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

Anonymous Authors[1]

1. 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 $t_d < t_{d+1}$.

1.1. 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 a to r 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. We first model each interaction-pattern $c \in [C]$ as a dicrete 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. Then, each topic-interaction pattern assignment c_d is then drawn from the

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

for $d \in [D]$.

1.2. 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. As in LDA, 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, we assume a interaction-pattern-topic distribution over K topics and the corresponding document-topic distribution

$$\begin{aligned} \boldsymbol{m}_c &\sim \text{Dirichlet}\Big(\gamma, (\frac{1}{K}, \ldots, \frac{1}{K})\Big), \\ \boldsymbol{\theta}_d &\sim \text{Dirichlet}(\alpha, \boldsymbol{m}_{c_d}), \end{aligned} \tag{4}$$

where γ and α are the concentration parameter and $\boldsymbol{m} = (m_1, \ldots, m_K)$ is the probability vector. 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.,

$$\begin{aligned} z_{dn} &\sim \text{Multinomial}(\boldsymbol{\theta}_d), \\ w_{dn} &\sim \text{Multinomial}(\phi_{z_{dn}}). \end{aligned} \tag{5}$$

1.3. 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 1.2), we assume the following steps of tie generating process. Much like in the SAOM (Snijders et al., 2010), we conceptualize tie generation as a process that is governed by senders acting in continuous time.

1.3.1. Latent Recipients

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

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

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

In the example of email networks, we form the covariate vector for recipients $\boldsymbol{x}_{adrc}$ using dynamic network statistics focused 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 (Perry & Wolfe, 2013), where the three time 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 1) are:

1. $\text{outdegree}(a, c, i) = \sum_{d':t_{d'}\in i} \delta(c_{d'} = c)\delta(a_{d'} = a).$

2. $\text{indegree}(r, c, i) = \sum_{d':t_{d'}\in i} \delta(c_{d'} = c)\delta(u_{d'r} = 1).$

3. $\text{send}(a, r, c, i) = \sum_{d':t_{d'}\in i} \delta(c_{d'} = c)\delta(a_{d'} = a)\delta(u_{d'r} = 1).$

4. $\text{receive}(a, r, c, i) = \text{send}(r, a, c, i).$

5. $\text{2-send}(a, r, c, i) = \sum_{\substack{i',i''\geq i:\\ i'=i \text{ or } i''=i}} \sum_{h\neq a,r} \text{send}(a, h, c, i')\text{send}(h, r, c, i'').$

6. $\text{2-receive}(a, r, c, i) = \sum_{\substack{i',i''\geq i:\\ i'=i \text{ or } i''=i}} \sum_{h\neq a,r} \text{send}(h, a, c, i')\text{send}(r, h, c, i'').$

6. $\text{sibling}(a, r, c, i) = \sum_{\substack{i',i''\geq i:\\ i'=i \text{ or } i''=i}} \sum_{h\neq a,r} \text{send}(h, a, c, i')\text{send}(h, r, c, i'').$

6. $\text{cosibling}(a, r, c, i) = \sum_{\substack{i',i''\geq i:\\ i'=i \text{ or } i''=i}} \sum_{h\neq a,r} \text{send}(a, h, c, i')\text{send}(r, h, c, i'').$

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.

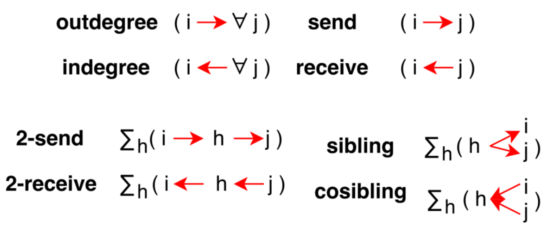

Figure 1. Eight dynamic network statistics used for the application to email networks.

We then compute the weighted average of $\{\nu_{adrc}\}_{c=1}^{C}$ and obtain the "recipient intensity"—the likelihood of document d being sent from a to r— using the the document's distribution over interaction patterns as mixture weights:

$$\lambda_{adr} = \nu_{adrc_d}. \tag{7}$$

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 (Fellows & Handcock, 2017)—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 $\boldsymbol{u}_{ad} = (u_{ad1}, \dots, u_{adA})$

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

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

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

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.

1.3.2. Latent Timestamps

Similarly, we hypothesize "If a were the author of document d, when would it be sent?" and define the "interaction-pattern-specific timing rate"

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

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 $\boldsymbol{y}_{adc}$ is a set of time-related covariates, which can be any feature that could affect timestamps of the document.

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, some 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.

Next, the "timing rate" for author i is then computed from the weighted average of $\{\xi_{adc}\}_{c=1}^{C}$

$$\mu_{ad} = g^{-1}(\xi_{adc_d}), \tag{11}$$

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 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 $g = I$..

1.3.3. 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 (Snijders, 1996; 2017):

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

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.

