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 14, 2017

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

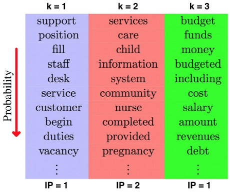
Interaction-Partitioned Topic Model (IPTM)

- Probablistic 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 $k = 1, ..., K$:
  1. Choose a topic-word distribution
  2. Choose a topic-interaction pattern assignment



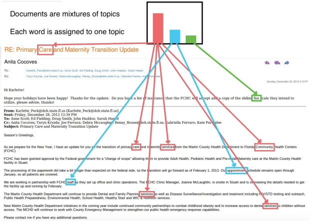
- For each document $d = 1, ..., D$:

  3-1. Choose a document-topic distribution

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

    (a) Choose a topic

    (b) Choose a word



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

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

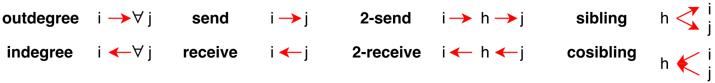
Dynamic Network Features (Perry and Wolfe, 2012)

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

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



Tie Generating Process: Latent Edges

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

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

where

- $\lambda_{ij}^{(d)} = \sum_{c=1}^{C} p_c^{(d)} \cdot \exp\Big\{\lambda_0^{(c)} + \boldsymbol{b}^{(c)T} \boldsymbol{x}_{t^{(d-1)}}^{(c)}(i,j)\Big\}$ is a stochastic intensity
- δ is a real-valued intercept controlling the recipient size
- $Z(\delta, \log(\lambda_i^{(d)})) = (\prod_{j \in \mathcal{A} \backslash i} (\exp\{\delta + \log(\lambda_{ij}^{(d)})\} + 1)) - 1$ 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

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

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

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

3. Set timestamp, sender, and receivers simultaneously:

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

i	1	2	3	4	……	A
1	0	1	0	1	……	1
2	1	0	1	0	……	0
…			……			
A	0	0	1	0	……	0

→

t_1
t_2
……
t_A

→

send: 2
receive: 1, 3
time: $t^{d-1} + t_2$

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

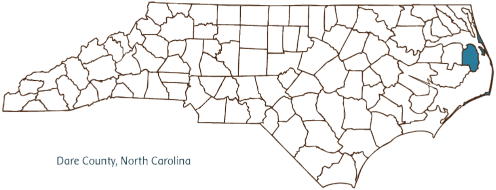
  Sample the interaction pattern assignments $\mathcal{C}$ via Gibbs sampling

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

  Sample the receiver size parameter δ via Metropolis-Hastings
end

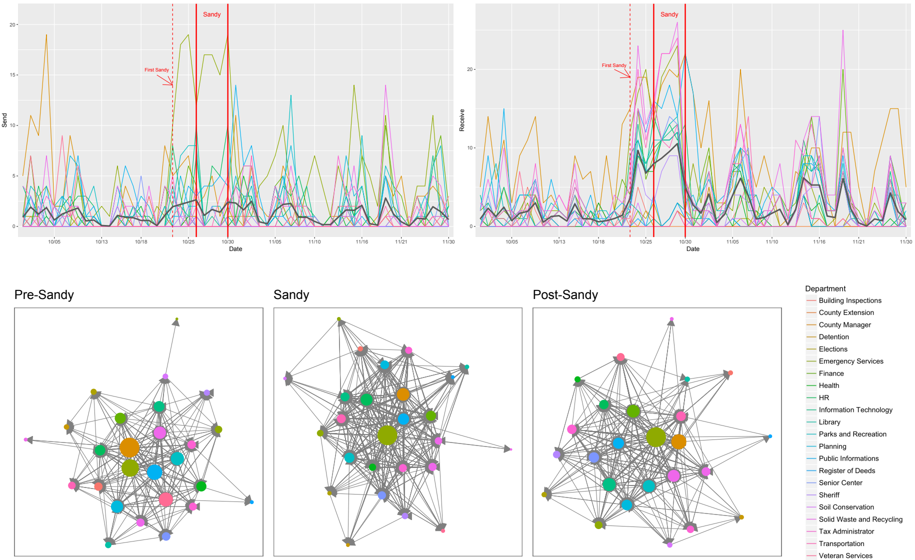
Data: North Carolina Dare county email data

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



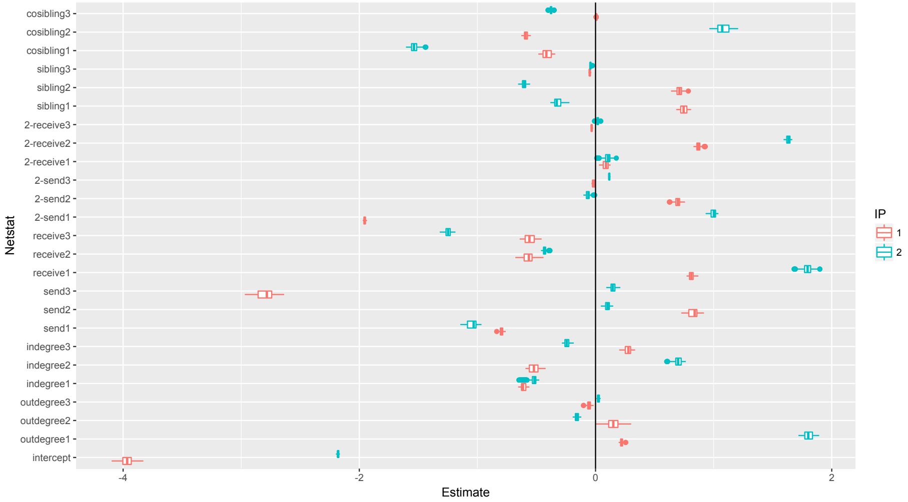
- Hurricane Sandy passed by NC: October 26 - October 30

Effect of Hurricane Sandy on Email Exchange



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