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

Anonymous Authors[1]

Abstract

We introduce the interaction-partitioned topic model (IPTM)—a probabilistic model for who communicates with whom about what, and when. Broadly speaking, the IPTM partitions time-stamped textual communications, according to both the network dynamics that they reflect and their content. To define the IPTM, we integrate a dynamic version of the exponential random graph model—a generative model for ties that tend toward structural features such as triangles—and latent Dirichlet allocation—a generative model for topic-based content. The IPTM assigns each document to an "interaction pattern"—a generative process for contents and ties that is governed by a topic distribution and a set of dynamic network features. We use the IPTM to analyze emails sent between department managers in Dare county government in North Carolina, and demonstrate that the model is effective at predicting and explaining continuous-time textual communications.

1. Introduction

In recent decades, real-time digitized textual communication has developed into a ubiquitous form of social and professional interaction (**????**). From the perspective of the computational social scientist, this has lead to a growing need for methods of modeling interactions that manifest as text exchanged in continuous time. A number of models that build upon topic modeling through Latent Dirichlet Allocation (**?**) to incorporate link data as well as textual content have been developed recently (**???**). These models are innovative in their extensions that incorporate network tie information. However, none of the models that are currently available in the literature integrate the rich random-graph structure offered by state of the art models for network structure—such as the exponential random graph model (ERGM) (**???**). The ERGM is the canonical model for modeling the structure of a static network. It is flexible enough to specify a generative model that accounts for nearly any pattern of tie formation (e.g., reciprocity, clustering, popularity effects) (**?**). Several models have been developed that handle time-stamped ties in which tie formation is governed by structural dynamics similar to those used in ERGMs (**???**). We develop the interaction-partitioned topic model (IPTM) which simultaneously models the network structural patterns that govern time-stamped tie formation, and the content in the communications.

The models on which we build, including the relational event model (**?**), the point process model of **?**, and most closely the stochastic actor oriented model (SAOM) (**?**), provide frameworks for explaining or predicting ties between nodes using the network sub-structures in which the two nodes are embedded (e.g., predict a tie is highly likely to form between two nodes if those two nodes have many shared partners). Models based on network structure have been used for many applications in which the ties between nodes are annotated with text. The text, despite providing rich information regarding the strength, scope, and character of the ties, has been largely excluded from these analyses, due to the inability of these network models to incorporate textual attributes of ties. These application domains include, among other applicaitons, the study of legislative networks in which networks reflect legislators' co-support of bills, but exclude bill text (**??**); the study of alliance networks in which networks reflect countries' co-signing of treaties, but exclude treaty text (**????**); the study of scientific co-authorship networks that exclude the text of the co-authored papers (**???**); and the study of text-based interaction on social media (e.g., users tied via 'mentions' on twitter) (**???**).

In defining and testing the IPTM we embed core conceptual property—interaction pattern—to link the content component of the model, and network component of the model such that knowing who is communicating with whom at what time (i.e., the network component) provides information about the content of communication, and vice versa (Section **??**). Figure **??** (plot needs to be replaced) illus-

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

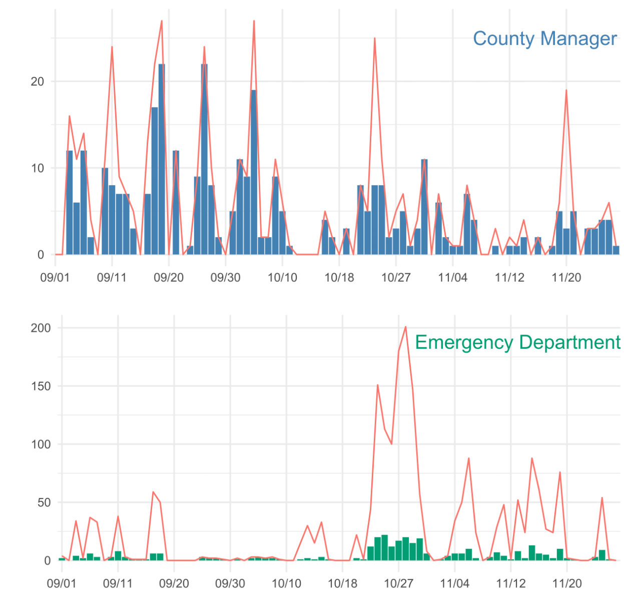

Figure 1. Sending behavior of two most active nodes in Dare County email data between 09/01/2012 and 11/30/2012. *Top*: the number of emails per day sent by County manager (blue bar) and the number of recipients from this person per day (red line). *Bottom*: the number of emails per day sent by emergency department official (green bar) and the number of recipients from this person per day (red line).

trates this structure. IPTM leads to an efficient MCMC inference algoritm (Section **??**) and acheives good predictive peformance (Section **??**). Finally, the IPTM discovers interesting and interpretable latent structure through application to email corpora of internal communications by government officials in Dare County, NC (Section **??**).

2. Interaction-partitioned Topic Model

Data generated under the IPTM consists of D unique documents. A single document, indexed by $d \in [D]$, is represented by the four components: the author $a_d \in [A]$, an indicator vector of recipients $\boldsymbol{r}_d = \{u_{dr}\}_{r=1}^{A}$, the timestamp $t_d \in (0, \infty)$, and a set of tokens $\boldsymbol{w}_d = \{w_{dn}\}_{n=1}^{N_d}$ that comprise the text of the document, where N_d denotes the total number of tokens in a document. For simplicity, we assume that documents are ordered by time such that $t_d < t_{d+1}$.

2.1. Interaction Patterns

They key idea that combines the IPTM component modeling "what" with the component modeling "who," "whom," and "when" is that different documents comes from the introduction of "interaction patterns". Each interaction pattern $c \in [C]$ is characterized by a set of dynamic network features—such as the number of messages sent from a to r in some time interval—and corresponding coefficients. We associate each document with the interaction pattern that best describes how people interact, and that is reflected to what people talk about via topic assignments. To be specific, we assume an interaction-pattern distribution over C unique interaction patterns

$$\boldsymbol{\psi} \sim \text{Dirichlet}\Big(\zeta, (\frac{1}{C}, \ldots, \frac{1}{C})\Big), \tag{1}$$

where ζ is the concentration parameter, and then each document $d \in [D]$ draws an interaction pattern c_d as below:

$$c_d \sim \text{Multinomial}(\boldsymbol{\psi}). \tag{2}$$

2.2. Content Generating Process

The words $\boldsymbol{w}_d$ are generated according to the cluster-based topic model (**?**), an extension of a well-known Bayesian topic model, latent Dirichlet allocation (LDA) (**?**). As in LDA, we generate the corpus-wide global variables that describe the content via topics. First, we model each topic $k \in [K]$ as a discrete distribution over V unique word types

