

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

Bomin Kim<sup>1</sup>, Aaron Schein<sup>3</sup>, Bruce Desmarais<sup>1</sup>, and Hanna Wallach<sup>2,3</sup>

<sup>1</sup>Pennsylvania State University

<sup>2</sup>Microsoft Research NYC

<sup>3</sup>University of Massachusetts Amherst

June 1, 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

- Paragraph framing the problem and previewing solution: Social interactions are often, and are increasingly so, dominated by interpersonal textual communication. Methods for modeling the evolution of relationships have developed rapidly in recent decades—and are now capable of making predictions regarding tie formation in real time that are based on highly complex characterizations of network dynamics.

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 led 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.

- Add paragraph: Discussion of potential applications of our model

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 applications, 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).

- Provide an overview of what we do in this paper, summarize our results.

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 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 tie information, 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.

Assume we have a collection of documents, consisting of  $D$  number of unique documents. A single email, indexed by  $d \in \{1, \dots, D\}$ , is represented by the four components  $(i^{(d)}, J^{(d)}, t^{(d)}, \mathbf{w}^{(d)})$ . The first two are the sender and receiver of the email: an integer  $i^{(d)} \in \{1, \dots, A\}$  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  $|J^{(d)}| \in \{1, \dots, A - 1\}$  denotes the total number of the receivers. Next,  $t^{(d)}$  is the (unix time-based) timestamp of the email  $d$ , and  $\mathbf{w}^{(d)} = \{w_m^{(d)}\}_{m=1}^{N^{(d)}}$  is a set of tokens that comprise the text of the email. In this section, we illustrate how the words  $\mathbf{w}^{(d)}$  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 simultaneously 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  $d = 1, \dots, D$ .

## 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 multinet network 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  $c = \{1, \dots, C\}$  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.  $\phi^{(k)} \sim \text{Dirichlet}(\beta, \mathbf{u})$  [See Algorithm 1]
  - A “topic”  $k$  is characterized by a discrete distribution over  $V$  word types with probability vector  $\phi^{(k)}$ . A symmetric Dirichlet prior  $\mathbf{u}$  with the concentration parameter  $\beta$  is placed.

There are  $C$  interaction patterns. Each interaction pattern consists of a vector of coefficients  $\mathbf{b}^{(c)}$  in  $\mathbf{R}^P$  and a vector of  $P$ -dimensional dynamic network statistics for directed edge  $(i, j)$  at time  $t$   $\mathbf{x}_t^{(c)}(i, j)$ ; however, we assume that our generative process is conditioned on these covariates.

2.  $\mathbf{b}^{(c)} \sim \text{Multivariate Normal}(\mu_{\mathbf{b}}, \Sigma_{\mathbf{b}})$  [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.  $P$ -variate multivariate Normal prior with mean vector  $\mu_{\mathbf{b}}$  and covariance matrix  $\Sigma_{\mathbf{b}}$  is assigned.

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.

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  $\boldsymbol{\theta}^{(d)} \sim \text{Dir}(\alpha, \mathbf{m})$
- 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 the generating process in 2.3. Assume that each document  $d \in \{1, \dots, D\}$  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  $\lambda_{ij}^{(d)}(t) = P\{\text{for document } d, i \rightarrow j \text{ occurs in time interval } [t, t+dt), \text{ given that it has not been sent until time } t\}$ .

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:

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

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:

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

where  $p_c^{(d)}$  is as defined above,  $\lambda_0^{(c)}$  is the baseline hazards for the interaction pattern  $c$ ,  $\mathbf{b}^{(c)}$  is an unknown vector of coefficients in  $\mathbf{R}^p$  corresponding to the interaction pattern  $c$ ,  $\mathbf{x}_t^{(c)}(i, j)$  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).

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  $\mathbf{b}^{(c)T} \mathbf{x}_t^{(c)}(i, j)$  terms across the receivers:

$$\lambda_{iJ}^{(d)}(t) = \sum_{c=1}^C p_c^{(d)} \cdot \exp\left\{\lambda_0^{(c)} + \frac{1}{|J|} \sum_{j \in J} \mathbf{b}^{(c)T} \mathbf{x}_t^{(c)}(i, j)\right\}. \quad (3)$$

### 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  $i \in \{1, \dots, A\}$ , create binary receiver vector of length  $A - 1$ ,  $J_i^{(d)}$ , by applying the non-empty Gibbs measure to every  $j \in \mathcal{A}_{\setminus i}$ .

$$P(J_i^{(d)}) = \frac{1}{Z(\delta, \log(\lambda_i^{(d)}))} \exp\left\{\log\left(\mathbb{I}\left(\sum_{j \in \mathcal{A}_{\setminus i}} J_{ij} > 0\right)\right) + \sum_{j \in \mathcal{A}_{\setminus i}} (\delta + \log(\lambda_{ij}^{(d)})) J_{ij}^{(d)}\right\}, \quad (4)$$

where  $\delta$  is real-valued intercept that controls the expected value of  $J_{ij}^{(d)}$  (or recipient size) with its prior distribution specified as  $\text{Nomral}(\mu_\delta, \sigma_\delta^2)$ ,  $\lambda_{ij}^{(d)}$  is a positive dyad-specific function of the attributes included in the model (as defined in 2.2), and  $\lambda_i^{(d)}$  is the vector of dyadic weights in which  $i$  is the sender. Note that  $\lambda_{ij}^{(d)}$  is evaluated at time  $t_+^{(d-1)}$ , where  $+$  denotes including the timepoint itself, meaning that  $\lambda_{ij}$  for  $d^{th}$  document is obtained using the history of interactions until and including the timestamp  $t^{(d-1)}$ .

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:

$$Z(\delta, \log(\lambda_i^{(d)})) = \left( \prod_{j \in \mathcal{A}_{\setminus i}} \left( \exp\{\delta + \log(\lambda_{ij}^{(d)})\} + 1 \right) \right) - 1. \quad (5)$$

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

2. For every sender  $i \in \mathcal{A}$ , generate the time increments

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

3. Set timestamp, sender, and receivers simultaneously (NOTE:  $t^{(0)} = 0$ ):

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

## 2.4 Dynamic covariates to measure network effects

For the network statistics  $\mathbf{x}_t^{(c)}(i, j)$  of Equation (3), we use 9 different effects as components of  $\mathbf{x}_t^{(c)}(i, j)$ , (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  $c = 1, \dots, C$ :

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

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  $l = 1, \dots, L - 1$  so that for  $t$  in this range  $\Delta_l$  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 use 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 effects are defined as

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

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

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

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

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

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

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

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

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

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

In this setting, the dimension of  $\mathbf{b}^{(c)}$  becomes  $P = 1 + 8 \times 3 = 25$  for each interaction pattern  $c = 1, \dots, C$  and the corresponding  $\mathbf{x}_t^{(c)}(i, j)$  consists of:

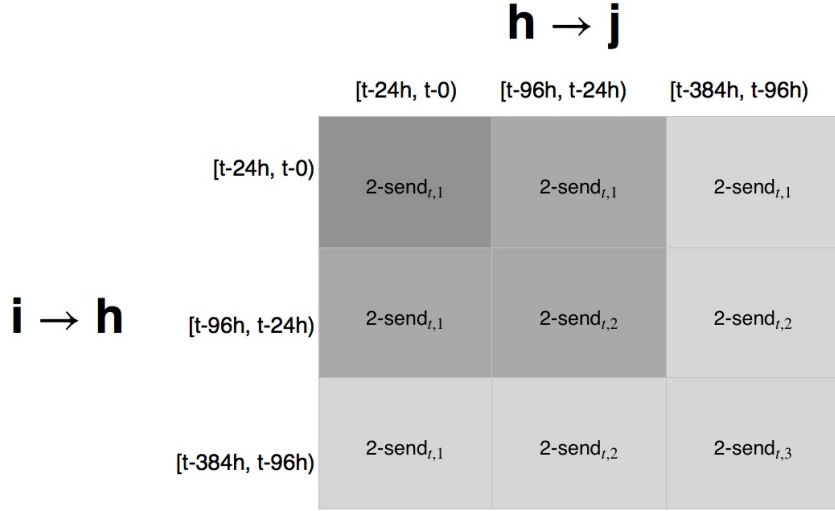


Figure 1: Example of 2-send statistic defined for each interval  $l = 1, \dots, 3$ . Cells with same shades sum up to one statistic, based on when the triads are “closed”.

```

"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"

