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

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

In defining and testing the IPTM we embed core conceptual property—interaction pattern—to link the content compo-

[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.

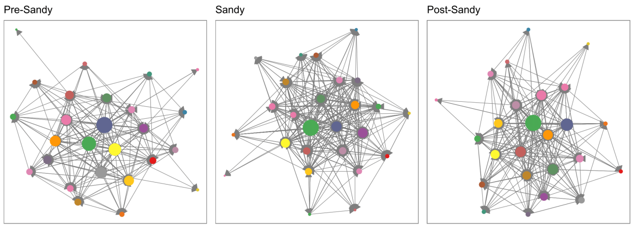

Figure 1. 95% credible intervals of the posterior estimates of $\{\boldsymbol{b}_c\}_{c=1}^{C}$ using Dare County data: (*left*) Recieve. (*right*) 2-Recieve.

nent 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 2). Figure 1 EDA plot should be added illustrates this structure. IPTM leads to an efficient MCMC inference algoritmIn (Section 3) and acheives good predictive peformance (Section 5). 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 6).

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 $\boldsymbol{r}_d = \{u_{dj}\}_{j=1}^{A}$, 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. Content Generating Process

The words $\boldsymbol{w}_d$ are generated according to latent Dirichlet allocation (LDA) (Blei et al., 2003), where we generate the corpus-wide global variables that describe the content via topics. LDA models each topic $k \in [K]$ as a discrete distribution over V unique word types

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

where β is the concentration parameter. Next, we assume a document- topic distribution over K topics

$$\boldsymbol{\theta}_d \sim \text{Dirichlet}(\alpha, \boldsymbol{m}), \tag{2}$$

where α is the concentration parameter and $\boldsymbol{m} = (m_1, \dots, m_K)$ is the probability vector. Given that 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 $n \in [N_d]$—i.e.,

$$\begin{aligned} z_{dn} &\sim \text{Multinomial}(\boldsymbol{\theta}_d), \\ w_{dn} &\sim \text{Multinomial}(\phi_{z_{dn}}). \end{aligned} \tag{3}$$

2.2. Interaction Patterns

They key idea that combines the IPTM component modeling "what" with the component modeling "who," "whom," and "when" is that different topics are associated with different 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 i to j in some time interval—and corresponding coefficients. We associate each topic with the interaction pattern that best describes how people interact when talking about that topic.

The topic-interaction pattern assignments are discrete-uniform distributed,

$$l_k \sim \text{Uniform}(1, C). \tag{4}$$

The content of each document is summarized as a distribution over interaction patterns:

$$\pi_{dc} = \frac{\sum_{k:l_k=c} N_{dk}}{N_d}, \tag{5}$$

where N_{dk} is the number of times topic k appears in document d. In other words, for each document and interaction pattern, we compute the fraction of tokens that were generated using a topic corresponding to that interaction pattern. We then use this to generate the tie components, which are discussed in the next section.

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 content (Section 2.1), we assume the following steps of tie generating process.

2.3.1. Latent Recipients

For every possible author–recipient pair $(i, j)_{i \neq j}$, we define the "interaction-pattern-specific recipient intensity":

$$\nu_{idjc} = {\boldsymbol{b}_c}^{\top} \boldsymbol{x}_{idjc}, \tag{6}$$

where $\boldsymbol{b}_c$ is P–dimensional vector of coefficients and $\boldsymbol{x}_{idjc}$ 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 $\boldsymbol{b}_c \sim N(\boldsymbol{\mu}_b, \Sigma_b)$.

We then compute the weighted average of $\{\nu_{idjc}\}_{c=1}^{C}$ and obtain the "recipient intensity"—the likelihood of document

d being sent from i to j— using the the document's distribution over interaction patterns as mixture weights:

$$\lambda_{idj} = \sum_{c=1}^{C} \pi_{dc}\, \nu_{idjc}. \tag{7}$$

Next, we hypothesize "If i 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 (Fellows & Handcock, 2017)—a probability measure we defined in order to 1) allow multiple recipients, 2) prevent from obtaining zero recipient, and 3) ensure tractable normalizing constant.

