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

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

[1] The Pennsylvania State University

[2] Microsoft Research NYC

[3] University of Massachusetts Amherst

June 9, 2017

Interaction-Partitioned Topic Model (IPTM)

- Probablistic model for time-stamped textual communications (e.g. emails, cosponsorship of bills, international sanctions)
- Integration of two generative models:
  - Latent Dirichlet allocation (LDA) for topic-based contents
  - Dynamic exponential random graph model (ERGM) for ties
- IPTM assigns each topic to an "interaction pattern," which is governed by a set of dynamic network features

"who communicates with whom about what, and when?"

Content Generating Process: LDA (Blei et al., 2003)

- For each topic $k = 1, ..., K$:

  1. Topic-word distribution $\boldsymbol{\phi}^{(k)} \sim$ Dirichlet$(\beta, \mathbf{u})$
     - A topic k is characterized by a discrete distribution over V word types with probability vector $\boldsymbol{\phi}^{(k)}$.

  2. Topic-IP distribution $c_k \sim$ Uniform$(1, C)$
     - Each topic is associated with a single interaction pattern.

- For each document $d = 1, ..., D$:

  3-1. Document-topic distribution $\boldsymbol{\theta}^{(d)} \sim$ Dirichlet$(\alpha, \boldsymbol{m})$
  - A document d is characterized by a discrete distribution over K topics with probability vector $\boldsymbol{\theta}^{(d)}$.

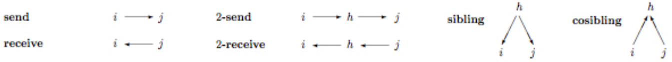
  3-2. For each word in a document $n = 1$ to $N^{(d)}$:
  (a) Choose a topic $z_n^{(d)} \sim$ Multinomial$(\boldsymbol{\theta}^{(d)})$
  (b) Choose a word $w_n^{(d)} \sim$ Multinomial$(\boldsymbol{\phi}^{(z_n^{(d)})})$

  3-3 Calculate the distribution of interaction patterns within a document:

$$p_c^{(d)} = \Big(\sum_{k:c_k=c} N^{(k|d)}\Big)/N^{(d)}, \tag{1}$$

Dynamic Network Features (Perry and Wolfe, 2012)

9 different effects as components of $\boldsymbol{x}_t^{(c)}(i,j)$, (intercept, outdegree, indegree, send, receive, 2-send, 2-receive, sibling, and cosibling) to measure popularity, centrality, reciprocity, and transitivity



$$\lambda_{ij}^{(d)}(t) = \sum_{c=1}^{C} p_c^{(d)} \cdot \exp\Big\{\lambda_0^{(c)} + \boldsymbol{b}^{(c)T}\boldsymbol{x}_t^{(c)}(i,j)\Big\}, \qquad (2)$$

$$\lambda_{iJ}^{(d)}(t) = \sum_{c=1}^{C} p_c^{(d)} \cdot \exp\Big\{\lambda_0^{(c)} + \frac{1}{|J|}\sum_{j\in J}\boldsymbol{b}^{(c)T}\boldsymbol{x}_t^{(c)}(i,j)\Big\}. \qquad (3)$$

Tie Generating Process

1. (Data augmentation) For each sender $i \in \{1, ..., A\}$, create binary receiver vector of length $A-1$, $J_i^{(d)}$, by applying the non-empty Gibbs measure to every $j \in \mathcal{A}_{\backslash i}$.

$$\mathsf{P}(J_i^{(d)}) = \frac{1}{Z(\delta, \log(\lambda_i^{(d)}))} \exp\Big\{\log\big(\mathsf{I}(\sum_{j \in \mathcal{A}_{\backslash i}} J_{ij}^{(d)} > 0)\big) + \sum_{j \in \mathcal{A}_{\backslash i}} (\delta + \log(\lambda_{ij}^{(d)}))\tag{4}$$

2. For every sender $i \in \mathcal{A}$, generate the time increments

$$\Delta T_{iJ_i} \sim \mathsf{Exp}(\lambda_{iJ_i}^{(d)}).\tag{5}$$

3. Set timestamp, sender, and receivers simultaneously (NOTE: $t^{(0)} = 0$):

$$\begin{aligned} t^{(d)} &= t^{(d-1)} + \mathsf{min}(\Delta T_{iJ_i}), \\ i^{(d)} &= i_{\mathsf{min}(\Delta T_{iJ_i})}, \\ J^{(d)} &= J_{i^{(d)}}. \end{aligned}\tag{6}$$

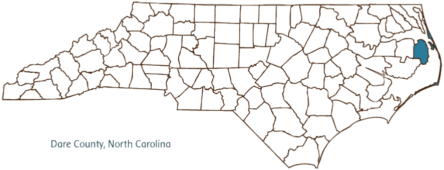
Inference - Pseudocode

Algorithm 1 MCMC

```
set initial values Z^(0), C^(0), and (B^(0), δ^(0))
for o=1 to O do
    for n=1 to n_1 do
        optimize α and m using hyperparameter optimization in [?]
    end
    for d=1 to D do
        for i ∈ A_{\i_o^(d)} do
            sample the augmented data J_i^(d)
        end
        for n=1 to N^(d) do
            draw of z_n^(d) ~ Multinomial(p^Z)
        end
    end
    for k=1 to K do
        draw C_k ~ Multinomial(p^C)
    end
    for n=1 to n_2 do
        sample B using M-H
```

Data: North Carolina Dare county email data

- $D = 1456$ emails between $A = 27$ county government managers, covering 2 month periods (October 1 - November 30) in 2013



Effect of Hurricane Sandy

IPTM Result

Conclusion

- Joint modeling of ties (sender, receiver, time) and contents
- Allowance of multicast – multiple senders and/or receivers
- Possible application to