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

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

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

September 15, 2017

Abstract

In this paper, we introduce the interaction-partitioned topic model (IPTM)—a probabilistic model for 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 these email corpora covers the Outer Banks during the time period surrounding Hurricane Sandy. Via this application, we demonstrate that the IPTM is effective at predicting and explaining continuous-time textual communications.

1 Introduction

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

*Prepared for presentation at the 2017 Annual Summer Meeting of the Society for Political Methodology. This work was supported in part by the University of Massachusetts Amherst Center for Intelligent Information Retrieval and in part by National Science Foundation grants DGE-1144860, SES-1619644, and CISE-1320219. Any opinions, findings, and conclusions or recommendations are those of the authors and do not necessarily reflect those of the sponsors.

ERGM, and models based on ERGM, provide a framework for explaining or predicting ties between nodes using the network sub-structures in which the two nodes are embedded (e.g., an ERGM specification may predict ties between two nodes that have many shared partners). ERGM-style models have been used for many applications in which the ties between nodes are annotated with text. The text, despite providing rich information regarding the strength, scope, and character of the ties, has been largely excluded from these analyses, due to the inability of ERGM-style models to incorporate textual attributes of ties. These application domains include, among other applicaitons, the study of legislative networks in which networks reflect legislators' co-support of bills, but exclude bill text (Bratton and Rouse, 2011; Alemán and Calvo, 2013); the study of alliance networks in which networks reflect countries' co-signing of treaties, but exclude treaty text (Camber Warren, 2010; Cranmer et al., 2012b,a; Kinne, 2016); the study of scientific co-authorship networks that exclude the text of the co-authored papers (Kronegger et al., 2011; Liang, 2015; Fahmy and Young, 2016); and the study of text-based interaction on social media (e.g., users tied via 'mentions' on twitter) (Yoon and Park, 2014; Peng et al., 2016; Lai et al., 2017).

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

2 IPTM: Model Definition and Derivation

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

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

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

2.1 Content Generating Process

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

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

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

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

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

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

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

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

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

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

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

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

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

2.2 Stochastic Intensity

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

First, the content of a document is reflected in the stochastic intensity via the distribution of interaction patterns, $\{p_c^{(d)}\}_{c=1}^C$. To calculate the distribution of interaction patterns within a document, we estimate the proportion of words in document d which are assigned to the topics corresponding to the interaction pattern c from Section 2.1:

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

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

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

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

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

2.3 Dynamic Network Statistics

We develop a suite of eight different effects to be used as the components of $\boldsymbol{x}_t^{(c)}(i,j)$, (intercept, outdegree, indegree, send, receive, 2-send, 2-receive, sibling, and cosibling), which are incorporated as in Equation (2). These statistics capture common network properties such as popularity, centrality, reciprocity, and transitivity. Each network statistic is calculated for each interaction pattern $c = 1,...,C$, which means that each interaction pattern can be understood in terms of the ways that network dynamics shape tie formation within the interaction pattern. Below are the specifications of the degree, dyadic, and triadic network statistics we use in this paper.

We follow Perry and Wolfe (2013) and define each network feature to have potentially different effects within a number of intervals of recency in the formation of the ties that contribute to the network feature. We partition the interval $[-\infty, t)$ into $L = 4$ sub-intervals with equal length in the log-scale, by setting $\Delta_l = (6 \text{ hours}) \times 4^l$ for $l = 1,...,L-1$ such that Δ_l takes the values 24 hours (=1 day), 96 hours (=4 days), 384 hours (=16 days):

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

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

In the application of the IPTM below, we do not include the last interval $I_t^{(4)}$, history before 16 days ago, since the time intervals covered by our datasets are only eight and twelve weeks in length. Although the specification of these dynamic network covariates could be reformulated based on the objectives of each study, in this paper, we define the degree and dyadic efffects for each $l = 1,...,L-1$ and $c = 1,...,C$ as

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

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

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

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

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

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

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

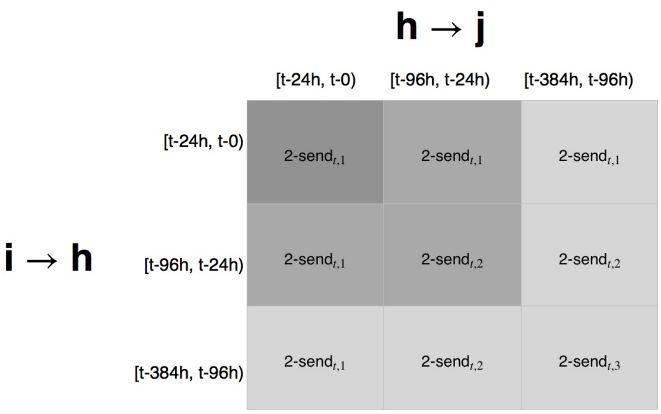

Figure 1: Example of 2-send statistic defined for each interval $l = 1, ..., 3$. Cells with same shades sum up to one statistic, based on when the triads are "closed".

2.4 Tie Generating Process

The tie generating process determines the sender, recipients, and timing $(i^{(d)}, J^{(d)}, t^{(d)})$ of the document. We assume the following tie generating process for each document d in a corpus of D documents:

1. For each sender $i \in \{1, ..., A\}$, we generate a binary receiver vector of length $A-1$, $J_i^{(d)}$, from the non-empty Gibbs measure (Fellows and Handcock, 2017) for every $j \in \mathcal{A}_{\setminus i}$.

$$\mathrm{P}(J_i^{(d)}) = \frac{1}{Z(\delta, \log(\lambda_i^{(d)}))} \exp\Big\{\log\big(\mathrm{I}(\|J_i^{(d)}\|_1 > 0)\big) + \sum_{j \in \mathcal{A}_{\setminus i}} (\delta + \log(\lambda_{ij}^{(d)}))J_{ij}^{(d)}\Big\}, \quad (3)$$

where δ is a real-valued intercept used to model the number of recipients—i.e, $\|J_i^{(d)}\|_1$, the ℓ_1-norm (or sum) of the binary recipient vector. The prior distribution for δ is specified as $\mathrm{Normal}(\mu_\delta, \sigma_\delta^2)$. As defined in Section 2.2, $\lambda_{ij}^{(d)}$ is a positive dyad-specific stochastic intensity included in the model, and we use $\lambda_i^{(d)} = \{\lambda_{ij}^{(d)}\}_{j\in\mathcal{A}_{\setminus i}}$ to denote the vector of dyadic weights in which i is the sender. Note that we omitted the notation (t) from Equation (2) and used $\lambda_{ij}^{(d)}$ instead, since the stochastic intensity $\lambda_{ij}^{(d)}$ is always evaluated at time $t_+^{(d-1)}$. The λ_{ij} for d^{th} document is obtained using the history of interactions up to and including the time when the previous document was sent, $t^{(d-1)}$.

The normalizing constant for the non-empty Gibbs measure $Z(\delta, \log(\lambda_i^{(d)}))$, which is the sum of $P(J_i^{(d)})$ over the entire support, can be simplified as:

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

Derivation of the normalizing constant is provided in Appendix A.

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

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

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

$$\lambda_{iJ_i}^{(d)}(t) = \sum_{c=1}^{C} p_c^{(d)} \cdot \exp\Big\{\lambda_0^{(c)} + \frac{1}{\|J_i^{(d)}\|_1} \sum_{j:J_{ij}^{(d)}=1} \boldsymbol{b}^{(c)T}\boldsymbol{x}_t^{(c)}(i,j)\Big\}. \tag{6}$$

Note that Equation (6) reduces to the stochastic intensity $\lambda_{ij}^{(d)}$ in Equation (2) in the case of a single receiver (i.e. $\|J_i^{(d)}\|_1 = 1$), so we can interpret this mean parameter as weighted stochastic intensity across the chosen receivers. When there are multiple chosen receivers (i.e. $\|J_i^{(d)}\|_1 > 1$), we call it a multicast interaction.

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

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

The intuition behind this choice is that all possible senders $i \in \mathcal{A}$ are competing against each other to send the next document, and correspondingly introduce the next modification to the history of interactions. Once that next document is sent, the actors in the network revise their plans for e-mail sending, considering the new entry in the history of interactions.

2.5 Joint Generative Process

The algorithms we present in this section form the generative process for D documents. This generative process integrates Sections 2.1 through 2.4.

Algorithm 1 Topic Word Distributions

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

Algorithm 2 Interaction Pattern Parameters

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

Algorithm 3 Topic Interaction Pattern Assginments

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

Algorithm 4 Recipient Size Parameter

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

Algorithm 5 Document Generating Process

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

3 Inference

We take a Bayesian approach to inferring the latent variables (i.e., parameters) in the IPTM. The likelihood function is implied by the generative process in Section 2.5. In this section, we derive the joint distribution over the variables $\Phi = \{\boldsymbol{\phi}^{(k)}\}_{k=1}^{K}, \Theta = \{\boldsymbol{\theta}^{(d)}\}_{d=1}^{D}, \mathcal{Z} = \{\boldsymbol{z}^{(d)}\}_{d=1}^{D}, \mathcal{C} = \{c_k\}_{k=1}^{K}, \mathcal{B} =$

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

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

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

The detailed derivation of sampling equations can be found in Appendix B.

To summarize the inference procedure outlined above, we provide pseudocode for Markov Chain Monte Carlo (MCMC) sampling. For better performance and interpretability of the topics we infer, we run n_1 iterations of the hyperparameter optimization technique called "new fixed-point iterations using the Digamma recurrence relation" in Wallach (2008), for every outer iteration o. Also, while we update the categorical variables $\mathcal{Z}$ and $\mathcal{C}$ once per outer iteration, we specify a larger number of inner iterations (n_2 and n_3) for the continuous variables $\mathcal{B}$ and δ, respectively. The continuous variables converge slower than the discrete variables since we sample the categorical variables using Gibbs sampling and the continuous variables using Metropolis-Hastings.. When summarizing model results, we only use the samples from the last (i.e., O^{th}) outer loop.

Algorithm 6 MCMC

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

4 Getting It Right (GiR) Test

Software development is integral to the objective of applying IPTM to real world data. Code review is a valuable process in any research computing context, and the prevalence of software bugs in statistical software is well documented (e.g., Altman et al., 2004; McCullough, 2009). With highly complex models such as IPTM, there are many ways in which software bugs can be introduced and go unnoticed. As such, we present a joint analysis of the integrity of our generative model, sampling equations, and software implementation.

Geweke (2004) introduced the "Getting it Right" (GiR) test—a joint distribution test of posterior simulators which can detect errors in sampling equations as well as coding errors. The test involves comparing the distributions of variables simulated from two joint distribution samplers, which we call "forward" and "backward" samples. The forward sampler draws unobservable variables from the prior and then generates the observable data conditional on the unobservables. The backward sampler alternates between the inference and an observables simulator, by running the inference code on observable data to obtain posterior estimates of the unobervable variables and then re-generating the observables given the inferred unobservables. The backward sampler is initialized by running an iteration of inference on observables drawn directly from the prior. Since the only information on which both the forward and backward samplers are based is the prior, if the sampling equations are correct and the code is implemented without bugs, each variable should have the same distribution in the forward and backward samples.

In the forward samples, both observable and unobservable variables are generated using Algorithm 5. In the backward samples, unobservable variables are generated using the sampling equations for inference, which we derived in Section 3. In order to generate observable variables in the backward samples (sender, recipients, timestamp), we use the collapsed-time generative process, which we presented in Section C.1. For each forward and backward sample that consists of D number of documents, we save the statistics below:

1. Mean of network effect parameters $(\boldsymbol{b}_p^{(1)}, ..., \boldsymbol{b}_p^{(C)})$ for every $p = 1, ..., P$,
2. Network statistic 'send' calculated for the last D^{th} document for every $l = 1, ..., 3$
3. δ value used to generate the samples
4. Mean of the recipient size $|J^{(d)}|$ across $d = 1, ..., D$,
5. Mean of time-increments $t^{(d)} - t^{(d-1)}$ across $d = 1, ..., D$,
6. Mean topic-interaction pattern assignment c_k across $k = 1, ..., K$,
7. Number of tokens in topics assigned to each interaction pattern $c = 1, ..., C$,
8. Number of tokens assigned to each topic $k = 1, ..., K$,
9. Number of tokens assigned to each unique word type $w = 1, ..., W$.