```

## 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.

---

### Algorithm 1 Topic Word Distributions

---

```

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

```

---



---

### Algorithm 2 Interaction Pattern Parameters

---

```

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

```

---



---

### Algorithm 3 Topic Interaction Pattern Assignments

---

```

for  $k=1$  to  $K$  do
  | draw  $C_k \sim \text{Uniform}(1, C)$ 
end

```

---



---

### Algorithm 4 Recipient Size Parameter

---

```

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

```

---

---

**Algorithm 5** Document Generating Process
 

---

```

for  $d=1$  to  $D$  do
  draw  $\bar{N}^{(d)} = \max(1, N^{(d)})$ 
  draw  $\boldsymbol{\theta}^{(d)} \sim \text{Dir}(\alpha, \mathbf{m})$ 
  for  $n=1$  to  $\bar{N}^{(d)}$  do
    draw  $z_n^{(d)} \sim \text{Multinomial}(\boldsymbol{\theta}^{(d)})$ 
    if  $N^{(d)} > 0$  then
      draw  $w_n^{(d)} \sim \text{Multinomial}(\boldsymbol{\phi}^{(z_n^{(d)})})$ 
    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  $\mathbf{x}_{t_+^{(d-1)}}^{(c)}(i, j)$ , the network statisitcs evaluated at time  $t_+^{(d-1)}$ 
        set  $\lambda_{ij}^{(d)} = \sum_{c=1}^C p_c^{(d)} \cdot \exp\left\{\lambda_0^{(c)} + \mathbf{b}^{(c)T} \mathbf{x}_{t_+^{(d-1)}}^{(c)}(i, j)\right\}$ 
      end
    end
    draw  $J_i^{(d)} \sim \text{Gibbs measure}(\{\lambda_{ij}^{(d)}\}_{j=1}^A, \delta)$ 
    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  $\Phi = \{\boldsymbol{\phi}^{(k)}\}_{k=1}^K$ ,  $\Theta = \{\boldsymbol{\theta}^{(d)}\}_{d=1}^D$ ,  $\mathcal{Z} = \{\mathbf{z}^{(d)}\}_{d=1}^D$ ,  $\mathcal{W} = \{\mathbf{w}^{(d)}\}_{d=1}^D$ ,  $\mathcal{C} = \{c_k\}_{k=1}^K$ ,  $\mathcal{B} = \{\mathbf{b}^{(c)}\}_{c=1}^C$ ,  $\delta$ ,  $\mathcal{J} = \{\{J_i^{(d)}\}_{i=1}^A\}_{d=1}^D$ ,  $\mathcal{T} = \{\{t_{iJ_i}^{(d)}\}_{i=1}^A\}_{d=1}^D$ , and  $\mathcal{P} = \{(i, J, t)^{(d)}\}_{d=1}^D$  given the observed  $\mathcal{P}' = \{(i, J, t)^{(d')}\}_{d'=1}^D$  and the hyperparameters  $(\beta, \mathbf{u}, \alpha, \mathbf{m}, \mu_{\mathbf{b}}, \Sigma_{\mathbf{b}}, \mu_{\delta}, \sigma_{\delta}^2)$ :

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

We can simplify this further by integreting out  $\Phi$  and  $\Theta$  using Dirichlet-multinomial conjugacy (see APPENDIX C):

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

Note that since  $p_c^{(d)}$  is a deterministic function of  $(\mathcal{Z}, \mathcal{C})$ ,  $\mathbf{x}_{t_+^{(d-1)}}^{(c)}$  is a deterministic function of  $(p_c^{(d)}, \mathcal{P}')$  and  $\boldsymbol{\lambda}^{(d)}$  is a deterministic function of  $(p_c^{(d)}, \mathbf{x}_{t_+^{(d-1)}}^{(c)}, \mathcal{B})$ , 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  $(\mathcal{Z}, \mathcal{C}, \mathcal{B})$ .

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:  $\mathcal{J}_a = \{\{J_i^{(d)}\}_{i \neq i_o^{(d)}}\}_{d=1}^D$ , and  $\mathcal{T}_a = \{\{t_{iJ_i}^{(d)}\}_{i \neq i_o^{(d)}}\}_{d=1}^D$  of the augmented data and  $\mathcal{I}_o = \{i_o^{(d)}\}_{d=1}^D$ ,  $\mathcal{J}_o = \{J_o^{(d)}\}_{d=1}^D$ , and  $\mathcal{T}_o = \{t^{(d)}\}_{d=1}^D$  from the observed data.

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

$$\begin{aligned} & P(\mathcal{Z}, \mathcal{C}, \mathcal{B}, \delta, \mathcal{J}_a, \mathcal{T}_a | \mathcal{W}, \mathcal{I}_o, \mathcal{J}_o, \mathcal{T}_o, \beta, \mathbf{u}, \alpha, \mathbf{m}, \mu_b, \Sigma_b, \mu_\delta, \sigma_\delta^2) \\ & \propto P(\mathcal{Z}, \mathcal{C}, \mathcal{B}, \delta, \mathcal{W}, \mathcal{J}_a, \mathcal{T}_a, \mathcal{I}_o, \mathcal{J}_o, \mathcal{T}_o | \beta, \mathbf{u}, \alpha, \mathbf{m}, \mu_b, \Sigma_b, \mu_\delta, \sigma_\delta^2) \\ & = P(\mathcal{Z} | \alpha, \mathbf{m}) P(\mathcal{C}) P(\mathcal{B} | \mathcal{C}, \mu_b, \Sigma_b) P(\delta | \mu_\delta, \sigma_\delta^2) P(\mathcal{W} | \mathcal{Z}, \beta, \mathbf{u}) P(\mathcal{J}_a, \mathcal{T}_a, \mathcal{I}_o, \mathcal{J}_o, \mathcal{T}_o | \mathcal{Z}, \mathcal{C}, \mathcal{B}, \delta) \end{aligned} \quad (10)$$

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 simple as other variables. Therefore, here we provide the detailed derivation steps for the last term of conditional probability:

$$\begin{aligned} & P(\mathcal{J}_a, \mathcal{T}_a, \mathcal{I}_o, \mathcal{J}_o, \mathcal{T}_o | \mathcal{Z}, \mathcal{C}, \mathcal{B}, \delta) \\ & = \prod_{d=1}^D P(\mathcal{J}_a^{(d)}, \mathcal{T}_a^{(d)}, i_o^{(d)}, J_o^{(d)}, t_o^{(d)} | \mathcal{I}_o^{(<d)}, \mathcal{J}_o^{(<d)}, \mathcal{T}_o^{(<d)}, \mathcal{Z}, \mathcal{C}, \mathcal{B}, \delta) \\ & = \prod_{d=1}^D P(\mathcal{J}_a^{(d)}, \mathcal{T}_a^{(d)}, i_o^{(d)}, J_o^{(d)}, t_o^{(d)} | \mathcal{I}_o^{(<d)}, \mathcal{J}_o^{(<d)}, \mathcal{T}_o^{(<d)}, \mathcal{Z}, \mathcal{C}, \mathcal{B}, \delta). \end{aligned} \quad (11)$$

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

Now we tackle the problem by deriving  $P(\mathcal{J}_a^{(d)}, \mathcal{T}_a^{(d)}, i_o^{(d)}, J_o^{(d)}, t_o^{(d)} | \mathcal{I}_o^{(<d)}, \mathcal{J}_o^{(<d)}, \mathcal{T}_o^{(<d)}, \mathcal{Z}, \mathcal{C}, \mathcal{B}, \delta)$  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  $(i_o^{(d)}, J_o^{(d)})$ , 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:

$$\begin{aligned} & P(\mathcal{J}_a^{(d)}, \mathcal{T}_a^{(d)}, i_o^{(d)}, J_o^{(d)}, t_o^{(d)} | \mathcal{I}_o^{(<d)}, \mathcal{J}_o^{(<d)}, \mathcal{T}_o^{(<d)}, \mathcal{Z}, \mathcal{C}, \mathcal{B}, \delta) \\ & = P(\text{latent receivers generation}) \times P(\text{latent time generation}) \times P(\text{choose the observed}) \\ & = \prod_{i \in \mathcal{A}} \left( J_i^{(d)} \sim \text{Gibbs measure}(\{\lambda_{ij}^{(d)}\}_{j=1}^A, \delta) \right) \times \prod_{i \in \mathcal{A}} \left( \Delta T_{iJ_i}^{(d)} \sim \text{Exp}(\lambda_{iJ_i}^{(d)}) \right) \times \prod_{i \in \mathcal{A}_{\setminus i_o^{(d)}}} P(\Delta T_{iJ_i}^{(d)} > \Delta T_{i_o^{(d)}J_o^{(d)}}^{(d)}) \\ & = \left( \prod_{i \in \mathcal{A}} \frac{1}{Z(\delta, \log(\lambda_i^{(d)}))} \exp \left\{ \log(\text{I}(\sum_{j \in \mathcal{A}_{\setminus i}} J_{ij} > 0)) + \sum_{j \in \mathcal{A}_{\setminus i}} (\delta + \log(\lambda_{ij}^{(d)})) J_{ij}^{(d)} \right\} \right) \\ & \quad \times \left( \prod_{i \in \mathcal{A}} \lambda_{iJ_i}^{(d)} e^{-\Delta T_{iJ_i}^{(d)} \lambda_{iJ_i}^{(d)}} \right) \times \left( \prod_{i \in \mathcal{A}_{\setminus i_o^{(d)}}} e^{-\Delta T_{i_o^{(d)}J_o^{(d)}}^{(d)} \lambda_{i_o^{(d)}J_o^{(d)}}^{(d)}} \right) \\ & \propto \left( \prod_{i \in \mathcal{A}} \frac{1}{\left( \prod_{j \in \mathcal{A}_{\setminus i}} (\exp\{\delta + \log(\lambda_{ij}^{(d)})\} + 1) \right) - 1} \exp \left\{ \sum_{j \in \mathcal{A}_{\setminus i}} (\delta + \log(\lambda_{ij}^{(d)})) J_{ij}^{(d)} \right\} \right) \\ & \quad \times \left( \lambda_{i_o^{(d)}J_o^{(d)}}^{(d)} e^{-\Delta T_{i_o^{(d)}J_o^{(d)}}^{(d)} \lambda_{i_o^{(d)}J_o^{(d)}}^{(d)}} \right) \times \left( \prod_{i \in \mathcal{A}_{\setminus i_o^{(d)}}} \lambda_{iJ_i}^{(d)} e^{-(\Delta T_{iJ_i}^{(d)} + \Delta T_{i_o^{(d)}J_o^{(d)}}^{(d)}) \lambda_{iJ_i}^{(d)}} \right), \end{aligned} \quad (12)$$



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

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

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

$$\begin{aligned}
& P(\mathcal{J}_a^{(d)}, i_o^{(d)}, J_o^{(d)}, t_o^{(d)} | \mathcal{I}_o^{(<d)}, \mathcal{J}_o^{(<d)}, \mathcal{T}_o^{(<d)}, \mathcal{Z}, \mathcal{C}, \mathcal{B}, \delta) \\
