Interaction-Partitioned Topic Models (IPTM) using a Point Process Approach

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

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

January 12, 2017

1 Ideas

Current CPME model does not involve any of temporal component, which plays a key role in email interactions. Intuitively, past interaction behaviors significantly influence future ones; for example, if an actor i sent an email to actor j, then j is highly likely to send an email back to i as a response (i.e. reciprocity). Moreover, the recency and frequency of past interactions can also be considered to effectively predict future interactions. Thus, as an exploratory data analysis, point process model for directional interaction is applied to the North Carolina email data. Starting from the existing framework focused on the analysis of content-partitioned subnetworks, I would suggest an extended approach to analyze the data using the timestamps in the email, aiming to develop a joint dynamic or longitudinal model of text-valued ties.

CPME model is a Bayesian framework using two well-known methods: Latent Dirichlet Allocation (LDA) and Latent Space Model (LSM). Basically, existence of edge depends on topic assignment k (LDA) and its corresponding interaction pattern c. Each topic $k = 1, \dots, K$ has one interaction pattern c=1,...,C, and each interaction pattern posits unique latent space (LSM), thus generating $A \times A$ matrix of probabilities $P^{(c)}$ that a message author a will include recipient r on the message, given that it is about a topic in cluster c. Incorporating point process approach, now assume that under each interaction pattern, we have $A \times A$ matrix of stochastic intensities at time t, $\boldsymbol{\lambda}^{(c)}(t)$, which depends on $\boldsymbol{x}_t^{(c)}(i, j)$, the history of interaction between the sender and receiver corresponding to the interaction pattern c. We will refer this as interaction-partitioned topic models (IPTM).

2 IPTM Model

In this section, we first introduce multiplicative Cox regression model for the edge formation process in a longitudinal communication network. Then, we illustrate the generative process of the model, as well as the specific dynamic network statistics used in our model. For concreteness, we frame our discussion of this model in terms of email data, although it is generally applicable to any similarly-structured communication data.

2.1 Point Process Framework

A single email, indexed by d, is represented by a set of tokens $w^{(d)} = \{w_m^{(d)}\}_{m=1}^{M^{(d)}}$ that comprise the text of that email, an integer $i^{(d)} \in \{1, ..., A\}$ indicating the identity of that email's sender, an integer $j^{(d)} \in \{1, ..., A\}$ indicating the identity of that email's receiver, and an integer $t^{(d)} \in [0, T]$ indicating the (unix time-based) timestamp of that email. To capture the relationship between the interaction patterns expressed in an email and that email's recipients, documents that share the interaction pattern c are associated with an $A \times A$ matrix of $\boldsymbol{\lambda}^{(c)}(t) = \{\{\lambda_{ij}^{(c)}(t)\}_{i=1}^{A}\}_{j=1}^{A}$, the stochastic intensity where $\lambda_{ij}^{(c)}(t)dt$=P{for interaction pattern c, $i \rightarrow j$ occurs in time interval $[t, t + dt)$}. We will model the counting process $\mathbf{N}^{(d|c)}(t)$ through $\boldsymbol{\lambda}^{(c)}(t)$ using a version of the Cox proportional intensity model, where $N_{ij}^{(d|c)}(t)$ denotes the number of edges (emails) for document d from actor i to actor j up to time t (from the starting point 0) given that the document corresponds to interaction pattern c. Since this counting proess $\mathbf{N}$ is document-based, each element is either 0 or 1, and only one element of the matrix is 1 while all the rests are 0 (assuming no multicast).

Combining the individual counting processes of all potential edges, $\mathbf{N}^{(d|c)}(t)$ is the multivariate counting process with $\mathbf{N}^{(d|c)}(t) = (N_{ij}^{(d|c)}(t) : i, j \in 1, ..., A, i \neq j)$. Here we make no assumption about the independence of individual edge counting process. As in Vu et al. (2011), we model the multivariate counting process via Doob-Meyer decomposition:

$$\mathbf{N}^{(d|c)}(t) = \int_0^t \boldsymbol{\lambda}^{(c)}(s)ds + \mathbf{M}(t) \tag{1}$$

where essentially $\boldsymbol{\lambda}^{(c)}(t)$ and $\mathbf{M}(t)$ may be viewed as the (deterministic) signal and (martingale) noise, respectively.

Following the multiplicative Cox model of the intensity process $\boldsymbol{\lambda}^{(c)}(t)$ given $\boldsymbol{H}_{t-}^{(c)}$, the entire past of the network corresponding to the interaction pattern c up to but not including time t, we consider for each potential directed edge (i, j) the intensity forms:

$$\lambda_{ij}^{(c)}(t|\boldsymbol{H}_{t-}^{(c)}) = \lambda_0 \cdot \exp\Big\{\boldsymbol{\beta}^{(c)T}\boldsymbol{x}_t^{(c)}(i, j)\Big\} \cdot 1\{j \in \mathcal{A}^{(c)}\} \tag{2}$$

where λ_0 is the common baseline hazards for the overall interaction, $\boldsymbol{\beta}^{(c)}$ is an unknown vector of coefficients in $\boldsymbol{R}^p$, $\boldsymbol{x}_t^{(c)}(i, j)$ is a vector of p statistics for directed edge (i, j) constructed based on $\boldsymbol{H}_{t-}^{(c)}$, and $\mathcal{A}^{(c)}$ is the predictable receiver set of sender i corresponding to the interaction pattern c within the set of all possible actors $\mathcal{A}$. Equivalently, by fixing $\lambda_0 = 1$, we can rewrite (2):

$$\lambda_{ij}^{(c)}(t|\boldsymbol{H}_{t-}^{(c)}) = \exp\Big\{\boldsymbol{\beta}^{(c)T}\boldsymbol{x}_t^{*(c)}(i, j)\Big\} \cdot 1\{j \in \mathcal{A}^{(c)}\} \tag{3}$$

where the first element of $\boldsymbol{\beta}^{(c)}$ corresponds to the deviation from λ_0, by setting $\boldsymbol{x}_t^{*(c)}(i, j) = (\mathbf{1}, \boldsymbol{x}_t^{(c)}(i, j))$.

Based on the framework illustrated so far, the likelihood we will use for inference procedure is that of Perry and Wolfe (2013). For each type of interaction pattern $c = 1, ..., C$, estimation for $\boldsymbol{\beta}^{(c)}$ proceeds by maximizing the so-called partial likelihood of Cox (1992):

$$PL_t(\boldsymbol{\beta}^{(c)}) = \prod_{d:c^{(d)}=c} \frac{\exp\{\boldsymbol{\beta}^{(c)T}x_{t^{(d)}}^{(c)}(i^{(d)}, j^{(d)})\}}{\sum_{j\in\mathcal{A}^{(c)}} \exp\{\boldsymbol{\beta}^{(c)T}x_{t^{(d)}}^{(c)}(i^{(d)}, j)\}}, \tag{4}$$

where $t^{(d)}$, $i^{(d)}$, and $j^{(d)}$ are the time, sender, and receiver of the dth document. For computational efficiency, we will use the log-partial likelihood:

$$\log PL_t(\boldsymbol{\beta}^{(c)}) = \sum_{d:c^{(d)}=c} \Big\{\boldsymbol{\beta}^{(c)T}x_{t^{(d)}}^{(c)}(i^{(d)}, j^{(d)}) - \log\big[\sum_{j\in\mathcal{A}^{(c)}} \exp\{\boldsymbol{\beta}^{(c)T}x_{t^{(d)}}^{(c)}(i^{(d)}, j)\}\big]\Big\}. \tag{5}$$

2.2 Generative Process

The generative process of this model follows the topic model (LDA) of Blei et al. (2003) and the author-topic model of Rosen-Zvi et al. (2004). Same as LDA, documents are represented as random mixtures over latent topics, where each topic is characterized by a distribution over words. However, one crucial difference is that each document is connected to one type of interaction pattern, and the topic distributions vary depending on the assigned interaction pattern.