$$\boldsymbol{\phi}_k \sim \text{Dirichlet}\Big(\beta, (\frac{1}{V}, \ldots, \frac{1}{V})\Big), \tag{3}$$

where β is the concentration parameter. Next, following the cluster-based topic model, document d has the document-topic distribution

$$\boldsymbol{\theta}_d \sim \text{Dirichlet}(\alpha, \boldsymbol{m}_{c_d}), \tag{4}$$

where α are the concentration parameter and $\boldsymbol{m} = (m_1, \ldots, m_K)$ is the base measure. In order to capture the overall prevalence of each topic in the corpus, we assume that each $\boldsymbol{m}_c$ is given Dirichlet priors with a single corpus-level base measure $\boldsymbol{m}$

$$\boldsymbol{m}_c \sim \text{Dirichlet}\Big(\alpha_1, \boldsymbol{m}\Big), \tag{5}$$

where α_1 is the concentration parameter determining the extent to which the group-specific base measures are affected by the corpus-level base measure. Finally, the corpus-level base measure is assumed to have Dirichlet prior with uniform base measure

$$\boldsymbol{m} \sim \text{Dirichlet}\Big(\alpha_0, (\frac{1}{K}, \ldots, \frac{1}{K})\Big). \tag{6}$$

Given that $\bar{N}_d = \max(1, N_d)$ where N_d is known, a topic z_{dn} is drawn from the document-topic distribution and then a word w_{dn} is drawn from the chosen topic for each $n \in [\bar{N}_d]$—i.e.,

$$\begin{aligned} z_{dn} &\sim \text{Multinomial}(\boldsymbol{\theta}_d), \\ w_{dn} &\sim \text{Multinomial}(\phi_{z_{dn}}). \end{aligned} \tag{7}$$

Pseudocode for content generating process is provided in the supplementary material.

2.3. Tie Generating Process

We generate ties—author a_d, recipients $\boldsymbol{r}_d$, and timestamp t_d—using a continuous-time process that depends on the interaction patterns' various features. Conditioned on the document-specific interaction pattern (Seciton **??**), we assume the following steps of tie generating process. Much like in the SAOM (**?**), we conceptualize tie generation as a process that is governed by senders acting in continuous time.

2.3.1. Latent Recipients

For every possible author–recipient pair $(a, r)_{a \neq r}$, we define the "recipient intensity", which is the likelihood of document d being sent from a to r:

$$\lambda_{adr} = {\boldsymbol{b}_{c_d}}^{\top} \boldsymbol{x}_{adrc_d}, \tag{8}$$

where $\boldsymbol{b}_c$ is P–dimensional vector of coefficients and $\boldsymbol{x}_{adrc}$ is a set of network features which vary depending on the hypotheses regarding canonical processes relevant to network theory such as popularity, reciprocity, and transitivity. We place a Normal prior $\boldsymbol{b}_c \sim N(\boldsymbol{\mu}_b, \Sigma_b)$.

In the example of email networks, we form the covariate vector for recipients $\boldsymbol{x}_{adrc}$ using dynamic network statistics on three time intervals prior to t^+_{d-1} (i.e., immediately after the previous document was sent). We compute eight network statistics within each time interval (**?**), where the intervals are $[t^+_{d-1} - 384h, t^+_{d-1} - 96h)$, $[t^+_{d-1} - 96h, t^+_{d-1} - 24h)$ and $[t^+_{d-1} - 24h, t^+_{d-1})$. We define the intervals to have equal length in the log-scale, and use $i = 1$ to denote the earliest interval—i.e., $[t^+_{d-1} - 384h, t^+_{d-1} - 96h)$—and i = 3 to denote the latest. The network statistics (illustrated in Figure **??**) are:

1. $\text{outdegree}(a, c, i) = \sum\limits_{d': t_{d'} \in i} I(c_{d'} = c) I(a_{d'} = a)$;
2. $\text{indegree}(r, c, i) = \sum\limits_{d': t_{d'} \in i} I(c_{d'} = c) I(u_{d'r} = 1)$;
3. $\text{send}(a, r, c, i) = \sum\limits_{d': t_{d'} \in i} I(c_{d'} = c) I(a_{d'} = a) I(u_{d'r} = 1)$;

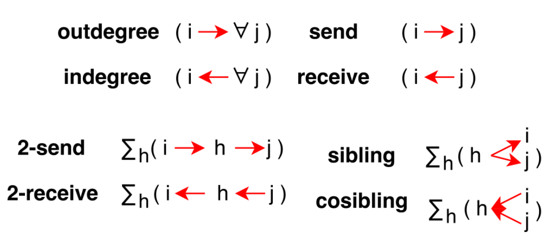

Figure 2. Eight dynamic network statistics used for the application to email networks.

4. $\text{receive}(a, r, c, i) = \text{send}(r, a, c, i)$;
5. $\text{2-send}(a, r, c, i) = \sum\limits_{\substack{i', i'' \geq i: \\ i' = i \text{ or } i'' = i}} \sum\limits_{h \neq a, r} \text{send}(a, h, c, i') \text{send}(h, r, c, i'')$;
6. $\text{2-receive}(a, r, c, i) = \sum\limits_{\substack{i', i'' \geq i: \\ i' = i \text{ or } i'' = i}} \sum\limits_{h \neq a, r} \text{send}(h, a, c, i') \text{send}(r, h, c, i'')$;
6. $\text{sibling}(a, r, c, i) = \sum\limits_{\substack{i', i'' \geq i: \\ i' = i \text{ or } i'' = i}} \sum\limits_{h \neq a, r} \text{send}(h, a, c, i') \text{send}(h, r, c, i'')$;
6. $\text{cosibling}(a, r, c, i) = \sum\limits_{\substack{i', i'' \geq i: \\ i' = i \text{ or } i'' = i}} \sum\limits_{h \neq a, r} \text{send}(a, h, c, i') \text{send}(r, h, c, i'')$;

where $I(\cdot)$ is an indicator function. Note that in order to obtain two-path statistics (i.e., 2-send, 2-receive, sibling, and cosibling) within a single time interval i, we compute the number of two-paths from a to r in interaction pattern c by summing over the pairs of intervals (i', i'') where the earlier email in the path was sent during interval i.

Next, we hypothesize "If a were the author of document d, who would be the recipent/recipients?" To do this, we draw each author's set of recipients from a non-empty Gibbs measure (**?**)—a probability measure we defined in order to 1) allow multiple recipients or "multicast", 2) prevent from obtaining zero recipient, and 3) ensure tractable normalizing constant.

