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

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

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

February 20, 2017

Abstract

In this paper, we introduce the interaction-partitioned topic model (IPTM)—a probabilistic model of who communicates with whom about what, and when. Broadly speaking, the IPTM partitions time-stamped textual communications, such as emails, according to both the network dynamics that they reflect and their content. To do this, it draws on the Cox multiplicative intensity model—a generative model for ties that tend toward structural features such as reciprocated dyads and triangles—and latent Dirichlet allocation—a generative model for topic-based content. The IPTM assigns each communication to an "interaction pattern," characterized by a set of dynamic network features and a distribution over a shared set of topics. We use the IPTM to analyze emails sent between department managers in two county governments in North Carolina; one of these email corpora covers the Outer Banks during the time period surrounding Hurricane Sandy. Via this application, we demonstrate that the IPTM is effective at predicting and explaining continuous-time textual communications.

1 IPTM Model

In this section, we first introduce the multiplicative Cox intensity model in the context of tie formation process in a continuous-time textual communication network. Then, we illustrate the generative process of the model which incorporates the generative process of stochastic actor-oriented models and latent Dirichlet allocation. Lastly, specification of the dynamic network statistics used is demonstrated. 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.

1.1 Multiplicative Cox Intensity Model

A single email, indexed by d, is represented by the four components $(i^{(d)}, J^{(d)}, t^{(d)}, W^{(d)})$. The first two are the sender and receiver of the email: an integer $i^{(d)} \in \{1, ..., A\}$ indicates the identity of the sender out of A number of actors (or nodes) and an integer vector $J^{(d)} = \{j_r^{(d)}\}_{r=1}^{|J^{(d)}|}$ indicates the identity of the receiver (or receivers) out of $A-1$ number of actors (by excluding the sender), where $|J^{(d)}| \in \{1, ..., A-1\}$ denotes the total number of the receivers. Next, $t^{(d)}$ is the (unix time-based) timestamp of the email, and $W^{(d)} = \{w_m^{(d)}\}_{m=1}^{M^{(d)}}$ is a set of tokens that comprise the text of the email. In this section, we only consider the first three , $(i^{(d)}, J^{(d)}, t^{(d)})$, and explain how we apply the basic survival analysis concepts and multiplicative Cox intensity model to the generating process of a document (or a tie).

Let T denote the survival time. In our context, survival times measure the time to send the document.

Following the typical survival analysis framework, the distribution of T is described or characterized by three functions, namely:

1. the probability density function $f(t)$:
   the probability of sending a document in a small interval per unit time

$$\begin{aligned} f(t) &= \lim_{\Delta t \to 0+} \frac{P\{\text{a document sent in the interval } (t, t+\Delta t)\}}{\Delta t} \\ &= \lim_{\Delta t \to 0+} \frac{P[T \in (t, t+\Delta t)]}{\Delta t} \end{aligned}$$

2. the survival function $S(t)$
   the probability that a document is not sent until time t

$$\begin{aligned} S(t) &= P(T > t) \\ &= \int_t^\infty f(u)du \\ &= 1 - F(t) \end{aligned}$$

3. the hazard rate function $\lambda(t)$
   the probability of sending a document in a very short interval t to $t + \Delta t$ per unit time, given that the document has not sent until time t

$$\begin{aligned} \lambda(t) &= \lim_{\Delta t \to 0+} \frac{P\{\text{a document sent in the interval } (t, t+\Delta t) | \text{not sent until time } t\}}{\Delta t} \\ &= \lim_{\Delta t \to 0+} \frac{P\{t \le T < t + \Delta t | T \ge t\}}{\Delta t} \end{aligned}$$

   Since our T is a continuous random variable, we have

$$\lambda(t) = \frac{f(t)}{S(t)} = \frac{f(t)}{1 - F(t)} = -\frac{d}{dt}\log[S(t)].$$

   The hazard function can alternatively be represented in terms of the cumulative hazard function

$$\Lambda(t) = \int_0^t \lambda(u)du = -\log[S(t)],$$

   which establishes another relationship

$$S(t) = \exp(-\Lambda(t)).$$

Given the equations above, we derive:

$$f(t) = -\frac{d}{dt}S(t) = -\frac{d}{dt}\exp(-\Lambda(t)) = \lambda(t)S(t) = \lambda(t)e^{-\int_0^t \lambda(u)du},$$

and thus following the dominant method for summarizing survival data, we will use the hazard function $\lambda(t)$ for the generative process of the document. Now we move to Cox multiplicative intensity model (Cox, 1992) using covariates that depend on the history of the process. Cox multiplicative intensity model, also known as the Cox proportional hazards (PH) model, is the most common approach to model covariate effects on survival.

The IPTM assigns each communication to an "interaction pattern," characterized by a set of dynamic network features and a distribution over a shared set of topics. Here we illustrate how a set of dynamic network features contribute uniquely identifies each interaction pattern. Assume that each interaction pattern $c \in \{1, ..., C\}$ has an $A \times A$ stochastic intensity (or hazard) matrix of $\boldsymbol{\lambda}^{(c)}(t) = \{\{\lambda_{ij}^{(c)}(t)\}_{i=1}^A\}_{j=1}^A$, where $\lambda_{ij}^{(c)}(t)$=P{for interaction pattern c, $i \to j$ occurs in time interval $[t, t+dt)$, given that it has not ben sent until time t}. There could be various static and dynamic

covariates of (i, j) that affects the stochastic intensity, however, we decide to use the covariates that depend on the history of the process, considering the strong recency and reciprocity effects of textual communications, especially emails. The detailed specifications of the dynamic network covariates are illustrated in Section 1.3.

