} \Bigg(\Big(\prod_{i \in \mathcal{A}} \frac{1}{\Big(\prod_{j \in \mathcal{A}_{\setminus i}} \Big(\exp\{\delta + \log(\lambda_{ij}^{(d)})\} + 1 \Big) \Big) - 1} \exp\Big\{ \sum_{j \in \mathcal{A}_{\setminus i}} (\delta + \log(\lambda_{ij}^{(d)})) J_{ij}^{(d)} \Big\} \Big) \\ &\qquad \times \Big(\lambda^{(d)}_{i_o^{(d)} J_o^{(d)}} \Big) \times \Big(e^{-\Delta T^{(d)}_{i_o^{(d)} J_o^{(d)}} \lambda^{(d)}_{i J_i^{(d)}}} \Big) \Bigg), \end{aligned} \tag{26}$$

with $c_k = c$ throughout.

B.5 Resampling $\mathcal{B}$

Next, we update $\mathcal{B} = \{\boldsymbol{b}^{(c)}\}_{c=1}^{C}$. For this, we use the Metropolis-Hastings algorithm with a proposal density Q being the multivariate Gaussian distribution, with a diagonal covariance matrix multiplied by σ_Q^2 (proposal distribution variance parameters set by the user), centered on the current values of $\mathcal{B} = \{\boldsymbol{b}^{(c)}\}_{c=1}^{C}$. Under the symmetric proposal distribution, we cancel out Q-ratio and then accept the new proposed value $\mathcal{B}' = \{\boldsymbol{b}'^{(c)}\}_{c=1}^{C}$ with probability equal to:

$$\text{Acceptance Probability} = \begin{cases} \frac{P(\mathcal{B}' | \mathcal{Z}, \mathcal{C}, \delta, \mathcal{W}, \mathcal{J}_{\mathrm{a}}, \mathcal{I}_{\mathrm{o}}, \mathcal{J}_{\mathrm{o}}, \mathcal{T}_{\mathrm{o}}, \beta, \boldsymbol{u}, \alpha, \boldsymbol{m}, \mu_{\boldsymbol{b}}, \Sigma_{\boldsymbol{b}}, \mu_\delta, \sigma_\delta^2)}{P(\mathcal{B} | \mathcal{Z}, \mathcal{C}, \delta, \mathcal{W}, \mathcal{J}_{\mathrm{a}}, \mathcal{I}_{\mathrm{o}}, \mathcal{J}_{\mathrm{o}}, \mathcal{T}_{\mathrm{o}}, \beta, \boldsymbol{u}, \alpha, \boldsymbol{m}, \mu_{\boldsymbol{b}}, \Sigma_{\boldsymbol{b}}, \mu_\delta, \sigma_\delta^2)} & \text{if} < 1 \\ 1 \quad \text{else} \end{cases} \tag{27}$$