Because the IPTM allows multicast, we draw a binary (0/1) vector $\boldsymbol{u}_{ad} = (u_{ad1}, \ldots, u_{adA})$

$$\boldsymbol{u}_{ad} \sim \text{Gibbs}(\delta, \boldsymbol{\lambda}_{ad}), \tag{9}$$

where δ is a real number controlling the average number of recipients and $\boldsymbol{\lambda}_{id} = \{\lambda_{adr}\}_{r=1}^{A}$. We place a Normal prior $\delta \sim N(\mu_\delta, \sigma^2_\delta)$. In particular, we define $\text{Gibbs}(\delta, \boldsymbol{\lambda}_{ad})$ as

$$p(\boldsymbol{u}_{ad} | \delta, \boldsymbol{\lambda}_{ad}) = \frac{\exp\Big\{\log\big(\mathrm{I}(\|\boldsymbol{u}_{ad}\|_1 > 0)\big) + \sum_{r \neq a} (\delta + \lambda_{adr}) u_{adr}\Big\}}{Z(\delta, \boldsymbol{\lambda}_{ad})}, \tag{10}$$

where $Z(\delta, \boldsymbol{\lambda}_{ad}) = \prod_{r \neq a} (\exp\{\delta + \lambda_{adr}\} + 1) - 1$ is the normalizing constant and $\|\cdot\|_1$ is the l_1–norm. We provide the derivation of the normalizing constant as a tractable form in the supplementary material.

2.3.2. Latent Timestamps

Similarly, we hypothesize "If a were the author of document d, when would it be sent?" and define the "timing rate" for author i

$$\mu_{ad} = g^{-1}(\boldsymbol{\eta}_{c_d}^{\top} \boldsymbol{y}_{adc_d}), \tag{11}$$

where $\boldsymbol{\eta}_c$ is Q–dimensional vector of coefficients with a Normal prior $\boldsymbol{\eta}_c \sim N(\boldsymbol{\mu}_\eta, \Sigma_\eta)$, $\boldsymbol{y}_{adc}$ is a set of time-related covariates, which can be any feature that could affect timestamps of the document, and $g(\cdot)$ is the appropriate link function such as identity, log, or inverse.

For example, the covariate vector for timestamps $\boldsymbol{y}_{adc}$ can include author-specific intercepts to account for individual differences in document-sending behavior. In addition, temporal features which possibly affect "when to send" can be added—e.g., an indicator of weekends/weekdays and an indicator of AM/PM when the previous document was sent.

In modeling "when", we do not directly model the timestamp t_d. Instead, we assume that each author's the time-increment or "time to next document" (i.e., $\tau_d = t_d - t_{d-1}$) is drawn from a specific distribution in the exponential family. We follow the generalized linear model framework:

$$\begin{aligned} E(\tau_{ad}) &= \mu_{ad}, \\ V(\tau_{ad}) &= V(\mu_{ad}), \end{aligned} \tag{12}$$

where τ_{ad} is a positive real number. Possible choices of distribution include Exponential, Weibull, Gamma, and lognormal[1] distributions, which are commonly used in time-to-event modeling. Based on the choice of distribution, we may introduce any additional parameter (e.g., σ_τ^2) to account for the variance.

Our preliminary analysis revealed that the Dare County email networks and the Enron data set showed the best fitting when we assume lognormal distribution on the observed time-increments—i.e., $\log(\tau_{a_d d}) \sim N(\mu_{a_d d}, \sigma_\tau^2)$—compared to Gamma or Weibull distributions. We also observed significant lack-of-fit for single parameter distribution (e.g., Exponential distribution) since it failed to capture the variance in time-increments. Therefore, we chose lognormal distribution by taking the log-transformation and apply $\mu = E(\log(\tau_{ad})) = \mu_{ad}$ and $\sigma_\tau^2 = V(\log(\tau_{ad})) = V(\mu_{ad})$, using identity link function $g = I$.

2.3.3. LATENT RECIPIENTS (SPARSE)

For every possible author–recipient pair $(a, r)_{a \neq r}$, we define the "recipient intensity", which is the likelihood of document d being sent from a to r:

$$\lambda_{adr} = \boldsymbol{b}_{c_d}{}^\top \boldsymbol{x}_{adrc_d}, \tag{13}$$

where $\boldsymbol{b}_c$ is P–dimensional vector of coefficients and $\boldsymbol{x}_{adrc}$ is a set of network features which vary depending on the hypotheses regarding canonical processes relevant to network theory such as popularity, reciprocity, and transitivity. We place a Normal prior $\boldsymbol{b}_c \sim N(\boldsymbol{\mu}_b, \Sigma_b)$.

[1] lognormal distribution is not exponential family but can be used via modeling of $\log(\tau_d)$.

In the example of email networks, we form the covariate vector for recipients $\boldsymbol{x}_{adrc}$ using dynamic network statistics on three time intervals prior to t_{d-1}^+ (i.e., immediately after the previous document was sent). We compute eight network statistics within each time interval (**?**), where the intervals are $[t_{d-1}^+ - 384h, t_{d-1}^+ - 96h)$, $[t_{d-1}^+ - 96h, t_{d-1}^+ - 24h)$ and $[t_{d-1}^+ - 24h, t_{d-1}^+)$. We define the intervals to have equal length in the log-scale, and use $i = 1$ to denote the earliest interval—i.e., $[t_{d-1}^+ - 384h, t_{d-1}^+ - 96h)$—and i = 3 to denote the latest. The network statistics (illustrated in Figure **??**) are:

1. outdegree$(a, c, i) = \sum_{d': t_{d'} \in i} I(c_{d'} = c) I(a_{d'} = a)$;
2. indegree$(r, c, i) = \sum_{d': t_{d'} \in i} I(c_{d'} = c) I(u_{d'r} = 1)$;
3. send(a, r, c, i) $= \sum_{d': t_{d'} \in i} I(c_{d'} = c) I(a_{d'} = a) I(u_{d'r} = 1)$;
4. receive$(a, r, c, i) =$ send(r, a, c, i);
5. 2-send(a, r, c, i) $= \sum_{\substack{i', i'' \geq i: \\ i' = i \text{ or } i'' = i}} \sum_{h \neq a, r}$ send(a, h, c, i')send(h, r, c, i'');
6. 2-receive(a, r, c, i) $= \sum_{\substack{i', i'' \geq i: \\ i' = i \text{ or } i'' = i}} \sum_{h \neq a, r}$ send(h, a, c, i')send(r, h, c, i'');
6. sibling(a, r, c, i) $= \sum_{\substack{i', i'' \geq i: \\ i' = i \text{ or } i'' = i}} \sum_{h \neq a, r}$ send(h, a, c, i')send(h, r, c, i'');
6. cosibling(a, r, c, i) $= \sum_{\substack{i', i'' \geq i: \\ i' = i \text{ or } i'' = i}} \sum_{h \neq a, r}$ send(a, h, c, i')send(r, h, c, i'');

where $I(\cdot)$ is an indicator function. Note that in order to obtain two-path statistics (i.e., 2-send, 2-receive, sibling, and cosibling) within a single time interval i, we compute the number of two-paths from a to r in interaction pattern

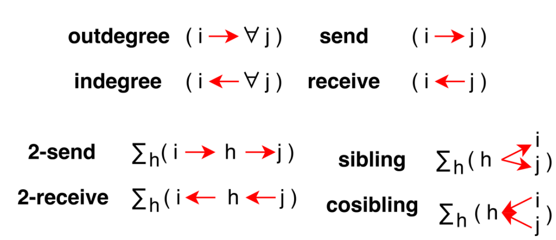

Figure 3. Eight dynamic network statistics used for the application to email networks.

c by summing over the pairs of intervals (i', i'') where the earlier email in the path was sent during interval i.

