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

Bomin Kim[1], Zachary Jones[1], Bruce Desmarais[1], and Hanna Wallach[2,3]

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

May 9, 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 networks consisting of text-valued edges that arise 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 in response to this need [**ASK HANNA FOR CITES**]. The exponential random graph model (ERGM) (Robins et al., 2007; Chatterjee et al., 2013; Hunter et al., 2008), which is the canonical model for network structure as it is flexible enough to account for nearly any pattern of tie formation (Desmarais and Cranmer, 2017), has yet to be adapted to the setting of time-stamped text-valued edges. 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.

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 (i.e., the content component), 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 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 [**popularity cites**], reciprocity [**reciprocity cites**], and transitivity [**transitivity cites**]. 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 Generative Process

Assume we have a collection of documents, consisting of D number of 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 receiver of the email: an integer $i^{(d)} \in \{1, ..., A\}$ indicates the identity of the sender out of A number of actors (or nodes) and an integer vector $J^{(d)} = \{j_r^{(d)}\}_{r=1}^{|J^{(d)}|}$ indicates the identity of the receiver (or receivers) out of $A-1$ number of actors (by excluding self-loop), where $|J^{(d)}| \in \{1, ..., A-1\}$ denotes the total number of the receivers. Next, $t^{(d)}$ is the (unix time-based) timestamp of the email d, and $\boldsymbol{w}^{(d)} = \{w_m^{(d)}\}_{n=1}^{N^{(d)}}$ is a set of tokens that comprise the text of the email. In this section, we illustrate how the words $\boldsymbol{w}^{(d)}$ are generated according to the variant of latent Dirichlet allocation (Blei et al., 2003), and then how the rest three components, $(i^{(d)}, J^{(d)}, t^{(d)})$, are simulatneously generated from the stochastic intensity and topic assignments of the document. 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 is a simple addition of the interaction pattern assignment to the existing generative process of Latent Dirichlet Allocation Blei et al. (2003). This concept is also very similar to the content-partitioned multinetwork embeddings (CPME) model [**ASK BRUCE FOR CITES**], in a way that each topic k is also associated with a cluster assignment c_k, where c_k can take one of $c = \{1, ..., C\}$ values.

First we generate the global (corpus-wide) variables. There are two main sets of global variables: those that describe the topics people talk about 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. $\boldsymbol{\phi}^{(k)} \sim \text{Dir}(\beta, \mathbf{u})$ [**See Algorithm 1**]
   - A "topic" k is characterized by a discrete distribution over V word types with probability vector $\phi^{(k)}$. A symmetric Dirichlet prior $\mathbf{u}$ with the concentration parameter β is placed.

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 $\boldsymbol{x}_t^{(c)}(i, j)$; however, we assume that our generative process is conditioned on these covariates.

2. $\boldsymbol{b}^{(c)} \sim \text{Normal}(\mathbf{0}, \sigma^2 I_P)$ [**See Algorithm 2**]:
   - The vector of coefficients depends on the interaction pattern c. This means that there is variation in the degree of influence from the dynamic network statistics.

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

3. $c_k \sim \text{Unif}(1, C)$ [**See Algorithm 3**]:
   - Each topic is associated with a single interaction pattern.

Then, we generate the local variables. 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

Here 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 tie generating process in 2.3. Assume that each document $d \in \{1, ..., D\}$ has an $A \times A$ stochastic intensity (or hazard) matrix of $\boldsymbol{\lambda}^{(d)}(t) = \{\{\lambda_{ij}^{(d)}(t)\}_{i=1}^{A}\}_{j=1}^{A}$, where $\lambda_{ij}^{(d)}(t)$=P{for document d, $i \to j$ occurs in time interval $[t, t+dt)$, given that it has not ben sent until time t}.

To calculate the distribution of interaction patterns within a document, we estimate the (empirical) proportion of the words in document d which are assigned the topics corresponding to the interaction pattern c from 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 shows up in the document d, and $\sum_{c=1}^{C} p_c^{(d)} = 1$.

Now, we define the $(i,j)^{th}$ element of the stochastic intensity matrix $\boldsymbol{\lambda}^{(d)}(t)$ forms:

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

where $p_c^{(d)}$ is as defined above, $\lambda_0^{(c)}$ is the baseline hazards for the interaction pattern c, $\boldsymbol{b}^{(c)}$ is an unknown vector of coefficients in $\boldsymbol{R}^p$ corresponding to the interaction pattern c, $\boldsymbol{x}_t^{*(c)}(i,j)$ is a vector of the p-dimensional dynamic network statistics for directed edge (i,j) at time t corresponding to the interaction pattern c, and $\mathcal{A}_{\setminus i}$ is the predictable receiver set of sender i within the set of all possible actors $\mathcal{A}$ (no self-loop). After we include intercept and set $\boldsymbol{x}_t^{(c)}(i,j) = (\mathbf{1}, \boldsymbol{x}_t^{*(c)}(i,j))$, we rewrite (2) as:

$$\lambda_{ij}^{(d)}(t) = \sum_{c=1}^{C} p_c^{(d)} \cdot \exp\Big\{\boldsymbol{b}^{(c)T} \boldsymbol{x}_t^{(c)}(i,j)\Big\} \cdot 1\{j \in \mathcal{A}_{\setminus i}\}, \tag{3}$$

where now the first element of $\boldsymbol{b}^{(c)}$ corresponds to the baseline intensity of interaction pattern c. This λ can be seen as the weighted average of stochastic intensities across the interaction patterns. Next, since multicast interactions—those involving a single sender but multiple receivers—are allowed for this model, we expand the rate of interaction between sender i and the receivers in a set J by taking the average of $\boldsymbol{b}^{(c)T} \boldsymbol{x}_t^{(c)}(i,j)$ terms across the receivers:

$$\lambda_{iJ}^{(d)}(t) = \sum_{c=1}^{C} p_c^{(d)} \cdot \exp\Big\{\frac{1}{|J|} \sum_{j \in J} \boldsymbol{b}^{(c)T} \boldsymbol{x}_t^{(c)}(i,j)\Big\} \cdot \prod_{j \in J} 1\{j \in \mathcal{A}_{\setminus i}\}. \tag{4}$$

In case of single receivers ($|J| = 1$), Equation (4) is reduced to Equation (3), thus in the following sections we use the Equation (4) of multicast cases as a general form of the stochastic intensity between the sender and receivers.

2.3 Tie Generating Process

We assume the following generative process for each document d in a corpus D:

1. (Data augmentation) For each sender $i \in \{1, ..., A\}$, create a list of receivers J_i by applying the Bernoulli probabilities to every $j \in \mathcal{A}_{\setminus i}$

$$I(i \to j) \sim \text{Ber}\Big(\frac{\delta\lambda_{ij}^{(d)}}{\delta\lambda_{ij}^{(d)} + 1}\Big), \tag{5}$$

   where the probability is from a inverse log-logit link function ($= \frac{1}{1+\exp(-\log(\delta\lambda_{ij}^{(d)}))}$) using $\lambda_{ij}^{(d)}$ evaluated at time $t_+^{(d-1)}$, and δ is a positive tuning parameter to control the number of recipients. Note that $_+$ denotes including the timepoint itself, meaning that λ_{ij} is obtained using the history of interactions until and including the timestamp $t^{(d-1)}$.
   (i.e. $\lambda_{ij}^{(d)} = \sum_{c=1}^{C} p_c^{(d)} \cdot \exp\Big\{\boldsymbol{b}^{(c)T}\boldsymbol{x}^{(c)}_{t_+^{(d-1)}}(i,j)\Big\} \cdot 1\{j \in \mathcal{A}_{\setminus i}\}$)

2. For sender with nonempty receiver sets $i \in \mathcal{A}_{\setminus J_i = \emptyset}$, generate the time increments

$$\Delta T_{iJ_i} \sim \text{Exp}(\lambda_{iJ_i}^{(d)}). \tag{6}$$

   where δ also works as a tuning parameter to controling the urgency of the document to be sent and $\lambda_{iJ_i}^{(d)} = \sum_{c=1}^{C} p_c^{(d)} \cdot \exp\Big\{\frac{1}{|J_i|}\sum_{j\in J_i} \boldsymbol{b}^{(c)T}\boldsymbol{x}^{(c)}_{t_+^{(d-1)}}(i,j)\Big\} \cdot \prod_{j\in J_i} 1\{j \in \mathcal{A}_{\setminus i}\}$ as before.

3. Set timestamp, sender, and receivers simultaneously:

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

   Note: $i^{(0)} = J^{(0)} = \emptyset$ and $t^{(0)} = 0$.

2.4 Dynamic covariates to measure network effects

For the network statistics $\boldsymbol{x}_t^{(c)}(i,j)$ of Equation (3), we use 9 different effects as components of $\boldsymbol{x}_t^{(c)}(i,j)$, (intercept, outdegree, indegree, send, receive, 2-send, 2-receive, sibling, and cosibling) to measure popularity, centrality, reciprocity, and transitivity. First, we have intercept term to estimate the baseline intensity for each interaction pattern $c = 1, ..., C$:

1. $\textbf{intercept}^{(c)} = 1$

Next, 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 by setting $\Delta_l = (6 \text{ hours}) \times 4^l$ for $l = 1, ..., L-1$ so that for t in this range 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 results in Perry and Wolfe (2013), we decided to not to include the iuse the last interval, history before 16 days ago, considering the diminishing effects of previous email interactions. Also, we do not include the binary indicator terms for each statistics, since those terms do not reflect the recency. The specification of these dynamic network covariates could be reformulated based on the objectives of each study. Thus, in this paper, the dyadic efffects are defined as

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

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

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

5. $\textbf{receive}_{t,l}^{(c)}(i,j) = \sum_{d:t^{(d)} \in I_t^{(l)}} p_c^{(d)} \cdot I\{j \to i\}$

for $l = 1, ..., L-1$ and for $c = 1, ..., C$.

For the triadic effects involve pairs of messages, we calculate 3×3 time windows but define the interval specific statistics based on the interval where the triads are closed. (Refer to Figure 1. below) Therefore, $c = 1, ..., C$ we define the triadic effects (NOTE: $l_1 = 1, ..., 3$ and $l_2 = 1, ..., 3$)

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

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

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

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

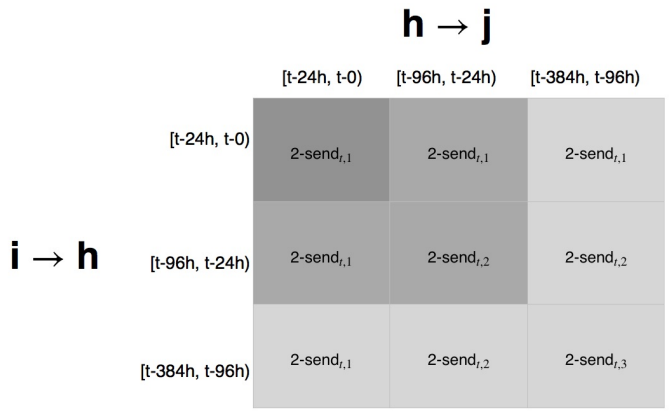

Figure 1: Example of 2-send statistic defined for each interval $l = 1, ..., 3$. Cells with same shades sum up to one statistic, based on when the triads are "closed".

In this setting, $\boldsymbol{x}_t^{(c)}(i,j)$ consists of:

```
"intercept"
"outdegree1"   "outdegree2"   "outdegree3"   "indegree1"    "indegree2"    "indegree3"
"send1"        "send2"        "send3"        "receive1"     "receive2"     "receive3"
"2-send1"      "2-send2"      "2-send3"      "2-receive1"   "2-receive2"   "2-receive3"
"sibling1"     "sibling2"     "sibling3"     "cosibling1"   "cosibling2"   "cosibling3"
```

and the corresponding dimension of $\boldsymbol{b}^{(c)}$ becomes $P = 1 + 8 \times 3 = 25$ for each interaction pattern $c = 1, ..., C$.