Because the IPTM allows multiple recipients, we draw a binary indicator vector $\boldsymbol{u}_{id} = (u_{id1}, \dots, u_{idA})$

$$\boldsymbol{u}_{id} \sim \text{Gibbs}(\delta, \boldsymbol{\lambda}_{id}), \tag{8}$$

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

$$p(\boldsymbol{u}_{id}|\delta, \boldsymbol{\lambda}_{id}) = \frac{\exp\Big\{\log\big(\mathrm{I}(\|\boldsymbol{u}_{id}\|_1 > 0)\big) + \sum_{j\neq i}(\delta + \lambda_{idj})u_{idj}\Big\}}{Z(\delta, \boldsymbol{\lambda}_{id})}, \tag{9}$$

where $Z(\delta, \boldsymbol{\lambda}_{id})$ 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 i were the author of document d, when would it be sent?" and define the "interaction-pattern-specific timing rate"

$$\xi_{idc} = \boldsymbol{\eta}_c^\top \boldsymbol{y}_{idc}, \tag{10}$$

where $\boldsymbol{\eta}_c$ is Q–dimensional vector of coefficients with a Normal prior $\boldsymbol{\eta}_c \sim N(\boldsymbol{\mu}_\eta, \Sigma_\eta)$, and $\boldsymbol{y}_{idc}$ is a set of time-related covariates, which can be any feature that could affect timestamps of the document. Some examples of the time-related covariates are described in Section 4.3.

The "timing rate" for author i is also computed from the weighted average of $\{\xi_{idc}\}_{c=1}^{C}$

$$\mu_{id} = \sum_{c=1}^{C} \pi_{dc} g^{-1}(\xi_{idc}), \tag{11}$$

where $g(\cdot)$ is the appropriate link function such as identity, log, or inverse.

In modeling "when", we do not directly model the timestamp t_d. Instead, we assume the time-increment $\tau_d = t_d - t_{d-1}$ is drawn from a specific distribution in the exponential family. We follow the generalized linear model framework:

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

where τ_{id} 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.

2.3.3. 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:

$$\begin{aligned} a_d &= \text{argmin}_i(\tau_{id}), \\ \boldsymbol{r}_d &= \boldsymbol{u}_{a_d d}, \\ t_d &= t_{d-1} + \tau_{a_d d}. \end{aligned} \tag{13}$$

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 $\alpha, \beta, \boldsymbol{m}, \boldsymbol{\mu}_b, \Sigma_b, \boldsymbol{\mu}_\eta, \Sigma_\eta, \mu_\delta, \sigma_\delta^2$. After integrating out Φ and Θ using Dirichlet-multinomial conjugacy (Griffiths & Steyvers, 2004), 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 $\boldsymbol{u}$ (Tanner & Wong, 1987). In this section, we outline a Metropolis-within-Gibbs sampling algorithm and each latent variable's conditional posterior.

Since u_{idj} is a binary random variable, new values may be sampled directly using

$$\begin{aligned} &P(u_{idj} = 1|\boldsymbol{u}_{id\backslash j}, \boldsymbol{z}, \boldsymbol{l}, \boldsymbol{b}, \delta, \boldsymbol{x}) \propto \exp\{\delta + \lambda_{idj}\} \\ &P(u_{idj} = 0|\boldsymbol{u}_{id\backslash j}, \boldsymbol{z}, \boldsymbol{l}, \boldsymbol{b}, \delta, \boldsymbol{x}) \propto \mathrm{I}(\|\boldsymbol{u}_{id\backslash j}\|_1 > 0), \end{aligned} \tag{14}$$

[1] lognormal is not exponential family but we can take the log-transformation and apply $\mu = E(\log(\tau_{id})) = \mu_{id}$ and $\sigma_\tau^2 = V(\log(\tau_{id})) = V(\mu_{id})$ using identity link function $g = I$.

which naturally prevent from the instances where the sender has no recipients to send the document.

The topic-interaction pattern assignment l_k is a discrete random variable and can be also directly sampled using

$$\begin{aligned}
&P(l_k = c|\boldsymbol{l}_{\backslash k}, \boldsymbol{z}, \boldsymbol{b}, \boldsymbol{\eta}, \delta, \boldsymbol{u}, \boldsymbol{a}, \boldsymbol{r}, \boldsymbol{t}, \boldsymbol{x}, \boldsymbol{y}) \\
&\propto \prod_{d=1}^{D} \Big(\prod_{i=1}^{A} \frac{\exp\Big\{\log\big(\mathrm{I}(\|\boldsymbol{u}_{id}\|_1 > 0)\big) + \sum_{j\neq i}(\delta + \lambda_{idj})u_{idj}\Big\}}{Z(\delta, \boldsymbol{\lambda}_{id})} \\
&\quad \times \varphi_\tau(\tau_{a_d d}; \mu_{a_d d}, \sigma_\tau^2) \times \prod_{i\neq a_d}\big(1 - \Phi_\tau(\tau_{a_d d}; \mu_{id}, \sigma_\tau^2)\big)\Big),
\end{aligned} \tag{15}$$

where φ_τ and Φ_τ are the probability density function (pdf) and cumulative distribution function (cdf) of the specified distribution of time-increments, respecitvely. The latter part of this conditional posterior reflects the fact that the minimum value of latent timestamps determines entire tie data (Section 2.3.3), thus can be interpreted as 'P(observed timestamp) $\times$ P(all latent timestamps greater than the observed time)' given the specified distribution of time-increment.