Next, we consider "If a were the author of document d, who would be the recipent/recipients?" To generate the prospective recipients for author a of document d, we distinguish between nodes that are "close" to a in the network of recent interactions ($\boldsymbol{r}_{ad}^{+}$) and those who are distant from a ($\boldsymbol{r}_{ad}^{-}$) in the network of recent interactions. If r is among $\boldsymbol{r}_{ad}^{+}$, we assume that the author considers the network structure of recent interactions (i.e., λ_{adr}) in determining whether to add r to the list of recipients. If r is among $\boldsymbol{r}_{ad}^{-}$, we assume that r is added to the list of recipients with low probability, and independent of the network history. We have two purposes in acknowledging this distinction. First, it reflects cognitive limitations of the author—that there are limits to the size of the network that an author can react to in formulating a list of recipients. Second, this distinction reduces the computational burden associated with accounting for network structure when drawing inferences regarding which recipients would have received an email if it were authored by someone other than the observed recipient. The distinction between close and distant prospective recipients should be formulated on an application-specific basis. One natural option is to consider all nodes within h hops of r in the network constructed from interactions within l time of document d as being in $\boldsymbol{r}_{ad}^{+}$ and all nodes outside of h hops of r in the network constructed from interactions within l time of document d as being in $\boldsymbol{r}_{ad}^{-}$.

Author a selects a set of recipients for document d by combining two sets selected according to two different probability distributions. The first set is selected by taking an independent Bernoulli draw corresponding to each recipient in $\boldsymbol{r}_{ad}^{-}$, and adding any recipients to the set for whom the Bernoulli draw resulted in success. The second set is drawn by selecting recipients from $\boldsymbol{r}_{ad}^{+}$ using either an empty or non-empty Gibbs measure (**?**)—a probability measure we defined in order to 1) allow multiple recipients or "multicast", 2) prevent from obtaining zero recipient, and 3) ensure tractable normalizing constant. The empty Gibbs measure includes the event of selecting no recipients from $\boldsymbol{r}_{ad}^{+}$. The non-empty Gibbs measure excludes the event of selecting no recipients from $\boldsymbol{r}_{ad}^{+}$. The non-empty Gibbs measure is used if and only if no recipients are selected from $\boldsymbol{r}_{ad}^{-}$.

Because the IPTM allows multicast, we draw a binary (0/1) vector $\boldsymbol{u}_{ad} = (u_{ad1}, \dots, u_{adA})$

$$\boldsymbol{u}_{ad} \sim \text{Gibbs}(\delta, \boldsymbol{\lambda}_{ad}), \tag{14}$$

where δ is a real number controlling the average number of recipients and $\boldsymbol{\lambda}_{id} = \{\lambda_{adr}\}_{r=1}^{A}$. We place a Normal prior $\delta \sim N(\mu_\delta, \sigma_\delta^2)$. In particular, we define Gibbs$(\delta, \boldsymbol{\lambda}_{ad})$ as

$$\begin{aligned} &p(\boldsymbol{u}_{ad}|\delta, \boldsymbol{\lambda}_{ad}) \\ &= \frac{\exp\left\{\log\big(\mathrm{I}(\|\boldsymbol{u}_{ad}\|_1 > 0)\big) + \sum_{r\neq a}(\delta + \lambda_{adr})u_{adr}\right\}}{Z(\delta, \boldsymbol{\lambda}_{ad})}, \end{aligned} \tag{15}$$

where $Z(\delta, \boldsymbol{\lambda}_{ad}) = \prod_{r\neq a}(\exp\{\delta + \lambda_{adr}\} + 1) - 1$ is the normalizing constant and $\|\cdot\|_1$ is the l_1–norm. We provide the derivation of the normalizing constant as a tractable form in the supplementary material.

2.3.4. Latent Timestamps

Similarly, we hypothesize "If a were the author of document d, when would it be sent?" and define the "timing rate" for author i

$$\mu_{ad} = g^{-1}(\boldsymbol{\eta}_{c_d}^{\top}\boldsymbol{y}_{adc_d}), \tag{16}$$

where $\boldsymbol{\eta}_c$ is Q–dimensional vector of coefficients with a Normal prior $\boldsymbol{\eta}_c \sim N(\boldsymbol{\mu}_\eta, \Sigma_\eta)$, $\boldsymbol{y}_{adc}$ is a set of time-related covariates, which can be any feature that could affect timestamps of the document, and $g(\cdot)$ is the appropriate link function such as identity, log, or inverse.

For example, the covariate vector for timestamps $\boldsymbol{y}_{adc}$ can include author-specific intercepts to account for individual differences in document-sending behavior. In addition, temporal features which possibly affect "when to send" can be added—e.g., an indicator of weekends/weekdays and an indicator of AM/PM when the previous document was sent.

In modeling "when", we do not directly model the timestamp t_d. Instead, we assume that each author's the time-increment or "time to next document" (i.e., $\tau_d = t_d - t_{d-1}$) is drawn from a specific distribution in the exponential family. We follow the generalized linear model framework:

$$\begin{aligned} E(\tau_{ad}) &= \mu_{ad}, \\ V(\tau_{ad}) &= V(\mu_{ad}), \end{aligned} \tag{17}$$

where τ_{ad} is a positive real number. Possible choices of distribution include Exponential, Weibull, Gamma, and lognormal[2] distributions, which are commonly used in time-to-event modeling. Based on the choice of distribution, we may introduce any additional parameter (e.g., σ_τ^2) to account for the variance.