Conditioned on the interaction pattern and their distributions over topics, the process by which a document is generated can be summarized as follows: first, an interaction pattern is chosen by multinomial for each document; next, a topic is sampled for each word from the distribution over topics associated with the interaction pattern of the document; finally, words themselves are sampled from the distribution over words associated with each topic. At the same time, the unique sender-recipient pair of the document is determined by the rate of intensities associated with the interaction pattern and history of interactions until the time the document is written. Below are the detailed generative process for each document in a corpus D and its plate notation (Figure 1), and Table 1 summarizes the notations used in this paper:

1. $\boldsymbol{\phi}^{(k)} \sim \text{Dir}(\delta, \mathbf{n})$ **[See Algorithm 1]**
   - A "topic" k is characterized by a discrete distribution over V word types with probability vector $\phi^{(k)}$. A symmetric Dirichlet prior with concentration parameter δ is placed.
2. For each of the C interaction patterns [**See Algorithm 2**]:
   (a) $\boldsymbol{\beta}^{(c)} \sim \text{Normal}(\mathbf{0}, \sigma^2 I_P)$
   - The vector of coefficients depends on the interaction pattern c. This means that there is variation in the degree of influence from the network statistics.

   (b) Update $\boldsymbol{x}_t^{*(c)}(i, j)$
   - Corpus are partitioned according to the assignment of interaction patterns, and the dynamic network statistics are calculated based on the documents of the same interaction pattern.

   (c) Using $\boldsymbol{\beta}^{(c)}$ in (a), update $\boldsymbol{\lambda}^{(c)}(t)$
   - Use the equation $\lambda_{ij}^{(c)}(t) = \exp\Big\{\boldsymbol{\beta}^{(c)T}\boldsymbol{x}_t^{*(c)}(i, j)\Big\} \cdot 1\{j \in \mathcal{A}^{(c)}\}$ for all $i \in \mathcal{A}, j \in \mathcal{A}, i \neq j$.

   (d) Set $\alpha^{(c)}$ and $\boldsymbol{m}^{(c)}$ using the hyperparmeter optimization step
   - The topic proportions for documents in the same cluster share the same parameters in the Dirichlet distribution, and how to obtain these parameters will be explained in Section X.X.
3. For each of the D documents [**See Algorithm 3**]:
   (a) $c^{(d)} \sim \text{Multinomial}(\boldsymbol{\gamma})$
   - Each document d is associated with one "interaction pattern" among C different types, with parameter $\boldsymbol{\gamma}$. Here, we assign the prior for the multinomial parameter $\boldsymbol{\gamma} \sim \text{Dir}(\eta, \boldsymbol{l})$

   (b) $\mathbf{N}^{(d|c^{(d)})}(t^{(d)}) \sim \text{CP}(\boldsymbol{\lambda}^{(c^{(d)})}(t^{(d)}))$
   - The actual update of the counting process $\mathbf{N}^{(d|c^{(d)})}(t)$ of the email d is $N_{i^{(d)}j^{(d)}}^{(d|c^{(d)})}(t^{(d)}) = 1$ and the rest $N_{(i,j)\neq(i^{(d)},j^{(d)})}^{(d|c^{(d)})}(t^{(d)}) = 0$.

   (c) $\boldsymbol{\theta}^{(d)} \sim \text{Dir}(\alpha^{(c^{(d)})}, \boldsymbol{m}^{(c^{(d)})})$
   - Each email has a discrete distribution over topics $\boldsymbol{\theta}^{(d)}$, since the topic proportions for documents in the same cluster share the same parameters in the Dirichlet distribution.
4. For each of the M words [**See Algorithm 4**]:
   (a) $z_m^{(d)} \sim \text{Multinomial}(\boldsymbol{\theta}^{(d)})$

   (b) $w_m^{(d)} \sim \text{Multinomial}(\phi^{(z_m^{(d)})})$

Algorithm 1 Topic Word Distributions

```
for k=1 to K do
    draw φ^(k) ~ Dir(δ, n)
end
```

$$\textbf{for } k=1 \textit{ to } K \textbf{ do} \quad \text{draw } \boldsymbol{\phi}^{(k)} \sim \text{Dir}(\delta, \mathbf{n}) \quad \textbf{end}$$

Algorithm 2 Interaction Patterns

for c=1 *to* C **do**
- draw $\boldsymbol{\beta}^{(c)} \sim \text{Normal}(\mathbf{0}, \sigma^2 I_P)$
- set $\boldsymbol{x}_t^{*(c)}(i,j)$ according to Section 2.3
- **for** i=1 *to* A **do**
  - **for** j=1 *to* A **do**
    - **if** $i \neq j$ **then**
      - set $\lambda_{ij}^{(c)}(t) = \exp\Big\{\boldsymbol{\beta}^{(c)T}\boldsymbol{x}_t^{*(c)}(i,j)\Big\} \cdot 1\{j \in \mathcal{A}^{(c)}\}$
    - **end**
    - **else**
      - set $\lambda_{ij}^{(c)}(t) = 0$
    - **end**
  - **end**
- **end**
- set $\alpha^{(c)}$ and $\boldsymbol{m}^{(c)}$ using the hyperparameter optimization

end

Algorithm 3 Document-Interaction Pattern Assignments

for d=1 *to* D **do**
- draw $c^{(d)} \sim \text{Multinomial}(\boldsymbol{\gamma})$
- draw $\mathbf{N}^{(d|c^{(d)})}(t^{(d)}) \sim \text{CP}(\boldsymbol{\lambda}^{(c^{(d)})}(t^{(d)}))$
- draw $\boldsymbol{\theta}^{(d)} \sim \text{Dir}(\alpha^{(c^{(d)})}, \boldsymbol{m}^{(c^{(d)})})$

end

Algorithm 4 Tokens

for d=1 *to* D **do**
- set $M^{(d)}$ = the number of words in document d
- **for** m=1 *to* $M^{(d)}$ **do**
  - draw $z_m^{(d)} \sim \text{Multinomial}(\boldsymbol{\theta}^{(d)})$
  - draw $w_m^{(d)} \sim \text{Multinomial}(\boldsymbol{\phi}^{(z_m^{(d)})})$
- **end**