The conditional posterior for topic assignment z_{dn} is derived by multiplying the two sampling equations of LDA:

$$\begin{aligned}
&p(z_{dn} = k|\boldsymbol{z}_{\backslash dn}, \boldsymbol{l}, \boldsymbol{b}, \boldsymbol{\eta}, \delta, \boldsymbol{u}, \boldsymbol{w}, \boldsymbol{a}, \boldsymbol{r}, \boldsymbol{t}, \alpha, \beta, \boldsymbol{m}) \\
&\propto (N_{dk,\backslash dn} + \alpha m_k) \times \frac{N_{w_{dn}k,\backslash dn} + \frac{\beta}{V}}{N_{k,\backslash dn} + \beta} \\
&\times \prod_{i=1}^{A} \frac{\exp\Big\{\log\big(\mathrm{I}(\|\boldsymbol{u}_{id}\|_1 > 0)\big) + \sum_{j\neq i}(\delta + \lambda_{idj})u_{idj}\Big\}}{Z(\delta, \boldsymbol{\lambda}_{id})} \\
&\quad \times \varphi_\tau(\tau_{a_d d}; \mu_{a_d d}, \sigma_\tau^2) \times \prod_{i\neq a_d}\big(1 - \Phi_\tau(\tau_{a_d d}; \mu_{id}, \sigma_\tau^2)\big),
\end{aligned} \tag{16}$$

where the subscript $\backslash dn$ denote the exclusion of document d and n^{th} element in document d, and $N_{w_{dn}k,\backslash dn}$ is the number of tokens assigned to topic k whose type is the same as that of w_{dn} (excluding w_{dn} itself).

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

$$\prod_{d=1}^{D}\prod_{i=1}^{A} \frac{\exp\Big\{\log\big(\mathrm{I}(\|\boldsymbol{u}_{id}\|_1 > 0)\big) + \sum_{j\neq i}(\delta + \lambda_{idj})u_{idj}\Big\}}{Z(\delta, \boldsymbol{\lambda}_{id})}, \tag{17}$$

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 shared conditional posterior is

$$\prod_{d=1}^{D}\Big(\varphi_\tau(\tau_{a_d d}; \mu_{a_d d}, \sigma_\tau^2) \times \prod_{i\neq a_d}\big(1 - \Phi_\tau(\tau_{a_d d}; \mu_{id}, \sigma_\tau^2)\big)\Big). \tag{18}$$

Although the IPTM has a lot of latent variables, the model can yield 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 2.3.1) and another from the latent timestamps (Section 2.3.2). In addition, for better performance and interpretability of the topics we infer, we adopt the hyperparameter optimization technique for α and $\boldsymbol{m}$ called "new fixed-point iterations using the Digamma recurrence relation" in (Wallach, 2008), for every outer iteration.

4. Applications to Email Networks

The IPTM is intended for any network with timestamped, text-valued ties, however, in our application of the model we focus on the analysis of email network—a canonical example of dynamic textual communication. Via this application, we demonstrate that the IPTM is effective at predicting and explaining continuous-time textual communications.

4.1. Data

Our data come from the North Carolina county government email dataset collected by (ben Aaron et al., 2017) 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, (1) in order to see whether and how communication networks surrounding a notable national emergency—Hurricane Sandy—differed from those surrounding other governmental functions, and (2) to 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 email data set (Klimt & Yang, 2004). For this dataset, 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 senders. Emails that were not sent to at least one other active actor were discarded, and also preprocessed to remove any stop words, URLs, quoted text, and signatures. These steps resulted in a total of 6,613 emails involving 30 actors.