2.5 Joint Generative Process of Document

Below are the joint generative process for each document in a corpus D, integrating 2.1, 2.2, 2.3, and 2.4. Figure 1 is the corresponding plate notation of IPTM.

Algorithm 1 Topic Word Distributions

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

Algorithm 2 Interaction Pattern Parameters

for *c=1 to C* **do**
  draw $\boldsymbol{b}^{(c)} \sim \text{Normal}(\mathbf{0}, \sigma^2 I_P)$
end

Algorithm 3 Topic Interaction Pattern Assginments

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

Algorithm 4 Document Generating Process

for *d=1 to D* **do**
  draw $\bar{N}^{(d)} = \max(1, N^{(d)})$
  draw $\boldsymbol{\theta}^{(d)} \sim \text{Dir}(\alpha, \boldsymbol{m})$
  for *n=1 to* $\bar{N}^{(d)}$ **do**
    draw $z_n^{(d)} \sim \text{Multinom}(\boldsymbol{\theta}^{(d)})$
    if $N^{(d)} > 0$ **then**
      draw $w_n^{(d)} \sim \text{Multinom}(\boldsymbol{\phi}^{(z_n^{(d)})})$
    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 $\boldsymbol{x}^{(c)}_{t_+^{(d-1)}}(i,j)$, the network statisitcs evaluated at time $t_+^{(d-1)}$
        set $\lambda_{ij}^{(d)} = \sum_{c=1}^{C} p_c^{(d)} \cdot \exp\left\{\boldsymbol{b}^{(c)T} \boldsymbol{x}^{(c)}_{t_+^{(d-1)}}(i,j)\right\} \cdot 1\{j \in \mathcal{A}_{\setminus i}\}$
        draw $\text{I}(i \to j) \sim \text{Ber}(\frac{\delta \lambda_{ij}^{(d)}}{\delta \lambda_{ij}^{(d)} + 1})$
      end
    end
    draw $\Delta T_{iJ_i} \sim \text{Exp}(\lambda_{iJ_i}^{(d)})$
  end
  set $t^{(d)} = t^{(d-1)} + \min(\Delta T_{iJ_i})$, $i^{(d)} = i_{\min(\Delta T_{iJ_i})}$, and $J^{(d)} = J_{i^{(d)}}$
end

2.6 Joint Generative Process of Document

Below are the joint generative process for each document in a corpus D, integrating 2.1, 2.2, 2.3, and 2.4. Figure 1 is the corresponding plate notation of IPTM.

Algorithm 5 Topic Word Distributions

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

Algorithm 6 Interaction Pattern Parameters

for *c=1 to C* **do**
  draw $\boldsymbol{b}^{(c)} \sim \text{Normal}(\mathbf{0}, \sigma^2 I_P)$
end

Algorithm 7 Topic Interaction Pattern Assginments

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

Algorithm 8 Document Generating Process

for *d=1 to D* **do**
  draw $\bar{N}^{(d)} = \max(1, N^{(d)})$
  draw $\boldsymbol{\theta}^{(d)} \sim \text{Dir}(\alpha, \boldsymbol{m})$
  for *n=1 to* $\bar{N}^{(d)}$ **do**
    draw $z_n^{(d)} \sim \text{Multinom}(\boldsymbol{\theta}^{(d)})$
    if $N^{(d)} > 0$ **then**
      draw $w_n^{(d)} \sim \text{Multinom}(\boldsymbol{\phi}^{(z_n^{(d)})})$
    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 $\boldsymbol{x}^{(c)}_{t_+^{(d-1)}}(i,j)$, the network statisitcs evaluated at time $t_+^{(d-1)}$
        set $\lambda_{ij}^{(d)} = \sum_{c=1}^{C} p_c^{(d)} \cdot \exp\left\{\boldsymbol{b}^{(c)T} \boldsymbol{x}^{(c)}_{t_+^{(d-1)}}(i,j)\right\} \cdot 1\{j \in \mathcal{A}_{\setminus i}\}$
        draw $\text{I}(i \to j) \sim \text{Ber}(\frac{\delta \lambda_{ij}^{(d)}}{\delta \lambda_{ij}^{(d)} + 1})$
      end
    end
    draw $\Delta T_{iJ_i} \sim \text{Exp}(\lambda_{iJ_i}^{(d)})$
  end
  set $t^{(d)} = t^{(d-1)} + \min(\Delta T_{iJ_i})$, $i^{(d)} = i_{\min(\Delta T_{iJ_i})}$, and $J^{(d)} = J_{i^{(d)}}$
end

3 Inference

The joint generative process implies a particular factorization of the joint distribution over the variables $\Phi = \{\boldsymbol{\phi}^{(k)}\}_{k=1}^{K}, \Theta = \{\boldsymbol{\theta}^{(d)}\}_{d=1}^{D}, \mathcal{Z} = \{\boldsymbol{z}^{(d)}\}_{d=1}^{D}, \mathcal{W} = \{\boldsymbol{w}^{(d)}\}_{d=1}^{D}, \mathcal{C} = \{c_k\}_{k=1}^{K}, \mathcal{B} = \{\boldsymbol{b}^{(c)}\}_{c=1}^{C}, \mathcal{J} = \{\{J_i^{(d)}\}_{i=1}^{A}\}_{d=1}^{D}, \mathcal{T} = \{\{t_{iJ_i}^{(d)}\}_{i=1}^{A}\}_{d=1}^{D}$, and $\mathcal{P} = \{(i,J,t)^{(d)}\}_{d=1}^{D}$ given the observed $\mathcal{P}' = \{\{(i,J,t)^{(d')}\}_{d'=1}^{d}\}_{d=1}^{D}$ and the hyperparameters $(\beta, \boldsymbol{u}, \alpha, \boldsymbol{m}, \sigma^2, \delta)$:

$$\begin{aligned} &P(\Phi, \Theta, \mathcal{Z}, \mathcal{W}, \mathcal{C}, \mathcal{B}, \mathcal{J}, \mathcal{T}, \mathcal{P} | \mathcal{P}', \beta, \boldsymbol{u}, \alpha, \boldsymbol{m}, \sigma^2, \delta) \\ &= P(\Phi|\beta, \boldsymbol{u}) P(\Theta|\alpha, \boldsymbol{m}) P(\mathcal{Z}|\Theta) P(\mathcal{W}|\mathcal{Z}, \Phi) P(\mathcal{C}) P(\mathcal{B}|\mathcal{C}, \sigma^2) \\ &\quad \times P(\mathcal{J}|\mathcal{Z}, \mathcal{C}, \mathcal{B}, \mathcal{P}', \delta) P(\mathcal{T}|\mathcal{Z}, \mathcal{C}, \mathcal{B}, \mathcal{J}, \mathcal{P}') P(\mathcal{P}|\mathcal{J}, \mathcal{T}, \mathcal{P}'). \end{aligned} \tag{8}$$

We can simplify this further by integreting out Φ and Θ using Dirichlet-multinomial conjugacy:

$$\begin{aligned} &P(\mathcal{Z}, \mathcal{W}, \mathcal{C}, \mathcal{B}, \mathcal{J}, \mathcal{T}, \mathcal{P} | \mathcal{P}', \beta, \boldsymbol{u}, \alpha, \boldsymbol{m}, \sigma^2, \delta) \\ &= P(\mathcal{Z}|\alpha, \boldsymbol{m}) P(\mathcal{W}|\mathcal{Z}, \beta, \boldsymbol{u}) P(\mathcal{C}) P(\mathcal{B}|\mathcal{C}, \sigma^2) P(\mathcal{J}|\mathcal{Z}, \mathcal{C}, \mathcal{B}, \mathcal{P}', \delta) P(\mathcal{T}|\mathcal{Z}, \mathcal{C}, \mathcal{B}, \mathcal{J}, \mathcal{P}') P(\mathcal{P}|\mathcal{J}, \mathcal{T}, \mathcal{P}'). \end{aligned} \tag{9}$$

Note that since $p_c^{(d)}$ is a deterministic function of $(\mathcal{Z}, \mathcal{C})$, $\boldsymbol{x}_{t_+^{(d-1)}}^{(c)}$ is a deterministic function of $(p_c^{(d)}, \mathcal{P}')$ and $\boldsymbol{\lambda}^{(d)}$ is a deterministic function of $(p_c^{(d)}, \boldsymbol{x}_{t_+^{(d-1)}}^{(c)}, \mathcal{B})$, we do not include them as variables in the joint distribution.

What we observe in the data is $\mathcal{W}$ and $\mathcal{P}$. We want to infer the values of all the latent variables $(\mathcal{Z}, \mathcal{C}, \mathcal{B})$ that most likely would have generated the data we observe, if we believe the generative process. Currently the variables $(\mathcal{J}, \mathcal{T}, \mathcal{P}, \mathcal{P}')$ are redundant, thus we re-define the variables considering the data augmentation: $\mathcal{J}_\text{a} = \{\{J_i^{(d)}\}_{i \neq i_o^{(d)}}\}_{d=1}^{D}$, and $\mathcal{T}_\text{a} = \{\{t_{iJ_i}^{(d)}\}_{i \neq i_o^{(d)}}\}_{d=1}^{D}$ of the augmented data and $\mathcal{I}_\text{o} = \{i_o^{(d)}\}_{d=1}^{D}$, $\mathcal{J}_\text{o} = \{J_o^{(d)}\}_{d=1}^{D}$, and $\mathcal{T}_\text{o} = \{t^{(d)}\}_{d=1}^{D}$ from the observed data.

Now, our inference goal is to draw samples from the posterior distribution

$$\begin{aligned} &P(\mathcal{Z}, \mathcal{C}, \mathcal{B}, \mathcal{J}_\text{a}, \mathcal{T}_\text{a} | \mathcal{W}, \mathcal{I}_\text{o}, \mathcal{J}_\text{o}, \mathcal{T}_\text{o}, \beta, \boldsymbol{u}, \alpha, \boldsymbol{m}, \sigma^2, \delta) \\ &\propto P(\mathcal{Z}, \mathcal{C}, \mathcal{B}, \mathcal{W}, \mathcal{J}_\text{a}, \mathcal{T}_\text{a}, \mathcal{I}_\text{o}, \mathcal{J}_\text{o}, \mathcal{T}_\text{o} | \beta, \boldsymbol{u}, \alpha, \boldsymbol{m}, \sigma^2, \delta) \\ &= P(\mathcal{Z}|\alpha, \boldsymbol{m}) P(\mathcal{C}) P(\mathcal{B}|\mathcal{C}, \sigma^2) P(\mathcal{W}|\mathcal{Z}, \beta, \boldsymbol{u}) P(\mathcal{J}_\text{a}, \mathcal{T}_\text{a}, \mathcal{I}_\text{o}, \mathcal{J}_\text{o}, \mathcal{T}_\text{o} | \mathcal{Z}, \mathcal{C}, \mathcal{B}, \delta) \end{aligned} \tag{10}$$