To keep the computational burden of re-running thousands of rounds of inference manageable, we run GiR using a relatively small artificial sample, consisting of 5 documents, 4 tokens per document, 4 actors, 5 unique word types, 2 interaction patterns, and 4 topics per each forward or backward samples. For detailed settings including the prior specifications, see Appendix C.4. We generated 10^5 sets of forward and backward samples, and then calculated 1,000 quantiles for each of the network effect statistics (1.), and 50 quantiles for the rest of the statistics. We also calculated t-test and Mann-Whitney test p-values in order to test for differences in the distributions generated in the forward and backward samples. Before we calculated these statistics, we thinned our samples by taking every 40th sample starting at the 10,000th sample for a resulting sample size of 1,000. Thinning reduces the autocorrelation in the Markov chains. In each case, if we observe a large p-value, this gives us evidence that the distributions generated under forward and backward sampling have the same locations. We depict the GiR results using probability-probability (PP) plots. To compare two samples with a PP-plot we calculate the empirical quantile in each sample of a set of values observed across the two samples, then plot the sets of quantiles in the two samples against each other. If the two samples are from equivalent distributions, the quantiles should line up on a line with zero y-intercept, and unit slope (i.e., a 45-degree line). The GiR results are depicted in Figure 2, which show that we pass the test on every statistic.

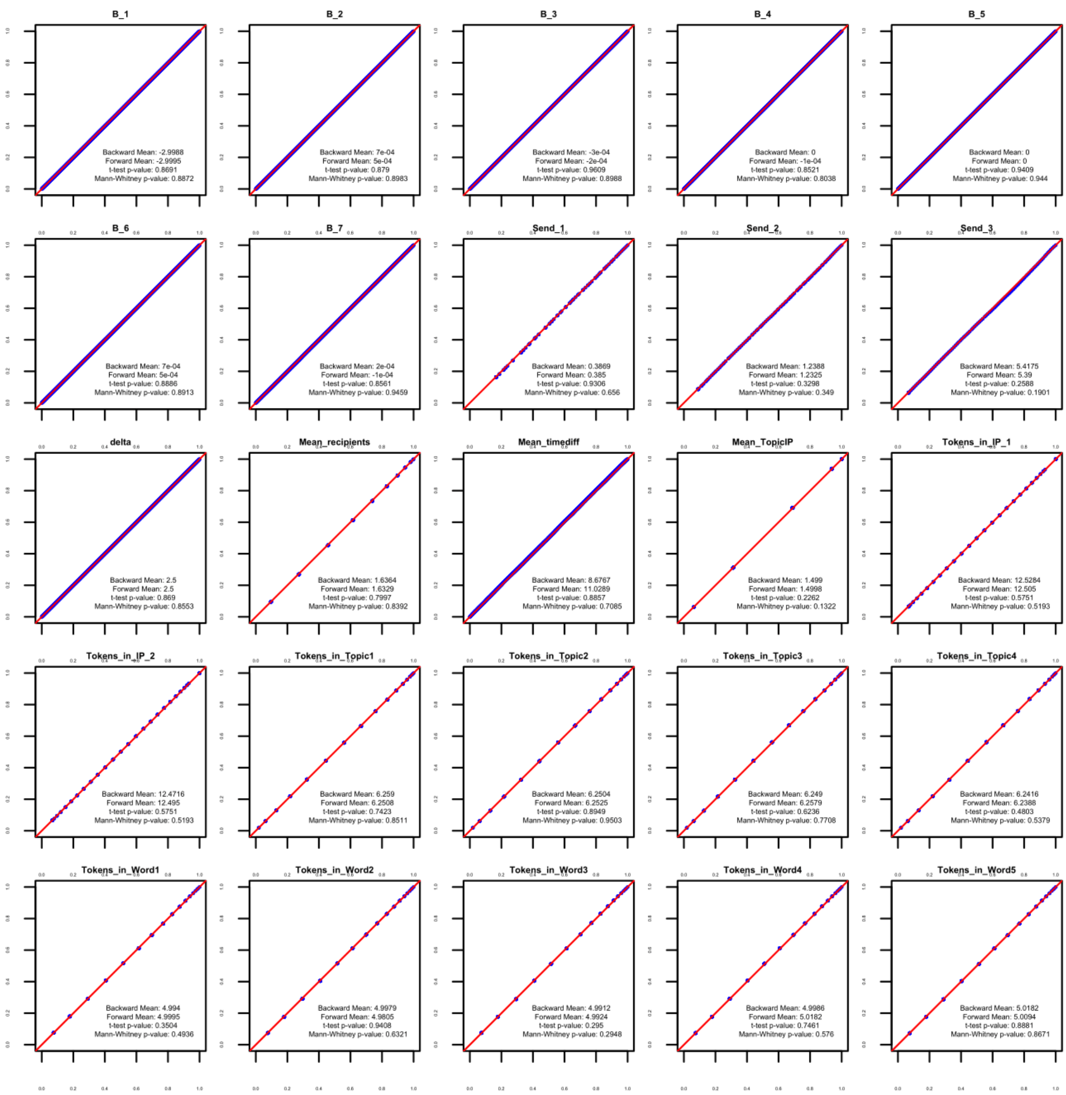

Figure 2: Probability-Probability plot for the 25 GiR test statistics.

5 Appliction: North Carolina County Government Email Communication During Hurricane Sandy

In our application of the IPTM, we use a subset of the North Carolina county government e-mail dataset collected by ben Aaron et al. (2017). This dataset includes internal e-mail corpora covering the inboxes and outboxes of managerial-level employees of North Carolina county governments. Each county corpus covers a three-month span in 2012. The full dataset covers over twenty counties, but we focus on two counties for which the time span included a notable national emergency—Hurricane Sandy (October 26, 2012—October 30, 2012). We chose these two counties, (1) in order to see whether and how communication networks surrounding Hurricane Sandy differed from those surrounding other governmental functions, and (2) to limit the scope of this initial application. Studying these two counties also presents a case study of how the geographic exposure to the storm relates to the prevalence of internal governmental communication regarding the storm. From Figure 4, we see that Dare County covers a large coastal region, including the Outer Banks, one of the most critically affected area by the hurricane. Vance County is much further inland than Dare. In this section we apply IPTM to both county corpora.

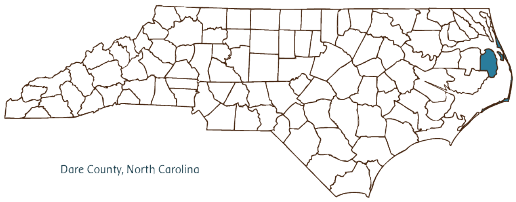

Figure 3: Geographical location of Vance County (left) and Dare County (right) in North Carolina

5.1 Exploratory Data Analysis

Before discussing the IPTM results, we present a set of exploratory analyses in which we examine characteristics of the data that are relevant to the prevalence of Hurricane Sandy in the two county government email networks. Based on this preliminary exploration, we see that Sandy received much more attention in Dare County. However, this basic descriptive finding does not shed light on whether, in one or both of the counties, the network structure of communication surrounding Sandy differed from the typical communication patterns.

5.1.1 Dare County

We ran the same exploratory analyses using the Dare County data, which spans October 1st to November 30th containing $D = 1,456$ emails between $A = 27$ actors from 22 departments, and with a vocabulary of size $W = 2907$. As we did with Vance County, we looked at three different plots—sending and receiving counts, networks, and word counts—to visualize the networks and content of the email data, with emphasis on the changes during hurricane Sandy.

In Dare County we saw considerable change in email sending/receiving behaviors during hurricane Sandy (See Figure 8). As the hurricane approaches on October 26th, the manager from the emergency services department sent significantly more emails than before, and at the same time there was dramatic rise in the receiving counts for almost every department. Further analysis demonstrated that emergency services department sent a lot of 'multicast' emails with a large number of receivers during the Sandy period.

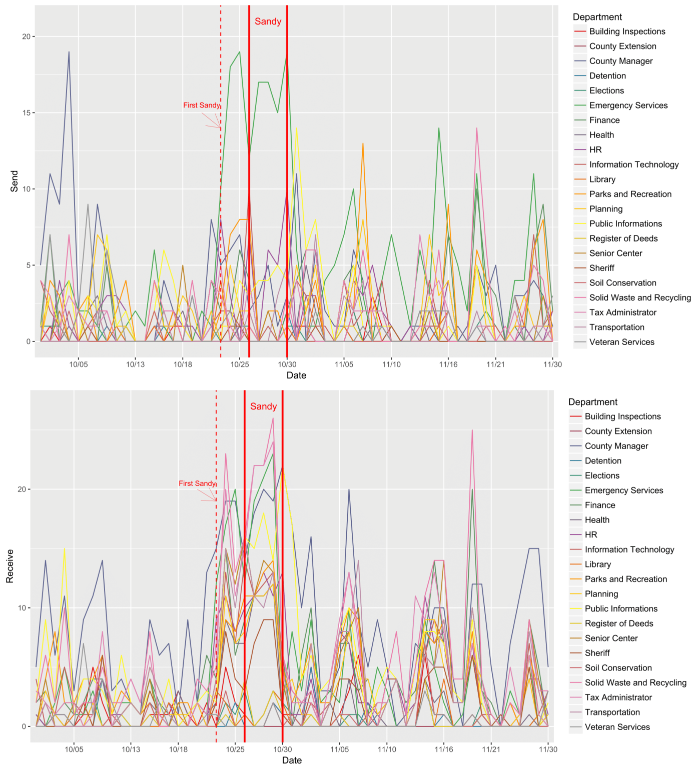

Figure 4: The number of emails sent to (upper) and received by (lower) department in Dare County

The network plots in Figure 9 illustrated same patterns we found in Figure 8. Again, the manager from the emergency services department became highly central in the network during the Sandy period, and it maintained this pattern after Sandy. Hurricane related conversations continued after Sandy passed the county, since there remained post-hurricane issues from the damage.

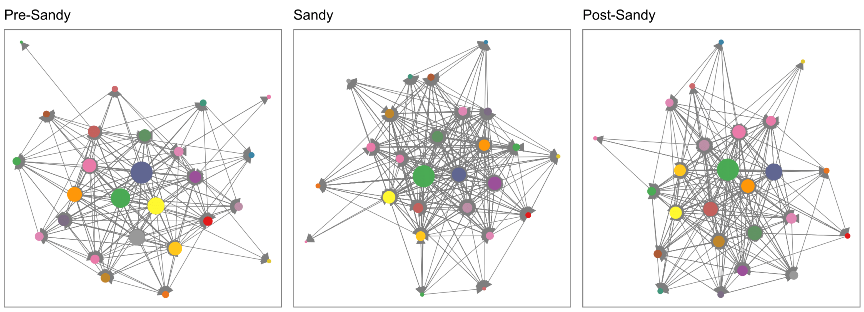

Figure 5: Network plot for three time windows: before Sandy (October 1—October 18), during Sandy (October 19—November 2), and after Sandy (November 3—November 30), in Dare County

Figure 10 reflects the hurricane's effects on email exchanges as well, and it matches our interpretations from the network aspects. Usage of the two words, 'hurricane' and 'Sandy', exploded starting a few days before Sandy arrived in Dare County, and multiple emails used the words again in November, implying the continuous discussions on hurricane-related topics.

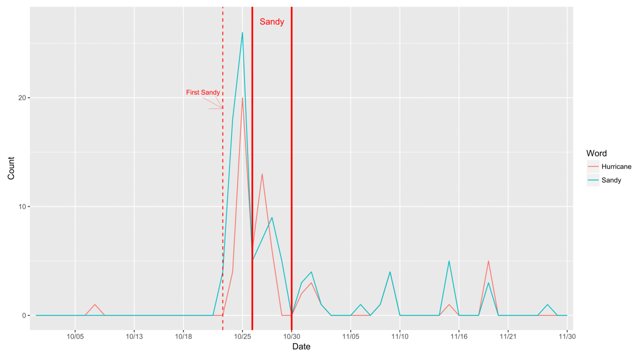

Figure 6: Frequency plot counting how many times the word 'hurricane' and 'sandy' appeared

5.2 IPTM Results

In this section we present the IPTM results. Researchers who use the IPTM can test hypotheses regarding network structure, which is a common—perhaps the most common—use of ERGM-stye models for networks. However, with the IPTM those hypotheses can be conditioned on the content area of communication. To provide an example, we articulate expectations regarding the content-conditional structural properties of the county government email networks. Information diffuses more efficiently in networks characterized by a lack of loops (Lin et al., 2010; Iribarren and Moro, 2011) and closed triangles (Roca et al., 2010; Tadić and Thurner, 2004). Assuming that the county governments' internal communication networks are characterized by efficient communication structure, we expect to see communication regarding the everyday business of the county characterized by negative reciprocity (i.e., 2-loop) effects, and negative triadic effects. The "receive" term captures reciprocity, and triadic effects are captured by 2-send, 2-receive, sibling and cosibling. Reciprocity and closed triangles are, however, common structural properties of social networks. We expect to see communication regarding personal/social matters to be characterized by positive reciprocity and triadic effects. Lastly, we have limited expectations regarding how communication surrounding Hurricane Sandy will be structured. We take an exploratory approach to the question of whether or not discussion surrounding Sandy forms an efficient communication network structure.