end

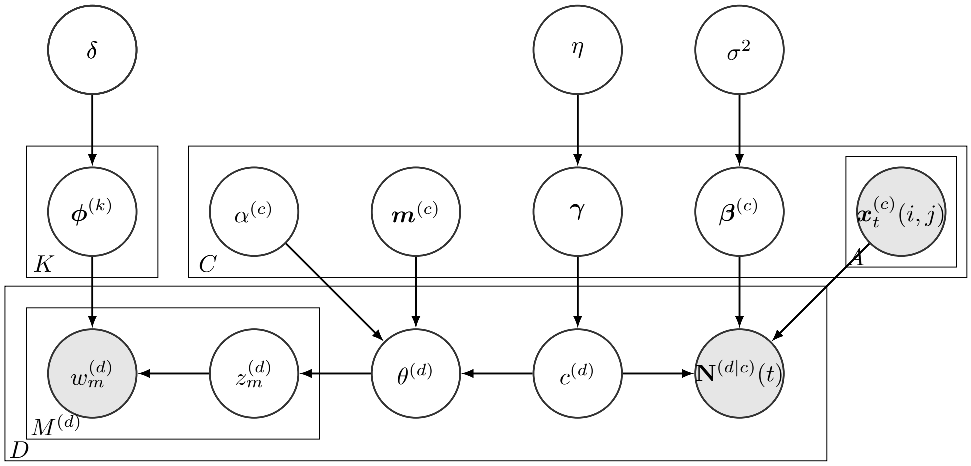

Figure 1: Plate notation of IPTM

Authors of the corpus	$\mathcal{A}$	Set
Authors of the corpus given interaction pattern c	$\mathcal{A}^{(c)}$	Set
Number of authors	A	Scalar
Number of documents	D	Scalar
Number of words in the d^{th} document	$M^{(d)}$	Scalar
Number of topics	K	Scalar
Vocabulary size	W	Scalar
Number of interaction patterns	C	Scalar
Number of words assigned to interaction pattern and topic	M^{CK}	Scalar
Number of words assigned to word and topic	M^{WK}	Scalar
Interaction pattern of the d^{th} document	$c^{(d)}$	Scalar
Time of the d^{th} document	$t^{(d)}$	Scalar
Words in the d^{th} document	$\boldsymbol{w}^{(d)}$	$M^{(d)}$-dimensional vector
m^{th} word in the d^{th} document	$w_m^{(d)}$	m^{th} component of $\boldsymbol{w}^{(d)}$
Topic assignments in the d^{th} document	$\boldsymbol{z}^{(d)}$	$M^{(d)}$-dimensional vector
Topic assignments for m^{th} word in the d^{th} document	$z_m^{(d)}$	m^{th} component of $\boldsymbol{z}^{(d)}$
Dirichlet concentration prior given interaction pattern c	$\alpha^{(c)}$	Scalar
Dirichlet base prior given interaction pattern c	$\boldsymbol{m}^{(c)}$	K-dimensional vector
Dirichlet concentration prior	δ	Scalar
Dirichlet base prior	$\boldsymbol{n}$	W-dimensional vector
Dirichlet concentration prior	η	Scalar
Dirichlet base prior	$\boldsymbol{l}$	C-dimensional vector
Multinomial prior	γ	C-dimensional vector
Variance of Normal prior	σ^2	Scalar
Probabilities of the words given topics	Φ	$W \times K$ matrix
Probabilities of the words given topic k	$\boldsymbol{\phi}^{(k)}$	W-dimensional vector
Probabilities of the topics	Θ	$K \times D$ matrix
Probabilities of the topics given the d^{th} document	$\boldsymbol{\theta}^{(d)}$	K-dimensional vector
Coefficient of the intensity process given interaction pattern c	$\boldsymbol{\beta}^{(c)}$	p-dimensional vector
Network statistics for directed edge (i,j) given interaction pattern c	$\boldsymbol{x}_t^{(c)}(i,j)$	p-dimensional vector
Counting process in the d^{th} document given interaction pattern	$\mathbf{N}^{(d\|c)}(t)$	$A \times A$ matrix

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

2.3 Dynamic covariates to measure network effects

The network statistics $\boldsymbol{x}_t^{(c)}(i,j)$ of Equation (2), corresponding to the ordered pair (i,j), can be time-invariant (such as gender) or time-dependent (such as the number of two-paths from i to j just before time t). Since time-invariant covariates can be easily specified in various manners (e.g. homophily or group-level effects), here we only consider specification of dynamic covariates.

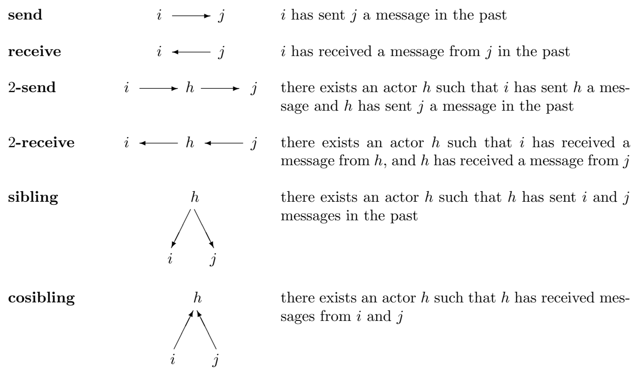

Fig. 3. Dynamic covariates to measure network effects

Following Perry and Wolfe (2013) (refer to Fig.3 of Perry and Wolfe (2013) attached above), we use 4 effects as components of $\boldsymbol{x}_t^{(c)}(i,j)$, including the intercept to estimate the baseline intensities. The two behaviors (send and receive) are dyadic, involving exactly two actors, while the one is triadic (sum of 2-send, 2-receive, sibling, and cosibling), involving exactly three actors. In addition to the ones from Perry and Wolfe (2013), we also include the indegree for receiver and outdegree for sender effects to measure the popularity and centrality. However, one different point from the existing specification is that we define the effects not to be based on finite sub-interval, which require large number of dimension. Instead, we create a single statistic for each effect by incorporating the recency of event into the statistic itself. As a result, all of the statistics can be seen as time-weighted dynamic network statistics.

1. $\text{intercept}_t(i,j) = 1$
2. $\text{send}_t(i,j) = \sum_{d:t^{(d)}<t} I\{i \to j\} \cdot g(t - t^{(d)})$
3. $\text{receive}_t(i,j) = \sum_{d:t^{(d)}<t} I\{j \to i\} \cdot g(t - t^{(d)})$
4. $\text{triangle}_t(i,j) = \sum_{h \neq i,j} \Big(\sum_{d:t^{(d)}<t} I\{i \to h \text{ or } h \to j\} \cdot g(t - t^{(d)}) \Big) \Big(\sum_{d:t^{(d)}<t} I\{j \to h \text{ or } h \to j\} \cdot g(t - t^{(d)}) \Big)$
5. $\text{outdegree}_t(i) = \sum_{j \neq i} \sum_{d:t^{(d)}<t} I\{i \to j\} \cdot g(t - t^{(d)})$
6. $\text{indegree}_t(j) = \sum_{i \neq j} \sum_{d:t^{(d)}<t} I\{j \to i\} \cdot g(t - t^{(d)})$

Here, $g(t - t^{(d)})$ reflects the difference between current time t and the timestamp of previous email $t^{(d)}$, thus measuring the recency. Inspired by the self-exciting Hawkes process, which is often used to model the temporal effect of email data, we can take the exponential kernel $g(t - t^{(d)}) = e^{-\lambda(t - t^{(d)})}$ where λ is the parameter of speed at which sender replies to emails, with larger values indicating faster response times. Indeed, λ^{-1} is the expected number of hours it takes to reply to a typical email. For simplicity, in our simulation we fixed $\lambda = 0.05$ (i.e. $g(t - t^{(d)}) = e^{-0.05(t - t^{(d)})}$), but this setup may vary based on the nature of document.

3 Inference

The inference for IPTM is similar to that of CPME. In this case, what we actually observe are the tokens $\mathcal{W} = \{\boldsymbol{w}^{(d)}\}_{d=1}^{D}$ and the sender, recipient, and timestamps of the email in the form of the counting process $\mathcal{N} = \{\boldsymbol{N}^{(d)}(t^{(d)})\}_{d=1}^{D}$. Next, $\mathcal{X} = \{\boldsymbol{x}_{t^{(d)}}^{(c)}(i,j)\}_{d=1}^{D}$ is the metadata, and the latent variables are $\Phi = \{\boldsymbol{\phi}^{(k)}\}_{k=1}^{K}, \Theta = \{\boldsymbol{\theta}^{(d)}\}_{d=1}^{D}, \mathcal{Z} = \{\boldsymbol{z}^{(d)}\}_{d=1}^{D}, \mathcal{C} = \{c^{(d)}\}_{d=1}^{D}$, and $\mathcal{B} = \{\boldsymbol{\beta}^{(c)}\}_{c=1}^{C}$.

Below is the the big joint distribution

$$\begin{aligned}
&P(\Phi, \Theta, \mathcal{W}, \mathcal{Z}, \mathcal{C}, \mathcal{B}, \mathcal{N} | \mathcal{X}, \delta, \boldsymbol{n}, \alpha, \boldsymbol{m}, \boldsymbol{\gamma}, \boldsymbol{\eta}, \sigma^2) \\
&= P(\mathcal{W}, \mathcal{Z}, \mathcal{C}, \mathcal{B}, \mathcal{N} | \Phi, \Theta, \mathcal{X}, \boldsymbol{\gamma}, \boldsymbol{\eta}, \sigma^2) P(\Phi, \Theta | \delta, \boldsymbol{n}, \alpha, \boldsymbol{m}) \\
&= P(\mathcal{W}|\mathcal{Z}, \Phi) P(\mathcal{Z}|\Theta) P(\mathcal{N}|\mathcal{C}, \mathcal{X}, \mathcal{B}) P(\mathcal{B}|\mathcal{C}, \sigma^2) P(\Phi|\delta, \boldsymbol{n}) P(\Theta|\mathcal{C}, \alpha, \boldsymbol{m}) P(\mathcal{C}|\boldsymbol{\gamma}) P(\boldsymbol{\gamma}|\boldsymbol{\eta})
\end{aligned} \tag{6}$$