Following the multiplicative Cox model of the intensity process $\boldsymbol{\lambda}^{(c)}(t)$ given $\boldsymbol{x}_t^{(c)}(i,j)$, the p-dimensional vector of time-dependent covariates corresponding to each pair of (i, j), the intensity forms:

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

where λ_0 is the common baseline hazards for the overall interaction (assume that λ_0 does not depend on t), $\boldsymbol{\beta}^{(c)}$ 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), and $\mathcal{A}_{\backslash i}$ is the predictable receiver set of sender i within the set of all possible actors $\mathcal{A}$ (no self-loop). Equivalently, by fixing $\lambda_0 = 1$, we can rewrite (1):

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

where the first element of $\boldsymbol{\beta}^{(c)}$ corresponds to the deviation from λ_0, by including the intercept term and setting $\boldsymbol{x}_t^{*(c)}(i,j) = (\mathbf{1}, \boldsymbol{x}_t^{(c)}(i,j))$. Since multicast interactions—those involving a single sender but multiple receivers—are allowed for this model, we expand the rate of interaction between sender i and receiver set J as:

$$\lambda_{iJ}^{(c)}(t|\boldsymbol{x}_t^{*(c)}(i,J)) = \exp\Big\{\sum_{j\in J}\boldsymbol{\beta}^{(c)T}\boldsymbol{x}_t^{*(c)}(i,j)\Big\} \cdot \prod_{j\in J} 1\{j \in \mathcal{A}_{\backslash i}\}. \tag{3}$$

Conditioned upon the existence of a unique document at some particular time t, the probability that the document is sent from i to j is

$$L_{ij}(\boldsymbol{\beta}^{(c)}) = \frac{\exp\Big\{\boldsymbol{\beta}^{(c)T}\boldsymbol{x}_t^{*(c)}(i,j)\Big\}}{\exp\Big\{\sum\limits_{j\in\mathcal{A}_{\backslash i}}\boldsymbol{\beta}^{(c)T}\boldsymbol{x}_t^{*(c)}(i,j)\Big\}},$$

and considering the multicasts and treating the documents being statistically independent, the joint probability, the full likelihood function is:

$$L(\boldsymbol{\beta}^{(c)}) = \prod_{d:c^{(d)}=c} \frac{\exp\{\sum\limits_{j\in J^{(d)}}\boldsymbol{\beta}^{(c)T}x_{t^{(d)}}^{(c)}(i^{(d)},j)\}}{\sum\limits_{\substack{J\subseteq\mathcal{A}_{\backslash i^{(d)}}\\|J|=|J^{(d)}|}}\exp\{\sum\limits_{j\in J}\boldsymbol{\beta}^{(c)T}x_{t^{(d)}}^{(c)}(i^{(d)},j)\}}. \tag{4}$$

This is exactly the form of the partial likelihood function of Cox (1992), however, in our case it is full likelihood since we do not have baseline hazard (incorporated into $\boldsymbol{x}$) in $\lambda_{ij}^{(c)}(t|\boldsymbol{x}_t^{*(c)}(i,j))$. Therefore, for interaction pattern $c = 1, ..., C$, estimation for $\boldsymbol{\beta}^{(c)}$ proceeds by maximizing the log-likelihood function:

$$\log PL(\boldsymbol{\beta}^{(c)}) = \sum_{d:c^{(d)}=c}\Big\{\sum_{j\in J^{(d)}}\boldsymbol{\beta}^{(c)T}\boldsymbol{x}_{t^{(d)}}^{(c)}(i^{(d)},j)\} - \log[\sum_{\substack{J\subseteq\mathcal{A}_{\backslash i^{(d)}}\\|J|=|J^{(d)}|}}\exp\{\sum_{j\in J}\boldsymbol{\beta}^{(c)T}\boldsymbol{x}_{t^{(d)}}^{(c)}(i^{(d)},j)\}]\Big\}, \tag{5}$$

where the risk set in the denominater is defined as all possible sets of recievers with the same cardinality as $|J^{(d)}|$. To prevent the bias in the parameter estimates from treating multicast interactions as well as achieve computational efficiency, we use the log-partial likelihood defined in Perry and Wolfe (2013):

$$\log\widetilde{PL}(\boldsymbol{\beta}^{(c)}) = \sum_{d:c^{(d)}=c}\Big\{\sum_{j\in J^{(d)}}\boldsymbol{\beta}^{(c)T}\boldsymbol{x}_{t^{(d)}}^{(c)}(i^{(d)},j) - |J^{(d)}|\log[\sum_{j\in\mathcal{A}_{\backslash i(d)}}\exp\{\boldsymbol{\beta}^{(c)T}\boldsymbol{x}_{t^{(d)}}^{(c)}(i^{(d)},j)\}]\Big\}, \tag{6}$$

where the approximation $\log PL_t(\boldsymbol{\beta}^{(c)}) \approx \log\widetilde{PL}_t(\boldsymbol{\beta}^{(c)})$ is suggested in Perry and Wolfe (2013) by replacing the sum over all sets of size $|J^{(d)}|$ in (5) with a sum over all multisets of size $|J^{(d)}|$ (i.e. allowing duplicate elements from $\mathcal{A}_{\backslash i^{(d)}}$) as below:

$$\begin{aligned}\log[\sum_{\substack{J\subseteq\mathcal{A}_{\backslash i^{(d)}}\\|J|=|J^{(d)}|}}\exp\{\sum_{j\in J}\boldsymbol{\beta}^{(c)T}\boldsymbol{x}^{(c)}_{t^{(d)}}(i^{(d)},j)\}] &\approx \log[(\sum_{j\in\mathcal{A}_{\backslash i^{(d)}}}\exp\{\sum_{j\in J}\boldsymbol{\beta}^{(c)T}\boldsymbol{x}^{(c)}_{t^{(d)}}(i^{(d)},j)\})^{|J^{(d)}|}] \\ &= |J^{(d)}|\times\log\big[\sum_{j\in\mathcal{A}_{\backslash i^{(d)}}}\exp\{\boldsymbol{\beta}^{(c)T}\boldsymbol{x}^{(c)}_{t^{(d)}}(i^{(d)},j)\}\big]\end{aligned} \tag{7}$$

1.2 Generative Process

The interaction-partitioned topic model (IPTM) is a probabilistic model of who communicates with whom about what, and when. The generative process of IPTM consists of two parts: 1) generation of the ties (i.e. 'who', 'whom', and 'when') and 2) generation of content (i.e. 'what'). The tie generating process resembles that of stochastic actor-oriented models (SAOMs) of Snijders (1996), and the content generating process directly follows latent Dirichlet allocation (LDA) of Blei et al. (2003). In this section, we illustrate the two generative process separately, and show how the two processes can jointly generate a document.