5.2.1 Dare County

For the IPTM application to the Dare County email data we used $C = 2$, $K = 20$, and $O = 100$[1]. Again, we applied hyperparameter optimization with $n_1 = 5$, while the inner iterations for $\mathcal{B}$ and δ were set as $n_2 = 15000$ and $n_3 = 1500$, respectively. This time, first 10000 and 500 iterations were discarded as a burn-in for inference on $\mathcal{B}$ and δ, and every 10^{th} and 5^{th} samples were taken as a thinning process for $\mathcal{B}$ and δ, respectively.

Same as Section 5.2.1, each interaction pattern is summarized with (a) Table 2: the top 15 most likely words to be generated in the topics within the interaction pattern, and (b) Figure 12: boxplots visualizing posterior estimates of dynamic network effects $\boldsymbol{b}^{(c)}$ within each interaction pattern.

We see significant differences in the contents related to each interaction pattern. Overall, 55.2% of the words were assigned to the topics in interaction pattern 1, and 44.8% were assigned to the topics in interaction pattern 2. Table 4 demonstrates 5 examples of topics from each interaction pattern, where each topic is summarized by the top 15 words. Similarly as what we saw in Vance County, interaction pattern 1 seems to represent topics commonly used by government managers. On the other hand, it is interesting that interaction pattern 2 included several topics (e.g. topic 2 and topic 18) with hurricane related words, such as 'storm', 'impacts' 'damage' and 'ocean'. In addition, one more impressive point is that topic 12 contained words related politics or election, e.g. 'survey', 'parties', and 'elections'. In summary, interaction pattern 1 represents usual administrative communications between managers

[1]Preliminary results with small number of outer iterations. Results subject to change.

in county government, and interaction pattern 2 to reflects temporary conversations driven by events or emergencies occurring in the County.

IP	**1**	**1**	**1**	**1**	**1**
Topic	**3** (0.078)	**5** (0.065)	**19** (0.064)	**13** (0.057)	**15** (0.053)
Word	water	planning	phone	questions	contact
	relocation	meter	collins	board	info
	location	room	drive	december	problem
	hills	asked	marshall	call	release
	utilities	needed	director	sheets	check
	mustian	sure	human	agenda	weather
	hydrant	afternoon	resources	nov	priority
	department	cheryl	manteo	hope	readings
	skyco	johnson	phr	item	rodanthe
	kill	issues	fax	weekly	top
	devil	case	box	management	collection
	road	letter	timesheets	internet	located
	lane	antennas	-lsb-	told	health
	tank	inspection	wanted	care	heads
	map	keep	touch	comp	ahead
IP	**2**	**2**	**2**	**2**	**2**
Topic	**14** (0.058)	**12** (0.047)	**2** (0.045)	**6** (0.044)	**18** (0.036)
Word	time	survey	road	library	status
	hours	voice	mirlo	week	system
	leave	copy	storm	working	area
	monday	discovery	beach	place	south
	administrative	regional	high	best	forecast
	employees	parties	coastal	start	track
	empoyee	disclosed	impacts	visit	pay
	work	elections	saturday	year	move
	day	pin	dot	albemarle	assessment
	friday	sending	night	librarian	opens
	october	prior	winds	web	damage
	storm	editor	hold	learning	well
	tomorrow	students	bridge	east	ocean
	hour	cost	expressed	holiday	operation
	question	residents	normal	system	addition

Table 1: Summary of topic-token assignments from Dare County data: top 15 words assigned to each topic, corresponding to interaction pattern assignments

In the Dare County analysis the two interaction patterns exhibited quite different network effects. The effects in interaction pattern 1 were generally greater in magnitude than those in interaction pattern 2, implying that topics related to interaction pattern 2 are less affected by previous email exchanges than those in interaction pattern 1. Most effects are greater in magnitude for the time interval $[t-24h, t)$ and the effect is diminishing as we move to older time intervals $[t-96h, t-24h)$ and $[t-384h, t-96h)$. On the contrary, the statistic 'send' had strongest effect in the time interval farthest in the past $[t-384h, t-96h)$. Furthermore, negative reciprocity, which we noted is expected in an efficient communication network, are found in the 'receive' statistics effect (except $[t-96h, t-24h)$), in interaction pattern 1. Most of the posterior distributions of network effects in interaction pattern 2 are centered close to zero, providing little evidence of complex network dynamics. The differences between interaction patterns can be explained by the content differences conveyed through Table 3. Since interaction pattern 2 consists of highly time-sensitive topics, ties are less likely to be effected by previous interactions.

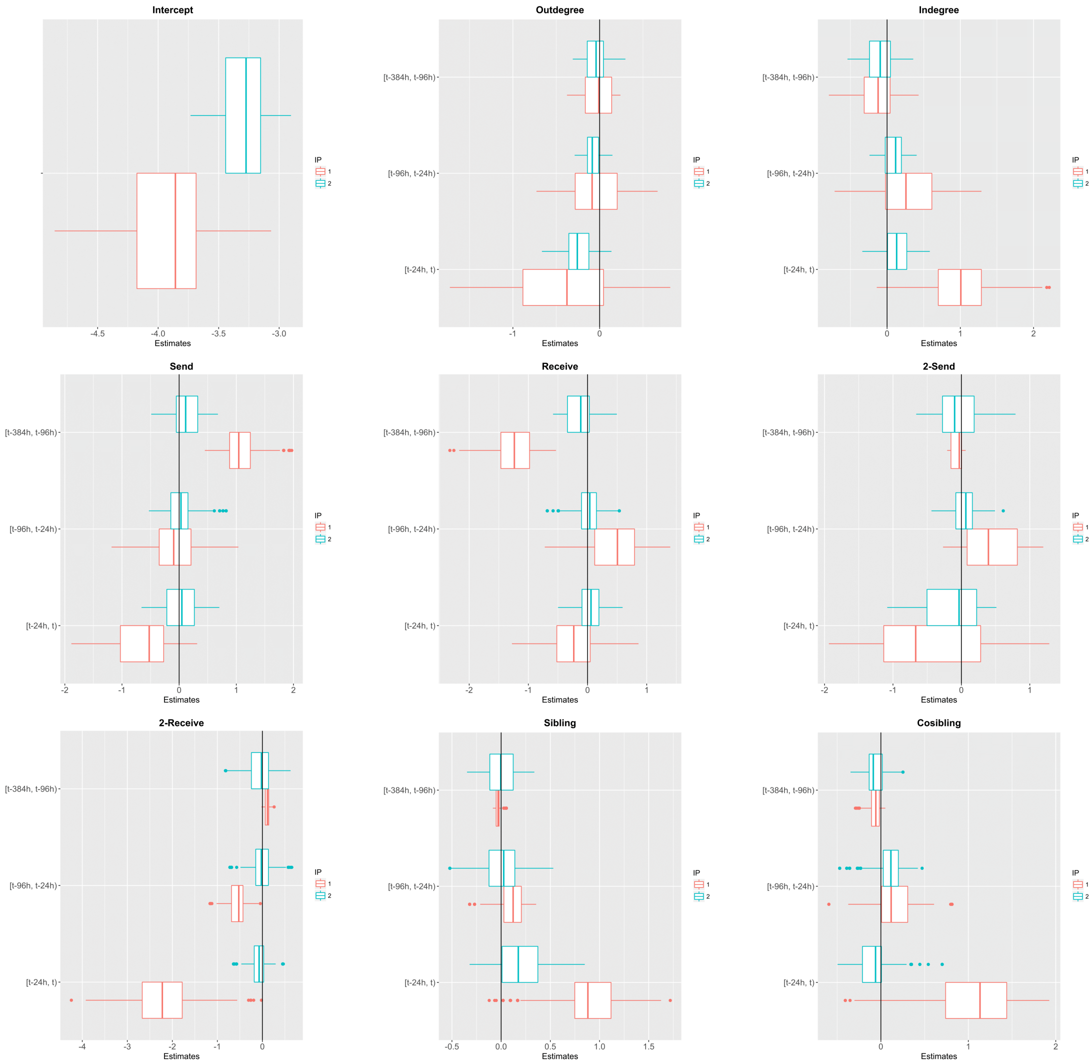

Table 2: 95% credible intervals of posterior estimates of the network effects $\boldsymbol{b}^{(c)}$: $c = 1$ (red) and $c = 2$ (green), using Dare County data

6 Topic Coherence

Topic coherence metrics Mimno et al. (2011) are often used to evaluate the semantic coherence in topic models. In order to demonstrate that incorporating network properties does not impair the ability of IPTM in modeling text, we compared the coherence of topics inferred using our model with the coherence of topics inferred using LDA. For each model, we varied the number of topics from 1 to 50 and drew three samples from the joint posterior distribution over the unobserved random variables in that model. We evaluated the topics resulting from each sample and averaged over the three samples, where the result is shown in Figure 10. Our model was not significantly different from LDA in terms of topic coherence, for any choice of the number of topics. This result, when combined with the results in Section 7 and Section 8, demonstrates that our model can achieve state-of-the-art predictive performance while producing coherent topics.

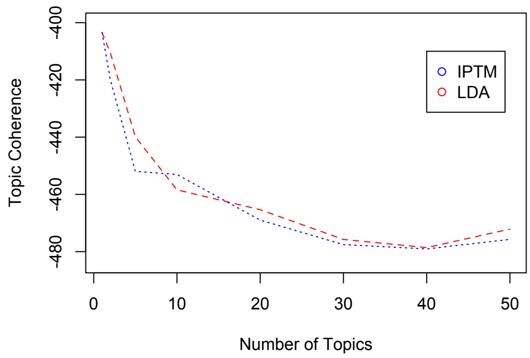

Figure 7: Average topic coherence scores for the Dare County email network (left) and the Enron data set (right).

7 Document Prediction Experiments

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

Algorithm 7 Predicting tie data for the next document (document D)

1. O, number of outer iterations of inference from which to generate predictions
2. D, the document to make a prediction
3. R, the number of iterations to sample predicted data within each outer iteration
4. $\boldsymbol{w}^D$, the observed words in the document to predict (D)
5. b, the number of burnin iterations of sampling predicted data within outer iteration
6. q, the thinning interval at which to keep predicted data within outer iteration

Run burn-in iterations for inference on documents 1 through $(D-1)$

for *o=1 to O* **do**

  run an outer iteration of inference on documents 1 through $(D-1)$

  obtain global latent variables ($\mathcal{B}$, $\mathcal{C}$, δ) and local variables ($\mathcal{Z}^{1:(D-1)}$, $\mathcal{J}_a^{1:(D-1)}$)

  initialize $N_{v|k}$ and N_k counts from the inference on $\mathcal{Z}^{1:(D-1)}$ and set $N_{k|D}=0$

  for n = *1 to* $|\boldsymbol{w}^D|$ **do**

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

      set $P(z_n^D=k|z_{1:n-1}^D, w_{1:n}^D, \mathcal{Z}^{1:(D-1)}, \boldsymbol{w}^{1:(D-1)}) = \frac{(N_{k|D}+\alpha m_k)}{(n-1+\alpha)} \times \frac{(N_{w_n^D|k}+\frac{\beta}{V})}{(N_k+\beta)}$

    end

    draw $z_n^D \sim P(z_n^D=k|z_{1:n-1}^D, w_{1:n}^D, \mathcal{Z}^{1:(D-1)}, \boldsymbol{w}^{1:(D-1)})$

    increment $N_{w_n^D|z_n^D}$ and $N_{z_n^D}$ and $N_{z_n^D|D}$

  end

  initialize $\mathcal{J}_a^{(D)}$ from the inference on $\mathcal{J}_a$ for documents 1 through $(D-1)$ as below:

  for i = *1 to* A **do**

    sample one document $d^* \sim$ Discrete-uniform$(1,(D-1))$

    set $\mathcal{J}_{a,i}^{(D)} = \mathcal{J}_{a,i}^{(d^*)}$

    (NOTE: can't directly sample from the generative process due to computational complexity when $A>10$)

  end

  calculate $p_c^{(D)}$ from $\mathcal{Z}^D$ – Equation (1)

  calculate $\lambda_{iJ_i}^{(D)}$ from $\mathcal{J}_a^{(D)}, p_c^{(D)}$ – Equation (6)

  for *r=1 to b+R* **do**

    sample $i_o^{(D)}$, $J_o^{(D)}$, and $t_o^{(D)}$ using the last 4 lines of Algorithm 11:

    - draw $i_o^{(D)} \sim$ multinomial$(\{\frac{\lambda_{iJ_i}^{(D)}}{\sum_{i\in\mathcal{A}}\lambda_{iJ_i}^{(D)}}\}_{i=1}^A)$

    - set $J_o^{(D)} = J_{a,i_o^{(D)}}^{(D)}$

    - draw $\Delta T \sim$ Exponential$(\sum_{i\in\mathcal{A}}\lambda_{iJ_i}^{(D)})$

    - set $t_o^{(D)} = t_o^{(D-1)} + \Delta T$

    for n = *1 to* $\bar{N}^{(D)}$ **do**

      Sample $z_n^{(D)}$ using Equation (27):

$$P(z_n^{(D)}=k|\mathcal{Z}_{\backslash D,n},\mathcal{C},\mathcal{B},\boldsymbol{w}^{(D)},\mathcal{J}_{\mathrm{a}},\mathcal{I}_{\mathrm{O}}^{(D)},\mathcal{J}_{\mathrm{O}}^{(D)},\mathcal{T}_{\mathrm{O}}^{(D)},\mathcal{Z}_{1:(D-1)},\boldsymbol{w}_{1:(D-1)},\beta,\boldsymbol{u},\alpha,\boldsymbol{m}).$$

    end

    Recalculate p_c^D from the updated $z^{(D)}$ **for** $i \neq i_o^{*(d)}$ **do**

      for $j \neq i$ **do**

        Sample $\mathcal{J}_{ij}^{(d)}$ using Equations (19) and (20):

$$P(\mathcal{J}_{ij}^{(D)}=J_{ij}^{(D)}|\mathcal{J}_{i\backslash j}^{(D)},\mathcal{Z},\mathcal{C},\mathcal{B},\delta,\boldsymbol{w}^{(D)},\mathcal{J}_{\mathrm{a},-i}^{(D)},\mathcal{I}_{\mathrm{O}}^{(D)},\mathcal{J}_{\mathrm{O}}^{(D)},\mathcal{T}_{\mathrm{O}}^{(D)}).$$

      end

    end

    Recalculate $\lambda_{iJ_i}^{(d)}$ from updated $\mathcal{J}_a^{(D)}$

    if $r>b$ & $r=qn$ (where n is an integer), store ($i_o^{(D)}$, $J_o^{(D)}$, $t_o^{(D)}$)

  end

end

7.1 Comparison Models

7.1.1 Ablation study

By setting the number of interaction pattern $nIP = 1$ (i.e. $\{p^d_c\}_{d=1}^D = 1$), the IPTM reduces to pure network model since the link between content and network dynamics are lost. However, this ablated model still is innovative network model in that we jointly make inference on sender, receiver, and timestamp of the document, according to the tie generating process in Section 2.4. Therefore, we first compare the ability of predicting tie data for the full IPTM vs. the ablated version of IPTM, and the main goal of this comparison will be to test whether including content information improves the predictibility on tie data.

7.1.2 Existing models

Next, we will compare the ablated verion of IPTM (in Section 7.1.1) to the comparative model that is built upon two regression models, in order to test if the ablated version of IPTM itself has a benefit over other existing models that does not jointly make inference on tie data, in terms of predictiibility. Below illustrates the comparison models we use to (separately) predict the sender, receiver, and time of d^{th} document.

The document-ahead forecasting described in Algorithm 7 provides predictions regarding (1) who sends the e-mail, (2) who receives the e-mail, and (3) when the e-mail is sent (specifically, the time between the last e-mail and the predicted e-mail). Three types of information are used in Algorithm 7 to predict these features of document D. First, Algorithm 7 uses model parameters inferred on documents 1 through $D-1$. Second, it uses the state of the history of interaction encoded in the network statistics at the time that document $D-1$ was sent. Third, it uses the content of document D. Though we are unaware of a comparison model that can be used to jointly predict the sender, receivers, and timing of an e-mail conditional upon interaction history and content, we can combine established existing models to make comparable predictions using the same inputs.

We will train three separate models to use in predicting document D recipients, sender, and timing, respectively. These models are selected to closely mirror the structure of the corresponding components of IPTM. In each model the documents used for training include documents 1 through $D-1$. The model of recipients will be a logistic regression model in which the dependent variable is the observed value of $J^{(d)}_{ij}$, where i indexes the observed sender, $i^{(d)}_o$. The network statistics, $\boldsymbol{x}^{(c)}_t(i,j)$, will be used as the covariates, with all $c = 1$ (i.e., all past interactions are within the same single interaction pattern in the comparison models). Specifically, the model used for predicting recipients of document D will have the form

$$P(J^{(D)}_{ij} = 1) = \frac{1}{1 + \exp\Big(-(b_0 + \boldsymbol{b}^T \boldsymbol{x}_t(i,j))\Big)}. \tag{9}$$

The model of when an email was sent is an exponential regression model in which the dependent variable is $t^{(d)}_i$. The single covariate in this model, assumed to have a coefficient of 1, is $\lambda^{(d)}_{iJ_i}(t)$, which is defined in Equation 6. We estimate an intercept term in this regression model in order to calibrate the scale of the model. Specifically, the probability density function of the model for the timing of the email is given by

$$f(t^{(d)}_i, \lambda^{(d)}_{iJ_i}(t), \eta) = \eta \lambda^{(d)}_{iJ_i}(t) \exp(-\eta \lambda^{(d)}_{iJ_i}(t) t^{(d)}_i), \tag{10}$$

where $\ln(\eta)$ is the intercept that calibrates the scale of the exponential distribution. The predicted sender of document D is determined by simulating $t^{(D)}_i$ for each i, and selecting the i that corresponds to the minimum value of $t^{(D)}_i$. The minimum value of $t^{(D)}_i$ is the predicted timing for document D.

Algorithm 8 Predicting tie data for document D

Input

1. $\{i_o^{(d)}\}_{d=1}^{D-1}$, Observed senders for documents 1 through $D-1$
2. $\{J_o^{(d)}\}_{d=1}^{D-1}$, Observed recipients for documents 1 through $D-1$
3. $\{t_o^{(d)}\}_{d=1}^{D-1}$, Observed timing for documents 1 through $D-1$
4. R, Number of predictions to make regarding document D

Train models defined by Equations (9) and (10) by maximum likelihood estimation using the input data.

```
for r = 1 to R do
    Draw b_0, b, and η from their sampling distributions conditional on the input data.
    for i=1 to A do
        set J_i^(D) to a vector of zeros
        while Σ_{j≠i} J_ij^(D) = 0 do
            for j ≠ i do
                Draw J_ij^(D) using Equation (9)
            end
        end
        Draw t_i^(D) using Equation (10)
    end
    Set s = min_{i∈1,2,..,A} t_i^(D)
    Store i_o^(D) = s
    Store t_o^(D) = t_s^*(D)
    Store J_o^(D) = J_s^*(D)
end
```

Running Algorithm 8 will produce R predictions regarding who sent document D, who received document D, and when document D was sent. We can then compare the IPTM and these predictions in terms of classification accuracy in predicting senders and receivers, as well as mean squared error in predicting document timing.

8 Posterior Predictive Checks

We perform posterior predictive checks Rubin et al. (1984) in order to evaluate the appropriateness of the model specification for Dare county. If the model is appropriate, the observed data should not be an outlier with respect to distributions of new data drawn from the posterior predictive distribution. To draw these comparisons, we first consider how to draw from the posterior predictive distribution. We produce a sample from the posterior predictive distribution by drawing new data conditional upon the parameters inferred in a single draw from the posterior distribution of the parameters, repeated over many draws from the posterior distribution. We already know how to generate new data conditional upon a set of inferred parameter values—that process is given by Algorithm 10, which describes how to generate data with backward sampling.

Formally, Algorithm 12 is used to draw new data, which we denote $(\mathcal{I}_O^*, \mathcal{J}_O^*, \mathcal{T}_O^*, \mathcal{W}^*)$.

$$P(\mathcal{I}_O^*, \mathcal{J}_O^*, \mathcal{T}_O^*, \mathcal{W}^* | \mathcal{Z}, \mathcal{C}, \mathcal{B}, \delta, \boldsymbol{u}, \alpha, \boldsymbol{m}, \mu_{\boldsymbol{b}}, \Sigma_{\boldsymbol{b}}, \mu_\delta, \sigma_\delta^2). \tag{11}$$

This is nearly the distribution of data that we require for posterior predictive checking, but we do not want to condition on the $\mathcal{Z}$ that came directly from inference on the observed data, as they represent document-specific (i.e., 'local') latent variables on which $\mathcal{I}_O^*, \mathcal{J}_O^*, \mathcal{T}_O^*$, and $\mathcal{W}^*$ are highly dependent. We need to somehow draw data conditional on $\mathcal{Z}$ that are independent of the observed data. This can be done via MCMC by iteratively drawing the new data, then the new latent recipients conditional upon the new data using Equations 18 and 15—which are required to draw token topic assignments, then the new token topic assignments conditional upon the new data and the new latent recipients. To recap what these equations give, Equation 25 is used to sample from

$$P(z_n^{(d)} = k|\mathcal{Z}_{\backslash d,n}, \mathcal{C}, \mathcal{B}, \delta, \mathcal{W}^*, \mathcal{J}_{\mathrm{a}}, \mathcal{I}_{\mathrm{O}}^*, \mathcal{J}_{\mathrm{O}}^*, \mathcal{T}_{\mathrm{O}}^*, \beta, \boldsymbol{u}, \alpha, \boldsymbol{m}, \mu_{\boldsymbol{b}}, \Sigma_{\boldsymbol{b}}, \mu_\delta, \sigma_\delta^2).$$

Equations 17 and 18 are used to sample from

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

We will need to run this procedure longer than we would if we marginalized the document-specific variables out of the big joint distribution to derive sampling equations for independent posterior predictive draws, but its an empirical question how long this would need to be run, and there is more than something to be said for re-using the math and code that we have already vetted so heavily. We present pseudocode in Algorithm 9 that can be used to take a single draw from the posterior predictive distribution conditional on a single outer iteration of inference.

Algorithm 9 Generate data from posterior predictive distribution

Input data :
1) $\mathcal{Z}$, estimates of token topic assignments from document 1 through D
2) $\mathcal{C}$, estimates of topic interaction pattern assignments for topic $k = 1, ..., K$
3) $\mathcal{B}$, estimates of interaction pattern parameters for intercation pattern $IP = 1, ..., nIP$
4) δ, receiver size parameter
Sample $(\mathcal{I}_{\mathrm{O}}^*, \mathcal{J}_{\mathrm{O}}^*, \mathcal{T}_{\mathrm{O}}^*, \mathcal{W}^*)$ from $P(\mathcal{I}_{\mathrm{O}}^*, \mathcal{J}_{\mathrm{O}}^*, \mathcal{T}_{\mathrm{O}}^*, \mathcal{W}^*|\mathcal{Z}, \mathcal{C}, \mathcal{B}, \delta, \boldsymbol{u}, \alpha, \boldsymbol{m})$ using Algorithm 11.
One difference from Algorithm 12 is that we initialize $\mathcal{W}^*$ as below:
Initialize NKV and NK counts from the inference on $\mathcal{Z}$ and observed $\mathcal{W}$ (instead of zeros)
Return $(\mathcal{I}_{\mathrm{O}}^*, \mathcal{J}_{\mathrm{O}}^*, \mathcal{T}_{\mathrm{O}}^*, \mathcal{W}^*)$ as a draw from the posterior distribution of the data

We would run Alogorithm 9 for O number of outer iterations, and obtain the empirical distribution of specific network properties of interest (e.g. degree distribution) and compare with the one calculated from observed data / network.

9 Conclusion

The IPTM is, to our knowledge, the first model to be capable of jointly modeling sender, receivers, time and contents in time stamped text valued networks. The IPTM incorporates innovative components, including the modeling of multicast tie formation and the conditioning of ERGM style network generative features on topic-based content. The application to North Carolina county government email data demonstrates, among other capabilities, the effectiveness at the IPTM in separating out both the content and relational structure underlying the normal day-to-day function of an organization and the management of a highly time-sensitive event—Hurricane Sandy. The IPTM is applicable to a variety of networks in which ties are attributed with textual documents. These include, for example, economic sanctions sent between countries and legislation attributed with sponsors and co-sponsors.

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.

Altman, M., Gill, J., and McDonald, M. P. (2004). *Numerical issues in statistical computing for the social scientist*, volume 508. John Wiley & Sons.

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

ben Aaron, J., Denny, M., Desmarais, B., and Wallach, H. (2017). Transparency by conformity: A field experiment evaluating openness in local governments. *Public Administration Review*, 77(1):68–77.

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

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

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

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

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

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

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

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

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

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

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

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

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

Griffiths, T. (2002). Gibbs sampling in the generative model of latent dirichlet allocation.