Our preliminary analysis revealed that the Dare County email networks and the Enron data set showed the best fitting when we assume lognormal distribution on the observed time-increments—i.e., $\log(\tau_{a_d d}) \sim N(\mu_{a_d d}, \sigma_\tau^2)$—compared to Gamma or Weibull distributions. We also observed significant lack-of-fit for single parameter distribution (e.g., Exponential distribution) since it failed to capture

[2]lognormal distribution is not exponential family but can be used via modeling of $\log(\tau_d)$.

the variance in time-increments. Therefore, we chose log-normal distribution by taking the log-transformation and apply $\mu = E(\log(\tau_{ad})) = \mu_{ad}$ and $\sigma_\tau^2 = V(\log(\tau_{ad})) = V(\mu_{ad})$, using identity link function $g = I$.

2.3.5. Actual Data

Finally, we choose the actual author, recipients, and timestamp—which will be observed—by selecting the author–recipient-set pair with the smallest time-increment (**?**):

$$\begin{aligned} a_d &= \text{argmin}_a(\tau_{ad}), \\ \boldsymbol{r}_d &= \boldsymbol{u}_{a_d d}, \\ t_d &= t_{d-1} + \tau_{a_d d}. \end{aligned} \tag{18}$$

Therefore, it is an author-driven process in that the author of a document determines its recipients and its timestamp, based on the author's urgency to send the document to chosen recipients.

3. Posterior Inference

Given that we only observe the authors, recipients, timestamps, and tokens $\{(a_d, \boldsymbol{r}_d, t_d, \boldsymbol{w}_d)\}_{d=1}^D$ in real-world, our inference goal is to invert the generative process to obtain the posterior distribution over the unknown parameters, conditioned on the observed data and hyperparamters $\alpha_0, \alpha_1, \alpha, \beta, \zeta, \boldsymbol{\mu}_b, \Sigma_b, \boldsymbol{\mu}_\eta, \Sigma_\eta, \mu_\delta, \sigma_\delta^2$. After integrating out Φ, Ψ, and Θ using Dirichlet-multinomial conjugacy (**?**), we draw the samples using Markov chain Monte Carlo (MCMC) methods, repeatedly resampling the value of each parameter from its conditional posterior given the observed data, hyperparamters, and the current values of the other parameters. We express each parameters conditional posterior in a closed form using the data augmentation schemes in $\boldsymbol{u}$ (**?**). In this section, we outline a Metropolis-within-Gibbs sampling algorithm and each latent variable's conditional posterior. Pseudocode is provided in the supplementary material.

First, since u_{adr} is a binary random variable, new values may be sampled directly using

$$\begin{aligned} &P(u_{adr} = 1 | \boldsymbol{u}_{ad\setminus r}, \boldsymbol{c}, \boldsymbol{b}, \delta, \boldsymbol{x}) \propto \exp\{\delta + \lambda_{adr}\}; \\ &P(u_{adr} = 0 | \boldsymbol{u}_{ad\setminus r}, \boldsymbol{c}, \boldsymbol{b}, \delta, \boldsymbol{x}) \propto \mathrm{I}(\|\boldsymbol{u}_{ad\setminus r}\|_1 > 0), \end{aligned} \tag{19}$$

where $I(\cdot)$ is the indicator function that is used to prevent from the instances where the author has no recipients to send the document.

Next, the conditional posterior for topic assignment z_{dn} is derived by multiplying the two sampling equations of the cluster-based topic model:

$$\begin{aligned} &p(z_{dn} = k | \boldsymbol{z}_{\setminus dn}, \boldsymbol{c}, \boldsymbol{w}, \alpha_0, \alpha, \beta) \\ &\propto \Big(\hat{N}_{dk,\setminus dn} + \alpha \frac{\hat{N}_{kc_d,\setminus dn} + \alpha_1 \frac{\hat{N}_{k,\setminus dn} + \frac{\alpha_0}{K}}{\hat{N}_{\setminus dn} + \alpha_0}}{\hat{N}_{c_d,\setminus dn} + \alpha_1}\Big) \\ &\quad \times \Big(\frac{\hat{N}_{w_{dn}k,\setminus dn} + \frac{\beta}{V}}{\hat{N}_{k,\setminus dn} + \beta}\Big), \end{aligned} \tag{20}$$

where $\hat{N}$ are defined according to the minimal path assumption (**?**). Specifically, $\hat{N}_{dk,\setminus dn}$ is the number of times topic k has been used in document d, $\hat{N}_{kc_d,\setminus dn}$ is the number of different documents belonging to c_d that use topic k, $\hat{N}_{k,\setminus dn}$ is the number of different interaction patterns in which k has been used, $\hat{N}_{w_{dn}k,\setminus dn}$ is the number of tokens assigned to topic k whose type is the same as that of w_{dn}, and the subscript $\setminus dn$ denotes the exclusion of n^{th} element in document d.

For each document $d \in [D]$, we sample interaction-pattern assignment from the discrete distribution over C interaction patterns using

$$\begin{aligned} &P(c_d = c | \boldsymbol{z}, \zeta, \boldsymbol{u}, \boldsymbol{a}, \boldsymbol{t}) \\ &\propto P(c_d = c | \boldsymbol{c}_{\setminus d}, \zeta) P(\boldsymbol{z}_d | \gamma, \alpha, c_d = c, \boldsymbol{c}_{\setminus d}, \boldsymbol{z}_{\setminus d}) \\ &\times P(a_d, t_d | c_d = c, \boldsymbol{\eta}, \sigma_\tau^2) P(\boldsymbol{u} | c_d = c, \boldsymbol{c}_{\setminus d}, \boldsymbol{b}, \delta) \\ &\propto (\hat{N}_{c,\setminus d} + \frac{\zeta}{C}) \\ &\times \prod_{n=1}^{\bar{N}_d} (\hat{N}_{dz_{dn},\setminus dn} + \alpha \frac{\hat{N}_{z_{dn}c,\setminus dn} + \alpha_1 \frac{\hat{N}_{z_{dn},\setminus dn} + \frac{\alpha_0}{K}}{\hat{N}_{\setminus dn} + \alpha_0}}{\hat{N}_{c,\setminus dn} + \alpha_1}) \\ &\times \varphi_\tau(\tau_d; \mu_{a_d d}, \sigma_\tau^2) \times \prod_{a \neq a_d} \big(1 - \Phi_\tau(\tau_d; \mu_{ad}, \sigma_\tau^2)\big) \\ &\times \prod_{a=1}^A \frac{\exp\Big\{\log\big(\mathrm{I}(\|\boldsymbol{u}_{ad}\|_1 > 0)\big) + \sum_{r \neq a} (\delta + \lambda_{adr}) u_{adr}\Big\}}{Z(\delta, \boldsymbol{\lambda}_{ad})}. \end{aligned} \tag{21}$$