1.2.1 Tie Generating Process

Motivated from SAOMs, we assume the following generative process for each document d in a corpus D:

1. Choose the interaction pattern $c^{(d)} \sim \text{Multinomial}(\gamma)$
2. Set $t = t^{(d-1)}$, $\boldsymbol{\beta} = \boldsymbol{\beta}^{(c^{(d)})}$ and $\boldsymbol{X} = \{\{\boldsymbol{x}^{*(c^{(d)})}_{t_+}(i,j)\}_{i=1}^{A}\}_{j=1}^{A}$ (NOTE: $\boldsymbol{x}^{*(c^{(d)})}_{t^{(d-1)}_+}(i,j) = \boldsymbol{x}^{*(c^{(d)})}_{t^{(d)}_-}(i,j)$)
3. Generate $\Delta T \sim \lambda(\boldsymbol{\beta}, \boldsymbol{X})$, where $\lambda(\boldsymbol{\beta}, \boldsymbol{X}) = \sum_{i=1}^{A}\sum_{j=1}^{A}\exp\Big\{\boldsymbol{\beta}^T\boldsymbol{X}_{ij}\Big\}$
4. Set the timestamp $t^{(d)} = t + \Delta T$
5. Select the sender $i^{(d)} \in 1, ..., A$ randomly using probabilities
$$\frac{\lambda_i(\boldsymbol{\beta}, \boldsymbol{X}_i)}{\lambda(\boldsymbol{\beta}, \boldsymbol{X})},$$
where $\lambda_i(\boldsymbol{\beta}, \boldsymbol{X}_i) = \sum_{j=1}^{A}\exp\Big\{\boldsymbol{\beta}^T\boldsymbol{X}_{ij}\Big\}$
6. Choose the number of receivers $|J^{(d)}| \sim \text{Poisson}(\xi) + 1$.
7. If $|J^{(d)}| = 1$, select the receiver $j^{(d)} \in \mathcal{A}_{\backslash i}$ randomly using probabilities
$$\frac{\lambda_{ij}(\boldsymbol{\beta}, \boldsymbol{X}_{ij})}{\lambda_i(\boldsymbol{\beta}, \boldsymbol{X}_i)},$$
where $\lambda_{ij}(\boldsymbol{\beta}, \boldsymbol{X}_{ij}) = \exp\Big\{\boldsymbol{\beta}^T\boldsymbol{X}_{ij}\Big\}$

7′. if $|J^{(d)}| > 1$, select the receiver set $J^{(d)} \subseteq \mathcal{A}_{\backslash i}$ randomly using probabilities
$$\frac{\lambda_{iJ}(\boldsymbol{\beta}, \boldsymbol{X}_{iJ})}{\sum_{\substack{J\subseteq\mathcal{A}_{\backslash i}\\|J|=|J^{(d)}|}}\lambda_{iJ}(\boldsymbol{\beta}, \boldsymbol{X}_{iJ})},$$
where $\lambda_{iJ}(\boldsymbol{\beta}, \boldsymbol{X}_{iJ}) = \exp\Big\{\sum_{j\in J}\boldsymbol{\beta}^T\boldsymbol{X}_{ij}\Big\}$ with $|J| = |J^{(d)}|$.

7″. (Using approximation in (7)) if $|J^{(d)}| > 1$, select the receiver set $J^{(d)} \subseteq \mathcal{A}_{\backslash i}$ randomly using probabilities

$$\frac{\lambda_{iJ}(\boldsymbol{\beta}, \boldsymbol{X}_{iJ})}{(\sum_{j \in \mathcal{A}_{\backslash i}} \lambda_{ij}(\boldsymbol{\beta}, \boldsymbol{X}_{ij}))^{|J^{(d)}|}},$$

where $\lambda_{iJ}(\boldsymbol{\beta}, \boldsymbol{X}_{iJ}) = \exp\Big\{ \sum_{j \in J} \boldsymbol{\beta}^T \boldsymbol{X}_{ij} \Big\}$ with $|J| = |J^{(d)}|$.

1.2.2 Content Generating Process

By simply adding the interaction pattern assignment of each doeument to LDA, we assume the following generative process for each document d in a corpus D:

1. Choose the interaction pattern $c^{(d)} \sim \text{Multinomial}(\gamma)$
2. Choose the number of words $N^{(d)} \sim \text{Poisson}(\zeta)$
3. Choose document-topic distribution $\boldsymbol{\theta}^{(d)} \sim \text{Dir}(\alpha^{(c^{(d)})}, \boldsymbol{m}^{(c^{(d)})})$
4. For each of the $N^{(d)}$ words $w_n^{(d)}$:
   (a) Choose a topic $z_n^{(d)} \sim \text{Multinomial}(\boldsymbol{\theta}^{(d)})$
   (b) Choose a word $w_n^{(d)} \sim \text{Multinomial}(\phi^{(z_n^{(d)})})$

1.2.3 Joint Generation Process of Document

Below are the detailed generative process for each document in a corpus D and its plate notation (Figure 1).

1. $\boldsymbol{\phi}^{(k)} \sim \text{Dir}(\beta, \mathbf{u})$ **[See Algorithm 1]**
   - A "topic" k is characterized by a discrete distribution over V word types with probability vector $\phi^{(k)}$. A symmetric Dirichlet prior with concentration parameter β is placed.
2. For the interaction pattern $c = 1, ..., C$, **[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) Set $\alpha^{(c)}$ and $\boldsymbol{m}^{(c)}$
   - The topic proportions for documents in the same cluster share the same parameters in the Dirichlet distribution, and how to choose these parameters will be explained in Section X.X.
3. For the document $d = 1, ..., D$ **[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) $N^{(d)} \sim \text{Poisson}(\zeta)$
   (c) Calculate $\boldsymbol{x}^{*(c^{(d)})}_{t_+^{(d-1)}}(i, j)$ and the corresponding $\boldsymbol{\lambda}^{(c^{(d)})}(t)$
   - The dynamic network statistics are calculated based on the documents of the same interaction pattern, using the history of interactions until the previous document.
   (d) Choose $t^{(d)}$, $i^{(d)}$, and $J^{(d)}$ following Section 1.2.1. (i.e. $\mathbf{N}^{(d|c^{(d)})}(t^{(d)}) \sim \text{CP}(\boldsymbol{\lambda}^{(c^{(d)})}(t_+^{(d-1)}))$)
   - $\mathbf{N}^{(d|c^{(d)})}(t^{(d)})$ is a $A \times A$ matrix where $(i^{(d)}, j)^{th}$ $(j \in J^{(d)})$ elements are 1 and the rest are 0.

(e) $\boldsymbol{\theta}^{(d)} \sim \text{Dir}(\alpha^{(c^{(d)})}, \boldsymbol{m}^{(c^{(d)})})$
- 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.

