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

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

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

    3-1. Document-topic distribution $\boldsymbol{\theta}^{(d)} \sim \text{Dirichlet}(\alpha, \boldsymbol{m})$
    - 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 $z_n^{(d)} \sim \text{Multinomial}(\boldsymbol{\theta}^{(d)})$
       (b) Choose a word $w_n^{(d)} \sim \text{Multinomial}(\boldsymbol{\phi}^{(z_n^{(d)})})$

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

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

Bomin Kim[1], Aaron Schein[3], Bruce Desmarais [1], A Network Model for Dynamic Textual Communication

Dynamic Network Features (Perry and Wolfe, 2012)

- $x_t^{(c)}(i, j)$ is the network statistics at time t, for interaction pattern c
  - Degree: outdegree and indegree
  - Dyadic: send and receive
  - Triadic: 2-send, 2-receive,
              sibling and cosibling
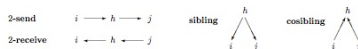
- Partition the interval $[-\infty, t)$ into 4 sub-intervals

$$[-\infty, t) = [-\infty, t - 16d) \cup [t - 16d, t - 3d) \cup [t - 3d, t - 24h) \cup [t - 24h, t - 0)$$

and define interval-based statistics such as
$$\textbf{outdegree}_{t,l}^{(c)}(i) = \sum_{d:t^{(d)} \in I_t^{(l)}} p_c^{(d)} \cdot I\{i \to \forall j\}$$

Stochastic Intensity

- Stochastic intensity $\lambda^{(d)}$

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

where $\boldsymbol{b}^{(c)}$ is a vector of coefficients in $\boldsymbol{R}^p$.

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

Tie Generating Process

1. For every sender $i \in \{1, ..., A\}$, choose a binary vector $J_i^{(d)}$ of length $(A - 1)$, by applying Gibbs measure (Fellows and Handcock, 2017)

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

where δ is a real-valued intercept controlling the recipient size and $Z(\delta, \log(\lambda_i^{(d)}))$ is the normalizing constant.

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

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

3. Set timestamp, sender, and receivers simultaneously:

$$t^{(d)} = t^{(d-1)} + \min(\Delta T_{iJ_i}),$$
$$i^{(d)} = i_{\min(\Delta T_{iJ_i})},$$
$$J^{(d)} = J_{i^{(d)}}.$$

Inference - Pseudocode

Algorithm 1 Markov Chain Monte Carlo (MCMC)

Set initial values $\mathcal{Z}^{(0)}, \mathcal{C}^{(0)}$, and $(\mathcal{B}^{(0)}, \delta^{(0)})$

for *o=1 to O* **do**

    for *d=1 to D* **do**

        for $i \in \mathcal{A}_{\backslash i_o^{(d)}}$ **do**

            for $j \in \mathcal{A}_{\backslash i}$ **do**

                Sample the latent edge $J_{ij}^{(d)}$ via Gibbs sampling

            end

        end

        for *n=1 to $N^{(d)}$* **do**

            Sample the topic assignments via Gibbs sampling

            $z_n^{(d)} \sim \text{Multinomial}(p^{\mathcal{Z}})$

        end

    end

    for *k=1 to K* **do**

        Sample the interaction pattern assignments via Gibbs sampling

        $C_k \sim \text{Multinomial}(p^{\mathcal{C}})$

    end

    for *n=1 to n_B* **do**

        Sample the interaction pattern parameters $\mathcal{B}$ via Metropolis-Hastings

    end

    for *n=1 to n_δ* **do**

        Sample the receiver size parameter δ via Metropolis-Hastings

    end

end

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
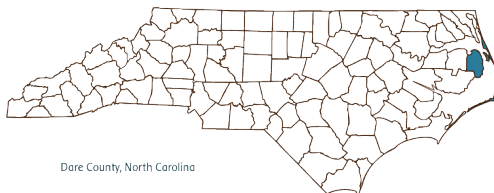
Dare County, North Carolina

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