& \propto \left( \prod_{i \in \mathcal{A}} \frac{1}{\left( \prod_{j \in \mathcal{A}_{i_i}} (\exp\{\delta + \log(\lambda_{ij}^{(d)})\} + 1) \right) - 1} \exp \left\{ \sum_{j \in \mathcal{A}_{i_i}} (\delta + \log(\lambda_{ij}^{(d)})) J_{ij}^{(d)} \right\} \right) \\
& \quad \times \left( \lambda_{i_o^{(d)}J_o^{(d)}}^{(d)} \right) \times \left( e^{-\Delta T_{i_o^{(d)}J_o^{(d)}}^{(d)} \sum_{i \in \mathcal{A}} \lambda_{iJ_i}^{(d)}} \right),
\end{aligned} \tag{14}$$

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:

$$\begin{aligned}
& \log \left( P(\mathcal{J}_a^{(d)}, i_o^{(d)}, J_o^{(d)}, t_o^{(d)} | \mathcal{I}_o^{(<d)}, \mathcal{J}_o^{(<d)}, \mathcal{T}_o^{(<d)}, \mathcal{Z}, \mathcal{C}, \mathcal{B}, \delta) \right) \\
& \propto \left( \sum_{i \in \mathcal{A}} \left( -\log \left( \left( \prod_{j \in \mathcal{A}_{i_i}} (\exp\{\delta + \log(\lambda_{ij}^{(d)})\} + 1) \right) - 1 \right) + \sum_{j \in \mathcal{A}_{i_i}} (\delta + \log(\lambda_{ij}^{(d)})) J_{ij}^{(d)} \right) \right) \\
& \quad + \left( \log(\lambda_{i_o^{(d)}J_o^{(d)}}^{(d)}) - \left( \Delta T_{i_o^{(d)}J_o^{(d)}}^{(d)} \sum_{i \in \mathcal{A}} \lambda_{iJ_i}^{(d)} \right) \right),
\end{aligned} \tag{15}$$

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

### 3.1 Resampling $\mathcal{J}_a$

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  $i \in \mathcal{A}_{i_o^{(d)}}$ . 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 latent 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:

$$P(\mathcal{J}_{ij}^{(d)} = J_{ij}^{(d)} | \mathcal{J}_{i \setminus j}^{(d)}, \mathcal{Z}, \mathcal{C}, \mathcal{B}, \delta, \mathcal{W}, \mathcal{J}_{a, -i}, \mathcal{I}_o, \mathcal{J}_o, \mathcal{T}_o, \beta, \mathbf{u}, \alpha, \mathbf{m}, \mu_b, \Sigma_b, \mu_\delta, \sigma_\delta^2), \tag{16}$$

which we can drop some independent terms and move to

$$\begin{aligned}
& P(\mathcal{J}_{ij}^{(d)} = J_{ij}^{(d)} | \mathcal{J}_{i \setminus j}^{(d)}, i_0^{(d)}, J_0^{(d)}, t_0^{(d)}, \mathcal{I}_0^{(<d)}, \mathcal{J}_0^{(<d)}, \mathcal{T}_0^{(<d)}, \mathcal{Z}, \mathcal{C}, \mathcal{B}, \delta) \\
& \propto P(\mathcal{J}_{ij}^{(d)} = J_{ij}^{(d)} | \mathcal{J}_{i \setminus j}^{(d)}, i_0^{(d)}, J_0^{(d)}, t_0^{(d)} | \mathcal{I}_0^{(<d)}, \mathcal{J}_0^{(<d)}, \mathcal{T}_0^{(<d)}, \mathcal{Z}, \mathcal{C}, \mathcal{B}, \delta) \\
& \propto \left( \frac{1}{\left( \prod_{j \in \mathcal{A}_{\setminus i}} \left( \exp\{\delta + \log(\lambda_{ij}^{(d)})\} + 1 \right) \right) - 1} \exp \left\{ \sum_{j \in \mathcal{A}_{\setminus i}} (\delta + \log(\lambda_{ij}^{(d)})) J_{ij}^{(d)} \right\} \right) \times \left( \lambda_{i_0^{(d)} J_0^{(d)}}^{(d)} \right) \times \left( e^{-\Delta T_{i_0^{(d)} J_0^{(d)}}^{(d)} \lambda_{i J_i^{(d)}}^{(d)}} \right) \\
& \propto \left( \exp \left\{ \sum_{j \in \mathcal{A}_{\setminus i}} (\delta + \log(\lambda_{ij}^{(d)})) J_{ij}^{(d)} \right\} \right) \times \left( e^{-\Delta T_{i_0^{(d)} J_0^{(d)}}^{(d)} \lambda_{i J_i^{(d)}}^{(d)}} \right),
\end{aligned} \tag{17}$$

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 (16) 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:

$$\begin{aligned}
& P(\mathcal{J}_{ij}^{(d)} = 1 | \mathcal{J}_{i \setminus j}^{(d)}, i_0^{(d)}, J_0^{(d)}, t_0^{(d)}, \mathcal{I}_0^{(<d)}, \mathcal{J}_0^{(<d)}, \mathcal{T}_0^{(<d)}, \mathcal{Z}, \mathcal{C}, \mathcal{B}, \delta) \\
& \propto \exp \left( \sum_{j \in \mathcal{A}_{\setminus i}} (\delta + \log(\lambda_{ij}^{(d)})) J_{i[+j]}^{(d)} - \Delta T_{i_0^{(d)} J_0^{(d)}}^{(d)} \lambda_{i J_{i[+j]}^{(d)}}^{(d)} \right) \\
& \propto \exp \left( \delta + \log(\lambda_{ij}^{(d)}) - \Delta T_{i_0^{(d)} J_0^{(d)}}^{(d)} \lambda_{i J_{i[+j]}^{(d)}}^{(d)} \right),
\end{aligned} \tag{18}$$

where  $J_{i[+j]}^{(d)}$  meaning that the  $j^{th}$  element of  $J_i^{(d)}$  is fixed as 1. On the other hand,

$$\begin{aligned}
& P(\mathcal{J}_{ij}^{(d)} = 0 | \mathcal{J}_{i \setminus j}^{(d)}, i_0^{(d)}, J_0^{(d)}, t_0^{(d)}, \mathcal{I}_0^{(<d)}, \mathcal{J}_0^{(<d)}, \mathcal{T}_0^{(<d)}, \mathcal{Z}, \mathcal{C}, \mathcal{B}, \delta) \\
& \propto \exp \left( \sum_{j \in \mathcal{A}_{\setminus i}} (\delta + \log(\lambda_{ij}^{(d)})) J_{i[-j]}^{(d)} - \Delta T_{i_0^{(d)} J_0^{(d)}}^{(d)} \lambda_{i J_{i[-j]}^{(d)}}^{(d)} \right) \\
& \propto \exp \left( - \Delta T_{i_0^{(d)} J_0^{(d)}}^{(d)} \lambda_{i J_{i[-j]}^{(d)}}^{(d)} \right),
\end{aligned} \tag{19}$$

where  $J_{i[-j]}^{(d)}$  meaning similarly that the  $j^{th}$  element of  $J_i^{(d)}$  is fixed as 0. Now we can use multinomial sampling using the two probabilities, Equation (17) and Equation (18), which is equivalent to Bernoulli sampling with probability  $\frac{P(\mathcal{J}_{ij}^{(d)}=1)}{P(\mathcal{J}_{ij}^{(d)}=0)+P(\mathcal{J}_{ij}^{(d)}=1)}$ . Again, we would calculate the probabilities in the log space to prevent from numerical underflow.

### 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:

$$\begin{aligned}
& P(z_n^{(d)} = k | \mathcal{Z}_{\setminus d, n}, \mathcal{C}, \mathcal{B}, \delta, \mathcal{W}, \mathcal{J}_a, \mathcal{I}_0, \mathcal{J}_0, \mathcal{T}_0, \beta, \mathbf{u}, \alpha, \mathbf{m}, \mu_b, \Sigma_b, \mu_\delta, \sigma_\delta^2) \\
& \propto P(z_n^{(d)} = k, w_n^{(d)}, \mathcal{J}_a^{(d)}, i_0^{(d)}, J_0^{(d)}, t_0^{(d)} | \mathcal{Z}_{\setminus d, n}, \mathcal{C}, \mathcal{B}, \delta, \mathcal{W}_{\setminus d, n}, \mathcal{I}_0^{(-d)}, \mathcal{J}_0^{(-d)}, \mathcal{T}_0^{(-d)}, \beta, \mathbf{u}, \alpha, \mathbf{m}) \\
& \propto P(z_n^{(d)} = k | \mathcal{Z}_{\setminus d, n}, \alpha, \mathbf{m}) P(w_n^{(d)} | z_n^{(d)} = k, \mathcal{W}_{\setminus d, n}, \mathcal{Z}_{\setminus d, n}, \beta, \mathbf{u}) \times P(\mathcal{J}_a^{(d)}, i_0^{(d)}, J_0^{(d)}, t_0^{(d)} | z_n^{(d)} = k, \mathcal{Z}_{\setminus d, n}, \mathcal{C}, \mathcal{B}, \delta)
\end{aligned} \tag{20}$$

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

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

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

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

where  $N_{\setminus d, n}^{(w_n^{(d)} | k)}$  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

$$\begin{aligned} & P(\mathcal{J}_a^{(d)}, i_o^{(d)}, J_o^{(d)}, t_o^{(d)} | z_n^{(d)} = k, \mathcal{Z}_{\setminus d, n}, \mathcal{C}, \mathcal{B}, \delta) \\ &= \left( \frac{1}{\left( \prod_{j \in \mathcal{A}_{\setminus i}} \left( \exp\{\delta + \log(\lambda_{ij}^{(d)})\} + 1 \right) \right) - 1} \exp \left\{ \sum_{j \in \mathcal{A}_{\setminus i}} (\delta + \log(\lambda_{ij}^{(d)})) J_{ij}^{(d)} \right\} \right) \times \left( \lambda_{i_o^{(d)} J_o^{(d)}}^{(d)} \right) \times \left( e^{-\Delta T_{i_o^{(d)} J_o^{(d)}}^{(d)} \lambda_{i J_i^{(d)}}^{(d)}} \right), \end{aligned} \quad (23)$$

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:

$$\begin{aligned} & P(z_n^{(d)} = k | \mathcal{Z}_{\setminus d, n}, \mathcal{C}, \mathcal{B}, \delta, \mathcal{W}, \mathcal{J}_a, \mathcal{I}_o, \mathcal{J}_o, \mathcal{T}_o, \beta, \mathbf{u}, \alpha, \mathbf{m}, \mu_b, \Sigma_b, \mu_\delta, \sigma_\delta^2) \\ & \propto (N_{\setminus d, n}^{(k|d)} + \alpha \mathbf{m}_k) \times \frac{N_{\setminus d, n}^{(w_n^{(d)} | k)} + \frac{\beta}{W}}{N_{\setminus d, n}^{(k)} + \beta} \times \left( \frac{1}{\left( \prod_{j \in \mathcal{A}_{\setminus i}} \left( \exp\{\delta + \log(\lambda_{ij}^{(d)})\} + 1 \right) \right) - 1} \exp \left\{ \sum_{j \in \mathcal{A}_{\setminus i}} (\delta + \log(\lambda_{ij}^{(d)})) J_{ij}^{(d)} \right\} \right) \\ & \quad \times \left( \lambda_{i_o^{(d)} J_o^{(d)}}^{(d)} \right) \times \left( e^{-\Delta T_{i_o^{(d)} J_o^{(d)}}^{(d)} \lambda_{i J_i^{(d)}}^{(d)}} \right), \end{aligned} \quad (24)$$