(f) For each of the $N^{(d)}$ words:

(f1) $z_n^{(d)} \sim \text{Multinomial}(\boldsymbol{\theta}^{(d)})$

(f2) $w_n^{(d)} \sim \text{Multinomial}(\boldsymbol{\phi}^{(z_n^{(d)})})$

Algorithm 1 Topic Word Distributions

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

Algorithm 2 Interaction Pattern-unique Parameters

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

Algorithm 3 Document Generating Process

for *d=1 to D* **do**
  draw $c^{(d)} \sim \text{Multinomial}(\boldsymbol{\gamma})$
  draw $N^{(d)} \sim \text{Poisson}(\zeta)$
  draw $(t^{(d)}, i^{(d)}, J^{(d)})$ using $\mathbf{N}^{(d|c^{(d)})}(t^{(d)}) \sim \text{CP}(\boldsymbol{\lambda}^{(c^{(d)})}(t_+^{(d-1)}))$
  draw $\boldsymbol{\theta}^{(d)} \sim \text{Dir}(\alpha^{(c^{(d)})}, \boldsymbol{m}^{(c^{(d)})})$
  for *n=1 to* $N^{(d)}$ **do**
    draw $z_n^{(d)} \sim \text{Multinomial}(\boldsymbol{\theta}^{(d)})$
    draw $w_n^{(d)} \sim \text{Multinomial}(\boldsymbol{\phi}^{(z_n^{(d)})})$
  end
end

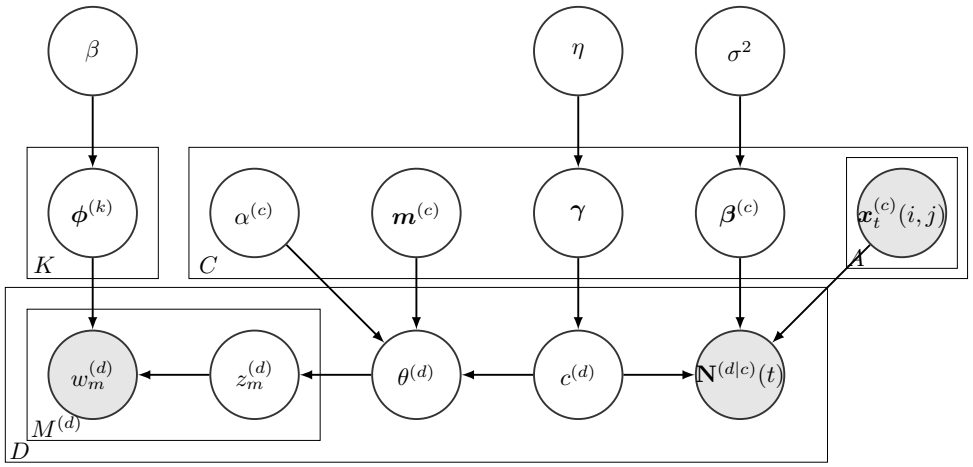

Figure 1: Plate notation of IPTM

1.3 Dynamic covariates to measure network effects

The network statistics $\boldsymbol{x}_t^{(c)}(i,j)$ of Equation (1), 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.

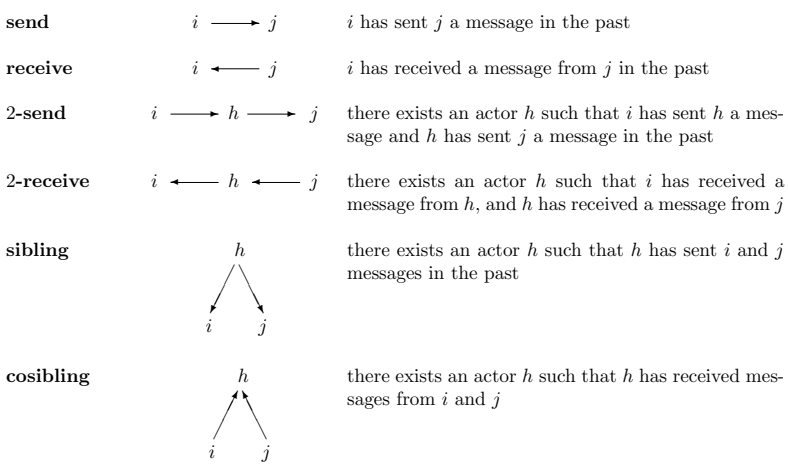

Fig. 3. Dynamic covariates to measure network effects

We use the statistics defined in Perry and Wolfe (2013) as the components of $\boldsymbol{x}_t^{(c)}(i,j)$ (refer to Fig.3 of Perry and Wolfe (2013) attached above), then additionaly use the intercept (to estimate the baseline intensities), outdegree, and indegree (to measure the popularity and centrality). For simplicity, we add the four triadic statistics (2-send, 2-receive, sibling, cosibling) and define it as 'triangle' statistic. Moreover, to maintain the consistent scale of the statistics, we divide by the total number of documents corresponding to the condition $c^{(d)} = c$ and $t^{(d)} < t$, which we refer to D^{ct}.

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

2 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

$$\begin{aligned}
&P(\Phi, \Theta, \mathcal{W}, \mathcal{Z}, \mathcal{C}, \mathcal{B}, \mathcal{N} | \mathcal{X}, \beta, \boldsymbol{u}, \alpha, \boldsymbol{m}, \boldsymbol{\gamma}, \boldsymbol{\eta}, \sigma^2) \\
&= P(\mathcal{W}, \mathcal{Z}, \mathcal{C}, \mathcal{B}, \mathcal{N} | \Phi, \Theta, \mathcal{X}, \boldsymbol{\gamma}, \boldsymbol{\eta}, \sigma^2) P(\Phi, \Theta | \beta, \boldsymbol{u}, \alpha, \boldsymbol{m}) \\
&= P(\mathcal{W}|\mathcal{Z}, \Phi) P(\mathcal{Z}|\Theta) P(\mathcal{N}|\mathcal{C}, \mathcal{X}, \mathcal{B}) P(\mathcal{B}|\mathcal{C}, \sigma^2) P(\Phi|\beta, \boldsymbol{u}) P(\Theta|\mathcal{C}, \alpha, \boldsymbol{m}) P(\mathcal{C}|\boldsymbol{\gamma}) P(\boldsymbol{\gamma}|\boldsymbol{\eta})
\end{aligned} \tag{8}$$