Now we can integrate out Φ and Θ in latent Dirichlet allocation by applying Dirichlet-multinomial conjugacy as we did in CPME. See APPENDIX A for the detailed steps. After integration, we obtain below:

$$\propto P(\mathcal{W}|\mathcal{Z}) P(\mathcal{Z}|\mathcal{C}, \delta, \boldsymbol{n}, \alpha, \boldsymbol{m}) P(\mathcal{N}|\mathcal{C}, \mathcal{B}, \mathcal{X}) P(\mathcal{B}|\mathcal{C}, \sigma^2) P(\mathcal{C}|\boldsymbol{\gamma}) \tag{7}$$

Then, we only have to perform inference over the remaining unobserved latent variables $\mathcal{Z}, \mathcal{C}$, and $\mathcal{B}$, using the equation below:

$$P(\mathcal{Z}, \mathcal{C}, \mathcal{B} | \mathcal{W}, \mathcal{N}, \mathcal{X}, \delta, \boldsymbol{n}, \alpha, \boldsymbol{m}, \boldsymbol{\gamma}, \boldsymbol{\eta}, \sigma^2) \propto P(\mathcal{W}, \mathcal{Z}, \mathcal{C}, \mathcal{B}, \mathcal{N} | \mathcal{X}, \delta, \boldsymbol{n}, \alpha, \boldsymbol{m}, \boldsymbol{\gamma}, \boldsymbol{\eta}, \sigma^2) \tag{8}$$

Either Gibbs sampling or Metropolis-Hastings algorithm is applied by sequentially resampling each latent variables from their respective conditional posterior.

3.1 Resampling $\mathcal{C}$

The first variable we are going to resample is the document-interaction pattern assignments, one document at a time. To obtain the Gibbs sampling equation, which is the posterior conditional probability for the interaction pattern $\mathcal{C}$ for d^{th} document, i.e. $P(c^{(d)} = c | \mathcal{W}, \mathcal{Z}, \mathcal{C}_{\backslash d}, \mathcal{B}, \mathcal{N}, \mathcal{X}, \delta, \boldsymbol{n}, \alpha, \boldsymbol{m}, \boldsymbol{\gamma}, \boldsymbol{\eta}, \sigma^2)$. We can derive the equation as below:

$$\begin{aligned}
&P(c^{(d)} = c | \mathcal{W}, \mathcal{Z}, \mathcal{C}_{\backslash d}, \mathcal{B}, \mathcal{N}, \mathcal{X}, \delta, \boldsymbol{n}, \alpha, \boldsymbol{m}, \boldsymbol{\gamma}, \boldsymbol{\eta}, \sigma^2) \\
&\propto P(c^{(d)} = c, \boldsymbol{w}^{(d)}, \boldsymbol{z}^{(d)}, \mathbf{N}^{(d)}(t^{(d)}) | \mathcal{W}_{\backslash d}, \mathcal{Z}_{\backslash d}, \mathcal{C}_{\backslash d}, \mathcal{B}, \mathcal{N}_{\backslash d}, \mathcal{X}, \delta, \boldsymbol{n}, \alpha, \boldsymbol{m}, \boldsymbol{\gamma}, \boldsymbol{\eta}, \sigma^2) \\
&\propto P(c^{(d)} = c | \mathcal{C}_{\backslash d}, \boldsymbol{\gamma}) P(\mathbf{N}^{(d)}(t^{(d)}) | c^{(d)} = c, \mathcal{C}_{\backslash d}, \mathcal{B}, \mathcal{N}_{\backslash d}, \mathcal{X}) P(\boldsymbol{w}^{(d)}, \boldsymbol{z}^{(d)} | c^{(d)} = c, \mathcal{W}_{\backslash d}, \mathcal{Z}_{\backslash d}, \mathcal{C}_{\backslash d}, \delta, \boldsymbol{n}, \alpha, \boldsymbol{m}),
\end{aligned} \tag{9}$$

where $P(c^{(d)} = c | \mathcal{C}_{\backslash d}, \boldsymbol{\gamma})$ comes from the multinomial prior γ and $P(\mathbf{N}^{(d)}(t^{(d)}) | c^{(d)} = c, \mathcal{C}_{\backslash d}, \mathcal{B}, \mathcal{N}_{\backslash d}, \mathcal{X})$ is the probability of observing a document with the sender, receiver, and time equal to $(i = i^{(d)}, j = j^{(d)}, t = t^{(d)})$, respectively, given a set of parameter values. We will replace this by the partial likelihood in Equation (4) (without the product term since resampling of c is document-specific). For the last term $P(\boldsymbol{w}^{(d)}, \boldsymbol{z}^{(d)} | c^{(d)} = c, \mathcal{W}_{\backslash d}, \mathcal{Z}_{\backslash d}, \mathcal{C}_{\backslash d}, \delta, \boldsymbol{n}, \alpha, \boldsymbol{m})$, we will follow typical LDA approach.

Using Bayes' theorem (See APPENDIX B for conditional probabilty of the last term), we have

$$= \Big[\gamma_c\Big] \times \Big[\frac{\exp\{\boldsymbol{\beta}^{(c)T} x_{t^{(d)}}^{(c)}(i^{(d)}, j^{(d)})\}}{\sum_{j \in \mathcal{A}^{(c)}} \exp\{\boldsymbol{\beta}^{(c)T} x_{t^{(d)}}^{(c)}(i^{(d)}, j)\}}\Big] \times \Big[\prod_{m=1}^{M^{(d)}} \frac{M_{z_m^{(d)}|d,\backslash d,m} + \alpha^{(c)} \boldsymbol{m}_{z_m^{(d)}}^{(c)}}{M_{\cdot|d} - 1 + \alpha^{(c)}}\Big], \tag{10}$$

where $M_{k|d}$ is the number of times topic k shows up in the document d. Furthermore, we can take the log of Equation (10) to avoid numerical issue from exponentiation and increase the speed of computation, which becomes:

$$\log(\gamma_c) + \Big(\boldsymbol{\beta}^{(c)T} x_{t^{(d)}}^{(c)}(i^{(d)}, j^{(d)}) - \log\Big[\sum_{j \in \mathcal{A}^{(c)}} \exp\{\boldsymbol{\beta}^{(c)T} x_{t^{(d)}}^{(c)}(i^{(d)}, j)\}\Big]\Big) + \sum_{m=1}^{M^{(d)}} \log\Big(\frac{M_{z_m^{(d)}|d,\backslash d,m} + \alpha^{(c)} \boldsymbol{m}_{z_m^{(d)}}^{(c)}}{M_{\cdot|d} - 1 + \alpha^{(c)}}\Big). \tag{11}$$

3.2 Resampling $\mathcal{Z}$

Next, the new values of $z_m^{(d)}$ are sampled for all of the token topic assignments (one token at a time), using the conditional posterior probability of being topic k as we derived in APPENDIX B:

$$\begin{aligned} &P(z_m^{(d)} = k|\mathcal{W}, \mathcal{Z}_{\backslash d,m}, \mathcal{C}, \mathcal{B}, \mathcal{N}, \mathcal{X}, \delta, \boldsymbol{n}, \alpha, \boldsymbol{m}, \boldsymbol{\gamma}, \boldsymbol{\eta}, \sigma^2) \\ &\propto P(z_m^{(d)} = k, w_m^{(d)}|\mathcal{W}_{\backslash d,m}, \mathcal{Z}_{\backslash d,m}, C, \delta, \boldsymbol{n}, \alpha, \boldsymbol{m}) \end{aligned} \tag{12}$$

where the subscript "$\backslash d, m$" denotes the exclsuion of position m in email d. In the last line of equation (10), it is the contribution of LDA, so similar to CPME we can write the conditional probability:

$$\propto (M_{k|d,\backslash d,m} + \alpha^{(c^{(d)})}\boldsymbol{m}_k^{(c^{(d)})}) \times \frac{M^{WK}_{w_m^{(d)}k,\backslash d,m} + \delta n_w}{\sum_{w=1}^{W} M^{WK}_{wk,\backslash d,m} + \delta} \tag{13}$$

which is the well-known form of collapsed Gibbs sampling equation for LDA.