As mentioned earlier in Section 2.3, 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:

$$\begin{aligned} &P(\mathcal{J}_\text{a}, \mathcal{T}_\text{a}, \mathcal{I}_\text{o}, \mathcal{J}_\text{o}, \mathcal{T}_\text{o} | \mathcal{Z}, \mathcal{C}, \mathcal{B}, \delta) \\ &= \prod_{d=1}^{D} P(\mathcal{J}_\text{a}^{(d)}, \mathcal{T}_\text{a}^{(d)}, i_\text{o}^{(d)}, J_\text{o}^{(d)}, t_\text{o}^{(d)} | \mathcal{I}_\text{o}^{(-d)}, \mathcal{J}_\text{o}^{(-d)}, \mathcal{T}_\text{o}^{(-d)}, \mathcal{Z}, \mathcal{C}, \mathcal{B}, \delta) \\ &= \prod_{d=1}^{D} P(\mathcal{J}_\text{a}^{(d)}, \mathcal{T}_\text{a}^{(d)}, i_\text{o}^{(d)}, J_\text{o}^{(d)}, t_\text{o}^{(d)} | \mathcal{I}_\text{o}^{(<d)}, \mathcal{J}_\text{o}^{(<d)}, \mathcal{T}_\text{o}^{(<d)}, \mathcal{Z}, \mathcal{C}, \mathcal{B}, \delta). \end{aligned} \tag{11}$$

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

Now we tackle the problem by deriving $P(\mathcal{J}_\text{a}^{(d)}, \mathcal{T}_\text{a}^{(d)}, i_\text{o}^{(d)}, J_\text{o}^{(d)}, t_\text{o}^{(d)} | \mathcal{I}_\text{o}^{(<d)}, \mathcal{J}_\text{o}^{(<d)}, \mathcal{T}_\text{o}^{(<d)}, \mathcal{Z}, \mathcal{C}, \mathcal{B}, \delta)$ 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 (5); second is the generation of the observed time increment $\Delta T^{(d)} = t^{(d)} - t^{(d-1)}$ from the observed sender-receiver pairs $(i_\text{o}^{(d)}, J_\text{o}^{(d)})$, which corresponds to the Exponential part of 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:

$$
\begin{aligned}
&P(\mathcal{J}_{\text{a}}^{(d)}, \mathcal{T}_{\text{a}}^{(d)}, i_{\text{o}}^{(d)}, J_{\text{o}}^{(d)}, t_{\text{o}}^{(d)} | \mathcal{I}_{\text{o}}^{(<d)}, \mathcal{J}_{\text{o}}^{(<d)}, \mathcal{T}_{\text{o}}^{(<d)}, \mathcal{Z}, \mathcal{C}, \mathcal{B}, \delta) \\
&= P\Big(\text{latent receivers generation}\Big) \times P\Big(\text{latent time generation}\Big) \times P\Big(\text{choose the observed}\Big) \\
&= P\Big(\text{from Equation (5)}\Big) \times P\Big(\text{from Equation (6)}\Big) \times P\Big(\text{from Equation (7)}\Big) \\
&= \prod_{i \in \mathcal{A}} \prod_{j \in \mathcal{A}_{\setminus i}} \Big(I(i \to j) \sim \text{Ber}(\frac{\delta \lambda_{ij}^{(d)}}{\delta \lambda_{ij}^{(d)} + 1}) \Big) \times \prod_{i \in \mathcal{A}} \Big(\Delta T_{iJ_i}^{(d)} \sim \text{Exp}(\delta \lambda_{iJ_i}^{(d)}) \Big) \times \prod_{i \in \mathcal{A}_{\setminus i_o^{(d)}}} P\Big(\Delta T_{iJ_i}^{(d)} > \Delta T_{i_o^{(d)} J_o^{(d)}} \Big) \\
&= \Big(\prod_{i \in \mathcal{A}} \prod_{j \in \mathcal{A}_{\setminus i}} (\frac{\delta \lambda_{ij}^{(d)}}{\delta \lambda_{ij}^{(d)} + 1})^{I(j \in J_i^{(d)})} (\frac{1}{\delta \lambda_{ij}^{(d)} + 1})^{1 - I(j \in J_i^{(d)})} \Big) \times \Big(\prod_{i \in \mathcal{A}} \lambda_{iJ_i}^{(d)} e^{-\Delta T_{iJ_i}^{(d)} \lambda_{iJ_i}^{(d)}} \Big) \times \Big(\prod_{i \in \mathcal{A}_{\setminus i_o^{(d)}}} e^{-\Delta T_{i_o^{(d)} J_o^{(d)}}^{(d)} \lambda_{iJ_i^{(d)}}^{(d)}} \Big) \\
&= \Big(\prod_{i \in \mathcal{A}} \prod_{j \in \mathcal{A}_{\setminus i}} \frac{(\delta \lambda_{ij}^{(d)})^{I(j \in J_i)}}{\delta \lambda_{ij}^{(d)} + 1} \Big) \times \Big(\lambda_{i_o^{(d)} J_o^{(d)}}^{(d)} e^{-\Delta T_{i_o^{(d)} J_o^{(d)}}^{(d)} \lambda_{i_o^{(d)} J_o^{(d)}}^{(d)}} \Big) \times \Big(\prod_{i \in \mathcal{A}_{\setminus i_o^{(d)}}} \lambda_{iJ_i}^{(d)} e^{-(\Delta T_{iJ_i}^{(d)} + \Delta T_{i_o^{(d)} J_o^{(d)}}^{(d)}) \lambda_{iJ_i}^{(d)}} \Big),
\end{aligned}
\tag{12}
$$

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

$$
\begin{aligned}
&\int_0^\infty \cdots \int_0^\infty \Big(\prod_{i \in \mathcal{A}_{\setminus i_o^{(d)}}} \lambda_{iJ_i}^{(d)} e^{-(\Delta T_{iJ_i}^{(d)} + \Delta T_{i_o^{(d)} J_o^{(d)}}^{(d)}) \lambda_{iJ_i}^{(d)}} \Big) d\Delta T_{1J_1}^{(d)} \cdots d\Delta T_{AJ_A}^{(d)} \\
&= \prod_{i \in \mathcal{A}_{\setminus i_o^{(d)}}} e^{-\Delta T_{i_o^{(d)} J_o^{(d)}}^{(d)} \lambda_{iJ_i}^{(d)}} \Big(\int_0^\infty \lambda_{iJ_i}^{(d)} e^{-\Delta T_{iJ_i}^{(d)} \lambda_{iJ_i}^{(d)}} d\Delta T_{iJ_i}^{(d)} \Big) \\
&= \prod_{i \in \mathcal{A}_{\setminus i_o^{(d)}}} e^{-\Delta T_{i_o^{(d)} J_o^{(d)}}^{(d)} \lambda_{iJ_i}^{(d)}} \Big(\Big[- e^{-\Delta T_{iJ_i}^{(d)} \lambda_{iJ_i}^{(d)}} \Big]_{\Delta T_{iJ_i}^{(d)} = 0}^{\infty} \Big) \\
&= e^{-\Delta T_{i_o^{(d)} J_o^{(d)}}^{(d)} \sum_{i \in \mathcal{A}_{\setminus i_o^{(d)}}} \lambda_{iJ_i}^{(d)}},
\end{aligned}
\tag{13}
$$

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

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

where $i \in \mathcal{A}_{\setminus J_i^{(d)} = \emptyset}$ implies the senders with non-empty receivers. If there is no latent receivers for the latent sender i ($|J_i| = 0$), the corresponding λ_{iJ_i} is zero, so we can ignore those cases. Note that the observed sender i_o always have non-empty receiver set J_o (i.e. $|J_o| > 0$). Finally for implementation, we need to compute these equations in log space to prevent underflow:

$$
\begin{aligned}
&\log\Big(P(\mathcal{J}_{\text{a}}^{(d)}, i_{\text{o}}^{(d)}, J_{\text{o}}^{(d)}, t_{\text{o}}^{(d)} | \mathcal{I}_{\text{o}}^{(<d)}, \mathcal{J}_{\text{o}}^{(<d)}, \mathcal{T}_{\text{o}}^{(<d)}, \mathcal{Z}, \mathcal{C}, \mathcal{B}, \delta) \Big) \\
&= \Big(\sum_{i \in \mathcal{A}} \sum_{j \in \mathcal{A}_{\setminus i}} I(j \in J_i) \cdot \log(\delta \lambda_{iJ_i}^{(d)}) - \log(\delta \lambda_{ij}^{(d)} + 1) \Big) + \Big(\log(\lambda_{i_o^{(d)} J_o^{(d)}}^{(d)}) \Big) - \Big(\Delta T_{i_o^{(d)} J_o^{(d)}}^{(d)} \sum_{i \in \mathcal{A}_{\setminus J_i^{(d)} = \emptyset}} \lambda_{iJ_i^{(d)}}^{(d)} \Big).
\end{aligned}
\tag{15}
$$

To actually perform inference, we therefore want to use Gibbs sampling, where we will sequentially resample the values of each of our latent variables from their posterior distribution, conditional on all of our other variables.

3.1 Resampling the augmented data $\mathcal{J}_{\mathtt{a}}$

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 $J_{\mathtt{a},i}^{(d)}$ be the $(A-1)$ length vector of indicators (0/1) representing the latent receivers corresponding to the sender i in the docuemnt d. Here, for each sender i and receiver $j \in \mathcal{A}_{\setminus i}$, we are going to resample each element $J_{\mathtt{a},ij}^{(d)}$, one at a time. For a latent sender $i \in \mathcal{A}_{\setminus i_o^{(d)}}$ and a latent receiver $j \in \mathcal{A}_{\setminus i}$, we derive the conditional probability:

$$\begin{aligned}
&P(\mathcal{J}_{\mathtt{a},ij}^{(d)} = J_{\mathtt{a},ij}^{(d)} | \mathcal{J}_{\mathtt{a},\setminus ij}^{(d)}, i_{\mathtt{o}}^{(d)}, J_{\mathtt{o}}^{(d)}, t_{\mathtt{o}}^{(d)}, \mathcal{I}_{\mathtt{o}}^{(<d)}, \mathcal{J}_{\mathtt{o}}^{(<d)}, \mathcal{T}_{\mathtt{o}}^{(<d)}, \mathcal{Z}, \mathcal{C}, \mathcal{B}, \delta) \\
&\propto P(\mathcal{J}_{\mathtt{a},ij}^{(d)} = J_{\mathtt{a},ij}^{(d)}, i_{\mathtt{o}}^{(d)}, J_{\mathtt{o}}^{(d)}, t_{\mathtt{o}}^{(d)} | \mathcal{J}_{\mathtt{a},\setminus ij}^{(d)}, \mathcal{I}_{\mathtt{o}}^{(<d)}, \mathcal{J}_{\mathtt{o}}^{(<d)}, \mathcal{T}_{\mathtt{o}}^{(<d)}, \mathcal{Z}, \mathcal{C}, \mathcal{B}, \delta) \\
&= P(\mathcal{J}_{\mathtt{a},ij}^{(d)} = J_{\mathtt{a},ij}^{(d)} | \mathcal{I}_{\mathtt{o}}^{(<d)}, \mathcal{J}_{\mathtt{o}}^{(<d)}, \mathcal{T}_{\mathtt{o}}^{(<d)}, \mathcal{Z}, \mathcal{C}, \mathcal{B}, \delta) \\
&\qquad \times P(i_{\mathtt{o}}^{(d)}, J_{\mathtt{o}}^{(d)}, t_{\mathtt{o}}^{(d)} | \mathcal{J}_{\mathtt{a},ij}^{(d)} = J_{\mathtt{a},ij}^{(d)}, \mathcal{J}_{\mathtt{a},\setminus ij}^{(d)}, \mathcal{I}_{\mathtt{o}}^{(<d)}, \mathcal{J}_{\mathtt{o}}^{(<d)}, \mathcal{T}_{\mathtt{o}}^{(<d)}, \mathcal{Z}, \mathcal{C}, \mathcal{B}, \delta) \\
&\propto (\delta \lambda_{ij}^{(d)})^{I(j \in J_{\mathtt{a},i}^{(d)})} \times e^{-\Delta T_{i_o^{(d)} J_o^{(d)}}^{(d)} \lambda_{iJ_{\mathtt{a},i}^{(d)}}^{(d)}}.
\end{aligned} \tag{16}$$