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

$$\propto P(\mathcal{W}|\mathcal{Z}) P(\mathcal{Z}|\mathcal{C}, \beta, \boldsymbol{u}, \alpha, \boldsymbol{m}) P(\mathcal{N}|\mathcal{C}, \mathcal{B}, \mathcal{X}) P(\mathcal{B}|\mathcal{C}, \sigma^2) P(\mathcal{C}|\boldsymbol{\gamma}) \tag{9}$$

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

$$P(\mathcal{Z}, \mathcal{C}, \mathcal{B}|\mathcal{W}, \mathcal{N}, \mathcal{X}, \beta, \boldsymbol{u}, \alpha, \boldsymbol{m}, \boldsymbol{\gamma}, \boldsymbol{\eta}, \sigma^2) \propto P(\mathcal{W}, \mathcal{Z}, \mathcal{C}, \mathcal{B}, \mathcal{N}|\mathcal{X}, \beta, \boldsymbol{u}, \alpha, \boldsymbol{m}, \boldsymbol{\gamma}, \boldsymbol{\eta}, \sigma^2) \tag{10}$$

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

2.1 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. $P(c^{(d)} = c|\mathcal{W}, \mathcal{Z}, \mathcal{C}_{\setminus d}, \mathcal{B}, \mathcal{N}, \mathcal{X}, \beta, \boldsymbol{u}, \alpha, \boldsymbol{m}, \boldsymbol{\gamma}, \boldsymbol{\eta}, \sigma^2)$. We can derive the equation as below:

$$\begin{aligned}
&P(c^{(d)} = c|\mathcal{W}, \mathcal{Z}, \mathcal{C}_{\setminus d}, \mathcal{B}, \mathcal{N}, \mathcal{X}, \beta, \boldsymbol{u}, \alpha, \boldsymbol{m}, \boldsymbol{\gamma}, \boldsymbol{\eta}, \sigma^2) \\
&\propto P(c^{(d)} = c, \boldsymbol{w}^{(d)}, \boldsymbol{z}^{(d)}, \mathbf{N}^{(d)}(t^{(d)})|\mathcal{W}_{\setminus d}, \mathcal{Z}_{\setminus d}, \mathcal{C}_{\setminus d}, \mathcal{B}, \mathcal{N}_{\setminus d}, \mathcal{X}, \beta, \boldsymbol{u}, \alpha, \boldsymbol{m}, \boldsymbol{\gamma}, \boldsymbol{\eta}, \sigma^2) \\
&\propto P(c^{(d)} = c|\mathcal{C}_{\setminus d}, \boldsymbol{\gamma}) P(\mathbf{N}^{(d)}(t^{(d)})|c^{(d)} = c, \mathcal{C}_{\setminus d}, \mathcal{B}, \mathcal{N}_{\setminus d}, \mathcal{X}) P(\boldsymbol{w}^{(d)}, \boldsymbol{z}^{(d)}|c^{(d)} = c, \mathcal{W}_{\setminus d}, \mathcal{Z}_{\setminus d}, \mathcal{C}_{\setminus d}, \beta, \boldsymbol{u}, \alpha, \boldsymbol{m}),
\end{aligned} \tag{11}$$

where $P(c^{(d)} = c|\mathcal{C}_{\setminus d}, \boldsymbol{\gamma})$ comes from the multinomial prior γ and $P(\mathbf{N}^{(d)}(t^{(d)})|c^{(d)} = c, \mathcal{C}_{\setminus d}, \mathcal{B}, \mathcal{N}_{\setminus d}, \mathcal{X})$ 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 $P(\boldsymbol{w}^{(d)}, \boldsymbol{z}^{(d)}|c^{(d)} = c, \mathcal{W}_{\setminus d}, \mathcal{Z}_{\setminus d}, \mathcal{C}_{\setminus d}, \beta, \boldsymbol{u}, \alpha, \boldsymbol{m})$, we will follow typical LDA approach.

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

$$= \Big[\gamma_c\Big] \times \Big[\frac{\exp\{\boldsymbol{\beta}^{(c)T} x^{(c)}_{t^{(d)}}(i^{(d)}, j^{(d)})\}}{\sum_{j \in \mathcal{A}^{(c)}} \exp\{\boldsymbol{\beta}^{(c)T} x^{(c)}_{t^{(d)}}(i^{(d)}, j)\}}\Big] \times \Big[\prod_{m=1}^{M^{(d)}} \frac{N_{z_m^{(d)}|d,\setminus d,m} + \alpha^{(c)} \boldsymbol{m}^{(c)}_{z_m^{(d)}}}{N_{\cdot|d} - 1 + \alpha^{(c)}}\Big], \tag{12}$$

where $N_{k|d}$ 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:

$$\log(\gamma_c) + \Big(\boldsymbol{\beta}^{(c)T} x^{(c)}_{t^{(d)}}(i^{(d)}, j^{(d)}) - \log\big[\sum_{j \in \mathcal{A}^{(c)}} \exp\{\boldsymbol{\beta}^{(c)T} x^{(c)}_{t^{(d)}}(i^{(d)}, j)\}\big]\Big) + \sum_{m=1}^{M^{(d)}} \log\Big(\frac{N_{z_m^{(d)}|d,\setminus d,m} + \alpha^{(c)} \boldsymbol{m}^{(c)}_{z_m^{(d)}}}{N_{\cdot|d} - 1 + \alpha^{(c)}}\Big). \tag{13}$$

2.2 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 C:

$$\begin{aligned}
&P(z_m^{(d)} = k|\mathcal{W}, \mathcal{Z}_{\setminus d,m}, \mathcal{C}, \mathcal{B}, \mathcal{N}, \mathcal{X}, \beta, \boldsymbol{u}, \alpha, \boldsymbol{m}, \boldsymbol{\gamma}, \boldsymbol{\eta}, \sigma^2) \\
&\propto P(z_m^{(d)} = k, w_m^{(d)}|\mathcal{W}_{\setminus d,m}, \mathcal{Z}_{\setminus d,m}, C, \beta, \boldsymbol{u}, \alpha, \boldsymbol{m})
\end{aligned} \tag{14}$$

where the subscript "$\setminus d, m$" 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:

$$\propto (N_{k|d,\setminus d,m} + \alpha^{(c^{(d)})} \boldsymbol{m}_k^{(c^{(d)})}) \times \frac{N^{WK}_{w_m^{(d)} k,\setminus d,m} + \beta n_w}{\sum_{w=1}^{W} N^{WK}_{wk,\setminus d,m} + \beta} \tag{15}$$