and if  $N^{(d)} = 0$ , then the first term becomes  $\alpha \mathbf{m}_k$  and disappears because it is a constant. The second term disappears since there are no tokens, thus we just have the last three terms,  $P(\mathcal{J}_a^{(d)}, i_o^{(d)}, J_o^{(d)}, t_o^{(d)} | z_n^{(d)} = k, \mathcal{Z}_{\setminus d, n}, \mathcal{C}, \mathcal{B}, \delta)$ , remaining.

$$\begin{aligned} & P(z_1^{(d)} = k | \mathcal{Z}_{\setminus d, 1} = \emptyset, \mathcal{C}, \mathcal{B}, \delta, \mathcal{W}, \mathcal{J}_a, \mathcal{I}_o, \mathcal{J}_o, \mathcal{T}_o, \beta, \mathbf{u}, \alpha, \mathbf{m}, \mu_b, \Sigma_b, \mu_\delta, \sigma_\delta^2) \\ & \propto \left( \frac{1}{\left( \prod_{j \in \mathcal{A}_{\setminus i}} \left( \exp\{\delta + \log(\lambda_{ij}^{(d)})\} + 1 \right) \right) - 1} \exp \left\{ \sum_{j \in \mathcal{A}_{\setminus i}} (\delta + \log(\lambda_{ij}^{(d)})) J_{ij}^{(d)} \right\} \right) \times \left( \lambda_{i_o^{(d)} J_o^{(d)}}^{(d)} \right) \times \left( e^{-\Delta T_{i_o^{(d)} J_o^{(d)}}^{(d)} \lambda_{i J_i^{(d)}}^{(d)}} \right). \end{aligned} \quad (25)$$

### 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, we derive the posterior conditional probability for the interaction pattern  $\mathcal{C}$  for  $k^{th}$  topic as below:

$$\begin{aligned} & P(c_k = c | \mathcal{Z}, \mathcal{C}_{\setminus k}, \mathcal{B}, \delta, \mathcal{W}, \mathcal{J}_a, \mathcal{I}_o, \mathcal{J}_o, \mathcal{T}_o, \beta, \mathbf{u}, \alpha, \mathbf{m}, \mu_b, \Sigma_b, \mu_\delta, \sigma_\delta^2) \\ & \propto P(c_k = c, \mathcal{J}_a, \mathcal{I}_o, \mathcal{J}_o, \mathcal{T}_o | \mathcal{Z}, \mathcal{C}_{\setminus k}, \mathcal{B}, \delta) \\ & \propto P(c_k = c) P(\mathcal{J}_a, \mathcal{I}_o, \mathcal{J}_o, \mathcal{T}_o | \mathcal{Z}, c_k = c, \mathcal{C}_{\setminus k}, \mathcal{B}, \delta) \end{aligned} \quad (26)$$

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

$$\begin{aligned} & P(c_k = c | \mathcal{Z}, \mathcal{C}_{\setminus k}, \mathcal{B}, \delta, \mathcal{W}, \mathcal{J}_a, \mathcal{I}_o, \mathcal{J}_o, \mathcal{T}_o, \beta, \mathbf{u}, \alpha, \mathbf{m}, \mu_b, \Sigma_b, \mu_\delta, \sigma_\delta^2) \\ & \propto P(\mathcal{J}_a, \mathcal{I}_o, \mathcal{J}_o, \mathcal{T}_o | \mathcal{Z}, c_k = c, \mathcal{C}_{\setminus k}, \mathcal{B}, \delta) \\ &= \prod_{d=1}^D \left( \left( \frac{1}{\left( \prod_{j \in \mathcal{A}_{\setminus i}} \left( \exp\{\delta + \log(\lambda_{ij}^{(d)})\} + 1 \right) \right) - 1} \exp \left\{ \sum_{j \in \mathcal{A}_{\setminus i}} (\delta + \log(\lambda_{ij}^{(d)})) J_{ij}^{(d)} \right\} \right) \right. \\ & \quad \left. \times \left( \lambda_{i_o^{(d)} J_o^{(d)}}^{(d)} \right) \times \left( e^{-\Delta T_{i_o^{(d)} J_o^{(d)}}^{(d)} \lambda_{i J_i^{(d)}}^{(d)}} \right) \right), \end{aligned} \quad (27)$$

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 $\mathcal{B}$

Finally, we sequentially update  $\mathcal{B} = \{\mathbf{b}^{(c)}\}_{c=1}^C$ . 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  $\sigma_Q^2$  (proposal distribution variance parameters set by the user), centered on the current values of  $\mathcal{B} = \{\mathbf{b}^{(c)}\}_{c=1}^C$ . Under the symmetric proposal distribution (multivariate Gaussian), we cancel out Q-ratio and then accept the new proposed value  $\mathcal{B}' = \{\mathbf{b}'^{(c)}\}_{c=1}^C$  with probability equal to:

$$\text{Acceptance Probability} = \begin{cases} \frac{P(\mathcal{B}'|\mathcal{Z}, \mathcal{C}, \delta, \mathcal{W}, \mathcal{J}_a, \mathcal{I}_o, \mathcal{J}_o, \mathcal{T}_o, \beta, \mathbf{u}, \alpha, \mathbf{m}, \mu_b, \Sigma_b, \mu_\delta, \sigma_\delta^2)}{P(\mathcal{B}|\mathcal{Z}, \mathcal{C}, \delta, \mathcal{W}, \mathcal{J}_a, \mathcal{I}_o, \mathcal{J}_o, \mathcal{T}_o, \beta, \mathbf{u}, \alpha, \mathbf{m}, \mu_b, \Sigma_b, \mu_\delta, \sigma_\delta^2)} & \text{if } < 1 \\ 1 & \text{else} \end{cases} \quad (28)$$

After factorization, we get

$$\begin{aligned} & \frac{P(\mathcal{B}'|\mathcal{Z}, \mathcal{C}, \delta, \mathcal{W}, \mathcal{J}_a, \mathcal{I}_o, \mathcal{J}_o, \mathcal{T}_o, \beta, \mathbf{u}, \alpha, \mathbf{m}, \mu_b, \Sigma_b, \mu_\delta, \sigma_\delta^2)}{P(\mathcal{B}|\mathcal{Z}, \mathcal{C}, \delta, \mathcal{W}, \mathcal{J}_a, \mathcal{I}_o, \mathcal{J}_o, \mathcal{T}_o, \beta, \mathbf{u}, \alpha, \mathbf{m}, \mu_b, \Sigma_b, \mu_\delta, \sigma_\delta^2)} \\ &= \frac{P(\mathcal{Z}, \mathcal{C}, \mathcal{B}', \delta, \mathcal{W}, \mathcal{J}_a, \mathcal{I}_o, \mathcal{J}_o, \mathcal{T}_o | \beta, \mathbf{u}, \alpha, \mathbf{m}, \mu_b, \Sigma_b, \mu_\delta, \sigma_\delta^2)}{P(\mathcal{Z}, \mathcal{C}, \mathcal{B}, \delta, \mathcal{W}, \mathcal{J}_a, \mathcal{I}_o, \mathcal{J}_o, \mathcal{T}_o | \beta, \mathbf{u}, \alpha, \mathbf{m}, \mu_b, \Sigma_b, \mu_\delta, \sigma_\delta^2)} \\ &= \frac{P(\mathcal{B}'|\mathcal{C}, \mu_b, \Sigma_b)P(\mathcal{J}_a, \mathcal{I}_o, \mathcal{J}_o, \mathcal{T}_o | \mathcal{Z}, \mathcal{C}, \mathcal{B}', \delta)}{P(\mathcal{B}|\mathcal{C}, \mu_b, \Sigma_b)P(\mathcal{J}_a, \mathcal{I}_o, \mathcal{J}_o, \mathcal{T}_o | \mathcal{Z}, \mathcal{C}, \mathcal{B}, \delta)}, \end{aligned} \quad (29)$$

where  $P(\mathcal{B}|\mathcal{C}, \mu_b, \Sigma_b)$  is calculated from the product of  $\mathbf{b}^{(c)} \sim \text{Multivariate Normal}(\mu_b, \Sigma_b)$  over the interaction patterns  $c \in \{1, \dots, C\}$  (as defined in Section 2) and  $P(\mathcal{J}_a, \mathcal{I}_o, \mathcal{J}_o, \mathcal{T}_o | \mathcal{Z}, \mathcal{C}, \mathcal{B}, \delta)$  is the same as Equation (27). Again, we take the log and obtain the log of acceptance ratio:

$$\begin{aligned} & \sum_{c=1}^C \log(\mathcal{N}(\mathbf{b}'^{(c)}; \mu_b, \Sigma_b)) - \sum_{c=1}^C \log(\mathcal{N}(\mathbf{b}^{(c)}; \mu_b, \Sigma_b)) \\ &+ \sum_{d=1}^D \left( \left( \sum_{i \in \mathcal{A}} \left( -\log \left( \left( \prod_{j \in \mathcal{A}_{\setminus i}} (\exp\{\delta + \log(\lambda_{ij}^{(d)})\} + 1) \right) - 1 \right) + \sum_{j \in \mathcal{A}_{\setminus i}} (\delta + \log(\lambda_{ij}^{(d)})) J_{ij}^{(d)} \right) \right. \right. \\ & \quad \left. \left. + \left( \log(\lambda_{i_o^{(d)} J_o^{(d)}}^{(d)}) - \Delta T_{i_o^{(d)} J_o^{(d)}}^{(d)} \sum_{i \in \mathcal{A}} \lambda_{i J_i^{(d)}}^{(d)} \right) \text{ given } \mathbf{b}' \right) \right) \\ &- \sum_{d=1}^D \left( \left( \sum_{i \in \mathcal{A}} \left( -\log \left( \left( \prod_{j \in \mathcal{A}_{\setminus i}} (\exp\{\delta + \log(\lambda_{ij}^{(d)})\} + 1) \right) - 1 \right) + \sum_{j \in \mathcal{A}_{\setminus i}} (\delta + \log(\lambda_{ij}^{(d)})) J_{ij}^{(d)} \right) \right. \right. \\ & \quad \left. \left. + \left( \log(\lambda_{i_o^{(d)} J_o^{(d)}}^{(d)}) - \Delta T_{i_o^{(d)} J_o^{(d)}}^{(d)} \sum_{i \in \mathcal{A}} \lambda_{i J_i^{(d)}}^{(d)} \right) \text{ given } \mathbf{b} \right) \right), \end{aligned} \quad (30)$$

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

$$\log(\text{Acceptance Probability}) = \min(\text{Equation (30)}, 0). \quad (31)$$

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) is less than the log-acceptance probability (30), we accept the proposal  $\mathbf{b}'$ . Otherwise, we reject.

### 3.5 Resampling $\delta$

After finishing the update of  $\mathcal{B}$ , we move on to the updates of  $\delta$ , which is very similar to the steps illustrated in Section 3.4. 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  $\sigma_\delta^2$  to ensure appropriate level of acceptance rate. Then, it follows that the simplified version of acceptance probability is