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

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

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

Iribarren, J. L. and Moro, E. (2011). Branching dynamics of viral information spreading. *Physical Review E*, 84(4):046116.

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

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

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

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

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

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

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

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

Lin, Y., Desouza, K. C., and Roy, S. (2010). Measuring agility of networked organizational structures via network entropy and mutual information. *Applied Mathematics and Computation*, 216(10):2824–2836.

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

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

McCullough, B. D. (2009). The accuracy of econometric software. *Handbook of computational econometrics*, pages 55–79.

Mimno, D., Wallach, H. M., Talley, E., Leenders, M., and McCallum, A. (2011). Optimizing semantic coherence in topic models. In *Proceedings of the conference on empirical methods in natural language processing*, pages 262–272. Association for Computational Linguistics.

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

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

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

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

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

Roca, C. P., Lozano, S., Arenas, A., and Sánchez, A. (2010). Topological traps control flow on real networks: The case of coordination failures. *PLoS One*, 5(12):e15210.

Rubin, D. B. et al. (1984). Bayesianly justifiable and relevant frequency calculations for the applied statistician. *The Annals of Statistics*, 12(4):1151–1172.

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

Tadić, B. and Thurner, S. (2004). Information super-diffusion on structured networks. *Physica A: Statistical Mechanics and its Applications*, 332:566–584.

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

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

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

Appendix

A Normalizing constant of non-empty Gibbs measure

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

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

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

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

This is straightforward to verify by looking at

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

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

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

Now, since

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

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

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

As such, we note that

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

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

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

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

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

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

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

B Sampling Equations

B.1 Joint distribution of latent and observed tie variables

As mentioned earlier in Section 2.4, we use data augmentation in the tie generating process. Since we should include both the observed and augmented data to make inferences on the related latent variables, the derivation steps for the contribution of tie data is not as smiple as other variables. Therefore, here we provide the detailed derivation steps for the last term of joint posterior distribution in Equation (8), starting from the likelihood before integriting out the latent time $\mathcal{T}_{\text{a}}$:

$$\begin{aligned}
&P(\mathcal{J}_{\text{a}},\mathcal{T}_{\text{a}},\mathcal{I}_{\text{o}},\mathcal{J}_{\text{o}},\mathcal{T}_{\text{o}}|\mathcal{Z},\mathcal{C},\mathcal{B},\delta)\\
&\propto P(\mathcal{J}_{\text{a}}^{(1)},...,\mathcal{J}_{\text{a}}^{(D)},\mathcal{T}_{\text{a}}^{(1)},...,\mathcal{T}_{\text{a}}^{(D)},i_{\text{o}}^{(1)},...,i_{\text{o}}^{(D)},J_{\text{o}}^{(1)},...,J_{\text{o}}^{(D)},t_{\text{o}}^{(1)},...,t_{\text{o}}^{(D)}|\mathcal{Z},\mathcal{C},\mathcal{B},\delta)\\
&\propto P(\mathcal{J}_{\text{a}}^{(D)},\mathcal{T}_{\text{a}}^{(D)},i_{\text{o}}^{(D)},J_{\text{o}}^{(D)},t_{\text{o}}^{(D)}|\mathcal{I}_{\text{o}}^{(<D)},\mathcal{J}_{\text{o}}^{(<D)},\mathcal{T}_{\text{o}}^{(<D)},\mathcal{Z},\mathcal{C},\mathcal{B},\delta).
\end{aligned} \tag{12}$$

Note that the joint likelihood of document $d = 1,...,D$ is proportional to the likelihood of D^{th} document, since the documents are not independent; all tie variables in D^{th} are conditioned on the earlier documents ($< D$) via network covariates $\boldsymbol{x}$ or past interaction history.

Now we tackle the problem of evaluating the last line of Equation (9) by deriving the likelihood for d^{th} document, $P(\mathcal{J}_{\text{a}}^{(d)},\mathcal{T}_{\text{a}}^{(d)},i_{\text{o}}^{(d)},J_{\text{o}}^{(d)},t_{\text{o}}^{(d)}|\mathcal{I}_{\text{o}}^{(<d)},\mathcal{J}_{\text{o}}^{(<d)},\mathcal{T}_{\text{o}}^{(<d)},\mathcal{Z},\mathcal{C},\mathcal{B},\delta)$. There are three steps involved. First is the generation of the latent receivers J_i for each i; second is the generation of the observed time increment $\Delta T^{(d)} = t^{(d)} - t^{(d-1)}$ from the observed sender-receiver pairs $(i_{\text{o}}^{(d)}, J_{\text{o}}^{(d)})$; and the last part is the simultaneous selection process of the observed sender, receivers, and timestamp, 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:

$$
\begin{aligned}
&P(\mathcal{J}_{\text{a}}^{(d)}, \mathcal{T}_{\text{a}}^{(d)}, i_{\text{o}}^{(d)}, J_{\text{o}}^{(d)}, t_{\text{o}}^{(d)} | \mathcal{I}_{\text{o}}^{(<d)}, \mathcal{J}_{\text{o}}^{(<d)}, \mathcal{T}_{\text{o}}^{(<d)}, \mathcal{Z}, \mathcal{C}, \mathcal{B}, \delta)\\
&= P\Big(\text{latent receivers generation}\Big) \times P\Big(\text{latent time generation}\Big) \times P\Big(\text{choose the observed}\Big)\\
&= \prod_{i\in\mathcal{A}} \Big(J_i^{(d)} \sim \text{Gibbs measure}(\{\lambda_{ij}^{(d)}\}_{j=1}^{A}, \delta)\Big) \times \prod_{i\in\mathcal{A}} \Big(\Delta T_{iJ_i}^{(d)} \sim \text{Exp}(\lambda_{iJ_i}^{(d)})\Big) \times \prod_{i\in\mathcal{A}_{\backslash i_o^{(d)}}} P\Big(\Delta T_{iJ_i}^{(d)} > \Delta T_{i_o^{(d)} J_o^{(d)}}\Big)\\
&= \Big(\prod_{i\in\mathcal{A}} \frac{1}{Z(\delta, \log(\lambda_i^{(d)}))} \exp\Big\{\log\big(\mathrm{I}(\sum_{j\in\mathcal{A}_{\backslash i}} J_{ij}^{(d)} > 0)\big) + \sum_{j\in\mathcal{A}_{\backslash i}} (\delta + \log(\lambda_{ij}^{(d)}))J_{ij}^{(d)}\Big\}\Big)\\
&\qquad\qquad \times \Big(\prod_{i\in\mathcal{A}} \lambda_{iJ_i}^{(d)} e^{-\Delta T_{iJ_i}^{(d)} \lambda_{iJ_i}^{(d)}}\Big) \times \Big(\prod_{i\in\mathcal{A}_{\backslash i_o^{(d)}}} e^{-\Delta T_{i_o^{(d)} J_o^{(d)}}^{(d)} \lambda_{iJ_i^{(d)}}^{(d)}}\Big)\\
&\propto \Big(\prod_{i\in\mathcal{A}} \frac{1}{\Big(\prod_{j\in\mathcal{A}_{\backslash i}} \Big(\exp\{\delta + \log(\lambda_{ij}^{(d)})\} + 1\Big)\Big) - 1} \exp\Big\{\sum_{j\in\mathcal{A}_{\backslash i}} (\delta + \log(\lambda_{ij}^{(d)}))J_{ij}^{(d)}\Big\}\Big)\\
&\qquad\qquad \times \Big(\lambda_{i_o^{(d)} J_o^{(d)}}^{(d)} e^{-\Delta T_{i_o^{(d)} J_o^{(d)}}^{(d)} \lambda_{i_o^{(d)} J_o^{(d)}}^{(d)}}\Big) \times \Big(\prod_{i\in\mathcal{A}_{\backslash i_o^{(d)}}} \lambda_{iJ_i}^{(d)} e^{-(\Delta T_{iJ_i}^{(d)} + \Delta T_{i_o^{(d)} J_o^{(d)}}^{(d)})\lambda_{iJ_i}^{(d)}}\Big),
\end{aligned}
\tag{13}
$$

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

$$
\begin{aligned}
&\int_0^\infty \cdots \int_0^\infty \Big(\prod_{i\in\mathcal{A}_{\backslash i_o^{(d)}}} \lambda_{iJ_i}^{(d)} e^{-(\Delta T_{iJ_i}^{(d)} + \Delta T_{i_o^{(d)} J_o^{(d)}}^{(d)})\lambda_{iJ_i}^{(d)}}\Big) d\Delta T_{1J_1}^{(d)} \cdots d\Delta T_{AJ_A}^{(d)}\\
&= \prod_{i\in\mathcal{A}_{\backslash i_o^{(d)}}} e^{-\Delta T_{i_o^{(d)} J_o^{(d)}}^{(d)} \lambda_{iJ_i}^{(d)}} \Big(\int_0^\infty \lambda_{iJ_i}^{(d)} e^{-\Delta T_{iJ_i}^{(d)} \lambda_{iJ_i}^{(d)}} d\Delta T_{iJ_i}^{(d)}\Big)\\
&= \prod_{i\in\mathcal{A}_{\backslash i_o^{(d)}}} e^{-\Delta T_{i_o^{(d)} J_o^{(d)}}^{(d)} \lambda_{iJ_i}^{(d)}} \Big(\Big[-e^{-\Delta T_{iJ_i}^{(d)} \lambda_{iJ_i}^{(d)}}\Big]_{\Delta T_{iJ_i}^{(d)}=0}^{\infty}\Big)\\
&= e^{-\Delta T_{i_o^{(d)} J_o^{(d)}}^{(d)} \sum_{i\in\mathcal{A}_{\backslash i_o^{(d)}}} \lambda_{iJ_i}^{(d)}},
\end{aligned}
\tag{14}
$$

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 (11) as below:

$$
\begin{aligned}
&P(\mathcal{J}_{\text{a}}^{(d)}, i_{\text{o}}^{(d)}, J_{\text{o}}^{(d)}, t_{\text{o}}^{(d)} | \mathcal{I}_{\text{o}}^{(<d)}, \mathcal{J}_{\text{o}}^{(<d)}, \mathcal{T}_{\text{o}}^{(<d)}, \mathcal{Z}, \mathcal{C}, \mathcal{B}, \delta)\\
&\propto \Big(\prod_{i\in\mathcal{A}} \frac{1}{\Big(\prod_{j\in\mathcal{A}_{\backslash i}} \Big(\exp\{\delta + \log(\lambda_{ij}^{(d)})\} + 1\Big)\Big) - 1} \exp\Big\{\sum_{j\in\mathcal{A}_{\backslash i}} (\delta + \log(\lambda_{ij}^{(d)}))J_{ij}^{(d)}\Big\}\Big)\\
&\qquad\qquad \times \Big(\lambda_{i_o^{(d)} J_o^{(d)}}^{(d)}\Big) \times \Big(e^{-\Delta T_{i_o^{(d)} J_o^{(d)}}^{(d)} \sum_{i\in\mathcal{A}} \lambda_{iJ_i^{(d)}}^{(d)}}\Big),
\end{aligned}
\tag{15}
$$

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

$$
\begin{aligned}
&\log\Big(P(\mathcal{J}_{\text{a}}^{(d)}, i_{\text{o}}^{(d)}, J_{\text{o}}^{(d)}, t_{\text{o}}^{(d)} | \mathcal{I}_{\text{o}}^{(<d)}, \mathcal{J}_{\text{o}}^{(<d)}, \mathcal{T}_{\text{o}}^{(<d)}, \mathcal{Z}, \mathcal{C}, \mathcal{B}, \delta)\Big)\\
&\propto \Big(\sum_{i\in\mathcal{A}} \Big(-\log\Big(\Big(\prod_{j\in\mathcal{A}_{\backslash i}} \Big(\exp\{\delta + \log(\lambda_{ij}^{(d)})\} + 1\Big)\Big) - 1\Big) + \sum_{j\in\mathcal{A}_{\backslash i}} (\delta + \log(\lambda_{ij}^{(d)}))J_{ij}^{(d)}\Big)\Big)\\
&\qquad\qquad + \Big(\log(\lambda_{i_o^{(d)} J_o^{(d)}}^{(d)})\Big) - \Big(\Delta T_{i_o^{(d)} J_o^{(d)}}^{(d)} \sum_{i\in\mathcal{A}} \lambda_{iJ_i^{(d)}}^{(d)}\Big).
\end{aligned}
\tag{16}
$$

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

First of all, for each document d, we sample the latent sender-receiver(s) pairs as in pseudocode (Algorithm 6). That is, given the observed sender of the document $i_o^{(d)}$, we sample the latent receivers for each sender $i \in \mathcal{A}_{\backslash i_o^{(d)}}$. Here we illustrate how each sender-receiver pair in the document d is updated.

