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 11, 2017

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. Topic-word distribution $\phi^{(k)} \sim \text{Dirichlet}(\beta, \mathbf{u})$

    2. Topic-IP distribution $c_k \sim \text{Uniform}(1, C)$

- For each document $d = 1, ..., D$:
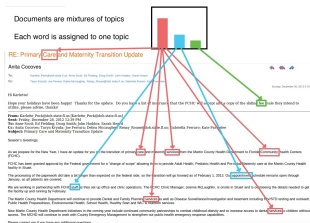
    3-1. Document-topic distribution:
    $\boldsymbol{\theta}^{(d)} \sim \text{Dirichlet}(\alpha, \boldsymbol{m})$

    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:

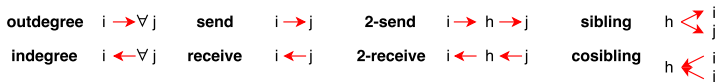$$p_c^{(d)} = \Big(\sum_{k:c_k=c} N^{(k|d)} \Big)/N^{(d)},$$

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

- $x_t^{(c)}(i, j)$ 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

outdegree	i → ∀ j	**send**	i → j	**2-send**	i → h → j	**sibling**	h ⭰ i j
indegree	i ← ∀ j	**receive**	i ← j	**2-receive**	i ← h ← j	**cosibling**	h ⭰ i j

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)

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

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

- $Z(\delta, \log(\lambda_i^{(d)}))$ 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

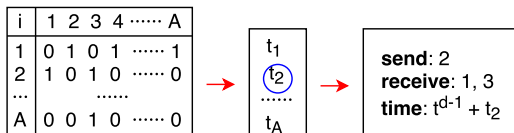$$\Delta T_{iJ_i} \sim \mathsf{Exp}(\lambda_{iJ_i}^{(d)}),$$

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

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

$\rightarrow$

t_1
$\textcircled{t_2}$
......
t_A

$\rightarrow$

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

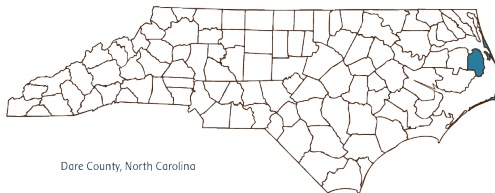   Sample the interaction pattern assignments $\mathcal{C}$ via Gibbs sampling

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

    Sample the receiver size parameter δ via Metropolis-Hastings
end

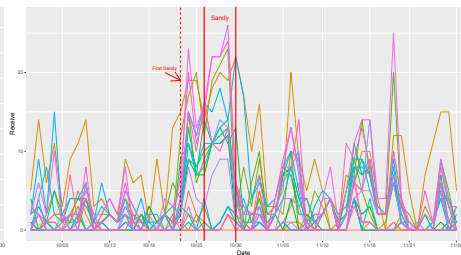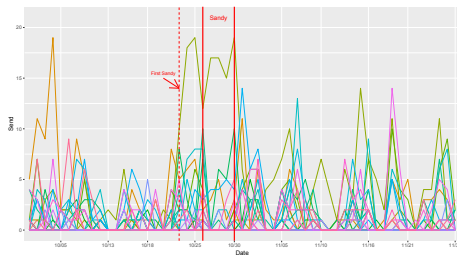---

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 2013
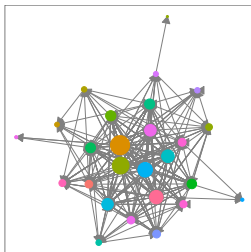

Dare County, North Carolina

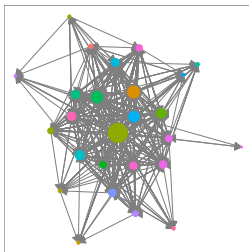- Hurricane Sandy passed by NC: October 26 - October 30
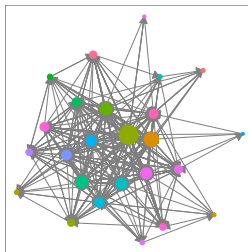
Effect of Hurricane Sandy on Email Exchange

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

$\hat{\boldsymbol{b}}_p^{(c)}$	IP = 1	IP = 2
intercept	-3.264	-7.217
outdegree$[t - 1d, t)$	0.025	1.520
outdegree$[t - 3d, t - 1d)$	0.538	-4.776
outdegree$[t - 16d, t - 3d)$	-0.167	0.255
indegree$[t - 1d, t)$	-1.435	-4.743
indegree$[t - 3d, t - 1d)$	0.952	-1.529
indegree$[t - 16d, t - 3d)$	-0.276	0.279
send$[t - 1d, t)$	1.639	-0.001
send$[t - 3d, t - 1d)$	0.054	-4.223
send$[t - 16d, t - 3d)$	0.972	3.765
receive$[t - 1d, t)$	-0.380	-4.940
receive$[t - 3d, t - 1d)$	-1.625	-1.076
receive$[t - 16d, t - 3d)$	-0.389	-2.490
2-send$[t - 1d, t)$	2.185	0.477
2-send$[t - 3d, t - 1d)$	0.919	2.364
2-send$[t - 16d, t - 3d)$	-0.071	0.154
2-receive$[t - 1d, t)$	1.020	1.189
2-receive$[t - 3d, t - 1d)$	-0.168	3.971
2-receive$[t - 16d, t - 3d)$	0.029	0.098
sibling$[t - 1d, t)$	-1.443	-0.608
sibling$[t - 3d, t - 1d)$	-1.289	-1.405
sibling$[t - 16d, t - 3d)$	-0.239	0.019
cosibling$[t - 1d, t)$	0.390	4.586
cosibling$[t - 3d, t - 1d)$	0.792	-2.063
cosibling$[t - 16d, t - 3d)$	-0.103	-0.693

Table: Effect of dynamic statistics on email exchange

Topic	IP = 1	IP = 2
	will	-7.217
	director	1.520
	manteo	-4.776
	-0.167	0.255
	-1.435	-4.743
	0.952	-1.529
	-0.276	0.279
	1.639	-0.001
	-0.071	0.154
	1.020	1.189
	-0.168	3.971

Table: Effect of dynamic statistics on email exchange

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
  e.g. cosponsorship of bills and international sanctions