4.2. Dynamic Network Features

In Section 2.3.1, we introduced the dynamic network features $\boldsymbol{x}_{idjc}$, which could be flexibly specified according to the researcher's interest. Here, we outline our specifications of the dynamic network statsitics, tailored for the Dare County email network. We follow roughly the same approach as (Perry & Wolfe, 2013), we employ a suite of eight different effects to be used as the components of $\boldsymbol{x}_{idjc}$—outdegree, indegree, send, receive, 2-send, 2-receive, sibling, and cosibling—to capture common network properties such as popularity, centrality, reciprocity, and transitivity. Visualization of each dynamic network statistics are described in Figure 2, where the upper four features are "dyadic", involving exactly two actors, while the lower four are "triadic", involving exactly three actors.

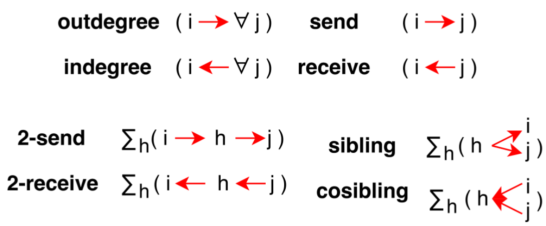

Figure 2. Eight dynamic network statistics used for the Dare County email network and Enron dataset.

Assuming that each network feature has potentially different effects within a number of time intervals (i.e., recency effect), we partition the interval $[-\infty, t_d)$ into 4 sub-intervals with equal length in the log-scale, and focus on three time intervals prior to just after the email's timestamp: 4–16 days, 1–4 days, and 0–1 day. We disregard the time interval before 16 days, considering that the Dare County email network only spans 12 weeks in length. We then compute each of the network feature within each time interval to obtain a set of 24 dynamic network features $\boldsymbol{x}_{idjc}$, specific to author i, recipient j, email d, and interaction pattern c. Detailed mathematical formulations and corresponding interpretations of the network statistics are provided in Appendix C.

4.3. Timestamp Specifications

Section 2.3.2 presented a set of covariates $\boldsymbol{y}_{idjc}$ which are used to predict the timestamps of documents. Similarly as dynamic network features, we exemplify our choice of time-related features that are used to analyze the Dare County email network. First of all, we include the set of 24 dynamic network features $\boldsymbol{x}_{idjc}$ defined in Section 4.2 as the component of $\boldsymbol{y}_{idjc}$, since "who talked to whom, how often and recent, and about what" could play a important role in determing "when to send" a document. Taking into account the fact that our data consists of government organizational emails as well as the exploratory results, we added two temporal features into $\boldsymbol{y}_{idjc}$ that stongly affects the timing of documents: the day of the week and time of the day when the previous document was sent.

Moreover, our exploratory analysis revealed that the Dare County email network shows the best fitting when we specify lognormal distribution on the observed timestamps (i.e., $\log(\tau_{a_d d}) \sim N(\mu_{a_d d}, V(\mu_{a_d d}) = \sigma_\tau^2))$, while we observed significant lack-of-fit for single parameter distributions such as Exponential distribution (i.e., $\tau_{a_d d} \sim \text{Exp}(\mu_{a_d d})$). Based on this result, we chose lognormal distribution.

5. Experiments

In this section, we conduct a set of posterior predictive experiments using the Dare County email network and the Enron dataset, to showcase the IPTM's predictive performance as compared to alternative modeling approaches.

5.1. Tie Prediction

For a randomly chosen document $d^* \in \{M, M+1, \ldots, D\}$, we fit the IPTM to the corpus consisting of the first $d = \{1, \ldots, d^* - 1\}$ documents, then use the inferred posterior distributions to generate a distribution of predicted tie data $(a_{d^*}, \boldsymbol{r}_{d^*}, t_{d^*})$ conditional on the content in the document $\boldsymbol{w}_{d^*}$, and compare the simulated ones to the observed data. We also compare the IPTM to the alternative model. Several models exist that could be used to model any of these three data types individually, but, to our knowledge, the literature does not offer any models that can be used to jointly generate all three types of tie data integrated into the IPTM. Thus, the alternative model is built upon two separate regression models for the recipients and timestamps, and test if the IPTM has any benefit over other existing models by jointly inferring the parameters that govern the generation of tie data—authors, recipients, and timestamps. Pseudocodes for generating predicted tie data using the IPTM and the regression model are demonstrated in Appendix D.