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

$$P(\mathcal{J}_{ij}^{(d)} = J_{ij}^{(d)} | \mathcal{J}_{i\backslash j}^{(d)}, \mathcal{Z}, \mathcal{C}, \mathcal{B}, \delta, \mathcal{W}, \mathcal{J}_{\mathrm{a},-i}, \mathcal{I}_\mathrm{O}, \mathcal{J}_\mathrm{O}, \mathcal{T}_\mathrm{O}, \beta, \boldsymbol{u}, \alpha, \boldsymbol{m}, \mu_{\boldsymbol{b}}, \Sigma_{\boldsymbol{b}}, \mu_\delta, \sigma_\delta^2), \tag{17}$$

which we can drop some independent terms and move to

$$\begin{aligned}
&P(\mathcal{J}_{ij}^{(d)} = J_{ij}^{(d)} | \mathcal{J}_{i\backslash j}^{(d)}, i_\mathrm{O}^{(d)}, J_\mathrm{O}^{(d)}, t_\mathrm{O}^{(d)}, \mathcal{I}_\mathrm{O}^{(<d)}, \mathcal{J}_\mathrm{O}^{(<d)}, \mathcal{T}_\mathrm{O}^{(<d)}, \mathcal{Z}, \mathcal{C}, \mathcal{B}, \delta)\\
&\propto P(\mathcal{J}_{ij}^{(d)} = J_{ij}^{(d)}, \mathcal{J}_{i\backslash j}^{(d)}, i_\mathrm{O}^{(d)}, J_\mathrm{O}^{(d)}, t_\mathrm{O}^{(d)} | \mathcal{I}_\mathrm{O}^{(<d)}, \mathcal{J}_\mathrm{O}^{(<d)}, \mathcal{T}_\mathrm{O}^{(<d)}, \mathcal{Z}, \mathcal{C}, \mathcal{B}, \delta)\\
&\propto \Big(\frac{1}{\Big(\prod_{j\in\mathcal{A}_{\backslash i}}\Big(\exp\{\delta + \log(\lambda_{ij}^{(d)})\} + 1\Big)\Big) - 1} \exp\Big\{\log\big(\mathrm{I}(\sum_{j\in\mathcal{A}_{\backslash i}} J_{ij}^{(d)} > 0)\big) + \sum_{j\in\mathcal{A}_{\backslash i}} (\delta + \log(\lambda_{ij}^{(d)}))J_{ij}^{(d)}\Big\}\Big)\\
&\qquad\qquad \times \Big(\lambda^{(d)}_{i_o^{(d)} J_o^{(d)}}\Big) \times \Big(e^{-\Delta T^{(d)}_{i_o^{(d)} J_o^{(d)}} \lambda^{(d)}_{iJ_i^{(d)}}}\Big)\\
&\propto \Big(\exp\Big\{\log\big(\mathrm{I}(\sum_{j\in\mathcal{A}_{\backslash i}} J_{ij}^{(d)} > 0)\big) + \sum_{j\in\mathcal{A}_{\backslash i}} (\delta + \log(\lambda_{ij}^{(d)}))J_{ij}^{(d)}\Big\}\Big) \times \Big(e^{-\Delta T^{(d)}_{i_o^{(d)} J_o^{(d)}} \lambda^{(d)}_{iJ_i^{(d)}}}\Big),
\end{aligned} \tag{18}$$

where we replace typical use of $(-d)$ to $(< d)$ on the right hand side, due to the fact that $d^{(th)}$ document only depends on the past documents. The last line of Equation (16) is obtained by dropping the terms that do not include $J_{ij}^{(d)}$, such as the normalizing constant of Gibbs measure.

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

$$\begin{aligned}
&P(\mathcal{J}_{ij}^{(d)} = 1 | \mathcal{J}_{i\backslash j}^{(d)}, i_\mathrm{O}^{(d)}, J_\mathrm{O}^{(d)}, t_\mathrm{O}^{(d)}, \mathcal{I}_\mathrm{O}^{(<d)}, \mathcal{J}_\mathrm{O}^{(<d)}, \mathcal{T}_\mathrm{O}^{(<d)}, \mathcal{Z}, \mathcal{C}, \mathcal{B}, \delta)\\
&\propto \exp\Big(\log(1) + \sum_{j\in\mathcal{A}_{\backslash i}} (\delta + \log(\lambda_{ij}^{(d)}))J_{i[+j]}^{(d)} - \Delta T^{(d)}_{i_o^{(d)} J_o^{(d)}} \lambda^{(d)}_{iJ_{i[+j]}^{(d)}}\Big)\\
&\propto \exp\Big(\delta + \log(\lambda_{ij}^{(d)}) - \Delta T^{(d)}_{i_o^{(d)} J_o^{(d)}} \lambda^{(d)}_{iJ_{i[+j]}^{(d)}}\Big),
\end{aligned} \tag{19}$$

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

$$\begin{aligned}
&P(\mathcal{J}_{ij}^{(d)} = 0 | \mathcal{J}_{i\backslash j}^{(d)}, i_\mathrm{O}^{(d)}, J_\mathrm{O}^{(d)}, t_\mathrm{O}^{(d)}, \mathcal{I}_\mathrm{O}^{(<d)}, \mathcal{J}_\mathrm{O}^{(<d)}, \mathcal{T}_\mathrm{O}^{(<d)}, \mathcal{Z}, \mathcal{C}, \mathcal{B}, \delta)\\
&\propto \exp\Big(\log\big(\mathrm{I}(\sum_{j\in\mathcal{A}_{\backslash i}} J_{i[-j]}^{(d)} > 0)\big) + \sum_{j\in\mathcal{A}_{\backslash i}} (\delta + \log(\lambda_{ij}^{(d)}))J_{i[-j]}^{(d)} - \Delta T^{(d)}_{i_o^{(d)} J_o^{(d)}} \lambda^{(d)}_{iJ_{i[-j]}^{(d)}}\Big)\\
&\propto \exp\Big(\log\big(\mathrm{I}(\sum_{j\in\mathcal{A}_{\backslash i}} J_{i[-j]}^{(d)} > 0)\big) - \Delta T^{(d)}_{i_o^{(d)} J_o^{(d)}} \lambda^{(d)}_{iJ_{i[-j]}^{(d)}}\Big),
\end{aligned} \tag{20}$$

where $J_{i[-j]}^{(d)}$ meaning similarly that the j^{th} element of $J_i^{(d)}$ is fixed as 0. In this case, we cannot guarantee that $\mathrm{I}(\sum_{j\in\mathcal{A}_{\backslash i}} J_{ij}^{(d)} > 0)$ is 0 or 1, so we have to leave the term. When it is zero, $\exp\{\log\big(\mathrm{I}(\sum_{j\in\mathcal{A}_{\backslash i}} J_{ij}^{(d)} > 0)\big)\} = 0$, thus we will sample 1 with probability 1. From this property of non-empty Gibbs measure, we prevent from the instances where the sender has no recipients to send the document. Now we can use multinomial sampling using the two probabilities, Equation (17) and Equation (18), which is equivalent to Bernoulli sampling with probability $\frac{P(\mathcal{J}_{ij}^{(d)}=1)}{P(\mathcal{J}_{ij}^{(d)}=0)+P(\mathcal{J}_{ij}^{(d)}=0)}$.

B.3 Resampling $\mathcal{Z}$

Second, we 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, and we derive the sampling equation by starting from the conditional distribution used in Latent Dirichlet allocation (Blei et al., 2003):

$$\begin{aligned} &P(\boldsymbol{w}^{(d)}, \boldsymbol{z}^{(d)} | \mathcal{W}_{\backslash d}, \mathcal{Z}_{\backslash d}, \beta, \boldsymbol{u}, \alpha, \boldsymbol{m}) \\ &\propto \prod_{n=1}^{N^{(d)}} P(z_n^{(d)} = k, w_n^{(d)} = w | \mathcal{W}_{\backslash d,n}, \mathcal{Z}_{\backslash d,n}, \beta, \boldsymbol{u}, \alpha, \boldsymbol{m}). \end{aligned} \tag{21}$$

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

$$\begin{aligned} &P(z_n^{(d)} = k, w_n^{(d)} = w | \mathcal{W}_{\backslash d,n}, \mathcal{Z}_{\backslash d,n}, \beta, \boldsymbol{u}, \alpha, \boldsymbol{m}) \\ &\propto \frac{P(\mathcal{W}, \mathcal{Z} | \beta, \boldsymbol{u}, \alpha, \boldsymbol{m})}{P(\mathcal{W}_{\backslash d,n}, \mathcal{Z}_{\backslash d,n} | \beta, \boldsymbol{u}, \alpha, \boldsymbol{m})} \\ &\propto \frac{\prod_{k=1}^{K} \frac{\prod_{w=1}^{W} \Gamma(N_{wk}^{WK} + \beta u_w)}{\Gamma(\sum_{w=1}^{W} N_{wk}^{WK} + \beta)} \times \prod_{k=1}^{K} \frac{\Gamma(N_{k|d} + \alpha m_k)}{\Gamma(N_{\cdot|d} + \alpha)}}{\prod_{k=1}^{K} \frac{\prod_{w=1}^{W} \Gamma(N_{wk,\backslash d,n}^{WK} + \beta u_w)}{\Gamma(\sum_{w=1}^{W} N_{wk,\backslash d,n}^{WK} + \beta)} \times \prod_{k=1}^{K} \frac{\Gamma(N_{k|d,\backslash d,n} + \alpha m_k)}{\Gamma(N_{\cdot|d,\backslash d,n} + \alpha)}} \\ &\propto \frac{N_{wk,\backslash d,n}^{WK} + \frac{\beta}{W}}{\sum_{w=1}^{W} N_{wk,\backslash d,n}^{WK} + \beta} \times \frac{N_{k|d,\backslash d,n} + \alpha m_k}{N^{(d)} - 1 + \alpha}. \end{aligned} \tag{22}$$

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

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

Now, considering the modeling framework of IPTM, we re-derive the sampling equation reflecting the network effects as well:

$$\begin{aligned} &P(z_n^{(d)} = k | \mathcal{Z}_{\backslash d,n}, \mathcal{C}, \mathcal{B}, \delta, \mathcal{W}, \mathcal{J}_{\mathrm{a}}, \mathcal{I}_{\mathrm{O}}, \mathcal{J}_{\mathrm{O}}, \mathcal{T}_{\mathrm{O}}, \beta, \boldsymbol{u}, \alpha, \boldsymbol{m}, \mu_{\boldsymbol{b}}, \Sigma_{\boldsymbol{b}}, \mu_\delta, \sigma_\delta^2) \\ &\propto P(z_n^{(d)} = k, w_n^{(d)}, \mathcal{J}_{\mathrm{a}}^{(\geq d)}, i_{\mathrm{O}}^{(\geq d)}, J_{\mathrm{O}}^{(\geq d)}, t_{\mathrm{O}}^{(\geq d)} | \mathcal{Z}_{\backslash d,n}, \mathcal{C}, \mathcal{B}, \delta, \mathcal{W}_{\backslash d,n}, \mathcal{I}_{\mathrm{O}}^{(<d)}, \mathcal{J}_{\mathrm{O}}^{(<d)}, \mathcal{T}_{\mathrm{O}}^{(<d)}, \beta, \boldsymbol{u}, \alpha, \boldsymbol{m}) \\ &\propto P(z_n^{(d)} = k | \mathcal{Z}_{\backslash d,n}, \alpha, \boldsymbol{m}) P(w_n^{(d)} | z_n^{(d)} = k, \mathcal{W}_{\backslash d,n}, \mathcal{Z}_{\backslash d,n}, \beta, \boldsymbol{u}) \times P(\mathcal{J}_{\mathrm{a}}^{(d^*)}, i_{\mathrm{O}}^{(d^*)}, J_{\mathrm{O}}^{(d^*)}, t_{\mathrm{O}}^{(d^*)} | z_n^{(d)} = k, \mathcal{Z}_{\backslash d,n}, \mathcal{C}, \mathcal{B}, \delta), \end{aligned} \tag{23}$$

where the subscript "$\backslash d, n$" denotes the exclsuion of position n in d^{th} document. Note that since selecting a topic for any token influences the histories acting on documents from the current one d to the future ones with the timepoints less than $t_o^{(d)} + 384$, we define $d^* = \text{argmax}_{d'}\{t_o^{(d')} \leq t_o^{(d)} + 384\}$ to correctly evaluate tie contribution part. From Equation (20), we know that:

$$P(z_n^{(d)} = k | \mathcal{Z}_{\backslash d,n}, \alpha, \boldsymbol{m}) = \frac{N_{\backslash d,n}^{(k|d)} + \alpha \boldsymbol{m}_k}{N^{(d)} - 1 + \alpha} \tag{24}$$