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

2.3 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 β_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:

$$\text{Acceptance Probability} = \begin{cases} \frac{P(\mathcal{B}'|\mathcal{W},\mathcal{Z},\mathcal{C},\mathcal{N},\mathcal{X})}{P(\mathcal{B}|\mathcal{W},\mathcal{Z},\mathcal{C},\mathcal{N},\mathcal{X})} & \text{if} < 1 \\ 1 \quad \text{else} \end{cases} \tag{16}$$

After factorization, we get

$$\begin{aligned} \frac{P(\mathcal{B}'|\mathcal{W},\mathcal{Z},\mathcal{C},\mathcal{N},\mathcal{X})}{P(\mathcal{B}|\mathcal{W},\mathcal{Z},\mathcal{C},\mathcal{N},\mathcal{X})} &= \frac{P(\mathcal{N}|\mathcal{B}',\mathcal{W},\mathcal{Z},\mathcal{C},\mathcal{X})P(\mathcal{B}')}{P(\mathcal{N}|\mathcal{B},\mathcal{W},\mathcal{Z},\mathcal{C},\mathcal{X})P(\mathcal{B})} \\ &= \frac{P(\mathcal{N}|\mathcal{C},\mathcal{X},\mathcal{B}')P(\mathcal{B}')}{P(\mathcal{N}|\mathcal{C},\mathcal{X},\mathcal{B})P(\mathcal{B})}, \end{aligned} \tag{17}$$

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:

$$\begin{aligned} &\log\Big(\phi_d(\boldsymbol{\beta}'^{(c)};\mathbf{0},\sigma^2 I_P)\Big) - \log\Big(\phi_d(\boldsymbol{\beta}'^{(c)};\mathbf{0},\sigma^2 I_P)\Big) \\ &+ \sum_{d:c^{(d)}=c} \Big\{\boldsymbol{\beta}'^{(c)T} x^{(c)}_{t^{(d)}}(i^{(d)},j^{(d)}) - \log\big[\sum_{j\in\mathcal{A}^{(c)}} \exp\{\boldsymbol{\beta}'^{(c)T} x^{(c)}_{t^{(d)}}(i^{(d)},j)\}\big]\Big\} \\ &- \sum_{d:c^{(d)}=c} \Big\{\boldsymbol{\beta}^{(c)T} x^{(c)}_{t^{(d)}}(i^{(d)},j^{(d)}) - \log\big[\sum_{j\in\mathcal{A}^{(c)}} \exp\{\boldsymbol{\beta}^{(c)T} x^{(c)}_{t^{(d)}}(i^{(d)},j)\}\big]\Big\}, \end{aligned} \tag{18}$$

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

$$\log(\text{Acceptance Probability}) = \min((16),\, 0) \tag{19}$$

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).

2.4 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 4 MCMC($I, n_1, n_2, n_3, \beta_B$)

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 $\mathcal{B}$

2.5 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

$$P(\mathcal{D}^{(c)}|\alpha^{(c)}\boldsymbol{m}^{(c)}) = \prod_{d:c^{(d)}=c} \frac{\Gamma(\alpha^{(c)})}{\Gamma(N_{\cdot|d} + \alpha^{(c)})} \prod_{k=1}^{K} \frac{\Gamma(N_{k|d} + \alpha^{(c)} m_k^{(c)})}{\Gamma(\alpha^{(c)} m_k^{(c)})} \tag{20}$$

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 $[\alpha^{(c)}m_k^{(c)}]^*$:

$$[\alpha^{(c)}m_k^{(c)}]^* = \alpha^{(c)}m_k^{(c)} \frac{\sum_{d:c^{(d)}=c} \Psi(N_{k|d} + \alpha^{(c)}m_k^{(c)}) - \Psi(\alpha^{(c)}m_k^{(c)})}{\sum_{d:c^{(d)}=c} \Psi(N_{\cdot|d} + \alpha^{(c)}) - \Psi(\alpha^{(c)})}, \tag{21}$$

where $\Psi(\cdot)$ is the first derivative of the log gamma function, known as the digamma function, and the quantity $N_{k|d}$ is the number of times that outcome k was observed in the document d. Moreover, the quantity $N_{\cdot|d} = \sum_{k=1}^{K} N_{k|d}$ 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 $C_k(n) = \sum_{d:c^{(d)}=c} \beta(N_{k|d} - n)$ and $C_\cdot(n) = \sum_{d:c^{(d)}=c} \beta(N_{\cdot|d} - n)$:

$$[\alpha^{(c)}m_k^{(c)}]^* = \alpha^{(c)}m_k^{(c)} \frac{\sum_{n=1}^{\max_d N_{k|d}} C_k(n)[\Psi(n + \alpha^{(c)}m_k^{(c)}) - \Psi(\alpha^{(c)}m_k^{(c)})]}{\sum_{n=1}^{\max_d N_{\cdot|d}} C_\cdot(n)[\Psi(n + \alpha^{(c)}) - \Psi(\alpha^{(c)})]}. \tag{22}$$

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:

$$[\alpha^{(c)}m_k^{(c)}]^* = \alpha^{(c)}m_k^{(c)} \frac{\sum_{n=1}^{\max_d N_{k|d}} C_k(n) \sum_{f=1}^{n} \frac{1}{f-1+\alpha^{(c)}m_k^{(c)}}}{\sum_{n=1}^{\max_d N_{\cdot|d}} C_\cdot(n) \sum_{f=1}^{n} \frac{1}{f-1+\alpha^{(c)}}}. \tag{23}$$

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).