For both data sets, the Dare County email network and Enron dataset, we randomly selected 200 documents from the later half of the corpus (i.e., $M = \frac{D}{2}$) and generated 100 samples of predicted tie data for every document d^*. We ran the same predictive experiments with 21 unique combinations of the number of interaction patterns ($C = 1, 2, 3$) and the number of topics ($K = 2, 5, 10, 25, 50, 75, 100$) as a grid-search based hyperparameter selection process. We compare the predictions in terms of classification accuracy in predicting the authors and recipients, as well as prediction error in the timestamps. Figure 3 presents the F_1 scores on author predictions, multiclass version of the area under the ROC curve (AUC) measure (Hand & Till, 2001) on reciptient predictions, and median absolute error (MAE) on timestamp predictions for each document we predicted, all averaged over the entire samples. The outcomes demonstrate the ability IPTM in predicting the author, recipient, and timetsamps of email. **Further comments after we conduct experiment again.**

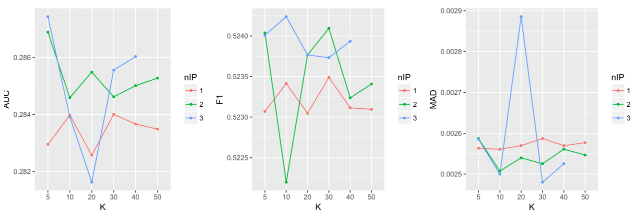

Figure 3. Average AUC, F1 score, MAE: (*top*) Dare County email network. (*bottom*) Enron dataset.

5.2. Topic Coherence

Topic coherence metrics (Mimno et al., 2011) are often used to evaluate the semantic coherence in topic models. In order to test whether 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 the latent dirichlet allocation (LDA). Instead of re-fit the data using standard LDA algorithms, we used the topic assignments from the IPTM with $C = 1$, which simply makes the IPTM reduced to LDA in terms of topic assignments by unlinking the text and networks. For each model, we varied the number of topics from 1 to 100 and draw five samples from the joint posterior distribution over the latent variable. We evaluated the topics resulting from each sample and averaged over the five samples, where the results are shown in Figure 4. Combined with the results in Section 5.1, this result demonstrates that the IPTM can achieve good predictive performance while producing coherent topics.

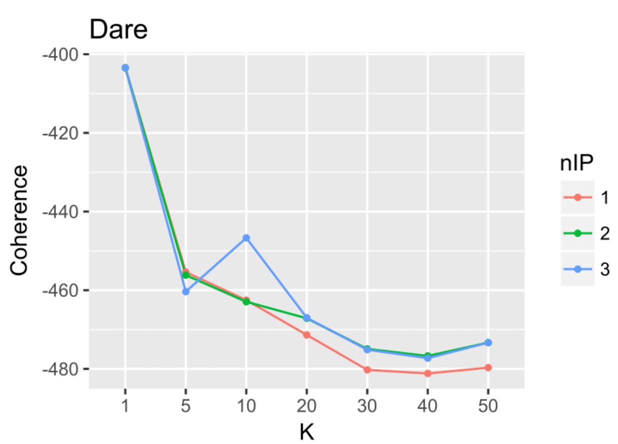

Figure 4. Average topic coherence scores: (*left*) Dare County email network. (*right*) Enron dataset.

5.3. Posterior Predictive Checks

We finally perform posterior predictive checks (Rubin et al., 1984) in order to evaluate the appropriateness of the model specification for the Dare County email network. Formally, we generated entirely new data, by simulating $\{(a_d, \boldsymbol{r}_d, t_d, \boldsymbol{w}_d)\}_{d=1}^{D}$ from the genenerative process in Section 2.1 and 2.3, conditional upon a set of inferred parameter values from the inference in Section 3 (see Appendix D for pseudocode). 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 networks: indegree distribution for author activities, outdegree distribution for recipient activities, recipient size distribution, document time-increment distributions, the edgewise shared partner distribution, and the geodesic distance distribution. For content-wise goodness-of-fit, we employed mutual information (MI) in (Mimno & Blei, 2011), which is often used to evaluate "bag of words" model assumptions. We then generated 1,000 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 and MI. If the model is appropriate, the observed data should not be an outlier with respect to distributions of new data drawn from the posterior predictive distribution.