3.3 Resampling $\mathcal{B}$

Finally, we wan to update the interaction pattern parameter $\boldsymbol{\beta}^{(c)}$, one interaction pattern at a time. For this, we will use the Metropolis-Hastings algorithm with a proposal density Q being the multivariate Gaussian distribution, with variance δ_B^2 (proposal distirbution variance parameters set by the user), centered on the current values of $\boldsymbol{\beta}^{(c)}$. Then we draw a proposal $\boldsymbol{\beta}'^{(c)}$ at each iteration. Under symmetric proposal distribution (such as multivariate Gaussian), we cancel out Q-ratio and obtain the acceptance probability equal to:

$$\text{Acceptance Probability} = \begin{cases} \frac{P(\mathcal{B}'|\mathcal{W},\mathcal{Z},\mathcal{C},\mathcal{N},\mathcal{X})}{P(\mathcal{B}|\mathcal{W},\mathcal{Z},\mathcal{C},\mathcal{N},\mathcal{X})} & \text{if} < 1 \\ 1 \quad \text{else} \end{cases} \tag{14}$$

After factorization, we get

$$\begin{aligned} \frac{P(\mathcal{B}'|\mathcal{W},\mathcal{Z},\mathcal{C},\mathcal{N},\mathcal{X})}{P(\mathcal{B}|\mathcal{W},\mathcal{Z},\mathcal{C},\mathcal{N},\mathcal{X})} &= \frac{P(\mathcal{N}|\mathcal{B}',\mathcal{W},\mathcal{Z},\mathcal{C},\mathcal{X})P(\mathcal{B}')}{P(\mathcal{N}|\mathcal{B},\mathcal{W},\mathcal{Z},\mathcal{C},\mathcal{X})P(\mathcal{B})} \\ &= \frac{P(\mathcal{N}|\mathcal{C},\mathcal{X},\mathcal{B}')P(\mathcal{B}')}{P(\mathcal{N}|\mathcal{C},\mathcal{X},\mathcal{B})P(\mathcal{B})}, \end{aligned} \tag{15}$$

where $P(\mathcal{N}|\mathcal{C},\mathcal{X},\mathcal{B})$ is the partial likelihood in Equation (4).

For $P(\mathcal{B})$, we select a multivarate Gaussian priors as mentioned earlier. Similar to what we did in Section 3.1, we can take the log and obtain the log of acceptance ratio as following:

$$\begin{aligned} &\log\Big(\phi_d(\boldsymbol{\beta}'^{(c)}; \mathbf{0}, \sigma^2 I_P)\Big) - \log\Big(\phi_d(\boldsymbol{\beta}'^{(c)}; \mathbf{0}, \sigma^2 I_P)\Big) \\ &+ \sum_{d:c^{(d)}=c} \Big\{\boldsymbol{\beta}'^{(c)T} x^{(c)}_{t^{(d)}}(i^{(d)}, j^{(d)}) - \log\big[\sum_{j\in\mathcal{A}^{(c)}} \exp\{\boldsymbol{\beta}'^{(c)T} x^{(c)}_{t^{(d)}}(i^{(d)}, j)\}\big]\Big\} \\ &- \sum_{d:c^{(d)}=c} \Big\{\boldsymbol{\beta}^{(c)T} x^{(c)}_{t^{(d)}}(i^{(d)}, j^{(d)}) - \log\big[\sum_{j\in\mathcal{A}^{(c)}} \exp\{\boldsymbol{\beta}^{(c)T} x^{(c)}_{t^{(d)}}(i^{(d)}, j)\}\big]\Big\}, \end{aligned} \tag{16}$$

where $\phi_d(\cdot; \mu, \Sigma)$ is the d-dimensional multivariate normal density. Then the log of acceptance ratio we have is:

$$\log(\text{Acceptance Probability}) = \min((16),\ 0) \tag{17}$$

To determine whether we accept the proposed update or not, we take the usual approach, by comparing the log of acceptance ratio we have to the log of a sample from uniform(0,1).

3.4 Pseudocode

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

Algorithm 5 MCMC($I, n_1, n_2, n_3, \delta_B$)

set initial values $\mathcal{C}^{(0)}, \mathcal{Z}^{(0)}$, and $\mathcal{B}^{(0)}$
for *i=1 to I* **do**
  for *n=1 to* n_1 **do**
    fix $\mathcal{Z} = \mathcal{Z}^{(i-1)}$ and $\mathcal{B} = \mathcal{B}^{(i-1)}$
    for *d=1 to D* **do**
      calculate $\boldsymbol{x}^{*(c)}_{t^{(d)}}(i^{(d)}, j)$ according to Section 2.3, for every $c = 1, ..., C$
      calculate $p^{\mathcal{C}}|\boldsymbol{z}^{(d)}, \boldsymbol{\beta}^{(c^{(d)})} = (p_1, ..., p_C)$, where $p_c = \exp$(Eq. (11) corresponding to c)
      draw $c^{(d)} \sim$ multinomial($p^{\mathcal{C}}$)
    end
  end
  for *n=1 to* n_2 **do**
    fix $\mathcal{C} = \mathcal{C}^{(i)}$ and $\mathcal{B} = \mathcal{B}^{(i-1)}$
    for *d=1 to D* **do**
      for *m=1 to* $M^{(d)}$ **do**
        calculate $p^{\mathcal{Z}}|\boldsymbol{c}^{(d)}, \alpha^{(c^{(d)})}, \boldsymbol{m}^{(c^{(d)})}, \boldsymbol{\beta}^{(c^{(d)})} = (p_1, ..., p_K)$, where $p_k = \exp$(Eq. (13) corresponding to k)
        draw of $z^{(d)}_m \sim$ multinomial($p^{\mathcal{Z}}$)
      end
    end
  end
  for *n=1 to* n_3 **do**
    fix $\mathcal{C} = \mathcal{C}^{(i)}$, $\mathcal{Z} = \mathcal{Z}^{(i)}$, and $\mathcal{B}^{(0)}$ = last value (n_3^{th}) of $\mathcal{B}^{(i-1)}$
    calculate $\mathcal{X} = \{\boldsymbol{x}^{*(c)}_{t^{(d)}}(i, j)\}_{d=1}^{D}$ according to Section 2.3, given fixed $\mathcal{C}$
    for *c=1 to C* **do**
      draw $\boldsymbol{\beta}^{(c)}|\mathcal{C}, \mathcal{Z}, \mathcal{B}^{(n-1)}$ using M-H algorithm in Section 3.3
    end
  end
end
summarize the results using:
the last value of $\mathcal{C}$, the last value of $\mathcal{Z}$, and the last n_3 length chain of $\mathcal{B}$

3.5 Asymmetric Dirichlet prior over Θ (topic distribution)

Wallach et al. (2009) demonstrated that the typical implementations of topic models using symmetric Dirichlet priors with fixed concentration parameters often result in less practical results, and the model fitting can be improved by applying an asymmetric Dirichlet prior over the document–topic distributions (i.e. Θ). Therefore, we assign an asymmetric Dirichlet prior over the interaction pattern-topic distributions, $\Theta = \{\boldsymbol{\theta}^{(d)}\}_{d=1}^{D}$, where $\boldsymbol{\theta}^{(d)}$ is drawn from Dir$(\alpha^{(c^{(d)})}, \boldsymbol{m}^{(c^{(d)})})$. While Wallach et al. (2009) illustrates two different methods, adding a hierarchy to Θ and optimizing the hyperparameters (α and $\boldsymbol{m}$), we choose to use hyperparameter optimization steps since it is computationally efficient and also sufficient to achieve the desired performance gains. Now, we assume $\boldsymbol{m}^{(c)}$ to be non-uniform base measures (while $\alpha^{(c)}$ is still a fixed concentration parameter), and implement the hyperparameter optimization technique called "new fixed-point iterations using the Digamma recurrence relation" in Wallach (2008) based on Minka's fixed-point iteration (Minka, 2000).