2. Posterior Inference

Now, we can marginalize out θ and ψ and then the big joint distribution becomes

$$\begin{aligned} & P(\boldsymbol{z}, \boldsymbol{c}, \boldsymbol{b}, \boldsymbol{\eta}, \delta, \boldsymbol{u}, \boldsymbol{w}, \boldsymbol{a}, \boldsymbol{r}, \boldsymbol{t} | \zeta, \gamma, \alpha, \beta, \boldsymbol{\mu}_b, \Sigma_b, \mu_\delta, \sigma_\delta^2) \\ & \propto P(\boldsymbol{z}|\gamma, \alpha, \boldsymbol{c}) P(\boldsymbol{w}|\boldsymbol{z}, \beta) P(\boldsymbol{c}|\zeta) P(\boldsymbol{b}|\boldsymbol{\mu}_b, \Sigma_b) P(\boldsymbol{\eta}|\boldsymbol{\mu}_\eta, \Sigma_\eta) \\ & \times P(\delta|\mu_\delta, \sigma_\delta^2) P(\boldsymbol{u}|\boldsymbol{c}, \boldsymbol{b}, \delta) P(\boldsymbol{a}, \boldsymbol{t}|\boldsymbol{c}, \boldsymbol{\eta}) P(\boldsymbol{r}|\boldsymbol{a}, \boldsymbol{u}), \end{aligned} \tag{14}$$

where and here we will sequentially update $\boldsymbol{z}, \boldsymbol{c}, \boldsymbol{b}, \boldsymbol{\eta}, \delta, \boldsymbol{u}$.

First, the conditional posterior for topic assignment z_{dn} is:

$$\begin{aligned} & p(z_{dn} = k | \boldsymbol{z}_{\backslash dn}, \boldsymbol{c}, \boldsymbol{w}, \gamma, \alpha, \beta) \\ & \propto P(z_{dn} = k | \boldsymbol{z}_{\backslash dn}, c_d, \gamma, \alpha) \times P(w_{dn} = w | z_{dn} = k, \boldsymbol{z}_{\backslash dn}, \boldsymbol{w}, \beta) \\ & \propto (N_{dk,\backslash dn} + \alpha \frac{N_{kc_d,\backslash dn} + \frac{\gamma}{K}}{N_{c_d,\backslash dn} + \gamma}) \times \frac{N_{w_{dn}k,\backslash dn} + \frac{\beta}{V}}{N_{k,\backslash dn} + \beta}, \end{aligned} \tag{15}$$

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

Although everything else stays the same (i.e., sampling equations for $\boldsymbol{b}, \boldsymbol{\eta}, \delta, \boldsymbol{u}$), we have quite complicated sampling equation for $\boldsymbol{c}$ now. For each document $d \in [D]$, we sample interaction-pattern assignment as below:

$$\begin{aligned} & P(c_d = c | \boldsymbol{z}, \zeta, \boldsymbol{u}, \boldsymbol{a}, \boldsymbol{t}) \\ & \propto P(c_d = c | \boldsymbol{c}_{\backslash d}, \zeta) P(\boldsymbol{z}_d | \gamma, \alpha, c_d = c, \boldsymbol{c}_{\backslash d}, \boldsymbol{z}_{\backslash d}) \\ & \times P(a_d, t_d | c_d = c, \boldsymbol{\eta}, \sigma_\tau^2) P(\boldsymbol{u} | c_d = c, \boldsymbol{c}_{\backslash d}, \boldsymbol{b}, \delta) \\ & \propto (N_{c,\backslash d} + \frac{\zeta}{C}) \times \prod_{n=1}^{N} (N_{dz_{dn},\backslash dn} + \alpha \frac{N_{z_{dn}c,\backslash dn} + \frac{\gamma}{K}}{N_{c,\backslash dn} + \gamma}) \\ & \times \varphi_\tau(\tau_d; \mu_{a_d d}, \sigma_\tau^2) \times \prod_{a \neq a_d} \big(1 - \Phi_\tau(\tau_d; \mu_{ad}, \sigma_\tau^2)\big) \\ & \times \prod_{d=d}^{d^*} \Big(\prod_{a=1}^{A} \frac{\exp\Big\{\log\big(\mathrm{I}(\|\boldsymbol{u}_{ad}\|_1 > 0)\big) + \sum_{r \neq a} (\delta + \lambda_{adr}) u_{adr}\Big\}}{Z(\delta, \boldsymbol{\lambda}_{ad})} \Big), \end{aligned} \tag{16}$$

where d^* is the last document that is affected by current document d (i.e., $t_d + 384 \approx t_{d^*}$).

So, all the computational burden (previously on z_{dn}) moved to c_d. For every c_d update, we need to re-calculate the history statistics given $c_d = c$, to evaluate the last line of Equation (16). If we use small C, I can imagine that this should be faster than the curent version. However, if we want to extend this to allow unknown number of latent cluster (such as cluster LDA), this may be worse.

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

References

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.

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

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.

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

Snijders, Tom AB. Stochastic actor-oriented models for network dynamics. *Annual Review of Statistics and Its Application*, (0), 2017.

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.