Figure 5 illustrates the result of posterior predictive checks, showing IPTM's goodness of fit for Dare County data. The results reveal that IPTM captures some important work features of the data, including spreadness and transitivity. **Further comments after we conduct PPC again.**

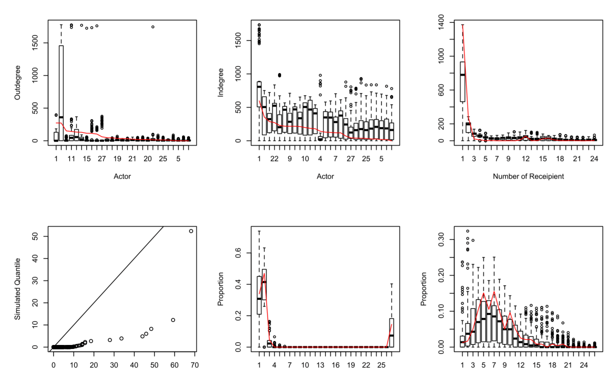

Figure 5. Posterior predictive checks for the Dare County email network: (a) outdegree, (b) indegree, (c) recipient size, (d) QQplot of time-increments, (e) geodesic distance, and (f) edgewise shared partners.

Table 1. Classification accuracies for naive Bayes and flexible Bayes on various data sets.

DATA SET	NAIVE	FLEXIBLE	BETTER?
BREAST	95.9± 0.2	96.7± 0.2	√
CLEVELAND	83.3± 0.6	80.0± 0.6	×
GLASS2	61.9± 1.4	83.8± 0.7	√
CREDIT	74.8± 0.5	78.3± 0.6	
HORSE	73.3± 0.9	69.7± 1.0	×
META	67.1± 0.6	76.5± 0.5	√
PIMA	75.1± 0.6	73.9± 0.5	
VEHICLE	44.9± 0.6	61.5± 0.4	√

6. Analysis

In order to demonstrate our model's novel ability to identify interaction-pattern-specific communications that exist in both the content and relational structure, we performed an exploratory analysis on the interaction patterns inferred from the Dare County email network using the IPTM. Our main focus 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

Table **??** and **??** show top ten words for the five topics that were most strongly associated with interaction pattern 1 and 2, respecitvely. It is pretty clear from the highlighted words that many of the topics in the interaction pattern 1 are about the hurricane. On the contrary, the topics most strongly associated with the interaction pattern 2 are about standard government activities. Very few of their top words are about the hurricane. Together, the assignment of hurricane-related topics to interaction pattern 1 and government-related topics to interaction pattern 2 provide support for our hypothesis that topics about Hurricane Sandy exhibit very a different interaction pattern to other topics.

6.2. Interaction Pattern Coefficients

In theory, we should be able to use the inferred coefficients to understand each interaction pattern's characteristics.

7. Conclusions

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

Table 2. Classification accuracies for naive Bayes and flexible Bayes on various data sets.

DATA SET	NAIVE	FLEXIBLE	BETTER?
BREAST	95.9± 0.2	96.7± 0.2	√
CLEVELAND	83.3± 0.6	80.0± 0.6	×
GLASS2	61.9± 1.4	83.8± 0.7	√
CREDIT	74.8± 0.5	78.3± 0.6	
HORSE	73.3± 0.9	69.7± 1.0	×
META	67.1± 0.6	76.5± 0.5	√
PIMA	75.1± 0.6	73.9± 0.5	
VEHICLE	44.9± 0.6	61.5± 0.4	√

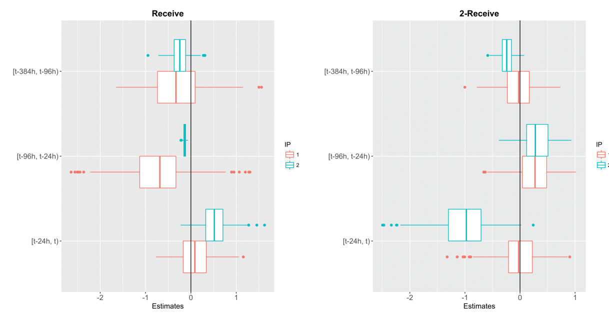

Figure 6. 95% credible intervals of the posterior estimates of $\{\boldsymbol{b}_c\}_{c=1}^{C}$ using Dare County data: (*left*) Recieve. (*right*) 2-Recieve.

of ERGM style network generative features on topic-based content. The application to North Carolina county government email data demonstrates, among other capabilities, the effectiveness at the IPTM in separating out both the content and relational structure underlying the normal day-to-day function of an organization and the management of a highly time-sensitive event—Hurricane Sandy. The IPTM is applicable to a variety of networks in which ties are attributed with textual documents. These include, for example, economic sanctions sent between countries and legislation attributed with sponsors and co-sponsors.

References

Alemán, 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.

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 and Blei, David. Bayesian checking for topic models. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pp. 227–237. Association for Computational Linguistics, 2011.

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.

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.

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.