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

$$\begin{aligned}
&P(\mathcal{J}_{\mathtt{a},ij}^{(d)} = 1 | \mathcal{J}_{\mathtt{a},\setminus ij}^{(d)} i_{\mathtt{o}}^{(d)}, J_{\mathtt{o}}^{(d)}, t_{\mathtt{o}}^{(d)}, \mathcal{I}_{\mathtt{o}}^{(<d)}, \mathcal{J}_{\mathtt{o}}^{(<d)}, \mathcal{T}_{\mathtt{o}}^{(<d)}, \mathcal{Z}, \mathcal{C}, \mathcal{B}, \delta) \\
&\propto \delta \lambda_{ij}^{(d)} \times e^{-\Delta T_{i_o^{(d)} J_o^{(d)}}^{(d)} \lambda_{iJ_{\mathtt{a},i[+j]}^{(d)}}^{(d)}},
\end{aligned} \tag{17}$$

where $\lambda_{iJ_{\mathtt{a},i[+j]}^{(d)}}^{(d)}$ meaning that $\lambda_{iJ_{\mathtt{a},i}^{(d)}}^{(d)} = \sum_{c=1}^{C} p_c^{(d)} \cdot \exp\Big\{ \frac{1}{|J_i|} \sum_{j \in J_i} \boldsymbol{b}^{(c)T} \boldsymbol{x}_{t_+^{(d-1)}}^{(c)}(i,j) \Big\} \cdot \prod_{j \in J_i} 1\{j \in \mathcal{A}_{\setminus i}\}$ is calculated with j^{th} element of $J_{\mathtt{a},i}^{(d)}$ fixed as 1. On the other hand,

$$\begin{aligned}
&P(\mathcal{J}_{\mathtt{a},ij}^{(d)} = 0 | \mathcal{J}_{\mathtt{a},\setminus ij}^{(d)} i_{\mathtt{o}}^{(d)}, J_{\mathtt{o}}^{(d)}, t_{\mathtt{o}}^{(d)}, \mathcal{I}_{\mathtt{o}}^{(<d)}, \mathcal{J}_{\mathtt{o}}^{(<d)}, \mathcal{T}_{\mathtt{o}}^{(<d)}, \mathcal{Z}, \mathcal{C}, \mathcal{B}, \delta) \\
&\propto e^{-\Delta T_{i_o^{(d)} J_o^{(d)}}^{(d)} \lambda_{iJ_{\mathtt{a},i[-j]}^{(d)}}^{(d)}},
\end{aligned} \tag{18}$$

where $\lambda_{iJ_{\mathtt{a},i[-j]}^{(d)}}^{(d)}$ meaning similarly that $\lambda_{iJ_{\mathtt{a},i}^{(d)}}^{(d)}$ is calculated with j^{th} element of $J_{\mathtt{a},i}^{(d)}$ fixed as 0.

Now we can use Gibbs sampling using the two probabilities, Equation (17) and Equation (18). Again, we would calculate the probabilities in the log space to prevent from numerical underflow. Other than Gibbs sampling, equivalently, we can apply a random sampling from the Bernoulli probability

$$\begin{aligned}
\frac{P(\mathcal{J}_{\mathtt{a},ij}^{(d)} = 1 | \cdot)}{P(\mathcal{J}_{\mathtt{a},ij}^{(d)} = 1 | \cdot) + P(\mathcal{J}_{\mathtt{a},ij}^{(d)} = 0 | \cdot)} &= \frac{\delta \lambda_{ij}^{(d)} \times e^{-\Delta T_{i_o^{(d)} J_o^{(d)}}^{(d)} \lambda_{iJ_{\mathtt{a},i[+j]}^{(d)}}^{(d)}}}{\delta \lambda_{ij}^{(d)} \times e^{-\Delta T_{i_o^{(d)} J_o^{(d)}}^{(d)} \lambda_{iJ_{\mathtt{a},i[+j]}^{(d)}}^{(d)}} + e^{-\Delta T_{i_o^{(d)} J_o^{(d)}}^{(d)} \lambda_{iJ_{\mathtt{a},i[-j]}^{(d)}}^{(d)}}} \\
&= \frac{1}{1 + \frac{1}{\delta \lambda_{ij}^{(d)}} e^{-\Delta T_{i_o^{(d)} J_o^{(d)}}^{(d)} \left(\lambda_{iJ_{\mathtt{a},i[-j]}^{(d)}}^{(d)} - \lambda_{iJ_{\mathtt{a},i[+j]}^{(d)}}^{(d)} \right)}}.
\end{aligned} \tag{19}$$

Moreover, to further protect from numerical underflow and overflow, we use the numerical approximation of $f(y) = \log(1 + e^y)$ as below:

$$f(y) = \begin{cases} y & y > 35 \\ e^y & y < -10 \\ \log(1 + e^y) & \text{otherwise,} \end{cases}$$

where $y = -\Delta T_{i_o^{(d)} J_o^{(d)}}^{(d)} \left(\lambda_{iJ_{\mathtt{a},i[-j]}^{(d)}}^{(d)} - \lambda_{iJ_{\mathtt{a},i[+j]}^{(d)}}^{(d)} \right) - \log(\delta \lambda_{ij}^{(d)})$ in this case.

3.2 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:

$$\begin{aligned}
&P(z_n^{(d)} = k|\mathcal{Z}_{\setminus d,n}, \mathcal{C}, \mathcal{B}, \mathcal{W}, \mathcal{I}_{\mathrm{a}}, \mathcal{J}_{\mathrm{a}}, \mathcal{I}_{\mathrm{o}}, \mathcal{J}_{\mathrm{o}}, \mathcal{T}_{\mathrm{o}}, \beta, \boldsymbol{u}, \alpha, \boldsymbol{m}, \sigma^2, \delta) \\
&\propto P(z_n^{(d)} = k, w_n^{(d)}, \mathcal{I}_{\mathrm{a}}^{(d)}, \mathcal{J}_{\mathrm{a}}^{(d)}, \mathcal{I}_{\mathrm{o}}^{(d)}, \mathcal{J}_{\mathrm{o}}^{(d)}, \mathcal{T}_{\mathrm{o}}^{(d)}|\mathcal{Z}_{\setminus d,n}, \mathcal{C}, \mathcal{B}, \mathcal{W}_{\setminus d,n}, \mathcal{I}_{\mathrm{o}}^{(-d)}, \mathcal{J}_{\mathrm{o}}^{(-d)}, \mathcal{T}_{\mathrm{o}}^{(-d)}, \beta, \boldsymbol{u}, \alpha, \boldsymbol{m}, \sigma^2, \delta) \\
&= P(z_n^{(d)} = k|\mathcal{Z}_{\setminus d,n}, \alpha, \boldsymbol{m}) P(w_n^{(d)}|z_n^{(d)} = k, \mathcal{W}_{\setminus d,n}, \mathcal{Z}_{\setminus d,n}, \beta, \boldsymbol{u}) \\
&\quad \times P(\mathcal{I}_{\mathrm{a}}^{(d)}, \mathcal{J}_{\mathrm{a}}^{(d)}, \mathcal{I}_{\mathrm{o}}^{(d)}, \mathcal{J}_{\mathrm{o}}^{(d)}, \mathcal{T}_{\mathrm{o}}^{(d)}|z_n^{(d)} = k, \mathcal{Z}_{\setminus d,n}, \mathcal{C}, \mathcal{B}, \sigma^2, \delta)
\end{aligned} \tag{20}$$

where the subscript "$\setminus d, n$" denotes the exclsuion of position n in d^{th} document and the subscript "$\setminus d$" and "$(-d)$" denotes the exclsuion of d^{th} document. We know that:

$$P(z_n^{(d)} = k|\mathcal{Z}_{\setminus d,n}, \alpha, \boldsymbol{m}) = \frac{N_{\setminus d,n}^{(k|d)} + \alpha \boldsymbol{m}_k}{N^{(d)} - 1 + \alpha} \tag{21}$$

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

$$P(w_n^{(d)}|z_n^{(d)} = k, \mathcal{W}_{\setminus d,n}, \mathcal{Z}_{\setminus d,n}, \beta, \boldsymbol{u}) = \frac{N_{\setminus d,n}^{(w_n^{(d)}|k)} + \frac{\beta}{W}}{N_{\setminus d,n}^{(k)} + \beta}, \tag{22}$$

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, in 3.1, we have shown that

$$\begin{aligned}
&P(\mathcal{I}_{\mathrm{a}}^{(d)}, \mathcal{J}_{\mathrm{a}}^{(d)}, \mathcal{I}_{\mathrm{o}}^{(d)}, \mathcal{J}_{\mathrm{o}}^{(d)}, \mathcal{T}_{\mathrm{o}}^{(d)}|z_n^{(d)} = k, \mathcal{Z}_{\setminus d,n}, \mathcal{C}, \mathcal{B}, \sigma^2, \delta) \\
&= \Big(\prod_{i \in \mathcal{A}} \prod_{j \in \mathcal{A}_{\setminus i}} \frac{(\delta \lambda_{ij}^{(d)})^{I(j \in J_i)}}{\delta \lambda_{ij}^{(d)} + 1} \Big) \times \Big(\lambda_{i_o^{(d)} J_o^{(d)}}^{(d)} \Big) \times \Big(e^{-\Delta T_{i_o^{(d)} J_o^{(d)}}^{(d)} \sum_{i \in \mathcal{A}_{\setminus J_i^{(d)} = \emptyset}} \lambda_{i J_i^{(d)}}^{(d)}} \Big).
\end{aligned} \tag{23}$$

Therefore, if $N^{(d)} > 0$, then

$$\begin{aligned}
&P(z_n^{(d)} = k|\mathcal{Z}_{\setminus d,n}, \mathcal{C}, \mathcal{B}, \mathcal{W}, \mathcal{I}_{\mathrm{a}}, \mathcal{J}_{\mathrm{a}}, \mathcal{I}_{\mathrm{o}}, \mathcal{J}_{\mathrm{o}}, \mathcal{T}_{\mathrm{o}}, \beta, \boldsymbol{u}, \alpha, \boldsymbol{m}, \sigma^2, \delta) \\
&\propto (N_{\setminus d,n}^{(k|d)} + \alpha \boldsymbol{m}_k) \times \frac{N_{\setminus d,n}^{(w_n^{(d)}|k)} + \frac{\beta}{W}}{N_{\setminus d,n}^{(k)} + \beta} \times \Big(\prod_{i \in \mathcal{A}} \prod_{j \in \mathcal{A}_{\setminus i}} \frac{(\delta \lambda_{ij}^{(d)})^{I(j \in J_i)}}{\delta \lambda_{ij}^{(d)} + 1} \Big) \times \Big(\lambda_{i_o^{(d)} J_o^{(d)}}^{(d)} \Big) \times \Big(e^{-\Delta T_{i_o^{(d)} J_o^{(d)}}^{(d)} \sum_{i \in \mathcal{A}_{\setminus J_i^{(d)} = \emptyset}} \lambda_{i J_i^{(d)}}^{(d)}} \Big),
\end{aligned} \tag{24}$$

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 last three terms, $P(\mathcal{I}_{\mathrm{a}}^{(d)}, \mathcal{J}_{\mathrm{a}}^{(d)}, \mathcal{I}_{\mathrm{o}}^{(d)}, \mathcal{J}_{\mathrm{o}}^{(d)}, \mathcal{T}_{\mathrm{o}}^{(d)}|z_n^{(d)} = k, \mathcal{Z}_{\setminus d,n}, \mathcal{C}, \mathcal{B}, \sigma^2, \delta)$, remaining.

