

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

Bomin Kim¹ Aaron Schein³
Bruce Desmarais¹ Hanna Wallach^{2,3}

¹ The Pennsylvania State University

² Microsoft Research NYC

³ University of Massachusetts Amherst

June 12, 2017

Work supported by NSF grants SES-1558661, SES-1619644, SES-1637089, and CISE-1320219)



Interaction-Partitioned Topic Model (IPTM)

- Probabilistic model for time-stamped textual communications
- Integration of two generative models:
 - Latent Dirichlet allocation (LDA) for topic-based contents
 - Dynamic exponential random graph model (ERGM) for ties

“who communicates with whom about what, and when?”

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

- For each topic $k = 1, \dots, K$:

- Topic-word distribution $\phi^{(k)} \sim \text{Dirichlet}(\beta, \mathbf{u})$
- Topic-IP distribution $c_k \sim \text{Uniform}(1, C)$

| | | |
|----------|-------------|-----------|
| support | services | budget |
| position | care | funds |
| fill | child | money |
| staff | information | budgeted |
| desk | system | including |
| service | community | cost |
| customer | nurse | salary |
| begin | completed | amount |
| duties | provided | revenues |
| vacancy | pregnancy | debt |
| ⋮ | ⋮ | ⋮ |

- For each document $d = 1, \dots, D$:

3-1. Document-topic distribution:

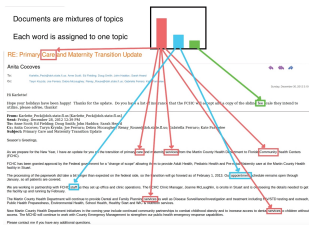
$$\theta^{(d)} \sim \text{Dirichlet}(\alpha, \mathbf{m})$$

3-2. For each word in a document $n = 1$ to $N^{(d)}$:

- Choose a topic $z_n^{(d)} \sim \text{Multinomial}(\theta^{(d)})$
- Choose a word $w_n^{(d)} \sim \text{Multinomial}(\phi^{(z_n^{(d)})})$

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

$$p_c^{(d)} = \left(\sum_{k: c_k = c} N^{(k|d)} \right) / N^{(d)},$$



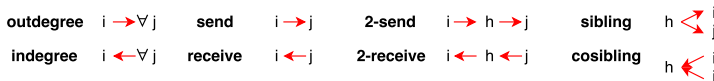
Dynamic Network Features (Perry and Wolfe, 2012)

- Partition the past 384 hours (=16 days) into 3 sub-intervals

$$[t - 384h, t) = [t - 384h, t - 96h) \cup [t - 96h, t - 24h) \cup [t - 24h, t),$$

then define the interval-based dynamic network statistics

- $\mathbf{x}_t^{(c)}(i, j)$ 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



Tie Generating Process: Latent Edges

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

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

where

- $\lambda_{ij}^{(d)} = \sum_{c=1}^C p_c^{(d)} \cdot \exp \left\{ \lambda_0^{(c)} + \mathbf{b}^{(c)T} \mathbf{x}_{t(d-1)}^{(c)}(i, j) \right\}$ is a stochastic intensity
- δ is a real-valued intercept controlling the recipient size
- $Z(\delta, \log(\lambda_i^{(d)}))$ is the normalizing constant

| i | 1 | 2 | 3 | 4 | | A |
|-----|---|---|---|---|-------|---|
| 1 | 0 | 1 | 0 | 1 | | 1 |
| 2 | 1 | 0 | 0 | 0 | | 0 |
| ... | | | | | | |
| A | 0 | 0 | 1 | 0 | | 0 |

Tie Generating Process: Observed

- For each sender $i \in \mathcal{A}$, generate the time increments

$$\Delta T_{iJ_i} \sim \text{Exp}(\lambda_{iJ_i}^{(d)}),$$

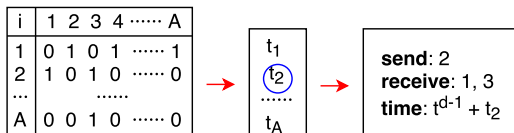
$$\text{where } \lambda_{iJ_i}^{(d)} = \sum_{c=1}^C p_c^{(d)} \cdot \exp\left\{\lambda_0^{(c)} + \frac{1}{|J_i|} \sum_{j \in J_i} \mathbf{b}^{(c)T} \mathbf{x}_{t^{(d-1)}}^{(c)}(i, j)\right\}.$$

- 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

- Bayesian Inference using Markov Chain Monte Carlo (MCMC)

Algorithm 1 MCMC

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

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

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

 Sample the topic assignments \mathcal{Z} via Gibbs sampling

 Sample the interaction pattern assignments \mathcal{C} via Gibbs sampling

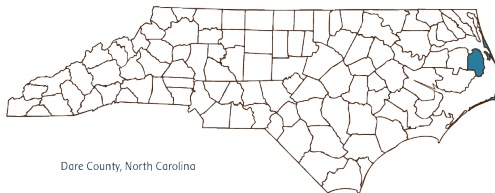
 Sample the interaction pattern parameters \mathcal{B} via Metropolis-Hastings