APPENDIX

APPENDIX A: Notations in IPTM

Authors of the corpus	$\mathcal{A}$	Set
Number of authors	A	Scalar
Number of documents	D	Scalar
Number of words in the d^{th} document	$N^{(d)}$	Scalar
Number of topics	K	Scalar
Vocabulary size	W	Scalar
Number of interaction patterns	C	Scalar
Number of words assigned to interaction pattern and topic	N^{CK}	Scalar
Number of words assigned to word and topic	N^{WK}	Scalar
Interaction pattern of the d^{th} document	$c^{(d)}$	Scalar
Time of the d^{th} document	$t^{(d)}$	Scalar
Number of documents corresponding to the condition $c^{(d)} = c$ and $t^{(d)} < t$	D^{ct}	Scalar
Words in the d^{th} document	$\boldsymbol{w}^{(d)}$	$N^{(d)}$-dimensional vector
n^{th} word in the d^{th} document	$w_n^{(d)}$	n^{th} component of $\boldsymbol{w}^{(d)}$
Topic assignments in the d^{th} document	$\boldsymbol{z}^{(d)}$	$N^{(d)}$-dimensional vector
Topic assignments for n^{th} word in the d^{th} document	$z_n^{(d)}$	n^{th} component of $\boldsymbol{z}^{(d)}$
Dirichlet concentration prior given interaction pattern c	$\alpha^{(c)}$	Scalar
Dirichlet base prior given interaction pattern c	$\boldsymbol{m}^{(c)}$	K-dimensional vector
Dirichlet concentration prior	β	Scalar
Dirichlet base prior	$\boldsymbol{u}$	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

APPENDIX B: Deriving the sampling equations for IPTM

$$
\begin{aligned}
&P(\Phi, \Theta, \mathcal{W}, \mathcal{Z}, \mathcal{C}, \mathcal{B}, \mathcal{N} | \mathcal{X}, \beta, \boldsymbol{n}, \alpha, \boldsymbol{m}, \boldsymbol{\gamma}, \boldsymbol{\eta}, \sigma^2) \\
&= P(\mathcal{W}, \mathcal{Z}, \mathcal{C}, \mathcal{B}, \mathcal{N} | \Phi, \Theta, \mathcal{X}, \boldsymbol{\gamma}, \boldsymbol{\eta}, \sigma^2) P(\Phi, \Theta | \beta, \boldsymbol{n}, \alpha, \boldsymbol{m}) \\
&= P(\mathcal{W} | \mathcal{Z}, \Phi) P(\mathcal{Z} | \Theta) P(\mathcal{N} | \mathcal{C}, \mathcal{B}, \mathcal{X}) P(\mathcal{B} | \mathcal{C}, \sigma^2) P(\Phi | \beta, \boldsymbol{n}) P(\Theta | \mathcal{C}, \alpha, \boldsymbol{m}) P(\mathcal{C} | \boldsymbol{\gamma}) P(\boldsymbol{\gamma} | \boldsymbol{\eta}) \\
&= \Big[\prod_{d=1}^{D} \prod_{m=1}^{M^{(d)}} P(w_m^{(d)} | \phi_{z_m^{(d)}}) \Big] \times \Big[\prod_{d=1}^{D} \prod_{m=1}^{M^{(d)}} P(z_m^{(d)} | \boldsymbol{\theta}^{(d)}) \Big] \times \Big[\prod_{d=1}^{D} P(\mathbf{N}^{(d)}(t^{(d)}) | c^{(d)}, \boldsymbol{x}^{(c^{(d)})}(t^{(d)}), \boldsymbol{\beta}^{(c)}) \Big] \\
&\quad \times \Big[\prod_{c=1}^{C} P(\boldsymbol{\beta}^{(c)} | \sigma^2) \Big] \times \Big[\prod_{k=1}^{K} P(\boldsymbol{\phi}^{(k)} | \beta, \boldsymbol{n}) \Big] \times \Big[\prod_{d=1}^{D} P(\boldsymbol{\theta}^{(d)} | \alpha^{(c^{(d)})}, \boldsymbol{m}^{(c^{(d)})}) \Big] \times \Big[\prod_{d=1}^{D} P(c^{(d)} | \boldsymbol{\gamma}) \Big] \times P(\boldsymbol{\gamma} | \boldsymbol{\eta})
\end{aligned}
\tag{24}
$$

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)}|\beta, \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}_{z_m^{(d)}}^{(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}\beta n_w)}{\prod_{w=1}^{W}\Gamma(\beta n_w)} \prod_{w=1}^{W} \phi_{wk}^{\beta 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}\beta n_w)}{\prod_{w=1}^{W}\Gamma(\beta 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} + \beta n_w - 1}\Big] \times \Big[\prod_{d=1}^{D}\prod_{k=1}^{K} (\boldsymbol{\theta}_k^{(d)})^{N_{k|d} + \alpha^{(c^{(d)})} m_k^{(c^{(d)})} - 1}\Big] \times \Big[\prod_{d=1}^{D} \gamma_{c^{(d)}}\Big]
\end{aligned}
\tag{25}
$$

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

$$
\begin{aligned}
&P(\mathcal{W}, \mathcal{Z}, \mathcal{C}, \mathcal{B}, \mathcal{N}|\mathcal{X}, \beta, \boldsymbol{n}, \alpha, \boldsymbol{m}, \boldsymbol{\gamma}, \boldsymbol{\eta}, \sigma^2) \\
&= \text{Const.} \int_{\Theta}\int_{\Phi} \Big[\prod_{k=1}^{K}\prod_{w=1}^{W} \phi_{wk}^{M_{wk}^{WK} + \beta n_w - 1}\Big]\Big[\prod_{d=1}^{D}\prod_{k=1}^{K} (\boldsymbol{\theta}_k^{(d)})^{N_{k|d} + \alpha^{(c^{(d)})} m_k^{(c^{(d)})} - 1}\Big] d\Phi d\Theta \\
&= \text{Const.} \Big[\prod_{k=1}^{K} \int_{\phi_{:k}} \prod_{w=1}^{W} \phi_{wk}^{M_{wk}^{WK} + \beta n_w - 1} d\phi_{:k}\Big] \times \Big[\prod_{d=1}^{D} \int_{\theta_{:d}} \prod_{k=1}^{K} (\boldsymbol{\theta}_k^{(d)})^{N_{k|d} + \alpha^{(c^{(d)})} m_k^{(c^{(d)})} - 1} d\theta_{:d}\Big] \\
&= \text{Const.} \Big[\prod_{k=1}^{K} \frac{\prod_{w=1}^{W}\Gamma(N_{wk}^{WK} + \beta n_w)}{\Gamma(\sum_{w=1}^{W} N_{wk}^{WK} + \beta)}\Big] \times \Big[\prod_{d=1}^{D} \frac{\prod_{k=1}^{K}\Gamma(N_{k|d} + \alpha^{(c^{(d)})} m_k^{(c^{(d)})})}{\Gamma(N_{\cdot|d} + \alpha^{(c^{(d)})})}\Big].
\end{aligned}
\tag{26}
$$