Here we summarize Chapter 2 of Wallach (2008) and its extension to our IPTM, to illustrate the basic steps and equations used for our optimization. Basically, we want to find the optimal hyperparameter $[\alpha\boldsymbol{m}]^*$ given the data $\mathcal{D}$ such that the probability of the data given the hyperparameters $P(\mathcal{D}|\alpha\boldsymbol{m})$ is maximized at $[\alpha\boldsymbol{m}]^*$. After incorporating the interaction pattern component, the evidence is now given by

$$P(\mathcal{D}^{(c)}|\alpha^{(c)}\boldsymbol{m}^{(c)}) = \prod_{d:c^{(d)}=c} \frac{\Gamma(\alpha^{(c)})}{\Gamma(M_{\cdot|d}+\alpha^{(c)})} \prod_{k=1}^{K} \frac{\Gamma(M_{k|d}+\alpha^{(c)}m_k^{(c)})}{\Gamma(\alpha^{(c)}m_k^{(c)})} \tag{18}$$

and is concave in $\alpha^{(c)}\boldsymbol{m}^{(c)}$, thus we will estimate $[\alpha^{(c)}\boldsymbol{m}^{(c)}]^*$ within each outer runs of MCMC.

First, the starting point is derived by Minka's fixed-point iteration which takes the derivative of the lower bound $B([\alpha^{(c)}\boldsymbol{m}^{(c)}]^*)$ of $\log P(\mathcal{D}^{(c)}|[\alpha^{(c)}\boldsymbol{m}^{(c)}]^*)$ with respect to $[\alpha^{(c)}m_k^{(c)}]^*$:

$$[\alpha^{(c)}m_k^{(c)}]^* = \alpha^{(c)}m_k^{(c)} \frac{\sum_{d:c^{(d)}=c} \Psi(M_{k|d}+\alpha^{(c)}m_k^{(c)}) - \Psi(\alpha^{(c)}m_k^{(c)})}{\sum_{d:c^{(d)}=c} \Psi(M_{\cdot|d}+\alpha^{(c)}) - \Psi(\alpha^{(c)})}, \tag{19}$$

where $\Psi(\cdot)$ is the first derivative of the log gamma function, known as the digamma function, and the quantity $M_{k|d}$ is the number of times that outcome k was observed in the document d. Moreover, the quantity $M_{\cdot|d} = \sum_{k=1}^{K} M_{k|d}$ is the total number of words in the document d. The value $\alpha^{(c)}m_k^{(c)}$ acts as an initial "pseudocount" for outcome k across the documents of interaction pattern c.

Next, Wallach's new method rewrites the equation above using the notation $C_k(n) = \sum\limits_{d:c^{(d)}=c} \delta(M_{k|d} - n)$ and $C_\cdot(n) = \sum\limits_{d:c^{(d)}=c} \delta(M_{\cdot|d} - n)$:

$$[\alpha^{(c)}m_k^{(c)}]^* = \alpha^{(c)}m_k^{(c)} \frac{\sum_{n=1}^{\max_d M_{k|d}} C_k(n)[\Psi(n+\alpha^{(c)}m_k^{(c)}) - \Psi(\alpha^{(c)}m_k^{(c)})]}{\sum_{n=1}^{\max_d M_{\cdot|d}} C_\cdot(n)[\Psi(n+\alpha^{(c)}) - \Psi(\alpha^{(c)})]}. \tag{20}$$

Finally, applying the digamma recurrence relation (for any positive integer n)

$$\Psi(n+z) - \Psi(z) = \sum_{f=1}^{n} \frac{1}{f-1+z},$$

we subtitute Equation (20) for below:

$$[\alpha^{(c)}m_k^{(c)}]^* = \alpha^{(c)}m_k^{(c)} \frac{\sum_{n=1}^{\max_d M_{k|d}} C_k(n) \sum_{f=1}^{n} \frac{1}{f-1+\alpha^{(c)}m_k^{(c)}}}{\sum_{n=1}^{\max_d M_{\cdot|d}} C_\cdot(n) \sum_{f=1}^{n} \frac{1}{f-1+\alpha^{(c)}}}. \tag{21}$$

This method is as accurate as Mika's fixed-point iteration method, but it acheives computational efficiency since the digamma recurrence relation reduces the number of new calculations required for each successive n to one. Pseudocode for the complete fixed-point iteration is given in algorithm 2.2 of Wallach (2008).

4 Appliction to North Carolina email data

To see the applicability of the model, we used the North Carolina email data using two counties, Vance county and Dare county, which are the two counties whose email corpus cover the date of Hurricane Sandy (October 22, 2012 – November 2, 2012). Exploratory analysis revealed that Dare county experienced significant change in the pattern of email exchanges; specifically, during the emergency period, email interactions significanty less rely on previous history of interactions, compared to the normal period. On the other hand, Vance county did not experience any distinctive change, and the possible reason for the difference is the locations of two counties. Here we apply IPTM to both data to see the differences in detail, in terms of the interaction patterns and topics of the corpus.

4.1 Vance county email data

After treating multicast emails (those involving a single sender but multiple receivers) as multiple distinct emails, Vance county data contains 269 emails (only count the email with the number of words greater than 0) between 18 actors, including 620 vocabulary in total. We used $K = 10$ topics and $C = 3$ interaction patterns, with the parameters fixed as $\alpha = 50/K = 50/10 = 5$ with $\boldsymbol{m} = (1, ..., 1)^T/K$ (α and $\boldsymbol{m}$ are fixed only at the initial step), $\delta = 0.01$, and $\eta = 50/C = 50/3 = 16.67$, following the common practice. The asymmetric model uses same δ and η, but $\alpha^{(c)}$ and $\boldsymbol{m}^{(c)}$ are optimized for every outer iteration, for every interaction pattern.

MCMC sampling was implemented based on the order and scheme illustrated in Section 3. We set the outer iteration number as $I = 1000$, the inner iteration numbers as $n_1 = 3, n_2 = 3$, and $n_3 = 3300$. First 50 outer iterations and first 300 iterations of third inner iteration were used as a burn-in, and every 3^{rd} sample was taken as a thinning process of third inner iteration. In addition, after some experimentation, δ_B was set as 0.5, to ensure sufficient acceptance rate (in our case, the average acceptance rate for $\boldsymbol{\beta}$ was 0.260). As demonstrated in Algorithm 5, the last value of $\mathcal{C}$, the last value of $\mathcal{Z}$, and the last n_3 length chain of $\mathcal{B}$ were taken as the final posterior samples. After these post-processing, MCMC diagnostic plots are attached in APPENDIX C, as well as the geweke test statistics.

Below are the summary of IP-topic-word assignments. Each interaction pattern is paired with (a) posterior estimates of dynamic network effects corresponding to the interaction pattern, (b) the top 3 topics most likely to be generated conditioned on the interaction pattern, and (c) the top 10 most likely words to have generated conditioned on the topic and interaction pattern.

4.2 Dare county email data

After treating multicast emails (those involving a single sender but multiple receivers) as multiple distinct emails, Dare county data contains 4845 emails (only count the email with the number of words greater than 0) between 27 actors, including 2907 vocabulary in total. Again, we used $K = 10$ topics and $C = 3$ interaction patterns, with the parameters fixed as$\alpha = 50/K = 50/10 = 5$ and $\eta = 50/C = 50/3 = 16.67$). For topic-word distributions, we assumed that ϕ follows symmectic Dirichlet distribution with the concentration parameter $\delta = 0.01$. MCMC sampling was implemented based on the order and scheme illustrated earlier. We set the outer iteration number as $I = 150$, and inner iteration numbers as $n_1 = 1, n_2 = 1$, and $n_3 = 3300$. In addition, after some experimentation, δ_B was set as 0.02, to ensure sufficient acceptance rate. In our case, the average acceptance rate for $\boldsymbol{\beta}$ was 0.277. As demonstrated in Algorithm 5, the last value of $\mathcal{C}$, the last value of $\mathcal{Z}$, and the last n_3 length chain of $\mathcal{B}$ were taken as the final posterior samples. Among the $\mathcal{B}$ samples, 300 were discarded as a burn-in and every 3rd samples were taken. After these post-processing, MCMC diagnostic plots are attached in APPENDIX D, as well as geweke test statistics.

