

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 15, 2017

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



Interaction-Partitioned Topic Model (IPTM)

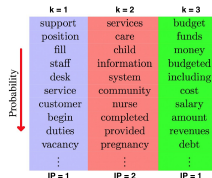
- Probabilistic model for time-stamped textual communications
- Integration of two generative models:
 - Latent Dirichlet allocation (LDA) for topic-based contents
 - Dynamic exponential random graph model (ERGM) for ties

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

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

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

1. Choose a topic-word distribution over the word types
2. Choose a topic-interaction pattern assignment



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

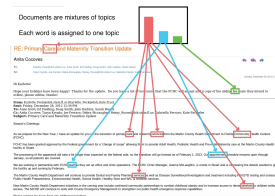
- ### 3-1. Choose a document-topic distribution

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

- (a) Choose a topic from document-topic distribution

- (b) Choose a word from topic-word distribution

- ### 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 ($l = 1, 2, 3$)

- $\mathbf{x}_{t,l}^{(c)}(i, j)$ 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

outdegree	$(i \rightarrow \forall j)$	send	$(i \rightarrow j)$	2-send	$\sum_h (i \rightarrow h \rightarrow j)$	sibling	$\sum_h (h \rightarrow i \rightarrow j)$
indegree	$(i \leftarrow \forall j)$	receive	$(i \leftarrow j)$	2-receive	$\sum_h (i \leftarrow h \leftarrow j)$	cosibling	$\sum_h (h \leftarrow i \leftarrow j)$

Tie Generating Process: Latent Edges

1. For each sender $i \in \{1, \dots, A\}$ and receiver $j \in \{1, \dots, A\}$, calculate the stochastic indensity between i and j :

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

which is a mixture of contents, baseline interaction rate, and network effects.

2. 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 Hancock, 2017)

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

where δ is a real-valued intercept controlling the recipient size

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

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

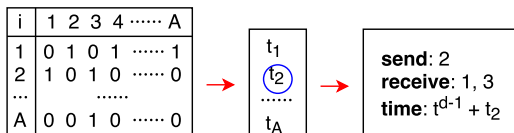
where $\lambda_{iJ_i}^{(d)} = \sum_{c=1}^C p_c^{(d)} \cdot \exp\left\{\lambda_0^{(c)} + \frac{1}{|J_i|} \sum_{j \in J_i} \mathbf{b}^{(c)T} \mathbf{x}_{t^{(d-1)}}^{(c)}(i, j)\right\}$ is the updated sender-specific stochastic intensity given the receivers.

- Set the observed sender, receivers and timestamp simultaneously:

$$i^{(d)} = i_{\min(\Delta T_{iJ_i})}$$

$$J^{(d)} = J_{i^{(d)}}$$

$$t^{(d)} = t^{(d-1)} + \min(\Delta T_{iJ_i})$$



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

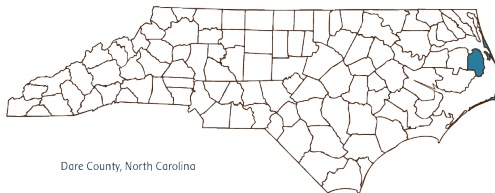
 Sample the interaction pattern parameters \mathcal{B} via Metropolis-Hastings

 Sample the receiver size parameter δ via Metropolis-Hastings

end

Data: North Carolina Dare county email data

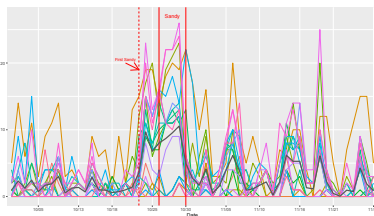
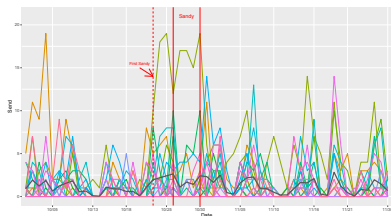
- $D = 1456$ emails between $A = 27$ county government managers, covering 2 month periods (October 1 - November 30) in 2012



Dare County, North Carolina

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

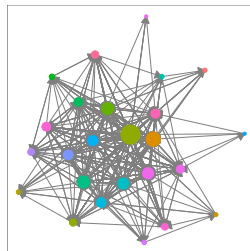
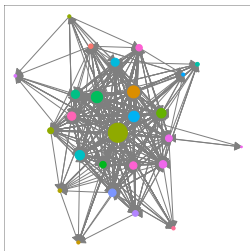
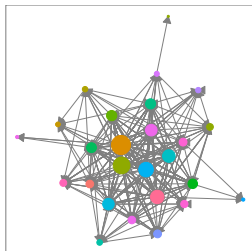
Exploratory Data Analysis: Effect of Sandy



Pre-Sandy

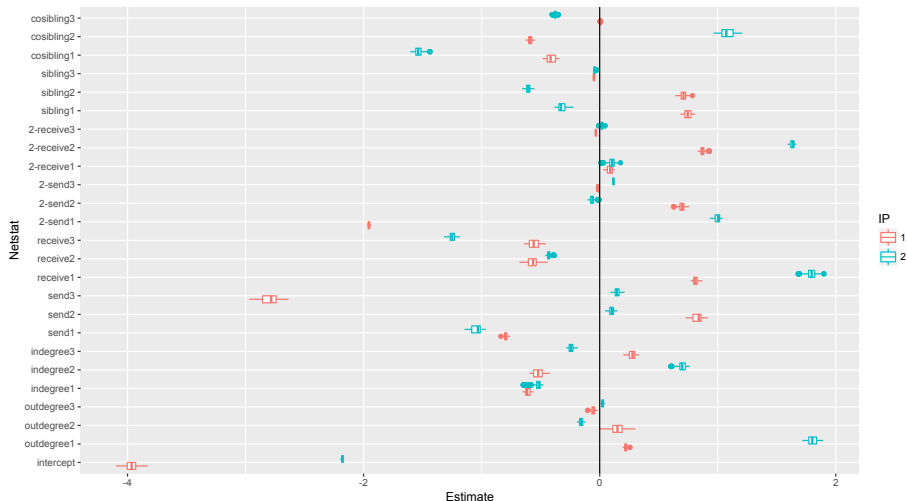
Sandy

Post-Sandy



IPTM Result: Dynamic Network Effects

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



*Preliminary results with small outer iterations. Model results subject to change.

IPTM Result: Contents

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

IP	1	1	1	2	2	2
Topic	15	1	5	10	20	14
Word	inclement east closed conditions coastal touching wind email cellular android-powered bobby surf tomorrow web side	winds hurricane changes inlet moday track sandy tuesday bridge forecast revision will tonight obx shore	report force water violation irene doc extend impacts view sandy thought flood color property outer	overtime update north personnel period outer office situation exam call moved comp well time carolina	late watned early request will rodanthe michelle evans sunday changing workcentre watch large communications planning	oct wednesday touch will breifing change night dot transportaion post collector monday cell hours point

[†]Preliminary results with small outer iterations. Model results subject to change.

Conclusion

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