APPENDIX C: Computing conditional probability

$$
\begin{aligned}
&P(\boldsymbol{w}^{(d)}, \boldsymbol{z}^{(d)}|c^{(d)} = c, \mathcal{W}_{\setminus d}, \mathcal{Z}_{\setminus d}, \mathcal{C}_{\setminus d}, \beta, \boldsymbol{n}, \alpha^{(c)}, \boldsymbol{m}^{(c)}) \\
&\propto \prod_{m=1}^{M^{(d)}} P(z_m^{(d)} = k, w_m^{(d)} = w|c^{(d)} = c, \mathcal{W}_{\setminus d,m}, \mathcal{Z}_{\setminus d,m}, \mathcal{C}_{\setminus d}, \beta, \boldsymbol{n}, \alpha^{(c)}, \boldsymbol{m}^{(c)})
\end{aligned}
\tag{27}
$$

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

$k\Gamma(k)$,

$$
\begin{aligned}
&P(z_m^{(d)}=k, w_m^{(d)}=w, c^{(d)}=c|\mathcal{W}_{\setminus d,m}, \mathcal{Z}_{\setminus d,m}, \mathcal{C}_{\setminus d}, \beta, \boldsymbol{n}, \alpha^{(c)}, \boldsymbol{m}^{(c)}) \\
&\propto \frac{P(\mathcal{W}, \mathcal{Z}, \mathcal{C}|\beta, \boldsymbol{n}, \alpha, \boldsymbol{m})}{P(\mathcal{W}_{\setminus d,m}, \mathcal{Z}_{\setminus d,m}, \mathcal{C}|\beta, \boldsymbol{n}, \alpha, \boldsymbol{m})} \\
&\propto \frac{\prod_{k=1}^{K} \frac{\prod_{w=1}^{W}\Gamma(N_{wk}^{WK}+\beta n_w)}{\Gamma(\sum_{w=1}^{W} N_{wk}^{WK}+\beta)} \times \prod_{k=1}^{K} \frac{\Gamma(N_{k|d}+\alpha^{(c)}m_k^{(c)})}{\Gamma(N_{\cdot|d}+\alpha^{(c)})}}{\prod_{k=1}^{K} \frac{\prod_{w=1}^{W}\Gamma(N_{wk,\setminus d,m}^{WK}+\beta n_w)}{\Gamma(\sum_{w=1}^{W} N_{wk,\setminus d,m}^{WK}+\beta)} \times \prod_{k=1}^{K} \frac{\Gamma(N_{k|d,\setminus d,m}+\alpha^{(c)}m_k^{(c)})}{\Gamma(N_{\cdot|d,\setminus d,m}+\alpha^{(c)})}} \\
&\propto \frac{N_{wk,\setminus d,m}^{WK}+\beta n_w}{\sum_{w=1}^{W} N_{wk,\setminus d,m}^{WK}+\beta} \times \frac{N_{k|d,\setminus d,m}+\alpha^{(c)}m_k^{(c)}}{N_{\cdot|d}-1+\alpha^{(c)}}
\end{aligned}
\tag{28}
$$

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:

$$
\begin{aligned}
&P(z_m^{(d)}=k|w_m^{(d)}=w, c^{(d)}=c, \mathcal{W}_{\setminus d,m}, \mathcal{Z}_{\setminus d,m}, \mathcal{C}_{\setminus d}, \beta, \boldsymbol{n}, \alpha^{(c)}, \boldsymbol{m}^{(c)}) \\
&\propto \frac{N_{k|d,\setminus d,m}+\alpha^{(c)}m_k^{(c)}}{N_{\cdot|d}-1+\alpha^{(c)}}
\end{aligned}
\tag{29}
$$

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:

$$
\begin{aligned}
&P(w_m^{(d)}=w|z_m^{(d)}=k, c^{(d)}=c, \mathcal{W}_{\setminus d,m}, \mathcal{Z}_{\setminus d,m}, \mathcal{C}_{\setminus d}, \beta, \boldsymbol{n}, \alpha^{(c)}, \boldsymbol{m}^{(c)}) \\
&\propto \frac{N_{wk,\setminus d,m}^{WK}+\beta n_w}{\sum_{w=1}^{W} N_{wk,\setminus d,m}^{WK}+\beta}
\end{aligned}
\tag{30}
$$

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

$$
\begin{aligned}
&\prod_{d=1}^{D}\prod_{m=1}^{M^{(d)}} P(z_m^{(d)}=k, w_m^{(d)}=w|\mathcal{W}_{\setminus d,m}, \mathcal{Z}_{\setminus d,m}, \mathcal{C}, \beta, \boldsymbol{n}, \alpha^{(c)}, \boldsymbol{m}^{(c)}) \\
&\propto \prod_{d=1}^{D}\prod_{m=1}^{M^{(d)}} \frac{N_{w_m^{(d)}z_m^{(d)},\setminus d,m}^{WK}+\beta n_{w_m^{(d)}}}{\sum_{w=1}^{W} N_{wz_m^{(d)},\setminus d,m}^{WK}+\beta} \times \frac{N_{k|d,\setminus d,m}+\alpha^{(c^{(d)})}m_{z_m^{(d)}}^{(c^{(d)})}}{N_{\cdot|d}-1+\alpha^{(c^{(d)})}},
\end{aligned}
\tag{31}
$$

$$
\begin{aligned}
&\sum_{d=1}^{D}\sum_{m=1}^{M^{(d)}} \log\Big(P(z_m^{(d)}=k, w_m^{(d)}=w|\mathcal{W}_{\setminus d,m}, \mathcal{Z}_{\setminus d,m}, \mathcal{C}, \beta, \boldsymbol{n}, \alpha^{(c)}, \boldsymbol{m}^{(c)})\Big) \\
&\propto \sum_{d=1}^{D}\sum_{m=1}^{M^{(d)}} \log\Big(N_{w_m^{(d)}z_m^{(d)},\setminus d,m}^{WK}+\beta n_{w_m^{(d)}}\Big) - \log\Big(\sum_{w=1}^{W} N_{wz_m^{(d)},\setminus d,m}^{WK}+\beta\Big) \\
&+ \log\Big(N_{k|d,\setminus d,m}+\alpha^{(c^{(d)})}m_{z_m^{(d)}}^{(c^{(d)})}\Big) - \log\Big(N_{\cdot|d}-1+\alpha^{(c^{(d)})}\Big)
\end{aligned}
\tag{32}
$$

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.

Snijders, T. A. (1996). Stochastic actor-oriented models for network change. *Journal of mathematical sociology*, 21(1-2):149–172.

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.