APPENDIX

APPENDIX A: Deriving the sampling equations for IPTM

$$
\begin{aligned}
&P(\Phi,\Theta,\mathcal{W},\mathcal{Z},\mathcal{C},\mathcal{B},\mathcal{N}|\mathcal{X},\delta,\boldsymbol{n},\alpha,\boldsymbol{m},\boldsymbol{\gamma},\boldsymbol{\eta},\sigma^2)\\
&=P(\mathcal{W},\mathcal{Z},\mathcal{C},\mathcal{B},\mathcal{N}|\Phi,\Theta,\mathcal{X},\boldsymbol{\gamma},\boldsymbol{\eta},\sigma^2)P(\Phi,\Theta|\delta,\boldsymbol{n},\alpha,\boldsymbol{m})\\
&=P(\mathcal{W}|\mathcal{Z},\Phi)P(\mathcal{Z}|\Theta)P(\mathcal{N}|\mathcal{C},\mathcal{B},\mathcal{X})P(\mathcal{B}|\mathcal{C},\sigma^2)P(\Phi|\delta,\boldsymbol{n})P(\Theta|\mathcal{C},\alpha,\boldsymbol{m})P(\mathcal{C}|\boldsymbol{\gamma})P(\boldsymbol{\gamma}|\boldsymbol{\eta})\\
&=\Big[\prod_{d=1}^{D}\prod_{m=1}^{M^{(d)}}P(w_m^{(d)}|\phi_{z_m^{(d)}})\Big]\times\Big[\prod_{d=1}^{D}\prod_{m=1}^{M^{(d)}}P(z_m^{(d)}|\boldsymbol{\theta}^{(d)})\Big]\times\Big[\prod_{d=1}^{D}P(\mathbf{N}^{(d)}(t^{(d)})|c^{(d)},\boldsymbol{x}^{(c^{(d)})}(t^{(d)}),\boldsymbol{\beta}^{(c)})\Big]\\
&\quad\times\Big[\prod_{c=1}^{C}P(\boldsymbol{\beta}^{(c)}|\sigma^2)\Big]\times\Big[\prod_{k=1}^{K}P(\boldsymbol{\phi}^{(k)}|\delta,\boldsymbol{n})\Big]\times\Big[\prod_{d=1}^{D}P(\boldsymbol{\theta}^{(d)}|\alpha^{(c^{(d)})},\boldsymbol{m}^{(c^{(d)})})\Big]\times\Big[\prod_{d=1}^{D}P(c^{(d)}|\boldsymbol{\gamma})\Big]\times P(\boldsymbol{\gamma}|\boldsymbol{\eta})
\end{aligned}
\tag{22}
$$

Since $P(\boldsymbol{\beta}^{(c)}|\sigma^2)$ is Normal$(\mathbf{0},\sigma^2)$ and $P(\boldsymbol{\gamma}|\boldsymbol{\eta})$ is Dirichlet$(\boldsymbol{\eta})$, we can drop the two terms out and further rewrite the equation (20) as below:

$$
\begin{aligned}
&\propto\Big[\prod_{d=1}^{D}\prod_{m=1}^{M^{(d)}}P(w_m^{(d)}|\phi_{z_m^{(d)}})\Big]\times\Big[\prod_{d=1}^{D}\prod_{m=1}^{M^{(d)}}P(z_m^{(d)}|\boldsymbol{\theta}^{(d)})\Big]\times\Big[\prod_{d=1}^{D}P(\mathbf{N}^{(d)}(t^{(d)})|c^{(d)},\boldsymbol{x}^{(c^{(d)})}(t^{(d)}),\boldsymbol{\beta}^{(c)})\Big]\\
&\quad\times\Big[\prod_{k=1}^{K}P(\boldsymbol{\phi}^{(k)}|\delta,\boldsymbol{n})\Big]\times\Big[\prod_{d=1}^{D}P(\boldsymbol{\theta}^{(d)}|\alpha^{(c^{(d)})},\boldsymbol{m}^{(c^{(d)})})\Big]\times\Big[\prod_{d=1}^{D}P(c^{(d)}|\boldsymbol{\gamma})\Big]\\
&=\Big[\prod_{d=1}^{D}\prod_{m=1}^{M^{(d)}}\phi_{w_m^{(d)}z_m^{(d)}}\Big]\times\Big[\prod_{d=1}^{D}\prod_{m=1}^{M^{(d)}}\boldsymbol{\theta}^{(d)}_{z_m^{(d)}}\Big]\times\Big[\prod_{d=1}^{D}\frac{\exp\{\boldsymbol{\beta}^{(c)T}x_{t^{(d)}}^{(c^{(d)})}(i^{(d)},j^{(d)})\}}{\sum_{j\in\mathcal{A}^{(c)}}\exp\{\boldsymbol{\beta}^{(c)T}x_{t^{(d)}}^{(c^{(d)})}(i^{(d)},j)\}}\Big]\\
&\quad\times\Big[\prod_{k=1}^{K}\Big(\frac{\Gamma(\sum_{w=1}^{W}\delta n_w)}{\prod_{w=1}^{W}\Gamma(\delta n_w)}\prod_{w=1}^{W}\phi_{wk}^{\delta n_w-1}\Big)\Big]\times\Big[\prod_{d=1}^{D}\Big(\frac{\Gamma(\sum_{k=1}^{K}\alpha^{(c^{(d)})}m_k^{(c^{(d)})})}{\prod_{k=1}^{K}\Gamma(\alpha^{(c^{(d)})}m_k^{(c^{(d)})})}\prod_{k=1}^{K}(\boldsymbol{\theta}_k^{(d)})^{\alpha^{(c^{(d)})}m_k^{(c^{(d)})}-1}\Big)\Big]\times\Big[\prod_{d=1}^{D}\gamma_c^{I(c^{(d)}=c)}\Big]\\
&=\Big[\frac{\Gamma(\sum_{w=1}^{W}\delta n_w)}{\prod_{w=1}^{W}\Gamma(\delta n_w)}\Big]^K\times\prod_{d=1}^{D}\Big[\frac{\Gamma(\sum_{k=1}^{K}\alpha^{(c^{(d)})}m_k^{(c^{(d)})})}{\prod_{k=1}^{K}\Gamma(\alpha^{(c^{(d)})}m_k^{(c^{(d)})})}\Big]\times\Big[\prod_{d=1}^{D}\frac{\exp\{\boldsymbol{\beta}^{(c)T}x_{t^{(d)}}^{(c^{(d)})}(i^{(d)},j^{(d)})\}}{\sum_{j\in\mathcal{A}^{(c)}}\exp\{\boldsymbol{\beta}^{(c)T}x_{t^{(d)}}^{(c^{(d)})}(i^{(d)},j)\}}\Big]\\
&\quad\times\Big[\prod_{k=1}^{K}\prod_{w=1}^{W}\phi_{wk}^{M_{wk}^{WK}+\delta n_w-1}\Big]\times\Big[\prod_{d=1}^{D}\prod_{k=1}^{K}(\boldsymbol{\theta}_k^{(d)})^{M_{k|d}+\alpha^{(c^{(d)})}m_k^{(c^{(d)})}-1}\Big]\times\Big[\prod_{d=1}^{D}\gamma_{c^{(d)}}\Big]
\end{aligned}
\tag{23}
$$

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

$$
\begin{aligned}
&P(\mathcal{W},\mathcal{Z},\mathcal{C},\mathcal{B},\mathcal{N}|\mathcal{X},\delta,\boldsymbol{n},\alpha,\boldsymbol{m},\boldsymbol{\gamma},\boldsymbol{\eta},\sigma^2)\\
&=\text{Const.}\int_{\Theta}\int_{\Phi}\Big[\prod_{k=1}^{K}\prod_{w=1}^{W}\phi_{wk}^{M_{wk}^{WK}+\delta n_w-1}\Big]\Big[\prod_{d=1}^{D}\prod_{k=1}^{K}(\boldsymbol{\theta}_k^{(d)})^{M_{k|d}+\alpha^{(c^{(d)})}m_k^{(c^{(d)})}-1}\Big]d\Phi d\Theta\\
&=\text{Const.}\Big[\prod_{k=1}^{K}\int_{\phi_{:k}}\prod_{w=1}^{W}\phi_{wk}^{M_{wk}^{WK}+\delta n_w-1}d\phi_{:k}\Big]\times\Big[\prod_{d=1}^{D}\int_{\theta_{:d}}\prod_{k=1}^{K}(\boldsymbol{\theta}_k^{(d)})^{M_{k|d}+\alpha^{(c^{(d)})}m_k^{(c^{(d)})}-1}d\theta_{:d}\Big]\\
&=\text{Const.}\Big[\prod_{k=1}^{K}\frac{\prod_{w=1}^{W}\Gamma(M_{wk}^{WK}+\delta n_w)}{\Gamma(\sum_{w=1}^{W}M_{wk}^{WK}+\delta)}\Big]\times\Big[\prod_{d=1}^{D}\frac{\prod_{k=1}^{K}\Gamma(M_{k|d}+\alpha^{(c^{(d)})}m_k^{(c^{(d)})})}{\Gamma(M_{\cdot|d}+\alpha^{(c^{(d)})})}\Big].
\end{aligned}
\tag{24}
$$