New values for continuous variables $\delta, \boldsymbol{b}$, and $\boldsymbol{\eta}$ and σ_τ^2 (if applicable) cannot be sampled directly from their conditional posteriors, but may instead be obtained using the Metropolis–Hastings algorithm. With uninformative priors (i.e., $N(0, \infty)$), the conditional posterior over δ and $\boldsymbol{b}$ is

$$\prod_{d=1}^D \prod_{a=1}^A \frac{\exp\Big\{\log\big(\mathrm{I}(\|\boldsymbol{u}_{ad}\|_1 > 0)\big) + \sum_{r \neq a} (\delta + \lambda_{adr}) u_{adr}\Big\}}{Z(\delta, \boldsymbol{\lambda}_{ad})}, \tag{22}$$

where the two variables share the conditional posterior and thus can be jointly sampled. Likewise, assuming uninformative priors on $\boldsymbol{\eta}$ (i.e., $N(0, \infty)$) and σ_τ^2 (i.e., half-

Cauchy(∞)), the conditional posterior is

$$\prod_{d=1}^{D}\Big(\varphi_\tau(\tau_d;\mu_{a_d d},\sigma_\tau^2)\times\prod_{a\neq a_d}\big(1-\Phi_\tau(\tau_d;\mu_{ad},\sigma_\tau^2)\big)\Big). \tag{23}$$

Although the IPTM is a highly complex model with a lot of latent variables, it yields an efficient inference algorithm by taking advantage of the two main parts of the likelihood repeatedly appear in the sampling equations—one from the latent recipients (Section **??**) and another from the latent timestamps (Section **??**).

4. Data

Our data come from the North Carolina county government email dataset collected by (**?**) that includes internal email corpora covering the inboxes and outboxes of managerial-level employees of North Carolina county governments. Out of over twenty counties, we chose Dare County to 1) see whether and how communication networks surrounding a notable national emergency—Hurricane Sandy—differed from those surrounding other governmental functions, and 2) limit the scope of this initial application. The Dare County email network contains 2,247 emails, sent and received by 27 department managers over a period of 3 months (September–November) in 2012.

To verify that our model is applicable beyond the Dare County email network, we also performed two validation experiments using the Enron data set (**?**). We took a subset of the original data such that we only include emails between actors who sent over 300 emails, and actors who received over 300 emails from the chosen authors. Emails that were not sent to at least one other active actor were discarded, which resulted in a total of 6,613 emails involving 30 actors. For the Enron data set, we changed the time unit from hour to day in modeling the timestamps.

5. Experiments

We conducted a set of posterior predictive experiments—1) out-of-sample tie predictions, 2) topic coherence, and 3) posterior predictive checks—to gauge the IPTM's predictive performance as compared to alternative modeling approaches.

5.1. Out-of-Sample Tie Predictions

We evaluated the IPTM's ability to predict ties in textual communications from either the Dare County email network or the Enron data set, conditioned on the text of those emails and "training" part of the data. We separately formed a test split of each three components—author, recipients, and timestamps—by randomly selecting "test" data with probability $p = 0.1$. Any missing variables were imputed by drawing samples from their conditional posterior distributions, given the observed data, estimates of latent variables, and current estimates of test data. The full conditional posterior distributions for "test" author, recipients, and timestamps are provided in the supplementary material.

We then run inference to update the latent variables given the imputed and observed data. We iterate the two steps—imputation and inference—multiple times to obtain enough number of estimates for "test" data. Algorithm **??** outlines this procedure.

Algorithm 1 Out-of-Sample Tie Predictions

Input: data $\{(a_d, \boldsymbol{r}_d, t_d, \boldsymbol{w}_d)\}_{d=1}^{D}$,
number of new data to generate R,
number of interaction patterns and topics (C, K),
hyperparameters $(\alpha_0, \alpha_1, \alpha, \beta, \zeta, \boldsymbol{\mu}_b, \Sigma_b, \boldsymbol{\mu}_\eta, \Sigma_\eta, \mu_\delta, \sigma^2_\delta)$

Test splits:
Draw test authors with $p = 0.1$ (out of D authors)
Draw test elements of recipient vector with $p = 0.1$ (out of $D \times (A-1)$ receipient indicators $\{\{\boldsymbol{r}_{dr}\}_{r \in [A]_{\backslash a_d}}\}_{d=1}^{D}$)
Draw test timestamps with $p = 0.1$ (out of D timestamps)
Set the "test" data as "missing" (NA)

Imputation and inference:
Initialize the parameters $(\boldsymbol{l}, \boldsymbol{z}, \boldsymbol{b}, \boldsymbol{\eta}, \delta, \boldsymbol{u})$
for $r = 1$ **to** R **do**
  for $d = 1$ **to** D **do**
    if $a_d =$ NA **then**
      for $a = 1$ **to** A **do**
        Compute π_a using $P(a_d = a|\cdot)$
      end for
      Draw $a_d \sim$ Multinomial(π)
    end if
    for $r \in [A]_{\backslash a_d}$ **do**
      if $r_{dr} =$ NA **then**
        Draw r_{dr} using $P(r_{dr} = 1|\cdot)$ and $P(r_{dr} = 0|\cdot)$
      end if
    end for
    if $t_d =$ NA **then**
      Draw proposal $\boldsymbol{\tau}_d^{new} \sim$ lognormal$(\mu_{a_d d}, \sigma^2_\tau)$
      Use Metropolis-Hastings to decide accept or reject using the probability

$$\frac{P(\boldsymbol{\tau}_d^{new}|\mu_{a_d d}, \sigma^2_\tau) P(\boldsymbol{\tau}_d^{new}|\cdot)}{P(\boldsymbol{\tau}_d^{old}|\mu_{a_d d}, \sigma^2_\tau) P(\boldsymbol{\tau}_d^{old}|\cdot)},$$

      where $\boldsymbol{\tau}_d^{old}$ is from earlier iteration.
    end if
    Run inference and update $(\boldsymbol{c}, \boldsymbol{z}, \boldsymbol{b}, \boldsymbol{\eta}, \delta, \boldsymbol{u})$ given the imputed and observed data
  end for
  Store the estimates for "test" data
end for

We compared the IPTM's performance with that of baseline—the IPTM with $C = 1$. This amounts to an ablation study (**??**), as a single interaction pattern breaks the link between text and network structure in the IPTM. The text and network structure are linked through the assignment of topics to different interaction patterns, and with one interaction pattern all topics are associated with the same network structure. We do not define any other baselines (i.e., other models'test" a fr machine learning literature) to which to compare the predictive performance of the IPTM. We omit comparison to baselines because we are unable to identify existing models that can predict the same form of social data that can be modeled by the IPTM—a form that includes one out of n authors, one through $n-1$ recipients, and a continuous and positive time point. Consider the prediction of e-mail recipients. As far as we are aware, the Gibbs measure model we derive is unique among existing methods in its ability to predict a set of one through $n-1$ (out of $n-1$) recipients of an e-mail. This is just the recipient component of the model—we are also not able to identify any other method that permits the prediction of the author, recipient multicast, and timing of ties. We could construct baseline models to compare in terms of predictive performance for each component of the social data (e.g., a regression model to predict e-mail timing, a multi-class classifier to predict author). However, that would be an arbitrary exercise, as it is not clear why we would select any particular baseline out of the dozens of candidates for each component of the social data modeled in the IPTM.