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

$$P(w_n^{(d)} | z_n^{(d)} = k, \mathcal{W}_{\backslash d,n}, \mathcal{Z}_{\backslash d,n}, \beta, \boldsymbol{u}) = \frac{N_{\backslash d,n}^{(w_n^{(d)}|k)} + \frac{\beta}{W}}{N_{\backslash d,n}^{(k)} + \beta}, \tag{25}$$

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 $N_{\backslash d,n}^{(k)} = \sum_{w=1}^{W} N_{\backslash d,n}^{(w_n^{(d)}|k)}$. We already have shown in Section B.1 that

$$\begin{aligned} &P(\mathcal{J}_{\mathrm{a}}^{(d)}, i_{\mathrm{O}}^{(d)}, J_{\mathrm{O}}^{(d)}, t_{\mathrm{O}}^{(d)} | z_n^{(d)} = k, \mathcal{Z}_{\backslash d,n}, \mathcal{C}, \mathcal{B}, \delta) \\ &= \Big(\prod_{i \in \mathcal{A}} \frac{1}{\Big(\prod_{j \in \mathcal{A}_{\backslash i}} \big(\exp\{\delta + \log(\lambda_{ij}^{(d)})\} + 1\big)\Big) - 1} \exp\Big\{\sum_{j \in \mathcal{A}_{\backslash i}} (\delta + \log(\lambda_{ij}^{(d)})) J_{ij}^{(d)}\Big\}\Big) \times \Big(\lambda_{i_o^{(d)} J_o^{(d)}}^{(d)}\Big) \times \Big(e^{-\Delta T_{i_o^{(d)} J_o^{(d)}}^{(d)} \lambda_{i J_i^{(d)}}^{(d)}}\Big), \end{aligned} \tag{26}$$

where every part includes $\lambda_{ij}^{(d)}$ such that we cannot simplify any further. Therefore, if $N^{(d)} > 0$, the conditional probability of n^{th} word in document d being topic k is:

$$\begin{aligned}
&P(z_n^{(d)} = k|\mathcal{Z}_{\backslash d,n}, \mathcal{C}, \mathcal{B}, \delta, \mathcal{W}, \mathcal{J}_{\text{a}}, \mathcal{I}_{\text{O}}, \mathcal{J}_{\text{O}}, \mathcal{T}_{\text{O}}, \beta, \boldsymbol{u}, \alpha, \boldsymbol{m}, \mu_{\boldsymbol{b}}, \Sigma_{\boldsymbol{b}}, \mu_\delta, \sigma_\delta^2) \\
&\propto (N_{\backslash d,n}^{(k|d)} + \alpha \boldsymbol{m}_k) \times \frac{N_{\backslash d,n}^{(w_n^{(d)}|k)} + \frac{\beta}{W}}{N_{\backslash d,n}^{(k)} + \beta} \times \\
&\Big(\prod_{i \in \mathcal{A}} \frac{1}{\Big(\prod_{j \in \mathcal{A}_{\backslash i}} \Big(\exp\{\delta + \log(\lambda_{ij}^{(d^*)})\} + 1\Big)\Big) - 1} \exp\Big\{\sum_{j \in \mathcal{A}_{\backslash i}} (\delta + \log(\lambda_{ij}^{(d^*)})) J_{ij}^{(d^*)}\Big\}\Big) \times \Big(\lambda_{i_o^{(d^*)} J_o^{(d^*)}}^{(d^*)} e^{-\Delta T_{i_o^{(d^*)} J_o^{(d^*)}}^{(d^*)} \lambda_{iJ_i^{(d^*)}}^{(d^*)}}\Big),
\end{aligned} \tag{27}$$

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

$$\begin{aligned}
&P(z_1^{(d)} = k|\mathcal{Z}_{\backslash d,1} = \emptyset, \mathcal{C}, \mathcal{B}, \delta, \mathcal{W}, \mathcal{J}_{\text{a}}, \mathcal{I}_{\text{O}}, \mathcal{J}_{\text{O}}, \mathcal{T}_{\text{O}}, \beta, \boldsymbol{u}, \alpha, \boldsymbol{m}, \mu_{\boldsymbol{b}}, \Sigma_{\boldsymbol{b}}, \mu_\delta, \sigma_\delta^2) \\
&\propto \Big(\prod_{i \in \mathcal{A}} \frac{1}{\Big(\prod_{j \in \mathcal{A}_{\backslash i}} \Big(\exp\{\delta + \log(\lambda_{ij}^{(d^*)})\} + 1\Big)\Big) - 1} \exp\Big\{\sum_{j \in \mathcal{A}_{\backslash i}} (\delta + \log(\lambda_{ij}^{(d^*)})) J_{ij}^{(d^*)}\Big\}\Big) \times \Big(\lambda_{i_o^{(d^*)} J_o^{(d^*)}}^{(d^*)} e^{-\Delta T_{i_o^{(d^*)} J_o^{(d^*)}}^{(d^*)} \lambda_{iJ_i^{(d^*)}}^{(d^*)}}\Big),
\end{aligned} \tag{28}$$

where $d^* = \text{argmax}_{d'}\{t_o^{(d')} \leq t_o^{(d)} + 384\}$.

B.4 Resampling $\mathcal{C}$

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

$$\begin{aligned}
&P(c_k = c|\mathcal{Z}, \mathcal{C}_{\backslash k}, \mathcal{B}, \delta, \mathcal{W}, \mathcal{J}_{\text{a}}, \mathcal{I}_{\text{O}}, \mathcal{J}_{\text{O}}, \mathcal{T}_{\text{O}}, \beta, \boldsymbol{u}, \alpha, \boldsymbol{m}, \mu_{\boldsymbol{b}}, \Sigma_{\boldsymbol{b}}, \mu_\delta, \sigma_\delta^2) \\
&\propto P(c_k = c, \mathcal{J}_{\text{a}}, \mathcal{I}_{\text{O}}, \mathcal{J}_{\text{O}}, \mathcal{T}_{\text{O}}|\mathcal{Z}, \mathcal{C}_{\backslash k}, \mathcal{B}, \delta) \\
&\propto P(c_k = c) P(\mathcal{J}_{\text{a}}, \mathcal{I}_{\text{O}}, \mathcal{J}_{\text{O}}, \mathcal{T}_{\text{O}}|\mathcal{Z}, c_k = c, \mathcal{C}_{\backslash k}, \mathcal{B}, \delta)
\end{aligned} \tag{29}$$

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

$$\begin{aligned}
&P(c_k = c|\mathcal{Z}, \mathcal{C}_{\backslash k}, \mathcal{B}, \delta, \mathcal{W}, \mathcal{I}_{\text{a}}, \mathcal{J}_{\text{a}}, \mathcal{I}_{\text{O}}, \mathcal{J}_{\text{O}}, \mathcal{T}_{\text{O}}, \beta, \boldsymbol{u}, \alpha, \boldsymbol{m}, \mu_{\boldsymbol{b}}, \Sigma_{\boldsymbol{b}}, \mu_\delta, \sigma_\delta^2) \\
&\propto P(\mathcal{J}_{\text{a}}, \mathcal{I}_{\text{O}}, \mathcal{J}_{\text{O}}, \mathcal{T}_{\text{O}}|\mathcal{Z}, c_k = c, \mathcal{C}_{\backslash k}, \mathcal{B}, \delta) \\
&= \Big(\prod_{i \in \mathcal{A}} \frac{1}{\Big(\prod_{j \in \mathcal{A}_{\backslash i}} \Big(\exp\{\delta + \log(\lambda_{ij}^{(D)})\} + 1\Big)\Big) - 1} \exp\Big\{\sum_{j \in \mathcal{A}_{\backslash i}} (\delta + \log(\lambda_{ij}^{(D)})) J_{ij}^{(D)}\Big\}\Big) \\
&\qquad \times \Big(\lambda_{i_o^{(D)} J_o^{(D)}}^{(D)}\Big) \times \Big(e^{-\Delta T_{i_o^{(D)} J_o^{(D)}}^{(D)} \lambda_{iJ_i^{(D)}}^{(D)}}\Big).
\end{aligned} \tag{30}$$

Note that for this one, $\mathcal{C}$, we need to consider all documents so we used the likelihood of D^{th} document.

B.5 Resampling $\mathcal{B}$

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

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

After factorization, we get

$$
\begin{aligned}
&\frac{P(\mathcal{B}'|\mathcal{Z},\mathcal{C},\delta,\mathcal{W},\mathcal{J}_{\text{a}},\mathcal{I}_{\text{O}},\mathcal{J}_{\text{O}},\mathcal{T}_{\text{O}},\beta,\boldsymbol{u},\alpha,\boldsymbol{m},\mu_{\boldsymbol{b}},\Sigma_{\boldsymbol{b}},\mu_\delta,\sigma_\delta^2)}{P(\mathcal{B}|\mathcal{Z},\mathcal{C},\delta,\mathcal{W},\mathcal{J}_{\text{a}},\mathcal{I}_{\text{O}},\mathcal{J}_{\text{O}},\mathcal{T}_{\text{O}},\beta,\boldsymbol{u},\alpha,\boldsymbol{m},\mu_{\boldsymbol{b}},\Sigma_{\boldsymbol{b}},\mu_\delta,\sigma_\delta^2)}\\
=&\frac{P(\mathcal{Z},\mathcal{C},\mathcal{B}',\delta,\mathcal{W},\mathcal{J}_{\text{a}},\mathcal{I}_{\text{O}},\mathcal{J}_{\text{O}},\mathcal{T}_{\text{O}}|\beta,\boldsymbol{u},\alpha,\boldsymbol{m},\mu_{\boldsymbol{b}},\Sigma_{\boldsymbol{b}},\mu_\delta,\sigma_\delta^2)}{P(\mathcal{Z},\mathcal{C},\mathcal{B},\delta,\mathcal{W},\mathcal{J}_{\text{a}},\mathcal{I}_{\text{O}},\mathcal{J}_{\text{O}},\mathcal{T}_{\text{O}}|\beta,\boldsymbol{u},\alpha,\boldsymbol{m},\mu_{\boldsymbol{b}},\Sigma_{\boldsymbol{b}},\mu_\delta,\sigma_\delta^2)}\\
=&\frac{P(\mathcal{B}'|\mathcal{C},\mu_{\boldsymbol{b}},\Sigma_{\boldsymbol{b}})P(\mathcal{J}_{\text{a}},\mathcal{I}_{\text{O}},\mathcal{J}_{\text{O}},\mathcal{T}_{\text{O}}|\mathcal{Z},\mathcal{C},\mathcal{B}',\delta)}{P(\mathcal{B}|\mathcal{C},\mu_{\boldsymbol{b}},\Sigma_{\boldsymbol{b}})P(\mathcal{J}_{\text{a}},\mathcal{I}_{\text{O}},\mathcal{J}_{\text{O}},\mathcal{T}_{\text{O}}|\mathcal{Z},\mathcal{C},\mathcal{B},\delta)},
\end{aligned}
\tag{32}
$$

where $P(\mathcal{B}|\mathcal{C},\mu_{\boldsymbol{b}},\Sigma_{\boldsymbol{b}})$ is calculated from the product of $\boldsymbol{b}^{(c)} \sim \text{Multivariate Normal}(\mu_{\boldsymbol{b}},\Sigma_{\boldsymbol{b}})$ over the interaction patterns $c \in \{1,...,C\}$ (as defined in Section 2) and $P(\mathcal{J}_{\text{a}},\mathcal{I}_{\text{O}},\mathcal{J}_{\text{O}},\mathcal{T}_{\text{O}}|\mathcal{Z},\mathcal{C},\mathcal{B},\delta)$ is the same as Equation (28). Again, we take the log and obtain the log of acceptance ratio:

$$
\begin{aligned}
&\sum_{c=1}^{C}\log\Big(\mathcal{N}(\boldsymbol{b}'^{(c)};\mu_{\boldsymbol{b}},\Sigma_{\boldsymbol{b}})\Big)-\sum_{c=1}^{C}\log\Big(\mathcal{N}(\boldsymbol{b}^{(c)};\mu_{\boldsymbol{b}},\Sigma_{\boldsymbol{b}})\Big)\\
&+\Bigg(\Big(\sum_{i\in\mathcal{A}}\Big(-\log\Big(\Big(\prod_{j\in\mathcal{A}_{\setminus i}}\Big(\exp\{\delta+\log(\lambda_{ij}^{(D)})\}+1\Big)\Big)-1\Big)+\sum_{j\in\mathcal{A}_{\setminus i}}(\delta+\log(\lambda_{ij}^{(D)}))J_{ij}^{(D)}\Big)\\
&\qquad+\Big(\log(\lambda^{(D)}_{i_o^{(D)}J_o^{(D)}})-\Delta T^{(D)}_{i_o^{(D)}J_o^{(D)}}\sum_{i\in\mathcal{A}}\lambda^{(D)}_{iJ_i^{(D)}}\Big)\text{ given }\boldsymbol{b}'\Bigg)\\
&-\Bigg(\Big(\sum_{i\in\mathcal{A}}\Big(-\log\Big(\Big(\prod_{j\in\mathcal{A}_{\setminus i}}\Big(\exp\{\delta+\log(\lambda_{ij}^{(D)})\}+1\Big)\Big)-1\Big)+\sum_{j\in\mathcal{A}_{\setminus i}}(\delta+\log(\lambda_{ij}^{(D)}))J_{ij}^{(D)}\Big)\\
&\qquad+\Big(\log(\lambda^{(D)}_{i_o^{(D)}J_o^{(D)}})-\Delta T^{(D)}_{i_o^{(D)}J_o^{(D)}}\sum_{i\in\mathcal{A}}\lambda^{(D)}_{iJ_i^{(D)}}\Big)\text{ given }\boldsymbol{b}\Bigg),
\end{aligned}
\tag{33}
$$

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

$$\log(\text{Acceptance Probability}) = \min(\text{Equation (31), 0}). \tag{34}$$

Use the log of acceptance ratio, if the log of a sample from Uniform(0,1) is less than the log-acceptance probability (31), we accept the proposal $\boldsymbol{b}'$. Otherwise, we reject.