After factorization, we get

$$\begin{aligned} &\frac{P(\mathcal{B}' | \mathcal{Z}, \mathcal{C}, \delta, \mathcal{W}, \mathcal{J}_{\mathrm{a}}, \mathcal{I}_{\mathrm{o}}, \mathcal{J}_{\mathrm{o}}, \mathcal{T}_{\mathrm{o}}, \beta, \boldsymbol{u}, \alpha, \boldsymbol{m}, \mu_{\boldsymbol{b}}, \Sigma_{\boldsymbol{b}}, \mu_\delta, \sigma_\delta^2)}{P(\mathcal{B} | \mathcal{Z}, \mathcal{C}, \delta, \mathcal{W}, \mathcal{J}_{\mathrm{a}}, \mathcal{I}_{\mathrm{o}}, \mathcal{J}_{\mathrm{o}}, \mathcal{T}_{\mathrm{o}}, \beta, \boldsymbol{u}, \alpha, \boldsymbol{m}, \mu_{\boldsymbol{b}}, \Sigma_{\boldsymbol{b}}, \mu_\delta, \sigma_\delta^2)} \\ =&\frac{P(\mathcal{Z}, \mathcal{C}, \mathcal{B}', \delta, \mathcal{W}, \mathcal{J}_{\mathrm{a}}, \mathcal{I}_{\mathrm{o}}, \mathcal{J}_{\mathrm{o}}, \mathcal{T}_{\mathrm{o}} | \beta, \boldsymbol{u}, \alpha, \boldsymbol{m}, \mu_{\boldsymbol{b}}, \Sigma_{\boldsymbol{b}}, \mu_\delta, \sigma_\delta^2)}{P(\mathcal{Z}, \mathcal{C}, \mathcal{B}, \delta, \mathcal{W}, \mathcal{J}_{\mathrm{a}}, \mathcal{I}_{\mathrm{o}}, \mathcal{J}_{\mathrm{o}}, \mathcal{T}_{\mathrm{o}} | \beta, \boldsymbol{u}, \alpha, \boldsymbol{m}, \mu_{\boldsymbol{b}}, \Sigma_{\boldsymbol{b}}, \mu_\delta, \sigma_\delta^2)} \\ =&\frac{P(\mathcal{B}' | \mathcal{C}, \mu_{\boldsymbol{b}}, \Sigma_{\boldsymbol{b}}) P(\mathcal{J}_{\mathrm{a}}, \mathcal{I}_{\mathrm{o}}, \mathcal{J}_{\mathrm{o}}, \mathcal{T}_{\mathrm{o}} | \mathcal{Z}, \mathcal{C}, \mathcal{B}', \delta)}{P(\mathcal{B} | \mathcal{C}, \mu_{\boldsymbol{b}}, \Sigma_{\boldsymbol{b}}) P(\mathcal{J}_{\mathrm{a}}, \mathcal{I}_{\mathrm{o}}, \mathcal{J}_{\mathrm{o}}, \mathcal{T}_{\mathrm{o}} | \mathcal{Z}, \mathcal{C}, \mathcal{B}, \delta)}, \end{aligned} \tag{28}$$

where $P(\mathcal{B} | \mathcal{C}, \mu_{\boldsymbol{b}}, \Sigma_{\boldsymbol{b}})$ is calculated from the product of $\boldsymbol{b}^{(c)} \sim$ Multivariate Normal$(\mu_{\boldsymbol{b}}, \Sigma_{\boldsymbol{b}})$ over the interaction patterns $c \in \{1, ..., C\}$ (as defined in Section 2) and $P(\mathcal{J}_{\mathrm{a}}, \mathcal{I}_{\mathrm{o}}, \mathcal{J}_{\mathrm{o}}, \mathcal{T}_{\mathrm{o}} | \mathcal{Z}, \mathcal{C}, \mathcal{B}, \delta)$ is the

same as Equation (27). Again, we take the log and obtain the log of acceptance ratio:

$$\begin{aligned}
&\sum_{c=1}^{C}\log\Big(\mathcal{N}(\boldsymbol{b}'^{(c)};\mu_{\boldsymbol{b}},\Sigma_{\boldsymbol{b}})\Big)-\sum_{c=1}^{C}\log\Big(\mathcal{N}(\boldsymbol{b}^{(c)};\mu_{\boldsymbol{b}},\Sigma_{\boldsymbol{b}})\Big)\\
&+\sum_{d=1}^{D}\Bigg(\Big(\sum_{i\in\mathcal{A}}\Big(-\log\Big(\Big(\prod_{j\in\mathcal{A}_{\setminus i}}\Big(\exp\{\delta+\log(\lambda_{ij}^{(d)})\}+1\Big)\Big)-1\Big)+\sum_{j\in\mathcal{A}_{\setminus i}}(\delta+\log(\lambda_{ij}^{(d)}))J_{ij}^{(d)}\Big)\\
&\qquad+\Big(\log(\lambda^{(d)}_{i_o^{(d)}J_o^{(d)}})-\Delta T^{(d)}_{i_o^{(d)}J_o^{(d)}}\sum_{i\in\mathcal{A}}\lambda^{(d)}_{iJ_i^{(d)}}\Big)\text{ given }\boldsymbol{b}'\Bigg)\\
&-\sum_{d=1}^{D}\Bigg(\Big(\sum_{i\in\mathcal{A}}\Big(-\log\Big(\Big(\prod_{j\in\mathcal{A}_{\setminus i}}\Big(\exp\{\delta+\log(\lambda_{ij}^{(d)})\}+1\Big)\Big)-1\Big)+\sum_{j\in\mathcal{A}_{\setminus i}}(\delta+\log(\lambda_{ij}^{(d)}))J_{ij}^{(d)}\Big)\\
&\qquad+\Big(\log(\lambda^{(d)}_{i_o^{(d)}J_o^{(d)}})-\Delta T^{(d)}_{i_o^{(d)}J_o^{(d)}}\sum_{i\in\mathcal{A}}\lambda^{(d)}_{iJ_i^{(d)}}\Big)\text{ given }\boldsymbol{b}\Bigg),
\end{aligned}\tag{29}$$