We varied the number of interaction patterns C from 1 to 3 and the number of topics K from 1 to 50 (Dare) or 100 (Enron) as a grid-search based hyperparameter selection process. For each combinations of C and K, predicted values of tie data were then compared to the true values to yield: F_1 scores for author predictions, multiclass version of the area under the ROC curve (AUC) measure (**?**) for reciptient predictions, and median absolute error (MAE) on timestamp predictions. We show the tie prediction results, averaged over five random test splits of each tie component, in Figure **??** (Plots to be updated). Although our model is intended for exploratory analysis, it achieves better link prediction performance than the baseline, validating our assumption that the IPTM acheives better predictive performance when topic-based contents are accounted to infer the parameters that govern the generation of tie data—authors, recipients, and timestamps.

5.2. Topic Coherence

Topic coherence metrics (**?**) are often used to evaluate the semantic coherence in topic models.To demonstrate that the IPTM's incorporation of network features improves the ability of modeling text, we compared the coherence of topics inferred using our model with the coherence of topics inferred using LDA. Instead of re-fitting the data using standard LDA algorithms, we used the topic assignments from the IPTM with $C = 1$, which reduces the IPTM to LDA in terms of topic assignments. We varied the number of interaction patterns and the number of topics as in Section **??**, and drew five samples from the joint posterior distribution over the latent variables. We evaluated the topics resulting from each sample and averaged over the five samples, where the results are shown in Figure **??**. Combined with the findings in Section **??**, this result demonstrates that the IPTM

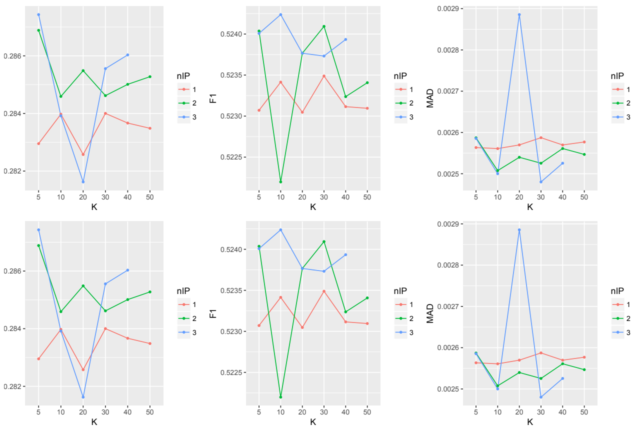

Figure 4. Average F1 score, AUC, MAE of out-of-sample tie predictions. *Top*: Dare County email network. *Bottom*: the Enron dataset.

can achieve good predictive performance while producing coherent topics.

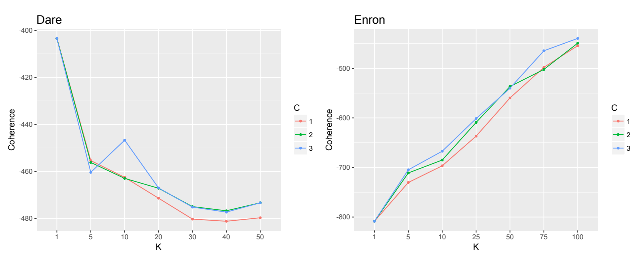

Figure 5. Average topic coherence scores: (*left*) Dare County email network. (*right*) the Enron data set.

5.3. Posterior Predictive Checks

Finally, we performed posterior predictive checks (**?**) to evaluate the appropriateness of the model specification for the Dare County email network. We formally generated entirely new data, by simulating ties and contents $\{(a_d, \boldsymbol{r}_d, t_d, \boldsymbol{w}_d)\}_{d=1}^{D}$ from the genenerative process in Section **??**, conditional upon a set of inferred parameter values from the inference in Section **??**. Pseudocode is provided in the supplementary material. We specified the number of interaction patterns as $C = ?$ and the number of topics as $K = ?$, which yielded the best performance in Section **??**. For the test of goodness-of-fit in terms of network dynamics, we defined multiple network statistics that summarize meaningful aspects of the Dare County email network: indegree distribution for author activities, outdegree distribution for recipient activities, recipient size distribution, document time-increments distribution, the edgewise shared partner distribution, and the geodesic distance distribution. We then generated 100 synthetic networks and texts from the posterior predictive distribution implied by the IPTM and Dare County email network. We applied each discrepancy function to each synthetic network to yield the distributions over the values of the six network statistics

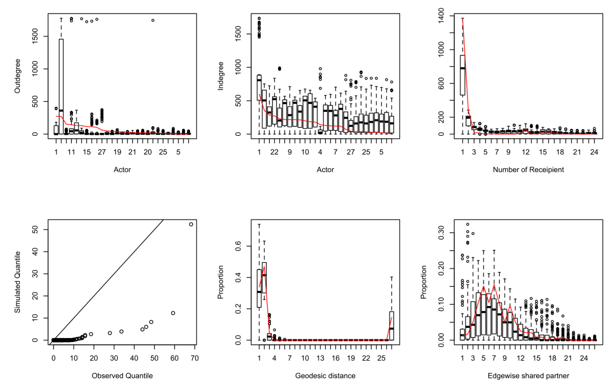

Figure 6. Posterior predictive checks for the Dare County email network: (a) outdegree, (b) indegree, (c) recipient size, (d) QQplot of time-increments, (e) geodesic distance, and (f) edgewise shared partners.

As shown in Figure **??** (Plots to be updated), the IPTM shows "good fit" for the Dare County email network in that the observed data is not an outlier with respect to the distributions of new data drawn from the posterior predictive distribution. The IPTM generated synthetic networks with indegree distribution, outdegree distribution, recipient size, document time-increments, and edgewise shared partners that are very similar to those of the Dare County email network, showing that the model captures some important work features of the data including spreadness and transitivity.

6. Exploratory Analysis

Our model is primarily intended as an exploratory analysis tool for time-stamped textual communication. Our main goal in this exploratory analysis was to test three hypotheses: 1) personal or social topics (if any) would exhibit strong reciprocity and transitivity in tie formation, 2) topics about dissemination of information would be characterized by a lack of reciprocity, and 3) topics about Hurricane Sandy would exhibit a very different interaction pattern from the normal day-to-day conversations.