$$\text{Acceptance Probability} = \begin{cases} \frac{P(\delta'|\mu_\delta, \sigma_\delta^2)P(\mathcal{J}_a, \mathcal{I}_o, \mathcal{J}_o, \mathcal{T}_o | \mathcal{Z}, \mathcal{C}, \mathcal{B}, \delta')}{P(\delta|\mu_\delta, \sigma_\delta^2)P(\mathcal{J}_a, \mathcal{I}_o, \mathcal{J}_o, \mathcal{T}_o | \mathcal{Z}, \mathcal{C}, \mathcal{B}, \delta)} & \text{if } < 1 \\ 1 & \text{else} \end{cases} \quad (32)$$

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

$$\begin{aligned} & \log(\mathcal{N}(\delta'; \mu_\delta, \sigma_\delta^2)) - \log(\mathcal{N}(\delta; \mu_\delta, \sigma_\delta^2)) \\ & + \sum_{d=1}^D \left( \sum_{i \in \mathcal{A}} \left( -\log \left( \left( \prod_{j \in \mathcal{A}_{\setminus i}} \left( \exp\{\delta' + \log(\lambda_{ij}^{(d)})\} + 1 \right) \right) - 1 \right) + \sum_{j \in \mathcal{A}_{\setminus i}} (\delta' + \log(\lambda_{ij}^{(d)})) J_{ij}^{(d)} \right) \right. \\ & \quad \left. - \sum_{i \in \mathcal{A}} \left( -\log \left( \left( \prod_{j \in \mathcal{A}_{\setminus i}} \left( \exp\{\delta + \log(\lambda_{ij}^{(d)})\} + 1 \right) \right) - 1 \right) + \sum_{j \in \mathcal{A}_{\setminus i}} (\delta + \log(\lambda_{ij}^{(d)})) J_{ij}^{(d)} \right) \right), \end{aligned} \quad (33)$$

and the corresponding log of acceptance ratio is

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

### 3.6 Pseudocode

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  $\delta$ , 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  $\alpha$  and  $\mathbf{m}$  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 3.1

**end**

**for**  $n=1$  to  $N^{(d)}$  **do**

            | calculate  $p^{\mathcal{Z}} | \text{others} = (p_1, \dots, p_K)$  following Section 3.2

            | draw of  $z_n^{(d)} \sim \text{Multinomial}(p^{\mathcal{Z}})$

**end**

**end**

**for**  $k=1$  to  $K$  **do**

        | calculate  $p^{\mathcal{C}} | \text{others} = (p_1, \dots, p_C)$  following Section 3.3

        | draw  $C_k \sim \text{Multinomial}(p^{\mathcal{C}})$

**end**

**for**  $n=1$  to  $n_2$  **do**

        | sample  $\mathcal{B}$  using M-H following Section 3.4

**end**

**for**  $n=1$  to  $n_3$  **do**

        | sample  $\delta$  using M-H following Section 3.5

**end**

**end**

Summarize the results with:

1. Last sample of  $\mathcal{C}$ ,
  2. Last sample of  $\mathcal{Z}$ ,
  3. Last  $n_2$  length chain of  $\mathcal{B}$ ,
  3. Last  $n_3$  length chain of  $\delta$
-

## 4 Getting It Right (GiR)

### 4.1 Collapsed-time tie generating process

Considering that we integrated out latent time  $\mathcal{T}_a$  in the inference, we develop the new generative process for (sender, recipients, timestamp) parts with latent time integrated out, which we call ‘collapsed-time document generating process’. Note that this is built upon the property of the minimum of independent Exponential random variables (the probability  $\Delta T_{iJ_i}$  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  $\mathbf{x}_{t_+^{(d-1)}}^{(c)}(i, j)$ , the network statistics evaluated at time  $t_+^{(d-1)}$ 
        set  $\lambda_{ij}^{(d)} = \sum_{c=1}^C p_c^{(d)} \cdot \exp\left\{\lambda_0^{(c)} + \mathbf{b}^{(c)T} \mathbf{x}_{t_+^{(d-1)}}^{(c)}(i, j)\right\} \cdot 1\{j \in \mathcal{A}_{\setminus i}\}$ 
        draw  $I(i \rightarrow j) \sim \text{Ber}\left(\frac{\delta \lambda_{ij}^{(d)}}{\delta \lambda_{ij}^{(d)} + 1}\right)$ 
      end
    end
    calculate  $\lambda_{iJ_i}^{(d)}$ 
  end
  choose  $i^{(d)} \sim \text{Multinomial}\left(\left\{\frac{\lambda_{iJ_i}^{(d)}}{\sum_{i \in \mathcal{A}} \lambda_{iJ_i}^{(d)}}\right\}_{i=1}^A\right)$ 
  set  $J^{(d)} = J_{i^{(d)}}$ 
  draw  $\Delta T_{i^{(d)}J^{(d)}} \sim \text{rexp}\left(\sum_{i \in \mathcal{A}} \lambda_{iJ_i}^{(d)}\right)$  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}_a^{(d)}, \mathcal{T}_a^{(d)}, i_o^{(d)}, J_o^{(d)}, t_o^{(d)} | \mathcal{T}_o^{(<d)}, \mathcal{J}_o^{(<d)}, \mathcal{T}_o^{(<d)}, \mathcal{Z}, \mathcal{C}, \mathcal{B}, \delta) \\
&= P(\text{latent receivers generation}) \times P(\text{choose the sender}) \times P(\text{observed minimum time generation}) \\
&= \prod_{i \in \mathcal{A}} \prod_{j \in \mathcal{A}_{\setminus i}} \left( I(i \rightarrow j) \sim \text{Ber}\left(\frac{\delta \lambda_{ij}^{(d)}}{\delta \lambda_{ij}^{(d)} + 1}\right) \right) \times \left( i_o^{(d)} \sim \text{Multinom}\left(\left\{\frac{\lambda_{iJ_i}^{(d)}}{\sum_{i \in \mathcal{A}} \lambda_{iJ_i}^{(d)}}\right\}_{i=1}^A\right) \right) \times \left( \Delta T_{i^{(d)}J^{(d)}} \sim \text{Exp}\left(\sum_{i \in \mathcal{A}} \lambda_{i_o^{(d)}J_o^{(d)}}^{(d)}\right) \right) \\
&= \left( \prod_{i \in \mathcal{A}} \prod_{j \in \mathcal{A}_{\setminus i}} \left( \frac{\delta \lambda_{ij}^{(d)}}{\delta \lambda_{ij}^{(d)} + 1} \right)^{I(j \in J_i^{(d)})} \left( \frac{1}{\delta \lambda_{ij}^{(d)} + 1} \right)^{1 - I(j \in J_i^{(d)})} \right) \times \left( \frac{\lambda_{i_o^{(d)}J_o^{(d)}}^{(d)}}{\sum_{i \in \mathcal{A}} \lambda_{iJ_i}^{(d)}} \right) \times \left( \left( \sum_{i \in \mathcal{A}} \lambda_{iJ_i}^{(d)} \right) e^{-\Delta T_{i_o^{(d)}J_o^{(d)}}^{(d)} \sum_{i \in \mathcal{A}} \lambda_{iJ_i}^{(d)}} \right) \\
&= \left( \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} \right) \times \left( \frac{\lambda_{i_o^{(d)}J_o^{(d)}}^{(d)}}{\sum_{i \in \mathcal{A}} \lambda_{iJ_i}^{(d)}} \right) \times \left( \left( \sum_{i \in \mathcal{A}} \lambda_{iJ_i}^{(d)} \right) e^{-\Delta T_{i_o^{(d)}J_o^{(d)}}^{(d)} \sum_{i \in \mathcal{A}} \lambda_{iJ_i}^{(d)}} \right) \\
&= \left( \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} \right) \times \left( \lambda_{i_o^{(d)}J_o^{(d)}}^{(d)} \right) \times \left( e^{-\Delta T_{i_o^{(d)}J_o^{(d)}}^{(d)} \sum_{i \in \mathcal{A}} \lambda_{iJ_i}^{(d)}} \right),
\end{aligned} \tag{35}$$

which is exactly the same as Equation (14).

Figure 2 shows the results from forward Getting it Right test, which compares the forward generative process in Section 2.5 and the new collapsed-time version of generative process. The new version follows the exactly same scheme for token generative part, while the tie generating part follows the steps in Algorithm 6.

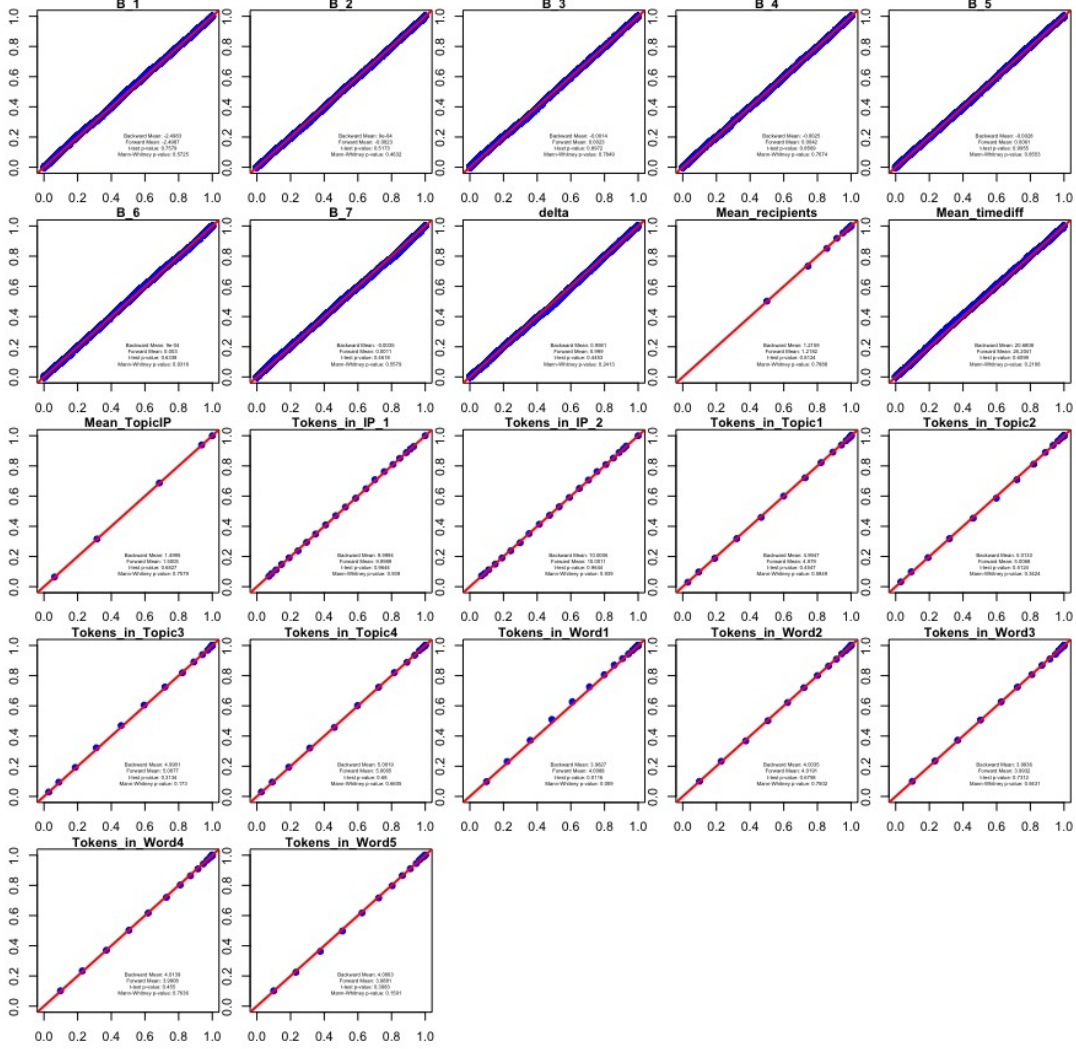


Figure 2: Forward Getting it Right test result (x-axis: forward, y-axis: backward)

## 4.2 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 (latent edges from data augmentation, token topic assignments, topic interaction pattern assignments, interaction pattern parameters, and tuning parameter  $\delta$ ) as inputs, and simply condition on these to draw new data. This “backward” version of the generative process is detailed below in Algorithm 10. Here, 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 the backward sampling, we do not generate latent edges, instead, we use inferred ones from the inference as an input. Only sample the receivers for previously observed sender ( $i_o^{(d)}$ ), which was not inferred, will be sampled according to the generative process, same as forward sampling (with inferred parameter values). The generative process for backward sampling including the token updates and collapsed-time tie generating process are demonstrated in Algorithm 7.

---

**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  $\{\mathbf{b}^{(c)}\}_{c=1}^C$ ,
- 5) tuning parameter  $\delta$ .