$$\begin{aligned}
&P(z_1^{(d)} = k|\mathcal{Z}_{\setminus d,1} = \emptyset, \mathcal{C}, \mathcal{B}, \mathcal{W}, \mathcal{I}_{\mathrm{a}}, \mathcal{J}_{\mathrm{a}}, \mathcal{I}_{\mathrm{o}}, \mathcal{J}_{\mathrm{o}}, \mathcal{T}_{\mathrm{o}}, \beta, \boldsymbol{u}, \alpha, \boldsymbol{m}, \sigma^2, \delta) \\
&\propto \Big(\prod_{i \in \mathcal{A}} \prod_{j \in \mathcal{A}_{\setminus i}} \frac{(\delta \lambda_{ij}^{(d)})^{I(j \in J_i)}}{\delta \lambda_{ij}^{(d)} + 1} \Big) \times \Big(\lambda_{i_o^{(d)} J_o^{(d)}}^{(d)} \Big) \times \Big(e^{-\Delta T_{i_o^{(d)} J_o^{(d)}}^{(d)} \sum_{i \in \mathcal{A}_{\setminus J_i^{(d)} = \emptyset}} \lambda_{i J_i^{(d)}}^{(d)}} \Big).
\end{aligned} \tag{25}$$

3.3 Resampling $\mathcal{C}$

The next variable we are going to resample is the topic-interaction pattern assignments, one topic at a time. To obtain the Gibbs sampling equation, which is the posterior conditional probability for the interaction pattern $\mathcal{C}$ for k^{th} topic. We can derive the equation as below:

$$\begin{aligned}
&P(c_k = c|\mathcal{Z}, \mathcal{C}_{\setminus k}, \mathcal{B}, \mathcal{W}, \mathcal{I}_{\mathrm{a}}, \mathcal{J}_{\mathrm{a}}, \mathcal{I}_{\mathrm{o}}, \mathcal{J}_{\mathrm{o}}, \mathcal{T}_{\mathrm{o}}, \beta, \boldsymbol{u}, \alpha, \boldsymbol{m}, \sigma^2, \delta) \\
&\propto P(c_k = c, \mathcal{I}_{\mathrm{a}}, \mathcal{J}_{\mathrm{a}}, \mathcal{I}_{\mathrm{o}}, \mathcal{J}_{\mathrm{o}}, \mathcal{T}_{\mathrm{o}}|\mathcal{Z}, \mathcal{C}_{\setminus k}, \mathcal{B}, \mathcal{W}, \beta, \boldsymbol{u}, \alpha, \boldsymbol{m}, \sigma^2, \delta) \\
&\propto P(c_k = c) P(\mathcal{I}_{\mathrm{a}}, \mathcal{J}_{\mathrm{a}}, \mathcal{I}_{\mathrm{o}}, \mathcal{J}_{\mathrm{o}}, \mathcal{T}_{\mathrm{o}}|\mathcal{Z}, c_k = c, \mathcal{C}_{\setminus k}, \mathcal{B}, \sigma^2, \delta)
\end{aligned} \tag{26}$$

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

$$\begin{aligned}
&P(c_k = c|\mathcal{Z}, \mathcal{C}_{\setminus k}, \mathcal{B}, \mathcal{W}, \mathcal{I}_\text{a}, \mathcal{J}_\text{a}, \mathcal{I}_\text{o}, \mathcal{J}_\text{o}, \mathcal{T}_\text{o}, \beta, \boldsymbol{u}, \alpha, \boldsymbol{m}, \sigma^2, \delta) \\
&\propto P(\mathcal{I}_\text{a}, \mathcal{J}_\text{a}, \mathcal{I}_\text{o}, \mathcal{J}_\text{o}, \mathcal{T}_\text{o}|\mathcal{Z}, c_k = c, \mathcal{C}_{\setminus k}, \mathcal{B}, \sigma^2, \delta) \\
&= \prod_{d=1}^{D} \Bigg(\Big(\prod_{i \in \mathcal{A}} \prod_{j \in \mathcal{A}_{\setminus i}} \frac{(\delta \lambda_{ij}^{(d)})^{I(j \in J_i)}}{\delta \lambda_{ij}^{(d)} + 1} \Big) \times \Big(\lambda_{i_o^{(d)} J_o^{(d)}}^{(d)} \Big) \times \Big(e^{-\Delta T_{i_o^{(d)} J_o^{(d)}}^{(d)} \sum_{i \in \mathcal{A}_{\setminus J_i^{(d)} = \emptyset}} \lambda_{i J_i^{(d)}}^{(d)}} \Big) \Bigg),
\end{aligned} \tag{27}$$

with $c_k = c$ throughout. Note that in the product over d, we only need to consider those emails that actually use topic k; the others will have no terms involving c_k.

3.4 Resampling $\mathcal{B}$ and δ

Finally, we sequentially update $\mathcal{B} = \{\boldsymbol{b}^{(c)}\}_{c=1}^{C}$ and δ. 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 $\mathcal{B} = \{\boldsymbol{b}^{(c)}\}_{c=1}^{C}$ and δ. Then we draw a proposal $\mathcal{B}' = \{\boldsymbol{b}'^{(c)}\}_{c=1}^{C}$ and δ'. Since the steps are identical for the two variables, here we only introduce the update scheme for $\mathcal{B}$.

Under symmetric proposal distribution (such as multivariate Gaussian), we cancel out Q-ratio and then accept the new proposal probability equal to:

$$\text{Acceptance Probability} = \begin{cases} \frac{P(\mathcal{B}'|\mathcal{Z}, \mathcal{C}, \mathcal{W}, \mathcal{I}_\text{a}, \mathcal{J}_\text{a}, \mathcal{I}_\text{o}, \mathcal{J}_\text{o}, \mathcal{T}_\text{o}, \beta, \boldsymbol{u}, \alpha, \boldsymbol{m}, \sigma^2, \delta)}{P(\mathcal{B}|\mathcal{Z}, \mathcal{C}, \mathcal{W}, \mathcal{I}_\text{a}, \mathcal{J}_\text{a}, \mathcal{I}_\text{o}, \mathcal{J}_\text{o}, \mathcal{T}_\text{o}, \beta, \boldsymbol{u}, \alpha, \boldsymbol{m}, \sigma^2, \delta)} & \text{if} < 1 \\ 1 \quad \text{else} \end{cases} \tag{28}$$

After factorization, we get

$$\begin{aligned}
&\frac{P(\mathcal{B}'|\mathcal{Z}, \mathcal{C}, \mathcal{W}, \mathcal{I}_\text{a}, \mathcal{J}_\text{a}, \mathcal{I}_\text{o}, \mathcal{J}_\text{o}, \mathcal{T}_\text{o}, \beta, \boldsymbol{u}, \alpha, \boldsymbol{m}, \sigma^2, \delta)}{P(\mathcal{B}|\mathcal{Z}, \mathcal{C}, \mathcal{W}, \mathcal{I}_\text{a}, \mathcal{J}_\text{a}, \mathcal{I}_\text{o}, \mathcal{J}_\text{o}, \mathcal{T}_\text{o}, \beta, \boldsymbol{u}, \alpha, \boldsymbol{m}, \sigma^2, \delta)} \\
&= \frac{P(\mathcal{Z}, \mathcal{C}, \mathcal{B}', \mathcal{W}, \mathcal{I}_\text{a}, \mathcal{J}_\text{a}, \mathcal{I}_\text{o}, \mathcal{J}_\text{o}, \mathcal{T}_\text{o}|\beta, \boldsymbol{u}, \alpha, \boldsymbol{m}, \sigma^2, \delta)}{P(\mathcal{Z}, \mathcal{C}, \mathcal{B}, \mathcal{W}, \mathcal{I}_\text{a}, \mathcal{J}_\text{a}, \mathcal{I}_\text{o}, \mathcal{J}_\text{o}, \mathcal{T}_\text{o}|\beta, \boldsymbol{u}, \alpha, \boldsymbol{m}, \sigma^2, \delta)} \\
&= \frac{P(\mathcal{B}'|\mathcal{C}, \sigma^2) P(\mathcal{I}_\text{a}, \mathcal{J}_\text{a}, \mathcal{I}_\text{o}, \mathcal{J}_\text{o}, \mathcal{T}_\text{o}|\mathcal{Z}, \mathcal{C}, \mathcal{B}', \sigma^2, \delta)}{P(\mathcal{B}|\mathcal{C}, \sigma^2) P(\mathcal{I}_\text{a}, \mathcal{J}_\text{a}, \mathcal{I}_\text{o}, \mathcal{J}_\text{o}, \mathcal{T}_\text{o}|\mathcal{Z}, \mathcal{C}, \mathcal{B}\prime, \sigma^2, \delta)},
\end{aligned} \tag{29}$$

where $P(\mathcal{B}|\mathcal{C}, \sigma^2)$ is calculated from the product of $\boldsymbol{b}^{(c)} \sim \text{Normal}(0, \sigma^2 I_P)$ (as defined in Section 2) and $P(\mathcal{I}_\text{a}, \mathcal{J}_\text{a}, \mathcal{I}_\text{o}, \mathcal{J}_\text{o}, \mathcal{T}_\text{o}|\mathcal{Z}, \mathcal{C}, \mathcal{B}', \sigma^2, \delta)$ is Equation (25). Again, we take the log and obtain the log of acceptance ratio:

$$\begin{aligned}
&\sum_{c=1}^{C} \log\Big(\text{dmvnorm}(\boldsymbol{b}'^{(c)}; \boldsymbol{0}, \sigma_{\boldsymbol{b}}^2 I_P)\Big) - \sum_{c=1}^{C} \log\Big(\text{dmvnorm}(\boldsymbol{b}^{(c)}; \boldsymbol{0}, \sigma_{\boldsymbol{b}}^2 I_P)\Big) \\
&+ \sum_{d=1}^{D} \Big\{ \Big(\sum_{i \in \mathcal{A}} \sum_{j \in \mathcal{A}_{\setminus i}} I(j \in J_i) \cdot \log(\delta \lambda_{iJ_i}^{(d)}) - \log(\delta \lambda_{ij}^{(d)} + 1) + \log(\lambda_{i_o^{(d)} J_o^{(d)}}^{(d)}) - \Delta T_{i_o^{(d)} J_o^{(d)}}^{(d)} \sum_{i \in \mathcal{A}_{\setminus J_i^{(d)} = \emptyset}} \lambda_{i J_i^{(d)}}^{(d)} \Big) \text{ given } \boldsymbol{b}' \Big\} \\
&- \sum_{d=1}^{D} \Big\{ \Big(\sum_{i \in \mathcal{A}} \sum_{j \in \mathcal{A}_{\setminus i}} I(j \in J_i) \cdot \log(\delta \lambda_{iJ_i}^{(d)}) - \log(\delta \lambda_{ij}^{(d)} + 1) + \log(\lambda_{i_o^{(d)} J_o^{(d)}}^{(d)}) - \Delta T_{i_o^{(d)} J_o^{(d)}}^{(d)} \sum_{i \in \mathcal{A}_{\setminus J_i^{(d)} = \emptyset}} \lambda_{i J_i^{(d)}}^{(d)} \Big) \text{ given } \boldsymbol{b} \Big\},
\end{aligned} \tag{30}$$

where dmvnorm is the multivariate normal density. Then the log of acceptance ratio we have is:

$$\log(\text{Acceptance Probability}) = \min(\text{Equation (28)}, 0). \tag{31}$$

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) (i.e. log(runif(1, 0, 1))) is less than the log-acceptance probability (29), we accept the proposal. Otherwise, we reject.