where $\mathcal{N}$ is the multivariate normal density. Then the log of acceptance ratio we have is:

$$\log(\text{Acceptance Probability})=\min(\text{Equation (30)},\,0).\tag{30}$$

To determine whether to accept the proposed update or not, we use the log of acceptance ratio; if the log of a sample from Uniform(0,1) is less than the log-acceptance probability (30), we accept the proposal $\boldsymbol{b}'$. Otherwise, we reject.

B.6 Resampling δ

Finally we move on to the updates of δ, which is very similar to the steps illustrated in Section B.5. Again we use Metropolis-Hastings algorithm with Normal proposal distribution such that we can cancel out the Q-ratio. We may change the proposal variance σ_δ^2 to ensure appropriate level of acceptance rate. Then, it follows that the simplified version of acceptance probability is

$$\text{Acceptance Probability}=\begin{cases}\frac{P(\delta'|\mu_\delta,\sigma_\delta^2)P(\mathcal{J}_{\mathrm{a}},\mathcal{I}_{\mathrm{O}},\mathcal{J}_{\mathrm{O}},\mathcal{T}_{\mathrm{O}}|\mathcal{Z},\mathcal{C},\mathcal{B},\delta')}{P(\delta|\mu_\delta,\sigma_\delta^2)P(\mathcal{J}_{\mathrm{a}},\mathcal{I}_{\mathrm{O}},\mathcal{J}_{\mathrm{O}},\mathcal{T}_{\mathrm{O}}|\mathcal{Z},\mathcal{C},\mathcal{B},\delta)} & \text{if} < 1\\ 1 \quad \text{else}\end{cases}\tag{31}$$

By taking the log, we obtain the log of acceptance ratio:

$$\begin{aligned}
&\log\Big(\mathcal{N}(\delta';\mu_\delta,\sigma_\delta^2)\Big)-\log\Big(\mathcal{N}(\delta;\mu_\delta,\sigma_\delta^2)\Big)\\
&+\sum_{d=1}^{D}\Bigg(\sum_{i\in\mathcal{A}}\Big(-\log\Big(\Big(\prod_{j\in\mathcal{A}_{\setminus i}}\Big(\exp\{\delta'+\log(\lambda_{ij}^{(d)})\}+1\Big)\Big)-1\Big)+\sum_{j\in\mathcal{A}_{\setminus i}}(\delta'+\log(\lambda_{ij}^{(d)}))J_{ij}^{(d)}\Big)\\
&\qquad-\sum_{i\in\mathcal{A}}\Big(-\log\Big(\Big(\prod_{j\in\mathcal{A}_{\setminus i}}\Big(\exp\{\delta+\log(\lambda_{ij}^{(d)})\}+1\Big)\Big)-1\Big)+\sum_{j\in\mathcal{A}_{\setminus i}}(\delta+\log(\lambda_{ij}^{(d)}))J_{ij}^{(d)}\Big)\Bigg),
\end{aligned}\tag{32}$$