B.6 Resampling δ

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

$$
\text{Acceptance Probability} = \begin{cases}\frac{P(\delta'|\mu_\delta,\sigma_\delta^2)P(\mathcal{J}_{\text{a}},\mathcal{I}_{\text{O}},\mathcal{J}_{\text{O}},\mathcal{T}_{\text{O}}|\mathcal{Z},\mathcal{C},\mathcal{B},\delta')}{P(\delta|\mu_\delta,\sigma_\delta^2)P(\mathcal{J}_{\text{a}},\mathcal{I}_{\text{O}},\mathcal{J}_{\text{O}},\mathcal{T}_{\text{O}}|\mathcal{Z},\mathcal{C},\mathcal{B},\delta)} & \text{if} < 1\\ 1 \quad \text{else}\end{cases}
\tag{35}
$$

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

$$
\begin{aligned}
&\log\Big(\mathcal{N}(\delta';\mu_\delta,\sigma_\delta^2)\Big)-\log\Big(\mathcal{N}(\delta;\mu_\delta,\sigma_\delta^2)\Big)\\
&+\Bigg(\sum_{i\in\mathcal{A}}\Big(-\log\Big(\Big(\prod_{j\in\mathcal{A}_{\setminus i}}\Big(\exp\{\delta'+\log(\lambda_{ij}^{(D)})\}+1\Big)\Big)-1\Big)+\sum_{j\in\mathcal{A}_{\setminus i}}(\delta'+\log(\lambda_{ij}^{(D)}))J_{ij}^{(D)}\Big)\\
&\quad-\sum_{i\in\mathcal{A}}\Big(-\log\Big(\Big(\prod_{j\in\mathcal{A}_{\setminus i}}\Big(\exp\{\delta+\log(\lambda_{ij}^{(D)})\}+1\Big)\Big)-1\Big)+\sum_{j\in\mathcal{A}_{\setminus i}}(\delta+\log(\lambda_{ij}^{(D)}))J_{ij}^{(D)}\Big)\Bigg),
\end{aligned}
\tag{36}
$$

and determine whether to accept or reject using the log of acceptance ratio

$$\log(\text{Acceptance Probability}) = \min(\text{Equation (34), 0}). \tag{37}$$

C Details on Getting It Right Test

C.1 Collapsed-time Tie Generating Process

Considering that we integrated out latent time $\mathcal{T}_{\text{a}}$ in the inference, we develop the new generative process for tie data with the latent time variable integrated out. Note that this is built upon the property of the minimum of independent Exponential random variables, where the probability ΔT_{iJ_i} being the minimum is $\frac{\lambda_{iJ_i}^{(d)}}{\sum_{i=1}^{A}\lambda_{iJ_i}^{(d)}}$. Details are illustrated in Algorithm 8.

Algorithm 10 Collapsed-time Tie Generating Process

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

With this generative process, the joint likelihood (comparable to Equation (10)) becomes:

$$
\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)\\
&= P\Big(\text{latent receivers generation}\Big) \times P\Big(\text{choose the sender}\Big) \times P\Big(\text{observed minimum time generation}\Big)\\
&= \prod_{i\in\mathcal{A}}\Big(J_i^{(d)} \sim \text{Gibbs measure}(\{\lambda_{ij}^{(d)}\}_{j=1}^{A}, \delta)\Big) \times \Big(i_o^{(d)} \sim \text{Multinom}(\{\frac{\lambda_{iJ_i}^{(d)}}{\sum\limits_{i\in\mathcal{A}}\lambda_{iJ_i}^{(d)}}\}_{i=1}^{A})\Big) \times \Big(\Delta T_{i^{(d)}J^{(d)}}^{(d)} \sim \text{Exp}(\sum_{i\in\mathcal{A}}\lambda_{i_o^{(d)}J_o^{(d)}}^{(d)})\Big)\\
&= \Big(\prod_{i\in\mathcal{A}}\frac{1}{Z(\delta, \log(\lambda_i^{(d)})}\exp\Big\{\sum_{j\in\mathcal{A}_{\setminus i}}(\delta + \log(\lambda_{ij}^{(d)}))J_{ij}^{(d)}\Big\}\Big) \times \Big(\frac{\lambda_{i_o^{(d)}J_o^{(d)}}^{(d)}}{\sum\limits_{i\in\mathcal{A}}\lambda_{iJ_i}^{(d)}}\Big) \times \Big((\sum_{i\in\mathcal{A}}\lambda_{iJ_i}^{(d)})e^{-\Delta T_{i_o^{(d)}J_o^{(d)}}^{(d)}\sum\limits_{i\in\mathcal{A}}\lambda_{iJ_i}^{(d)}}\Big)\\
&= \Big(\prod_{i\in\mathcal{A}}\frac{1}{Z(\delta, \log(\lambda_i^{(d)})}\exp\Big\{\sum_{j\in\mathcal{A}_{\setminus i}}(\delta + \log(\lambda_{ij}^{(d)}))J_{ij}^{(d)}\Big\}\Big) \times \Big(\frac{\lambda_{i_o^{(d)}J_o^{(d)}}^{(d)}}{\sum\limits_{i\in\mathcal{A}}\lambda_{iJ_i}^{(d)}}\Big) \times \Big((\sum_{i\in\mathcal{A}}\lambda_{iJ_i}^{(d)})e^{-\Delta T_{i_o^{(d)}J_o^{(d)}}^{(d)}\sum\limits_{i\in\mathcal{A}}\lambda_{iJ_i}^{(d)}}\Big)\\
&\propto \Big(\prod_{i\in\mathcal{A}}\frac{1}{\Big(\prod_{j\in\mathcal{A}_{\setminus i}}\Big(\exp\{\delta + \log(\lambda_{ij}^{(d)})\} + 1\Big)\Big) - 1}\exp\Big\{\sum_{j\in\mathcal{A}_{\setminus i}}(\delta + \log(\lambda_{ij}^{(d)}))J_{ij}^{(d)}\Big\}\Big)\\
&\qquad\qquad \times \Big(\lambda_{i_o^{(d)}J_o^{(d)}}^{(d)}\Big) \times \Big(e^{-\Delta T_{i_o^{(d)}J_o^{(d)}}^{(d)}\sum\limits_{i\in\mathcal{A}}\lambda_{iJ_i}^{(d)}}\Big),
\end{aligned}
\tag{38}
$$

which is exactly the same as Equation (12), thus we will use this collpased-time generative process as a forward/backward generative process in Geweke's "Getting it Right" test in Section C.2.

C.2 Backward Generating Process

For backward sampling, we let NKV be a $V \times K$ dimensional matrix where each entry will record the count of the number of tokens of word-type v that are currently assigned to topic k. Also let NK be a K dimensional vector recording the total count of tokens currently assigned to topic k. Word-assignments are implemented via collapsed Gibbs sampling (Griffiths, 2002), while the generation of tie data directly follows the generating process in Section 2.4, only with the latent time integrated out as following Algorithm 8 (in order to save computing time). This "backward" version of the generative process is detailed below in Algorithm 9.

Algorithm 11 Generate data with backward sampling

Input:
1) token topic assignments $\{\{z_n^{(d)}\}_{n=1}^{N^{(d)}}\}_{d=1}^{D}$,
2) topic interaction pattern assignments, $\{C_k\}_{k=1}^{K}$,
3) interaction pattern parameters $\{\boldsymbol{b}^{(c)}\}_{c=1}^{C}$,
4) receiver size parameter δ.

Set $NKV = 0$ and $NK = 0$
for *d=1 to D* **do**
  for *n=1 to* $\bar{N}^{(d)}$ **do**
    for *v=1 to V* **do**
      token-word-type-distribution$_n^{(d)}[v] = \frac{NKV_{v,z_n^{(d)}}+\beta \boldsymbol{u}_v}{NK_{z_n^{(d)}}+\beta}$
    end
    draw $w_n^{(d)} \sim$ (token-word-type-distribution$_n^{(d)}$)
    $NKV_{w_n^{(d)},z_n^{(d)}} += 1$
    $NK_{z_n^{(d)}} += 1$
  end
  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 $\lambda_{ij}^{(d)} = \sum\limits_{c=1}^{C} p_c^{(d)} \cdot \exp\Big\{\lambda_0^{(c)} + \boldsymbol{b}^{(c)T}\boldsymbol{x}^{(c)}_{t_+^{(d-1)}}(i,j)\Big\} \cdot 1\{j \in \mathcal{A}_{\setminus i}\}$
      end
    end
    draw $J_i^{(d)} \sim$ Gibbs measure$(\{\lambda_{ij}^{(d)}\}_{j=1}^{A}, \delta)$
  end
  draw $i^{(d)} \sim$ Multinomial$(\{\frac{\lambda_{iJ_i}^{(d)}}{\sum_{i\in\mathcal{A}}\lambda_{iJ_i}^{(d)}}\}_{i=1}^{A})$
  set $J^{(d)} = J_{i^{(d)}}$
  draw $\Delta T_{i^{(d)}J^{(d)}} \sim$ Exponential$(\sum\limits_{i\in\mathcal{A}} \lambda_{iJ_i}^{(d)})$
  set $t^{(d)} = t^{(d-1)} + \Delta T_{i^{(d)}J^{(d)}}$
end

C.3 Intitialization of History $\boldsymbol{x}_t^{(c)}$

Considering that our network statistics $\boldsymbol{x}_t^{(c)}$ are generated as a function of the network history, 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, we cannot guarantee the same number of documents used for the inference, since only the documents with its timestamp greater than 384 hours 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 some 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 fixed number of documents given the initial $\boldsymbol{x}_{t^{(0)}=384}^{(c)}$ so that we can acheive consistency in the generated number of documents with $t^{(d)} > 384$.

C.4 GiR Implementation Details

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

- D (number of documents) = 5
- $N^{(d)}$ (tokens per document) = 4
- A (number of actors)= 4
- W (unique word types) = 5
- C (number of interaction patterns) = 2
- K (number of topics) = 4
- α (Dirichlet concentration prior) = 2
- $\boldsymbol{m}$ (Dirichlet base prior) = $\boldsymbol{u}$
- β (Dirichlet concentration prior)= 2
- $\boldsymbol{n}$ (Dirichlet base prior) = $\boldsymbol{u}$
- netstat = "intercept" and "dyadic"
- prior for $\boldsymbol{b}^{(c)}$: $\mu_{\boldsymbol{b}^{(c)}} = (-3, \mathbf{0}_6), \Sigma_{\boldsymbol{b}^{(c)}} = 0.05 \times I_7$
- prior for δ: $\mu_\delta = 2.5$, $\sigma_\delta^2 = 0.0001$
- I (outer iteration) = 5
- n_1 (hyperparamter optimization) = 0
- n_2 (M-H sampling iteration of $\mathcal{B}$) = 5500
- burn (M-H sampling burn-in of $\mathcal{B}$)= 500
- thin (M-H sampling thinning of $\mathcal{B}$)= 5
- σ_{Q1}^2 (proposal variance for $\mathcal{B}$) = 0.1
- n_3 (M-H sampling iteration of δ) = 500
- σ_{Q2}^2 (proposal variance for δ) = 0.0002