After finishing the updates of $\mathcal{B}$, we move on to the updates of δ. Again, we use the Metropolis-Hastings algorithm with a proposal density Q being the Normal distribution, with variance equals to σ_δ^2 (proposal distribution variance parameters set by the user). However, this time we need a proposal distribution with support on the unit interval, since $\delta \sim \text{Beta}(a, b)$. Now we define an auxiliary variable η which come from the proposal distribution (in the real line) and take the transformation:

$$\delta = \Phi(\eta),$$

where Φ is any link function that maps a value in the real line to the value in the unit length $[0, 1]$. Here, for simplicity, we choose standard Gaussian cumulative distribution function. Given that still we cancel out the symmetric part $\phi(\eta' - \eta) = \phi(\eta - \eta')$ (where $\phi(\cdot)$ is Normal pdf), it follows that the correct acceptance ratio is

$$\text{Acceptance Probability} = \begin{cases} \frac{P(\eta'|\mathcal{Z},\mathcal{C},\mathcal{B},\mathcal{W},\mathcal{I}_\text{a},\mathcal{J}_\text{a},\mathcal{I}_\text{o},\mathcal{J}_\text{o},\mathcal{T}_\text{o},\beta,\boldsymbol{u},\alpha,\boldsymbol{m},\sigma^2)}{P(\eta|\mathcal{Z},\mathcal{C},\mathcal{B},\mathcal{W},\mathcal{I}_\text{a},\mathcal{J}_\text{a},\mathcal{I}_\text{o},\mathcal{J}_\text{o},\mathcal{T}_\text{o},\beta,\boldsymbol{u},\alpha,\boldsymbol{m},\sigma^2)} & \text{if} < 1 \\ 1 \quad \text{else} \end{cases} \tag{32}$$

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

$$\begin{aligned} &\log\Big(\text{dnorm}(\eta'; 0, \sigma_\eta^2)\Big) - \log\Big(\text{dnorm}(\eta; 0, \sigma_\eta^2)\Big) \\ &+ \sum_{d=1}^{D} \Big\{ \sum_{i \in \mathcal{A}} \sum_{j \in \mathcal{A}_{\setminus i}} I(j \in J_i) \Big(\log(\delta' \lambda_{iJ_i}^{(d)}) - \log(\delta \lambda_{iJ_i}^{(d)}) \Big) - \Big(\log(\delta' \lambda_{ij}^{(d)} + 1) - \log(\delta \lambda_{ij}^{(d)} + 1) \Big) \Big\}. \end{aligned} \tag{33}$$

Then the log of acceptance ratio we have is:

$$\log(\text{Acceptance Probability}) = \min(\text{Equation (33)}, 0). \tag{34}$$

3.5 Pseudocode

To implement the inference procedure outlined above, we provide a pseudocode for Markov Chain Monte Carlo (MCMC) sampling. Note that we use two loops, outer iteration and inner iteration, in order to avoid the label switching problem (Jasra et al., 2005), which is an issue caused by the nonidentifiability of the components under symmetric priors in Bayesian mixture modeling. When summarizing model results, we will only use the values from the last I^{th} outer loop because there is no label switching problem within the inner iteration.

Algorithm 9 MCMC

set initial values $\mathcal{Z}^{(0)}, \mathcal{C}^{(0)}$, and $(\mathcal{B}^{(0)}, \delta^{(0)})$
for *i=1 to I* **do**
  optimize α and $\boldsymbol{m}$ using hyperparameter optimization in Wallach (2008)
  for *d=1 to D* **do**
    for $i \in \mathcal{A}_{\setminus i^{(d)}}$ **do**
      sample the augmented data (i, J_i)
    end
  end
  fix $\mathcal{C} = \mathcal{C}^{(i-1)}$ and $\mathcal{B} = \mathcal{B}^{(i-1)}$
  for *n=1 to* n_1 **do**
    for *d=1 to D* **do**
      for *n=1 to* $N^{(d)}$ **do**
        calculate $p^{\mathcal{Z}}|\text{others} = (p_1, ..., p_K)$ using Equation (20) or Equation (21)
        draw of $z_n^{(d)} \sim \text{multinomial}(p^{\mathcal{Z}})$
      end
    end
  end
  fix $\mathcal{Z} = \mathcal{Z}^{(i)}$ and $\mathcal{B} = \mathcal{B}^{(i-1)}$
  for *n=1 to* n_2 **do**
    for *k=1 to K* **do**
      calculate $p^{\mathcal{C}}|\text{others} = (p_1, ..., p_C)$ using Equation (23)
      draw $C_k \sim \text{multinomial}(p^{\mathcal{C}})$
    end
  end
  fix $\mathcal{C} = \mathcal{C}^{(i)}$, $\mathcal{Z} = \mathcal{Z}^{(i)}$, and $\mathcal{B}^{(0)}$ = last value (n_3^{th}) of $\mathcal{B}^{(i-1)}$
  for *n=1 to* n_3 **do**
    sample $\mathcal{B}$ from proposal distribution
    accept-reject using Equation (28) and (29)
  end
  sample δ from proposal distribution
  accept-reject using Equation (33) and (34)
end
summarize the results using the last value of $\mathcal{C}$, the last value of $\mathcal{Z}$, and the last n_3 length chain of $\mathcal{B}$

4 Getting It Right (GiR)

4.1 Backward generative process

We need to define a "backward" generative process in order to perform Geweke's "Getting it Right" test (Geweke, 2004) 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 (token topic assignments, topic interaction pattern assignments, and interaction pattern parameters) as input , and simply condition on these to draw new data. This "backward" version of the generative process is detailed below in Algorithm 10.

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. These data structures, along with the latent variables we got by running our inference procedure (token topic assignments, topic interaction pattern assignments, interaction pattern parameters, and tuning parameter δ) are passed in to the backward generative process, which then outputs new documents with the generated token word types and (sender, recipients, timestamp) pairs.

Algorithm 10 Generate data with backward sampling

for *d=1 to D* **do**
  set $\bar{N}^{(d)} = \max(1, N^{(d)})$, where $N^{(d)}$ is known
  for *n=1 to* $\bar{N}^{(d)}$ **do**
    if $N^{(d)} > 0$ **then**
      for *v=1 to V* **do**
        $\phi_n^{(d)}[v] = \frac{NKV_{v,z_n^{(d)}} + \beta \boldsymbol{u}_v}{NK_k + \beta}$
      end
      draw $w_n^{(d)} \sim \phi_n^{(d)}$
      $NKV_{w_n^{(d)}, z_n^{(d)}} += 1$
    end
  end
  for *i=1 to A* **do**
    for *j=1 to A* **do**
      if $j \neq i$ **then**
        calculate $\boldsymbol{x}^{(c)}_{t_+^{(d-1)}}(i,j)$, the network statisitcs evaluated at time $t_+^{(d-1)}$
        set $\lambda_{ij}^{(d)} = \sum_{c=1}^{C} p_c^{(d)} \cdot \exp\Big\{ \boldsymbol{b}^{(c)T} \boldsymbol{x}^{(c)}_{t_+^{(d-1)}}(i,j) \Big\} \cdot 1\{j \in \mathcal{A}_{\setminus i}\}$
        draw $\mathrm{I}(i \to j) \sim \mathrm{Ber}(\frac{\delta \lambda_{ij}^{(d)}}{\delta \lambda_{ij}^{(d)} + 1})$
      end
    end
    draw $\Delta T_{iJ_i} \sim \mathrm{Exp}(\lambda_{iJ_i}^{(d)})$
  end
  set $t^{(d)} = t^{(d-1)} + \min(\Delta T_{iJ_i})$, $i^{(d)} = i_{\min(\Delta T_{iJ_i})}$, and $J^{(d)} = J_{i^{(d)}}$
end

The generative process for backward sampling and the corresponding implementation is built upon those of CPME. Further details on the process of Getting It Right test and how we compare the forward and backward samples can be found in the CPME document.

4.2 Issue of intitial state of $\boldsymbol{x}_t^{(c)}$

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 10 (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 $t^{(d)} > 384$.

4.3 GiR implementaion details

While we tried a number of different parameter combinations in the course of testing, we outline our standard setup below.

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 (network statistics) = "dyadic"
- prior for $\boldsymbol{b}^{(c)} \sim \text{MVN}_P(\boldsymbol{0}, I_P)$
- prior for $\delta \sim \text{Beta}(1, 1)$
- I (outer iteration) = 1
- n_1 (Gibbs sampling iteration of $\mathcal{Z}$) =1
- n_2 (Gibbs sampling iteration of $\mathcal{C}$) =1
- n_3 (M-H sampling iteration of $\mathcal{B}$) = 50
- burn (M-H sampling burn-in of $\mathcal{B}$)= 0
- thin (M-H sampling thinning of $\mathcal{B}$)= 1
- σ^2_Q (proposal variance for $\mathcal{B}$) = 0.25
- n_4 (M-H sampling iteration of δ) = 50
- σ^2_Q (proposal variance for δ) = 0.25

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 $(\boldsymbol{b}_p^{(1)}, ..., \boldsymbol{b}_p^{(C)})$ for every $p = 1, ..., P$,
2. Mean number of recipients (indirect measure of δ),
3. Mean of time-increments $t^{(d)} - t^{(d-1)}$ for every $d = 2, ...,$ nDocs,
4. Mean topic-interaction pattern assignment (for interaction patterns indexed by 1 and 2),
5. Number of tokens in topics assigned to each interaction pattern $c = 1, ..., C$, ?
6. Number of tokens assigned to each topic $k = 1, ..., K$,
7. Number of tokens assigned to each unique word type $w = 1, ..., W$.

4.4 GiR Results

Having saved a set of samples (collected 10^5 forward and backward samples), we compare them. To do so, we generated PP (Probability-Probability) plots for each of the 21 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. The reason we calculate a greater number of quantiles for the interaction pattern statistics is because they take on continuous values, while the other statistics can only take on a relatively small number of distinct values.

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.

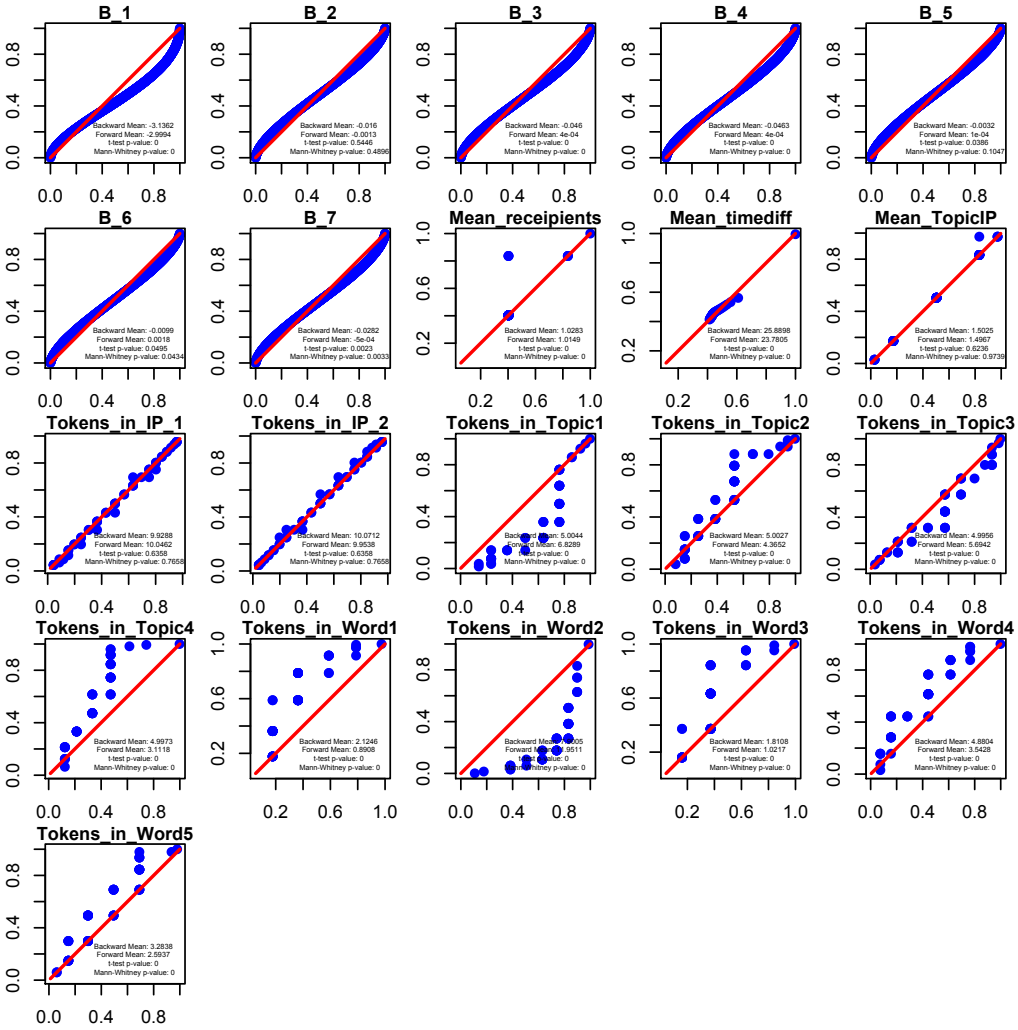

Figure 2: Getting it Right test result

Possible failure reasons:

- Baseline intensity (B_1) not identifiable when every pair of nodes havee some sort of history (this is for sure when we have small number of nodes as in GiR case) and it may have affected all other B values.
- Tokens in word esitmate seems to be all failure. No idea at all. Need to double check for possible coding errors.