and the corresponding log of acceptance ratio is

$$\log(\text{Acceptance Probability})=\min(\text{Equation (33)},\,0).\tag{33}$$

C Conditional probability of $\mathcal{Z}$

$$\begin{aligned}
&P(\boldsymbol{w}^{(d)},\boldsymbol{z}^{(d)}|\mathcal{W}_{\setminus d},\mathcal{Z}_{\setminus d},\beta,\boldsymbol{u},\alpha,\boldsymbol{m})\\
&\propto\prod_{n=1}^{N^{(d)}}P(z_n^{(d)}=k,w_n^{(d)}=w|\mathcal{W}_{\setminus d,n},\mathcal{Z}_{\setminus d,n},\beta,\boldsymbol{u},\alpha,\boldsymbol{m})
\end{aligned}$$

To obtain the Gibbs sampling equation, we need to obtain an expression for $P(z_n^{(d)} = k, w_n^{(d)} = w|\mathcal{W}_{\setminus d,n}, \mathcal{Z}_{\setminus d,n}, \beta, \boldsymbol{u}, \alpha, \boldsymbol{m})$, From Bayes' theorem and Gamma identity $\Gamma(k+1) = k\Gamma(k)$,

$$
\begin{aligned}
&P(z_n^{(d)} = k, w_n^{(d)} = w|\mathcal{W}_{\setminus d,n}, \mathcal{Z}_{\setminus d,n}, \beta, \boldsymbol{u}, \alpha, \boldsymbol{m}) \\
&\propto \frac{P(\mathcal{W}, \mathcal{Z}|\beta, \boldsymbol{u}, \alpha, \boldsymbol{m})}{P(\mathcal{W}_{\setminus d,n}, \mathcal{Z}_{\setminus d,n}|\beta, \boldsymbol{u}, \alpha, \boldsymbol{m})} \\
&\propto \frac{\prod_{k=1}^{K} \frac{\prod_{w=1}^{W} \Gamma(N_{wk}^{WK} + \beta u_w)}{\Gamma(\sum_{w=1}^{W} N_{wk}^{WK} + \beta)} \times \prod_{k=1}^{K} \frac{\Gamma(N_{k|d} + \alpha m_k)}{\Gamma(N_{\cdot|d} + \alpha)}}{\prod_{k=1}^{K} \frac{\prod_{w=1}^{W} \Gamma(N_{wk,\setminus d,n}^{WK} + \beta u_w)}{\Gamma(\sum_{w=1}^{W} N_{wk,\setminus d,n}^{WK} + \beta)} \times \prod_{k=1}^{K} \frac{\Gamma(N_{k|d,\setminus d,n} + \alpha m_k)}{\Gamma(N_{\cdot|d,\setminus d,n} + \alpha)}} \\
&\propto \frac{N_{wk,\setminus d,n}^{WK} + \frac{\beta}{W}}{\sum_{w=1}^{W} N_{wk,\setminus d,n}^{WK} + \beta} \times \frac{N_{k|d,\setminus d,n} + \alpha m_k}{N^{(d)} - 1 + \alpha}
\end{aligned}
$$

Then, same as for LDA, we also know that the topic assignment $z_n^{(d)} = k$ is obtained by:

$$
P(z_n^{(d)} = k|w_n^{(d)} = w, \mathcal{W}_{\setminus d,n}, \mathcal{Z}_{\setminus d,n}, \beta, \boldsymbol{u}, \alpha, \boldsymbol{m}) \propto \frac{N_{k|d,\setminus d,n} + \alpha m_k}{N^{(d)} - 1 + \alpha}
$$

In addition, the conditional probability that a new word generated in the document would be $w_n^{(d)} = w$, given that it is generated from topic $z_n^{(d)} = k$ is obtained by:

$$
P(w_m^{(d)} = w|z_m^{(d)} = k, \mathcal{W}_{\setminus d,n}, \mathcal{Z}_{\setminus d,nm}, \beta, \boldsymbol{u}, \alpha, \boldsymbol{m}) \propto \frac{N_{wk,\setminus d,n}^{WK} + \frac{\beta}{W}}{\sum_{w=1}^{W} N_{wk,\setminus d,n}^{WK} + \beta}
$$