6.1. Topic Assignments

6.2. Interaction Pattern Coefficients

7. Summary

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

References

Alemán, Eduardo and Calvo, Ernesto. Explaining policy ties in presidential congresses: A network analysis of bill initiation data. *Political Studies*, 61(2):356–377, 2013.

ben Aaron, James, Denny, Matthew, Desmarais, Bruce, and Wallach, Hanna. Transparency by conformity: A field experiment evaluating openness in local governments. *Public Administration Review*, 77(1):68–77, 2017.

Bilgic, Mustafa, Mihalkova, Lilyana, and Getoor, Lise. Active learning for networked data. In *Proceedings of the 27th international conference on machine learning (ICML-10)*, pp. 79–86, 2010.

Blei, David M., Ng, Andrew Y., and Jordan, Michael I. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022, March 2003. ISSN 1532-4435.

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

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

Butts, Carter T. A relational event framework for social action. *Sociological Methodology*, 38(1):155–200, 2008. ISSN 1467-9531. doi: 10.1111/j.1467-9531.2008.00203.x.

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

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

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

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

Desmarais, Bruce A. and Cranmer, Skyler J. Statistical inference in political networks research. In Victor, Jennifer Nicoll, Montgomery, Alexander H., and Lubell, Mark (eds.), *The Oxford Handbook of Political Networks*. Oxford University Press, 2017.

Fahmy, Chantal and Young, Jacob TN. Gender inequality and knowledge production in criminology and criminal justice. *Journal of Criminal Justice Education*, pp. 1–21, 2016.

Fellows, Ian and Handcock, Mark. Removing phase transitions from gibbs measures. In *Artificial Intelligence and Statistics*, pp. 289–297, 2017.

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

Hand, David J and Till, Robert J. A simple generalisation of the area under the roc curve for multiple class classification problems. *Machine learning*, 45(2):171–186, 2001.

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

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

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

Klimt, Bryan and Yang, Yiming. Introducing the enron corpus. In *CEAS*, 2004.

Krafft, Peter, Moore, Juston, Desmarais, Bruce, and Wallach, Hanna M. Topic-partitioned multinetwork embeddings. In Pereira, F., Burges, C.J.C., Bottou, L., and Weinberger, K.Q. (eds.), *Advances in Neural Information*

Processing Systems 25, pp. 2807–2815. Curran Associates, Inc., 2012.

Kronegger, Luka, Mali, Franc, Ferligoj, Anuška, and Doreian, Patrick. Collaboration structures in slovenian scientific communities. *Scientometrics*, 90(2):631–647, 2011.

Lai, Chih-Hui, She, Bing, and Tao, Chen-Chao. Connecting the dots: A longitudinal observation of relief organizations' representational networks on social media. *Computers in Human Behavior*, 74:224–234, 2017.

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

Lim, Kar Wai, Chen, Changyou, and Buntine, Wray. Twitter-network topic model: A full bayesian treatment for social network and text modeling. In *NIPS2013 Topic Model workshop*, pp. 1–5, 2013.

McCallum, Andrew, Corrada-Emmanuel, Andrés, and Wang, Xuerui. The author-recipient-topic model for topic and role discovery in social networks, with application to enron and academic email. In *Workshop on Link Analysis, Counterterrorism and Security*, pp. 33, 2005.

Mimno, David, Wallach, Hanna M, Talley, Edmund, Leenders, Miriam, and McCallum, Andrew. Optimizing semantic coherence in topic models. In *Proceedings of the conference on empirical methods in natural language processing*, pp. 262–272. Association for Computational Linguistics, 2011.

Peng, Tai-Quan, Liu, Mengchen, Wu, Yingcai, and Liu, Shixia. Follower-followee network, communication networks, and vote agreement of the us members of congress. *Communication Research*, 43(7):996–1024, 2016.

Perry, Patrick O. and Wolfe, Patrick J. Point process modelling for directed interaction networks. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 75(5):821–849, 2013. ISSN 1467-9868. doi: 10.1111/rssb.12013.

Pew, Research Center. Social media fact sheet. *Accessed on 03/07/17*, 2016.

Richardson, Matthew, Prakash, Amit, and Brill, Eric. Beyond pagerank: machine learning for static ranking. In *Proceedings of the 15th international conference on World Wide Web*, pp. 707–715. ACM, 2006.

Robins, Garry, Pattison, Pip, Kalish, Yuval, and Lusher, Dean. An introduction to exponential random graph (p*) models for social networks. *Social networks*, 29(2):173–191, 2007.

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

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

Snijders, Tom AB, Van de Bunt, Gerhard G, and Steglich, Christian EG. Introduction to stochastic actor-based models for network dynamics. *Social networks*, 32(1):44–60, 2010.

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

Tanner, Martin A and Wong, Wing Hung. The calculation of posterior distributions by data augmentation. *Journal of the American statistical Association*, 82(398):528–540, 1987.

Wallach, Hanna Megan. *Structured topic models for language*. PhD thesis, University of Cambridge, 2008.

Yoon, Ho Young and Park, Han Woo. Strategies affecting twitter-based networking pattern of south korean politicians: social network analysis and exponential random graph model. *Quality & Quantity*, pp. 1–15, 2014.

8. Psuedocode for posterior predictive checks

Algorithm 2 Collapsed Content Generating Process

Input: number of new data to generate R,
observed text data $\{\boldsymbol{w}_d\}_{d=1}^{D}$,
estimated latent variables $(\boldsymbol{u}, \boldsymbol{l}, \boldsymbol{z}, \boldsymbol{b}, \delta, \boldsymbol{\eta}, \sigma_\tau^2)$,
hyperparameters $(\alpha, \beta, \boldsymbol{m})$,
number of vocabularies V

for $r = 1$ **to** R **do**
  Initialize N_{vk} and N_k from $\boldsymbol{z}$ and $\boldsymbol{w}$
  for $d = 1$ **to** D **do**
    if $N_d > 0$ **then**
      for $n = 1$ **to** N_d **do**
        Draw w_{dn} from $P(w_{dn} = v) = \frac{N_{vz_{dn}} + \frac{\beta}{V}}{N_{z_{dn}} + \beta}$
        Increment $N_{w_{dn}z_{dn}}$ and $N_{z_{dn}}$
      end for
    end if
    Compute $\boldsymbol{x}_d$ given $\{(a_d, \boldsymbol{r}_d, t_d)\}_{[1:(d-1)]}$
    Draw (a_d, $\boldsymbol{r}_d$, t_d) following Section 2.3
  end for
  Store every r^{th} new data $\{(a_d, \boldsymbol{r}_d, t_d, \boldsymbol{w}_d)\}_{d=1}^{D}$
end for