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

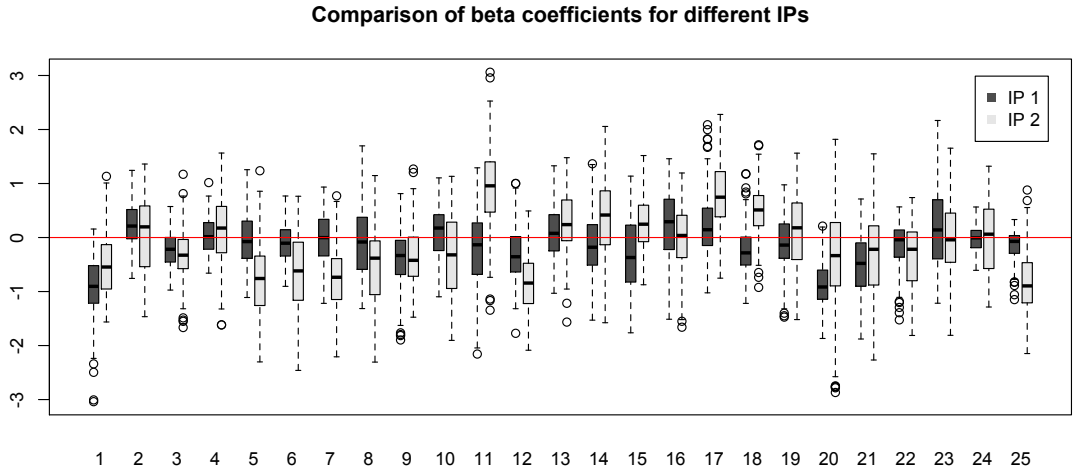

Figure 3: Posterior distribution of $\boldsymbol{b}^{(c)}$ for Vance county emails

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?

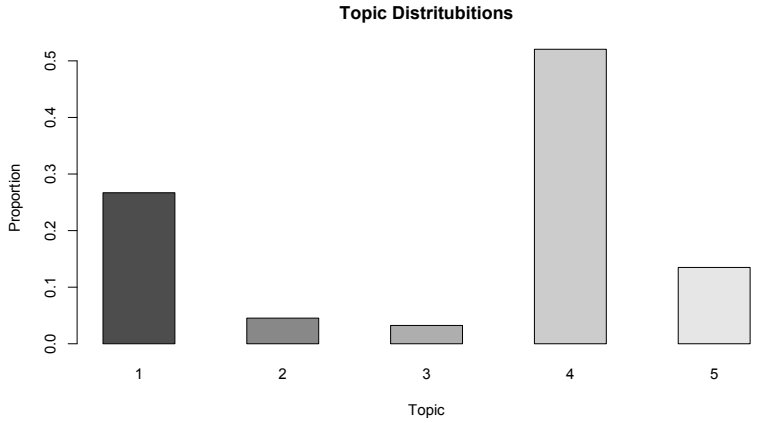

Figure 4: Posterior distribution of $\mathcal{Z}$ for Vance county emails

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.

References

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

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.

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

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.

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

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.

Jasra, A., Holmes, C., and Stephens, D. (2005). Markov chain monte carlo methods and the label switching problem in bayesian mixture modeling. *Statistical Science*, pages 50–67.

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

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

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*.

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.

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

Wallach, H. M., Mimno, D. M., and McCallum, A. (2009). Rethinking lda: Why priors matter. In *Advances in neural information processing systems*, pages 1973–1981.

APPENDIX

A Notations in IPTM

Sender of the d^{th} document	$i^{(d)}$	Scalar
Receivers of the d^{th} document	$J^{(d)}$	$\lvert J^{(d)}\rvert$-dimensional vector
Individual receiver of the d^{th} document	$j^{(d)}$	Scalar
Time of the d^{th} document	$t^{(d)}$	Scalar
Authors of the corpus	$\mathcal{A}$	Set
Number of authors	A	Scalar
Number of documents	D	Scalar
Number of words in the d^{th} document	$N^{(d)}$	Scalar
Number of topics	K	Scalar
Vocabulary size	W	Scalar
Number of interaction patterns	C	Scalar
Number of words assigned to word and topic	N^{WK}	Scalar
Interaction pattern of the k^{th} topic	C_k	Scalar
Tuning parameter in tie generative process	δ	Scalar
Poisson parameter for number of words $N^{(d)}$	ζ	Scalar
Words in the d^{th} document	$\boldsymbol{w}^{(d)}$	$N^{(d)}$-dimensional vector
n^{th} word in the d^{th} document	$w_n^{(d)}$	n^{th} component of $\boldsymbol{w}^{(d)}$
Topic assignments in the d^{th} document	$\boldsymbol{z}^{(d)}$	$N^{(d)}$-dimensional vector
Topic assignments for n^{th} word in the d^{th} document	$z_n^{(d)}$	n^{th} component of $\boldsymbol{z}^{(d)}$
Dirichlet concentration prior for document topic distribution	α	Scalar
Dirichlet base prior for document topic distribution	$\boldsymbol{m}$	K-dimensional vector
Dirichlet concentration prior for topic word distribution	β	Scalar
Dirichlet base prior for topic word distribution	$\boldsymbol{u}$	W-dimensional vector
Variance of Normal prior	σ^2	Scalar
Probabilities of the words given topic k	$\boldsymbol{\phi}^{(k)}$	W-dimensional vector
Probabilities of the topics given the d^{th} document	$\boldsymbol{\theta}^{(d)}$	K-dimensional vector
Coefficient of the intensity process given interaction pattern c	$\boldsymbol{b}^{(c)}$	p-dimensional vector
Network statistics for (i,j) at time t given interaction pattern c	$\boldsymbol{x}_t^{(c)}(i,j)$	p-dimensional vector
Stochastic intensity of document d at time t	$\boldsymbol{\lambda}^{(d)}(t)$	$A \times A$ matrix
Stochastic intensity of (i,j) in document d at time t	$\lambda_{ij}^{(d)}(t)$	Scalar
Proportion of topics in document d corresponding to interaction pattern c	$p_c^{(d)}$	Scalar
Time increments associated to (i,J) pair	ΔT_{iJ}	Scalar

Table 3: Symbols associated with IPTM, as used in this paper

B Integrating out Φ and Θ in latent Dirichlet allocation

$$
\begin{aligned}
&P(\Phi,\Theta,\mathcal{Z},\mathcal{C},\mathcal{B},\mathcal{W},\mathcal{I}_{\mathrm{a}},\mathcal{J}_{\mathrm{a}},\mathcal{T}_{\mathrm{a}},\mathcal{I}_{\mathrm{o}},\mathcal{J}_{\mathrm{o}},\mathcal{T}_{\mathrm{o}}|\beta,\boldsymbol{u},\alpha,\boldsymbol{m},\sigma^2,\delta)\\
&\propto P(\mathcal{Z},\mathcal{C},\mathcal{B},\mathcal{W},\mathcal{I}_{\mathrm{a}},\mathcal{J}_{\mathrm{a}},\mathcal{T}_{\mathrm{a}},\mathcal{I}_{\mathrm{o}},\mathcal{J}_{\mathrm{o}},\mathcal{T}_{\mathrm{o}}|\Phi,\Theta,\sigma^2,\delta)P(\Phi,\Theta|\beta,\boldsymbol{u},\alpha,\boldsymbol{m})\\
&\propto P(\mathcal{Z}|\mathcal{W},\Theta)P(\mathcal{C})P(\mathcal{B}|\mathcal{C},\sigma^2)P(\mathcal{W}|\Phi)P(\Phi|\beta,\boldsymbol{u})P(\Theta|\alpha,\boldsymbol{m})P(\mathcal{I}_{\mathrm{a}},\mathcal{J}_{\mathrm{a}},\mathcal{T}_{\mathrm{a}},\mathcal{I}_{\mathrm{o}},\mathcal{J}_{\mathrm{o}},\mathcal{T}_{\mathrm{o}}|\mathcal{Z},\mathcal{C},\mathcal{B},\delta)\\
&=\Big[\prod_{d=1}^{D}\prod_{n=1}^{N^{(d)}}P(z_n^{(d)}|w_n^{(d)},\boldsymbol{\theta}^{(d)})\Big]\times\Big[\prod_{k=1}^{K}P(C_k)\Big]\times\Big[\prod_{c=1}^{C}P(\boldsymbol{b}^{(c)}|\sigma^2)\Big]\times\Big[\prod_{d=1}^{D}\prod_{n=1}^{N^{(d)}}P(w_n^{(d)}|\phi_{z_n^{(d)}})\Big]\times\Big[\prod_{k=1}^{K}P(\boldsymbol{\phi}^{(k)}|\beta,\boldsymbol{u})\Big]\\
&\quad\times\Big[\prod_{d=1}^{D}P(\boldsymbol{\theta}^{(d)}|\alpha,\boldsymbol{m})\Big]\times\Big[\prod_{d=1}^{D}P(\mathcal{I}_{\mathrm{a}}^{(d)},\mathcal{J}_{\mathrm{a}}^{(d)},\mathcal{T}_{\mathrm{a}}^{(d)},i_{\mathrm{o}}^{(d)},J_{\mathrm{o}}^{(d)},t_{\mathrm{o}}^{(d)}|\mathcal{I}_{\mathrm{o}}^{(<d)},\mathcal{J}_{\mathrm{o}}^{(<d)},\mathcal{T}_{\mathrm{o}}^{(<d)},\mathcal{Z},\mathcal{C},\mathcal{B},\delta)\Big]
\end{aligned}
\tag{35}
$$