```

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 $_n^{(d)}[v] = \frac{NKV_{v,z_n^{(d)}} + \beta \mathbf{u}_v}{NK_{z_n^{(d)}} + \beta}$ 
    end
    draw  $w_n^{(d)} \sim (\text{token-word-type-distribution}_n^{(d)})$ 
     $NKV_{w_n^{(d)}, z_n^{(d)}} += 1$ 
     $NK_{z_n^{(d)}} += 1$ 
  end
  only for the previously observed sender  $i_o^{(d)}$ ,
  for  $j \neq i_o^{(d)}$  do
    calculate  $\mathbf{x}_{t_{+}^{(d-1)}}^{(c)}(i_o^{(d)}, j)$ , the network statistics evaluated at time  $t_{+}^{(d-1)}$ 
    set  $\lambda_{i_o^{(d)}j}^{(d)} = \sum_{c=1}^C p_c^{(d)} \cdot \exp\left\{\lambda_0^{(c)} + \mathbf{b}^{(c)T} \mathbf{x}_{t_{+}^{(d-1)}}^{(c)}(i_o^{(d)}, j)\right\} \cdot 1\{j \in \mathcal{A}_{\setminus i_o^{(d)}}\}$ 
    draw  $I(i_o^{(d)} \rightarrow j) \sim \text{Ber}\left(\frac{\delta \lambda_{i_o^{(d)}j}^{(d)}}{\delta \lambda_{i_o^{(d)}j}^{(d)} + 1}\right)$ 
  end
  choose  $i^{(d)} \sim \text{Multinomial}\left(\left\{\frac{\lambda_{iJ_i}^{(d)}}{\sum_{i \in \mathcal{A}} \lambda_{iJ_i}^{(d)}}\right\}_{i=1}^A\right)$ 
  set  $J^{(d)} = J_{i^{(d)}}$ 
  draw  $\Delta T_{i^{(d)}J^{(d)}} \sim \text{rexp}\left(\sum_{i \in \mathcal{A}} \lambda_{iJ_i}^{(d)}\right)$  and set  $t^{(d)} = t^{(d-1)} + \Delta T_{i^{(d)}J^{(d)}}$ 
end

```

---

### 4.3 Issue of initial state of $\mathbf{x}_t^{(c)}$

Considering that our network statistics  $\mathbf{x}_t^{(c)}$  are generated by the network itself, it is necessary to use the same initial value of  $\mathbf{x}_t^{(c)}$  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  $\forall \mathbf{x}_t^{(c)}(i, j) = 0$ .

Therefore, we fix the initial state of  $\mathbf{x}_t^{(c)}$  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  $\mathbf{x}_{t^{(0)}=384}^{(c)}$  so that we can achieve consistency in the generated number of documents with  $t^{(d)} > 384$ .

#### 4.4 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) = 10
- 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
- $\alpha$  (Dirichlet concentration prior) = 2
- $\mathbf{m}$  (Dirichlet base prior) =  $\mathbf{u}$
- $\beta$  (Dirichlet concentration prior)= 2
- $\mathbf{n}$  (Dirichlet base prior) =  $\mathbf{u}$
- netstat = “intercept” and “dyadic”
- prior for  $\mathbf{b}^{(c)} \sim \text{MVN}_7((-3, \mathbf{0}_6), 0.1 \times I_7)$
- prior for  $\delta \sim \text{Gamma}(20, 2)$
- $I$  (outer iteration) = 3
- $n_1$  (iteration for hyperparamter optimization) = 0
- $n_2$  (M-H sampling iteration of  $\mathcal{B}$ ) = 100
- burn (M-H sampling burn-in of  $\mathcal{B}$ )= 0
- thin (M-H sampling thinning of  $\mathcal{B}$ )= 5
- $\sigma_Q^2$  (proposal variance for  $\mathcal{B}$ ) = 1
- $n_3$  (M-H sampling iteration of  $\delta$ ) = 20
- $\sigma_Q^2$  (proposal variance for  $\delta$ ) = 2.5

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 switches across the updates. Therefore, at each iteration, we calculate and save the statistics below:

1. Mean of interaction pattern parameters ( $\mathbf{b}_p^{(1)}, \dots, \mathbf{b}_p^{(C)}$ ) for every  $p = 1, \dots, P$ ,
2. Mean number of recipients,
3. Mean of time-increments  $t^{(d)} - t^{(d-1)}$  for every  $d = 2, \dots, \text{nDocs}$ ,
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  $c = 1, \dots, C$ ,
6. Number of tokens assigned to each topic  $k = 1, \dots, K$ ,
7. Number of tokens assigned to each unique word type  $w = 1, \dots, W$ .

#### 4.5 GiR Results

Having saved a set of samples (collected  $10^5$  forward and backward samples), 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. 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.

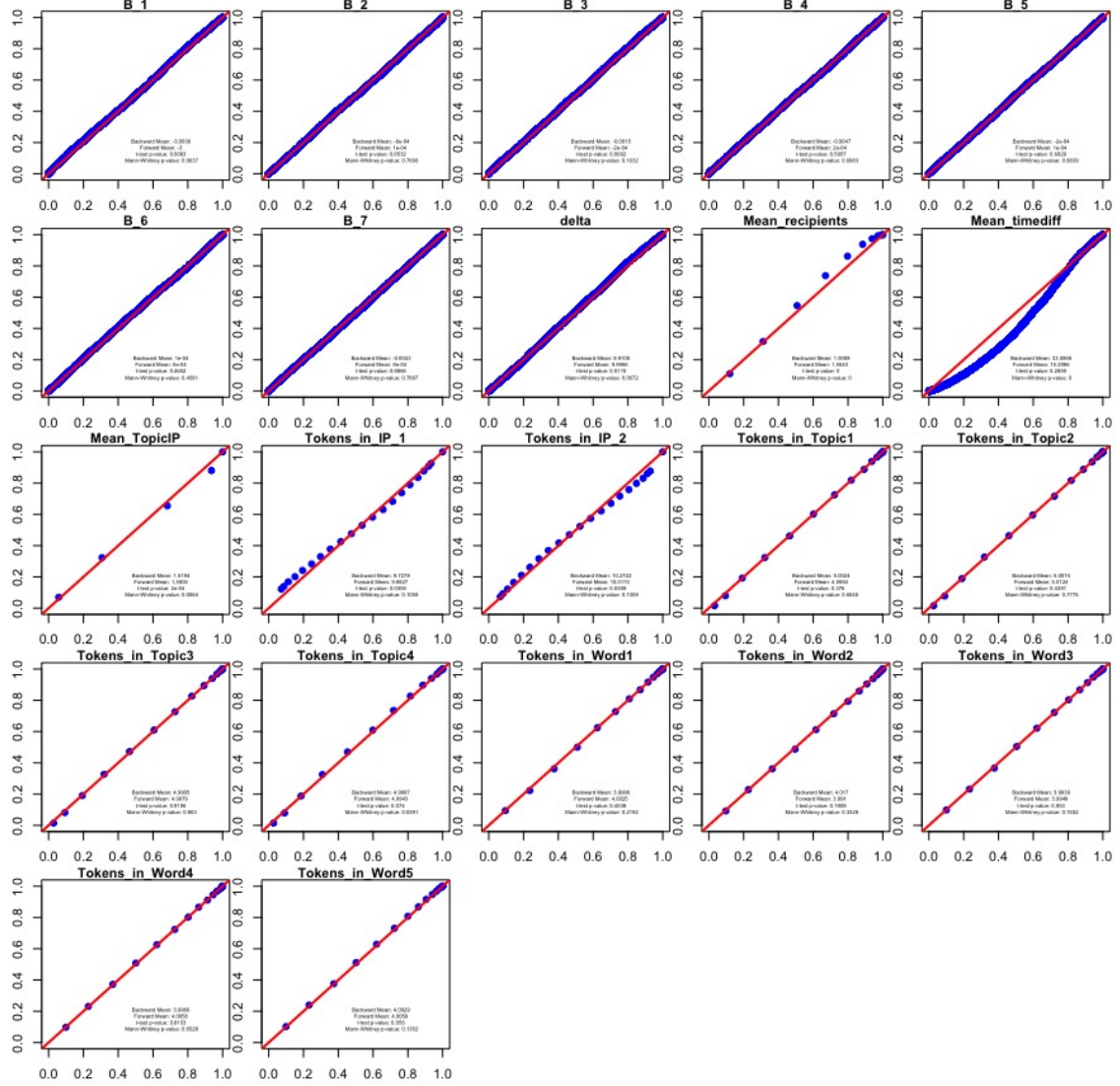


Figure 3: Getting it Right test result

## 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.
- 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.

- 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.
- 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.
- 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.
- 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.
- 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*.

- 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.
- 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 Notations in IPTM

Sender of the $d^{th}$ document	$i^{(d)}$	Scalar
Receivers of the $d^{th}$ document	$J^{(d)}$	$ J^{(d)} $ -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	$\delta$	Scalar
Poisson parameter for number of words $N^{(d)}$	$\zeta$	Scalar
Words in the $d^{th}$ document	$\mathbf{w}^{(d)}$	$N^{(d)}$ -dimensional vector
$n^{th}$ word in the $d^{th}$ document	$w_n^{(d)}$	$n^{th}$ component of $\mathbf{w}^{(d)}$
Topic assignments in the $d^{th}$ document	$\mathbf{z}^{(d)}$	$N^{(d)}$ -dimensional vector
Topic assignments for $n^{th}$ word in the $d^{th}$ document	$z_n^{(d)}$	$n^{th}$ component of $\mathbf{z}^{(d)}$
Dirichlet concentration prior for document topic distribution	$\alpha$	Scalar
Dirichlet base prior for document topic distribution	$\mathbf{m}$	$K$ -dimensional vector
Dirichlet concentration prior for topic word distribution	$\beta$	Scalar
Dirichlet base prior for topic word distribution	$\mathbf{u}$	$W$ -dimensional vector
Variance of Normal prior	$\sigma^2$	Scalar
Probabilities of the words given topic $k$	$\phi^{(k)}$	$W$ -dimensional vector
Probabilities of the topics given the $d^{th}$ document	$\theta^{(d)}$	$K$ -dimensional vector
Coefficient of the intensity process given interaction pattern $c$	$\mathbf{b}^{(c)}$	$p$ -dimensional vector