 Sample the receiver size parameter δ via Metropolis-Hastings

end

Data: North Carolina Dare county email data

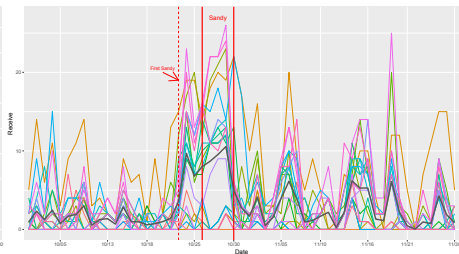
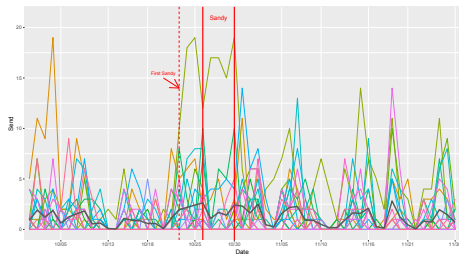
- $D = 1456$ emails between $A = 27$ county government managers, covering 2 month periods (October 1 - November 30) in 2013



Dare County, North Carolina

- Hurricane Sandy passed by NC: October 26 - October 30

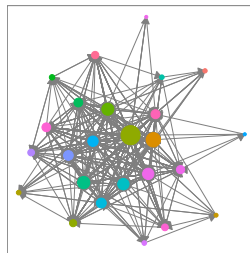
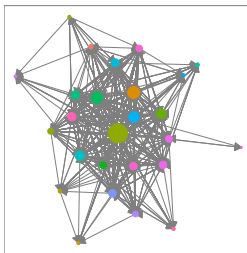
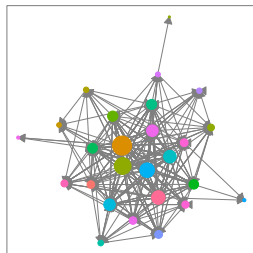
Effect of Hurricane Sandy on Email Exchange



Pre-Sandy

Sandy

Post-Sandy

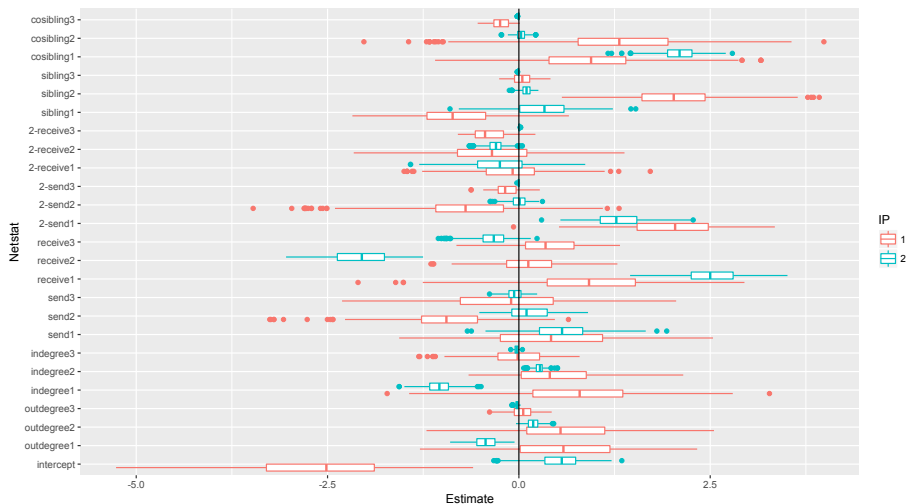


Department

- Building Inspections
- County Extension
- County Manager
- Detention
- Elections
- Emergency Services
- Finance
- Health
- HR
- Information Technology
- Library
- Parks and Recreation
- Planning
- Public Informations
- Register of Deeds
- Senior Center
- Sheriff
- Soil Conservation
- Solid Waste and Recycling
- Tax Administrator
- Transportation
- Veteran Services

IPTM Result: Dynamic Network Effects

- IPTM result with $C = 2$, $K = 5$ and $O = 20$:



IPTM Result: Contents

- IPTM result with $C = 2$, $K = 5$ and $O = 20$:

| IP | 2 | 2 | 1 | 2 | 2 |
|-------|---|---|--|--|--|
| Topic | 5 | 1 | 2 | 3 | 4 |
| Word | tim request services report tax northwest michelle evans tonnage coastal | forecast force today rip race moderate app summary operations late | updates amount mph exam machine esi view dangerous curves north | parcels billed real ocean continues watched duration early situation wash | overtime east scheduled library comp count expected human period administrative |

Conclusion

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