Dropping the terms independent of tokens out, we further rewrite the equation (28) as below:

$$
\begin{aligned}
&\propto \Big[\prod_{d=1}^{D}\prod_{n=1}^{N^{(d)}} P(z_n^{(d)}|w_n^{(d)},\boldsymbol{\theta}^{(d)})\Big] \times \Big[\prod_{d=1}^{D}\prod_{n=1}^{N^{(d)}} P(w_n^{(d)}|\phi_{z_n^{(d)}})\Big] \times \Big[\prod_{k=1}^{K} P(\boldsymbol{\phi}^{(k)}|\beta,\boldsymbol{u})\Big] \times \Big[\prod_{d=1}^{D} P(\boldsymbol{\theta}^{(d)}|\alpha,\boldsymbol{m})\Big] \\
&\propto \Big[\prod_{d=1}^{D}\prod_{n=1}^{N^{(d)}} \phi_{w_n^{(d)} z_n^{(d)}}\Big] \times \Big[\prod_{d=1}^{D}\prod_{n=1}^{N^{(d)}} \boldsymbol{\theta}_{z_n^{(d)}}^{(d)}\Big] \\
&\quad \times \Big[\prod_{k=1}^{K}\Big(\frac{\Gamma(\sum_{w=1}^{W}\beta u_w)}{\prod_{w=1}^{W}\Gamma(\beta u_w)}\prod_{w=1}^{W}\phi_{wk}^{\beta u_w-1}\Big)\Big] \times \Big[\prod_{d=1}^{D}\Big(\frac{\Gamma(\sum_{k=1}^{K}\alpha m_k)}{\prod_{k=1}^{K}\Gamma(\alpha m_k)}\prod_{k=1}^{K}(\boldsymbol{\theta}_k^{(d)})^{\alpha m_k-1}\Big)\Big] \\
&\propto \Big[\frac{\Gamma(\sum_{w=1}^{W}\beta u_w)}{\prod_{w=1}^{W}\Gamma(\beta u_w)}\Big]^{K} \times \prod_{d=1}^{D}\Big[\frac{\Gamma(\sum_{k=1}^{K}\alpha m_k)}{\prod_{k=1}^{K}\Gamma(\alpha m_k)}\Big] \times \Big[\prod_{k=1}^{K}\prod_{w=1}^{W}\phi_{wk}^{N_{wk}^{WK}+\beta u_w-1}\Big] \times \Big[\prod_{d=1}^{D}\prod_{k=1}^{K}(\boldsymbol{\theta}_k^{(d)})^{N_{k|d}+\alpha m_k-1}\Big]
\end{aligned}
\tag{36}
$$

where N_{wk}^{WK} is the number of times the w^{th} word in the vocabulary is assigned to topic k, and $N_{k|d}$ is the number of times topic k shows up in the document d. By looking at the forms of the terms involving Θ and Φ in Equation (21), we integrate out the random variables Θ and Φ, making use of the fact that the Dirichlet distribution is a conjugate prior of multinomial distribution. Applying the well-known formula $\int\prod_{n=1}^{M}[x_m^{k_m-1}dx_m] = \frac{\prod_{n=1}^{M}\Gamma(k_m)}{\Gamma(\sum_{n=1}^{M}k_m)}$ to (22), we have:

$$
\begin{aligned}
&P(\mathcal{Z},\mathcal{C},\mathcal{B},\mathcal{W},\mathcal{I}_{\mathrm{a}},\mathcal{J}_{\mathrm{a}},\mathcal{T}_{\mathrm{a}},\mathcal{I}_{\mathrm{o}},\mathcal{J}_{\mathrm{o}},\mathcal{T}_{\mathrm{o}}|\beta,\boldsymbol{u},\alpha,\boldsymbol{m},\sigma^2,\delta) \\
&= \text{Const.}\int_{\Theta}\int_{\Phi}\Big[\prod_{k=1}^{K}\prod_{w=1}^{W}\phi_{wk}^{N_{wk}^{WK}+\beta u_w-1}\Big]\Big[\prod_{d=1}^{D}\prod_{k=1}^{K}(\boldsymbol{\theta}_k^{(d)})^{N_{k|d}+\alpha m_k-1}\Big]d\Phi d\Theta \\
&= \text{Const.}\Big[\prod_{k=1}^{K}\int_{\phi_{:k}}\prod_{w=1}^{W}\phi_{wk}^{N_{wk}^{WK}+\beta u_w-1}d\phi_{:k}\Big] \times \Big[\prod_{d=1}^{D}\int_{\theta_{:d}}\prod_{k=1}^{K}(\boldsymbol{\theta}_k^{(d)})^{N_{k|d}+\alpha m_k-1}d\theta_{:d}\Big] \\
&= \text{Const.}\Big[\prod_{k=1}^{K}\frac{\prod_{w=1}^{W}\Gamma(N_{wk}^{WK}+\frac{\beta}{W})}{\Gamma(\sum_{w=1}^{W}N_{wk}^{WK}+\beta)}\Big] \times \Big[\prod_{d=1}^{D}\frac{\prod_{k=1}^{K}\Gamma(N_{k|d}+\alpha m_k)}{\Gamma(N_{\cdot|d}+\alpha)}\Big].
\end{aligned}
\tag{37}
$$

C Conditional probability of $\mathcal{Z}$

$$
\begin{aligned}
&P(\boldsymbol{w}^{(d)},\boldsymbol{z}^{(d)}|\mathcal{W}_{\setminus d},\mathcal{Z}_{\setminus d},\beta,\boldsymbol{u},\alpha,\boldsymbol{m}) \\
&\propto \prod_{n=1}^{N^{(d)}} P(z_n^{(d)}=k, w_n^{(d)}=w|\mathcal{W}_{\setminus d,n},\mathcal{Z}_{\setminus d,n},\beta,\boldsymbol{u},\alpha,\boldsymbol{m})
\end{aligned}
\tag{38}
$$

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 $\Gamma(k+1)=k\Gamma(k)$,

$$
\begin{aligned}
&P(z_n^{(d)}=k, w_n^{(d)}=w|\mathcal{W}_{\setminus d,n},\mathcal{Z}_{\setminus d,n},\beta,\boldsymbol{u},\alpha,\boldsymbol{m}) \\
&\propto \frac{P(\mathcal{W},\mathcal{Z}|\beta,\boldsymbol{u},\alpha,\boldsymbol{m})}{P(\mathcal{W}_{\setminus d,n},\mathcal{Z}_{\setminus d,n}|\beta,\boldsymbol{u},\alpha,\boldsymbol{m})} \\
&\propto \frac{\prod_{k=1}^{K}\frac{\prod_{w=1}^{W}\Gamma(N_{wk}^{WK}+\beta u_w)}{\Gamma(\sum_{w=1}^{W}N_{wk}^{WK}+\beta)} \times \prod_{k=1}^{K}\frac{\Gamma(N_{k|d}+\alpha m_k)}{\Gamma(N_{\cdot|d}+\alpha)}}{\prod_{k=1}^{K}\frac{\prod_{w=1}^{W}\Gamma(N_{wk,\setminus d,n}^{WK}+\beta u_w)}{\Gamma(\sum_{w=1}^{W}N_{wk,\setminus d,n}^{WK}+\beta)} \times \prod_{k=1}^{K}\frac{\Gamma(N_{k|d,\setminus d,n}+\alpha m_k)}{\Gamma(N_{\cdot|d,\setminus d,n}+\alpha)}} \\
&\propto \frac{N_{wk,\setminus d,n}^{WK}+\frac{\beta}{W}}{\sum_{w=1}^{W}N_{wk,\setminus d,n}^{WK}+\beta} \times \frac{N_{k|d,\setminus d,n}+\alpha m_k}{N^{(d)}-1+\alpha}
\end{aligned}
\tag{39}
$$

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

$$
P(z_n^{(d)}=k|w_n^{(d)}=w,\mathcal{W}_{\setminus d,n},\mathcal{Z}_{\setminus d,n},\beta,\boldsymbol{u},\alpha,\boldsymbol{m}) \propto \frac{N_{k|d,\setminus d,n}+\alpha m_k}{N^{(d)}-1+\alpha} \tag{40}
$$

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:

$$P(w_m^{(d)} = w | z_m^{(d)} = k, \mathcal{W}_{\setminus d,n}, \mathcal{Z}_{\setminus d,nm}, \beta, \boldsymbol{u}, \alpha, \boldsymbol{m}) \propto \frac{N^{WK}_{wk,\setminus d,n} + \frac{\beta}{W}}{\sum_{w=1}^{W} N^{WK}_{wk,\setminus d,n} + \beta} \tag{41}$$

NOTE: Using Equation (28), the unnormalized constant we use to check the model convergence and the corresponding log-constant is,

$$\begin{aligned} &\prod_{d=1}^{D} \prod_{n=1}^{N^{(d)}} P(z_n^{(d)} = k, w_n^{(d)} = w | \mathcal{W}_{\setminus d,n}, \mathcal{Z}_{\setminus d,m}, \beta, \boldsymbol{u}, \alpha, \boldsymbol{m}) \\ &\propto \prod_{d=1}^{D} \prod_{n=1}^{N^{(d)}} \frac{N^{WK}_{w_n^{(d)} z_n^{(d)},\setminus d,n} + \frac{\beta}{W}}{\sum_{w=1}^{W} N^{WK}_{w z_n^{(d)},\setminus d,n} + \beta} \times \frac{N_{k|d,\setminus d,n} + \alpha m_{z_n^{(d)}}}{N^{(d)} - 1 + \alpha}, \end{aligned} \tag{42}$$

$$\begin{aligned} &\sum_{d=1}^{D} \sum_{n=1}^{N^{(d)}} \log\Big(P(z_n^{(d)} = k, w_n^{(d)} = w | \mathcal{W}_{\setminus d,n}, \mathcal{Z}_{\setminus d,n}, \beta, \boldsymbol{u}, \alpha, \boldsymbol{m}) \Big) \\ &\propto \sum_{d=1}^{D} \sum_{n=1}^{N^{(d)}} \log\Big(N^{WK}_{w_m^{(d)} z_m^{(d)},\setminus d,n} + \frac{\beta}{W} \Big) - \log\Big(\sum_{w=1}^{W} N^{WK}_{w z_m^{(d)},\setminus d,n} + \beta \Big) \\ &+ \log\Big(N_{k|d,\setminus d,n} + \alpha m_{z_m^{(d)}} \Big) - \log\Big(N^{(d)} - 1 + \alpha \Big) \end{aligned} \tag{43}$$

D More results on Vance county

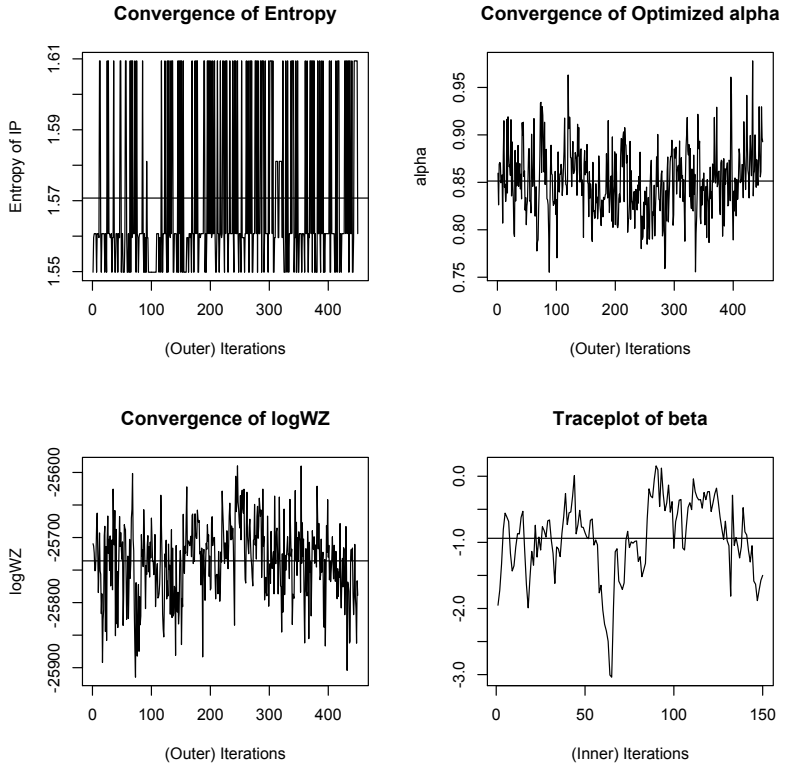

Figure 5: Convergence plot