Network statistics for $(i, j)$ at time $t$ given interaction pattern $c$	$\mathbf{x}_t^{(c)}(i, j)$	$p$ -dimensional vector
Stochastic intensity of document $d$ at time $t$	$\lambda^{(d)}(t)$	$A \times A$ 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	$\Delta T_{i,j}$	Scalar

Table 1: Symbols associated with IPTM, as used in this paper

## B Non-empty Gibbs measure

The probability that vertex  $i$  selects the binary receiver vector of length  $n - 1$ ,  $J_i^{(d)}$  is given by

$$P(J_i^{(d)}) = \frac{1}{Z(\delta, \lambda_i^{(d)})} \exp \left[ \ln \left( \mathbb{I} \left( \left( \sum_{j \neq i} J_{ij} \right) > 0 \right) \right) + \sum_{j \neq i} (\delta + \lambda_{ij}^{(d)}) J_{ij}^{(d)} \right],$$

where  $\delta$  is real-valued intercept that controls the expected value of  $J_{ij}^{(d)}$ , and  $\lambda_{ij}^{(d)}$  is a positive dyad-specific function of the network histories of  $i$  and  $j$ , and any additional attributes included in the model (note, it may fit the data better to include lambda as  $\ln(\lambda_{ij}^{(d)})$ , and  $\lambda_i^{(d)}$  is the vector of dyadic weights in which  $i$  is the sender.

To use this distribution efficiently, we need to derive a closed-form expression for  $\frac{1}{Z(\delta, \lambda_i^{(d)})}$  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  $P(J_{ij}^{(d)}=1)$  was given by  $\text{logit}(\delta + \lambda_{ij}^{(d)})$ ,

then

$$P(J_i^{(d)}) \propto \exp \left[ \sum_{j \neq i} (\delta + \lambda_{ij}^{(d)}) J_{ij}^{(d)} \right].$$

This is straightforward to verify by looking at

$$P(J_{ij}^{(d)} = 1 | J_{i,-j}) = \frac{\exp(\delta + \lambda_{ij}) \exp \left[ \sum_{h \neq i, j} (\delta + \lambda_{ih}^{(d)}) J_{ih}^{(d)} \right]}{\exp(\delta + \lambda_{ij}) \exp \left[ \sum_{h \neq i, j} (\delta + \lambda_{ih}^{(d)}) J_{ih}^{(d)} \right] + \exp(0) \exp \left[ \sum_{h \neq i, j} (\delta + \lambda_{ih}^{(d)}) J_{ih}^{(d)} \right]},$$

$$P(J_{ij}^{(d)} = 1 | J_{i,-j}) = \frac{\exp(\delta + \lambda_{ij})}{\exp(\delta + \lambda_{ij}) + 1}.$$

We denote the logistic-Bernoulli normalizing constant as

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

Now, since

$$\exp \left[ \ln \left( \mathbb{I} \left( \left( \sum_{j \neq i} J_{ij} \right) > 0 \right) \right) + \sum_{j \neq i} (\delta + \lambda_{ij}^{(d)}) J_{ij}^{(d)} \right] = \exp \left[ \sum_{j \neq i} (\delta + \lambda_{ij}^{(d)}) J_{ij}^{(d)} \right],$$

except when  $\sum_{j \neq i} J_{ij} = 0$ , in which case

$$\exp \left[ \ln \left( \mathbb{I} \left( \left( \sum_{j \neq i} J_{ij} \right) > 0 \right) \right) + \sum_{j \neq i} (\delta + \lambda_{ij}^{(d)}) J_{ij}^{(d)} \right] = 0.$$

As such, we note that

$$Z(\delta, \lambda_i^{(d)}) = Z^l(\delta, \lambda_i^{(d)}) - \exp \left[ \sum_{j \neq i, J_{ij}=0} (\delta + \lambda_{ij}^{(d)}) J_{ij}^{(d)} \right],$$

$$Z(\delta, \lambda_i^{(d)}) = Z^l(\delta, \lambda_i^{(d)}) - 1.$$

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

$$\frac{\exp \left[ \sum_{j \neq i, J_{ij}=0} (\delta + \lambda_{ij}^{(d)}) J_{ij}^{(d)} \right]}{Z^l(\delta, \lambda_i^{(d)})} = \prod_{i \neq j} \frac{\exp(-(\delta + \lambda_{ij}))}{\exp(-(\delta + \lambda_{ij})) + 1},$$

$$\frac{1}{Z^l(\delta, \lambda_i^{(d)})} = \prod_{i \neq j} \frac{\exp(-(\delta + \lambda_{ij}))}{\exp(-(\delta + \lambda_{ij})) + 1},$$

$$Z^l(\delta, \lambda_i^{(d)}) = \frac{1}{\prod_{i \neq j} \frac{\exp(-(\delta + \lambda_{ij}))}{\exp(-(\delta + \lambda_{ij})) + 1}}.$$

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

$$Z(\delta, \lambda_i^{(d)}) = \frac{1}{\prod_{i \neq j} \frac{\exp(-(\delta + \lambda_{ij}))}{\exp(-(\delta + \lambda_{ij})) + 1}} - 1.$$

## C Integrating out $\Phi$ and $\Theta$ in latent Dirichlet allocation

$$\begin{aligned}
& P(\Phi, \Theta, \mathcal{Z}, \mathcal{C}, \mathcal{B}, \mathcal{W}, \mathcal{I}_a, \mathcal{J}_a, \mathcal{T}_a, \mathcal{I}_o, \mathcal{J}_o, \mathcal{T}_o | \beta, \mathbf{u}, \alpha, \mathbf{m}, \sigma^2, \delta) \\
& \propto P(\mathcal{Z}, \mathcal{C}, \mathcal{B}, \mathcal{W}, \mathcal{I}_a, \mathcal{J}_a, \mathcal{T}_a, \mathcal{I}_o, \mathcal{J}_o, \mathcal{T}_o | \Phi, \Theta, \sigma^2, \delta) P(\Phi, \Theta | \beta, \mathbf{u}, \alpha, \mathbf{m}) \\
& \propto P(\mathcal{Z} | \mathcal{W}, \Theta) P(\mathcal{C} | \mathcal{B}, \sigma^2) P(\mathcal{W} | \Phi) P(\Phi | \beta, \mathbf{u}) P(\Theta | \alpha, \mathbf{m}) P(\mathcal{I}_a, \mathcal{J}_a, \mathcal{T}_a, \mathcal{I}_o, \mathcal{J}_o, \mathcal{T}_o | \mathcal{Z}, \mathcal{C}, \mathcal{B}, \delta) \\
& = \left[ \prod_{d=1}^D \prod_{n=1}^{N^{(d)}} P(z_n^{(d)} | w_n^{(d)}, \boldsymbol{\theta}^{(d)}) \right] \times \left[ \prod_{k=1}^K P(C_k) \right] \times \left[ \prod_{c=1}^C P(\mathbf{b}^{(c)} | \sigma^2) \right] \times \left[ \prod_{d=1}^D \prod_{n=1}^{N^{(d)}} P(w_n^{(d)} | \phi_{z_n^{(d)}}) \right] \times \left[ \prod_{k=1}^K P(\phi^{(k)} | \beta, \mathbf{u}) \right] \\
& \quad \times \left[ \prod_{d=1}^D P(\boldsymbol{\theta}^{(d)} | \alpha, \mathbf{m}) \right] \times \left[ \prod_{d=1}^D P(\mathcal{I}_a^{(d)}, \mathcal{J}_a^{(d)}, \mathcal{T}_a^{(d)}, i_o^{(d)}, J_o^{(d)}, t_o^{(d)} | \mathcal{I}_o^{(<d)}, \mathcal{J}_o^{(<d)}, \mathcal{T}_o^{(<d)}, \mathcal{Z}, \mathcal{C}, \mathcal{B}, \delta) \right]
\end{aligned} \tag{36}$$

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

$$\begin{aligned}