APPENDIX B: Computing conditional probability

$$
\begin{aligned}
&P(\boldsymbol{w}^{(d)}, \boldsymbol{z}^{(d)} | c^{(d)} = c, \mathcal{W}_{\backslash d}, \mathcal{Z}_{\backslash d}, \mathcal{C}_{\backslash d}, \delta, \boldsymbol{n}, \alpha^{(c)}, \boldsymbol{m}^{(c)}) \\
&\propto \prod_{m=1}^{M^{(d)}} P(z_m^{(d)} = k, w_m^{(d)} = w | c^{(d)} = c, \mathcal{W}_{\backslash d,m}, \mathcal{Z}_{\backslash d,m}, \mathcal{C}_{\backslash d}, \delta, \boldsymbol{n}, \alpha^{(c)}, \boldsymbol{m}^{(c)})
\end{aligned} \tag{25}
$$

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

$$
\begin{aligned}
&P(z_m^{(d)} = k, w_m^{(d)} = w, c^{(d)} = c | \mathcal{W}_{\backslash d,m}, \mathcal{Z}_{\backslash d,m}, \mathcal{C}_{\backslash d}, \delta, \boldsymbol{n}, \alpha^{(c)}, \boldsymbol{m}^{(c)}) \\
&\propto \frac{P(\mathcal{W}, \mathcal{Z}, \mathcal{C} | \delta, \boldsymbol{n}, \alpha, \boldsymbol{m})}{P(\mathcal{W}_{\backslash d,m}, \mathcal{Z}_{\backslash d,m}, \mathcal{C} | \delta, \boldsymbol{n}, \alpha, \boldsymbol{m})} \\
&\propto \frac{\prod_{k=1}^{K} \frac{\prod_{w=1}^{W} \Gamma(M_{wk}^{WK} + \delta n_w)}{\Gamma(\sum_{w=1}^{W} M_{wk}^{WK} + \delta)} \times \prod_{k=1}^{K} \frac{\Gamma(M_{k|d} + \alpha^{(c)} m_k^{(c)})}{\Gamma(M_{\cdot|d} + \alpha^{(c)})}}{\prod_{k=1}^{K} \frac{\prod_{w=1}^{W} \Gamma(M_{wk,\backslash d,m}^{WK} + \delta n_w)}{\Gamma(\sum_{w=1}^{W} M_{wk,\backslash d,m}^{WK} + \delta)} \times \prod_{k=1}^{K} \frac{\Gamma(M_{k|d,\backslash d,m} + \alpha^{(c)} m_k^{(c)})}{\Gamma(M_{\cdot|d,\backslash d,m} + \alpha^{(c)})}} \\
&\propto \frac{M_{wk,\backslash d,m}^{WK} + \delta n_w}{\sum_{w=1}^{W} M_{wk,\backslash d,m}^{WK} + \delta} \times \frac{M_{k|d,\backslash d,m} + \alpha^{(c)} m_k^{(c)}}{M_{\cdot|d} - 1 + \alpha^{(c)}}
\end{aligned} \tag{26}
$$

Then, the conditional probability that a novel word generated in the document of interaction pattern $c^{(d)} = c$ would be assigned to topic $z_m^{(d)} = k$ is obtained by:

$$
\begin{aligned}
&P(z_m^{(d)} = k | w_m^{(d)} = w, c^{(d)} = c, \mathcal{W}_{\backslash d,m}, \mathcal{Z}_{\backslash d,m}, \mathcal{C}_{\backslash d}, \delta, \boldsymbol{n}, \alpha^{(c)}, \boldsymbol{m}^{(c)}) \\
&\propto \frac{M_{k|d,\backslash d,m} + \alpha^{(c)} m_k^{(c)}}{M_{\cdot|d} - 1 + \alpha^{(c)}}
\end{aligned} \tag{27}
$$

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

$$
\begin{aligned}
&P(w_m^{(d)} = w | z_m^{(d)} = k, c^{(d)} = c, \mathcal{W}_{\backslash d,m}, \mathcal{Z}_{\backslash d,m}, \mathcal{C}_{\backslash d}, \delta, \boldsymbol{n}, \alpha^{(c)}, \boldsymbol{m}^{(c)}) \\
&\propto \frac{M_{wk,\backslash d,m}^{WK} + \delta n_w}{\sum_{w=1}^{W} M_{wk,\backslash d,m}^{WK} + \delta}
\end{aligned} \tag{28}
$$

NOTE: Using Equation (26), the unnormalized constant we use to check the model convergence and the corresponding log-constant are,

$$
\begin{aligned}
&\prod_{d=1}^{D} \prod_{m=1}^{M^{(d)}} P(z_m^{(d)} = k, w_m^{(d)} = w | \mathcal{W}_{\backslash d,m}, \mathcal{Z}_{\backslash d,m}, \mathcal{C}, \delta, \boldsymbol{n}, \alpha^{(c)}, \boldsymbol{m}^{(c)}) \\
&\propto \prod_{d=1}^{D} \prod_{m=1}^{M^{(d)}} \frac{M_{w_m^{(d)} z_m^{(d)},\backslash d,m}^{WK} + \delta n_{w_m^{(d)}}}{\sum_{w=1}^{W} M_{w z_m^{(d)},\backslash d,m}^{WK} + \delta} \times \frac{M_{k|d,\backslash d,m} + \alpha^{(c^{(d)})} m_{z_m^{(d)}}^{(c^{(d)})}}{M_{\cdot|d} - 1 + \alpha^{(c^{(d)})}},
\end{aligned} \tag{29}
$$

$$
\begin{aligned}
&\sum_{d=1}^{D} \sum_{m=1}^{M^{(d)}} \log\Big(P(z_m^{(d)} = k, w_m^{(d)} = w | \mathcal{W}_{\backslash d,m}, \mathcal{Z}_{\backslash d,m}, \mathcal{C}, \delta, \boldsymbol{n}, \alpha^{(c)}, \boldsymbol{m}^{(c)})\Big) \\
&\propto \sum_{d=1}^{D} \sum_{m=1}^{M^{(d)}} \log\Big(M_{w_m^{(d)} z_m^{(d)},\backslash d,m}^{WK} + \delta n_{w_m^{(d)}}\Big) - \log\Big(\sum_{w=1}^{W} M_{w z_m^{(d)},\backslash d,m}^{WK} + \delta\Big) \\
&+ \log\Big(M_{k|d,\backslash d,m} + \alpha^{(c^{(d)})} m_{z_m^{(d)}}^{(c^{(d)})}\Big) - \log\Big(M_{\cdot|d} - 1 + \alpha^{(c^{(d)})}\Big)
\end{aligned} \tag{30}
$$

References

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

Cox, D. R. (1992). Regression models and life-tables. In *Breakthroughs in statistics*, pages 527–541. Springer.

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

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

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

Rosen-Zvi, M., Griffiths, T., Steyvers, M., and Smyth, P. (2004). The author-topic model for authors and documents. In *Proceedings of the 20th conference on Uncertainty in artificial intelligence*, pages 487–494. AUAI Press.

Vu, D. Q., Hunter, D., Smyth, P., and Asuncion, A. U. (2011). Continuous-time regression models for longitudinal networks. In Shawe-Taylor, J., Zemel, R., Bartlett, P., Pereira, F., and Weinberger, K., editors, *Advances in Neural Information Processing Systems 24*, pages 2492–2500. Curran Associates, Inc.

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

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