& \propto \left[ \prod_{d=1}^D \prod_{n=1}^{N^{(d)}} P(z_n^{(d)} | w_n^{(d)}, \boldsymbol{\theta}^{(d)}) \right] \times \left[ \prod_{d=1}^D \prod_{n=1}^{N^{(d)}} P(w_n^{(d)} | \phi_{z_n^{(d)}}) \right] \times \left[ \prod_{k=1}^K P(\phi^{(k)} | \beta, \mathbf{u}) \right] \times \left[ \prod_{d=1}^D P(\boldsymbol{\theta}^{(d)} | \alpha, \mathbf{m}) \right] \\
& \propto \left[ \prod_{d=1}^D \prod_{n=1}^{N^{(d)}} \phi_{w_n^{(d)} z_n^{(d)}} \right] \times \left[ \prod_{d=1}^D \prod_{n=1}^{N^{(d)}} \boldsymbol{\theta}^{(d)} \right] \\
& \quad \times \left[ \prod_{k=1}^K \left( \frac{\Gamma(\sum_{w=1}^W \beta u_w)}{\prod_{w=1}^W \Gamma(\beta u_w)} \prod_{w=1}^W \phi_{wk}^{\beta u_w - 1} \right) \right] \times \left[ \prod_{d=1}^D \left( \frac{\Gamma(\sum_{k=1}^K \alpha m_k)}{\prod_{k=1}^K \Gamma(\alpha m_k)} \prod_{k=1}^K (\boldsymbol{\theta}_k^{(d)})^{\alpha m_k - 1} \right) \right] \\
& \propto \left[ \frac{\Gamma(\sum_{w=1}^W \beta u_w)}{\prod_{w=1}^W \Gamma(\beta u_w)} \right]^K \times \prod_{d=1}^D \left[ \frac{\Gamma(\sum_{k=1}^K \alpha m_k)}{\prod_{k=1}^K \Gamma(\alpha m_k)} \right] \times \left[ \prod_{k=1}^K \prod_{w=1}^W \phi_{wk}^{N_{wk}^{WK} + \beta u_w - 1} \right] \times \left[ \prod_{d=1}^D \prod_{k=1}^K (\boldsymbol{\theta}_k^{(d)})^{N_{k|d} + \alpha m_k - 1} \right]
\end{aligned} \tag{37}$$

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  $\Theta$  and  $\Phi$  in Equation (21), we integrate out the random variables  $\Theta$  and  $\Phi$ , 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:

$$\begin{aligned}
& P(\mathcal{Z}, \mathcal{C}, \mathcal{B}, \mathcal{W}, \mathcal{I}_a, \mathcal{J}_a, \mathcal{T}_a, \mathcal{I}_o, \mathcal{J}_o, \mathcal{T}_o | \beta, \mathbf{u}, \alpha, \mathbf{m}, \sigma^2, \delta) \\
& = \text{Const.} \int_{\Theta} \int_{\Phi} \left[ \prod_{k=1}^K \prod_{w=1}^W \phi_{wk}^{N_{wk}^{WK} + \beta u_w - 1} \right] \left[ \prod_{d=1}^D \prod_{k=1}^K (\boldsymbol{\theta}_k^{(d)})^{N_{k|d} + \alpha m_k - 1} \right] d\Phi d\Theta \\
& = \text{Const.} \left[ \prod_{k=1}^K \int_{\phi:k} \prod_{w=1}^W \phi_{wk}^{N_{wk}^{WK} + \beta u_w - 1} d\phi_{:k} \right] \times \left[ \prod_{d=1}^D \int_{\theta:d} \prod_{k=1}^K (\boldsymbol{\theta}_k^{(d)})^{N_{k|d} + \alpha m_k - 1} d\theta_{:d} \right] \\
& = \text{Const.} \left[ \prod_{k=1}^K \frac{\prod_{w=1}^W \Gamma(N_{wk}^{WK} + \frac{\beta}{W})}{\Gamma(\sum_{w=1}^W N_{wk}^{WK} + \beta)} \right] \times \left[ \prod_{d=1}^D \frac{\prod_{k=1}^K \Gamma(N_{k|d} + \alpha m_k)}{\Gamma(N_{\cdot|d} + \alpha)} \right].
\end{aligned} \tag{38}$$

## D Conditional probability of $\mathcal{Z}$

$$\begin{aligned}
& P(\mathbf{w}^{(d)}, \mathbf{z}^{(d)} | \mathcal{W}_{\setminus d}, \mathcal{Z}_{\setminus d}, \beta, \mathbf{u}, \alpha, \mathbf{m}) \\
& \propto \prod_{n=1}^{N^{(d)}} P(z_n^{(d)} = k, w_n^{(d)} = w | \mathcal{W}_{\setminus d, n}, \mathcal{Z}_{\setminus d, n}, \beta, \mathbf{u}, \alpha, \mathbf{m})
\end{aligned} \tag{39}$$

To obtain the Gibbs sampling equation, we need to obtain an expression for  $P(z_n^{(d)} = k, w_n^{(d)} =$

$w|\mathcal{W}_{\setminus d,n}, \mathcal{Z}_{\setminus d,n}, \beta, \mathbf{u}, \alpha, \mathbf{m}$ ), From Bayes' theorem and Gamma identity  $\Gamma(k+1) = k\Gamma(k)$ ,

$$\begin{aligned}
P(z_n^{(d)} = k, w_n^{(d)} = w | \mathcal{W}_{\setminus d,n}, \mathcal{Z}_{\setminus d,n}, \beta, \mathbf{u}, \alpha, \mathbf{m}) \\
&\propto \frac{P(\mathcal{W}, \mathcal{Z} | \beta, \mathbf{u}, \alpha, \mathbf{m})}{P(\mathcal{W}_{\setminus d,n}, \mathcal{Z}_{\setminus d,n} | \beta, \mathbf{u}, \alpha, \mathbf{m})} \\
&\propto \frac{\prod_{k=1}^K \frac{\prod_{w=1}^W \Gamma(N_{wk}^{WK} + \beta u_w)}{\Gamma(\sum_{w=1}^W N_{wk}^{WK} + \beta)}}{\prod_{k=1}^K \frac{\prod_{w=1}^W \Gamma(N_{wk, \setminus d,n}^{WK} + \beta u_w)}{\Gamma(\sum_{w=1}^W N_{wk, \setminus d,n}^{WK} + \beta)}} \times \frac{\prod_{k=1}^K \frac{\Gamma(N_{k|d} + \alpha m_k)}{\Gamma(N_{\cdot|d} + \alpha)}}{\prod_{k=1}^K \frac{\Gamma(N_{k|d, \setminus d,n} + \alpha m_k)}{\Gamma(N_{\cdot|d, \setminus d,n} + \alpha)}} \\
&\propto \frac{N_{wk, \setminus d,n}^{WK} + \frac{\beta}{W}}{\sum_{w=1}^W N_{wk, \setminus d,n}^{WK} + \beta} \times \frac{N_{k|d, \setminus d,n} + \alpha m_k}{N^{(d)} - 1 + \alpha}
\end{aligned} \tag{40}$$

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

$$P(z_n^{(d)} = k | w_n^{(d)} = w, \mathcal{W}_{\setminus d,n}, \mathcal{Z}_{\setminus d,n}, \beta, \mathbf{u}, \alpha, \mathbf{m}) \propto \frac{N_{k|d, \setminus d,n} + \alpha m_k}{N^{(d)} - 1 + \alpha} \tag{41}$$

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:

$$P(w_m^{(d)} = w | z_m^{(d)} = k, \mathcal{W}_{\setminus d,n}, \mathcal{Z}_{\setminus d,n}, \beta, \mathbf{u}, \alpha, \mathbf{m}) \propto \frac{N_{wk, \setminus d,n}^{WK} + \frac{\beta}{W}}{\sum_{w=1}^W N_{wk, \setminus d,n}^{WK} + \beta} \tag{42}$$

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

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

$$\begin{aligned}
&\sum_{d=1}^D \sum_{n=1}^{N^{(d)}} \log \left( P(z_n^{(d)} = k, w_n^{(d)} = w | \mathcal{W}_{\setminus d,n}, \mathcal{Z}_{\setminus d,n}, \beta, \mathbf{u}, \alpha, \mathbf{m}) \right) \\
&\propto \sum_{d=1}^D \sum_{n=1}^{N^{(d)}} \log \left( N_{w_n^{(d)} z_n^{(d)}, \setminus d,n}^{WK} + \frac{\beta}{W} \right) - \log \left( \sum_{w=1}^W N_{w z_n^{(d)}, \setminus d,n}^{WK} + \beta \right) \\
&\quad + \log \left( N_{k|d, \setminus d,n} + \alpha m_{z_n^{(d)}} \right) - \log \left( N^{(d)} - 1 + \alpha \right)
\end{aligned} \tag{44}$$