

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

July 11, 2017

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



Motivation

- In many networks, ties are attributed with text
 - International treaties
 - International sanctions
 - Legislative cosponsorship
 - Discussion networks on social media
- Network models can't model text
- Models for text either...
 - Are not designed for networks
 - Include simplistic network structure

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

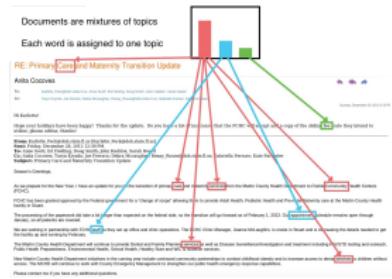
| | k = 1 | k = 2 | k = 3 |
|---------------|---------------|--------------|---------------|
| support | services | | budget |
| position | care | | funds |
| fill | child | | money |
| staff | information | | budgeted |
| desk | system | | including |
| service | community | | cost |
| customer | nurse | | salary |
| begin | completed | | amount |
| duties | provided | | revenues |
| vacancy | pregnancy | | debt |
| : | : | | : |
| IP = 1 | IP = 2 | | IP = 1 |

- 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:

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



Network Model Components

- Models real time ties
- Ties predicted using recent network structure
 - Vertex attributes
 - Popularity
 - Reciprocity
 - Transitivity
- Sender selects vector of recipients and timing
- Innovative modeling of multicasts

Dynamic Network Features (Perry and Wolfe, 2012)

Current network features modeled

- memory
- reciprocity
- popularity and activity
- transitivity

outdegree ($i \rightarrow \forall j$) **send** ($i \rightarrow j$)

indegree ($i \leftarrow \forall j$) **receive** ($i \leftarrow j$)

2-send $\sum_h (i \rightarrow h \rightarrow j)$ **sibling** $\sum_h (h \xleftrightarrow{i} j)$

2-receive $\sum_h (i \leftarrow h \leftarrow j)$ **cosibling** $\sum_h (h \leftarrow i \leftarrow j)$

Conditioning features on recency

- Network features conditioned on degree of recency
- Partition the past 384 hours ($=16$ days) into 3 sub-intervals

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

- $x_{t,l}^{(c)}(i,j)$ is the network statistics at time t , for interaction pattern c

$\mathbf{h} \rightarrow \mathbf{j}$

| | [t-24h, t-0) | [t-96h, t-24h) | [t-384h, t-96h) |
|-----------------|-----------------------|-----------------------|-----------------------|
| [t-24h, t-0) | 2-send _{i,1} | 2-send _{i,1} | 2-send _{i,1} |
| [t-96h, t-24h) | 2-send _{i,1} | 2-send _{i,2} | 2-send _{i,2} |
| [t-384h, t-96h) | 2-send _{i,1} | 2-send _{i,2} | 2-send _{i,3} |

$\mathbf{i} \rightarrow \mathbf{h}$

| | [t-96h, t-24h) | [t-384h, t-96h) |
|--|-----------------------|-----------------------|
| | 2-send _{i,1} | 2-send _{i,1} |

Tie Generating Process: Receivers

- For each sender $i \in \{1, \dots, A\}$ and receiver $j \in \{1, \dots, A\}$ ($i \neq j$), calculate the stochastic intensity between i and j :

$$\lambda_{ij}^{(d)} = \sum_{c=1}^C p_c^{(d)} \cdot \exp \left\{ \mathbf{b}_0^{(c)} + \mathbf{b}^{(c)T} \mathbf{x}_{t^{(d-1)}}^{(c)}(i, j) \right\},$$

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

- 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)

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

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

The diagram illustrates the inputs to the Tie Generating Process. On the left, three components are shown: 'contents' (represented by a blue dashed box labeled 'IP'), 'stochastic intensity' (represented by a red arrow pointing to a column in the matrix), and 'history of interactions' (represented by a red arrow pointing to a row in the matrix). These three components are connected by red arrows to a large square matrix on the right. The matrix has a header row 'i | 1 2 3 4 A'. The first column is labeled 'i'. The rows are labeled 1, 2, ..., A. The matrix contains binary values (0 or 1) representing the tie strength between senders and receivers. For example, row 1 (sender 1) has values [0, 1, 0, 1, ..., 1]. Row 2 (sender 2) has values [1, 0, 0, 0, ..., 0]. Row A (sender A) has values [0, 0, 1, 0, ..., 0]. Ellipses indicate intermediate senders.

| 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: Sender and Time

3. For each sender $i \in \{1, \dots, A\}$, generate the time increments for document d

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

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

4. Set the observed sender, receivers and timestamp simultaneously:

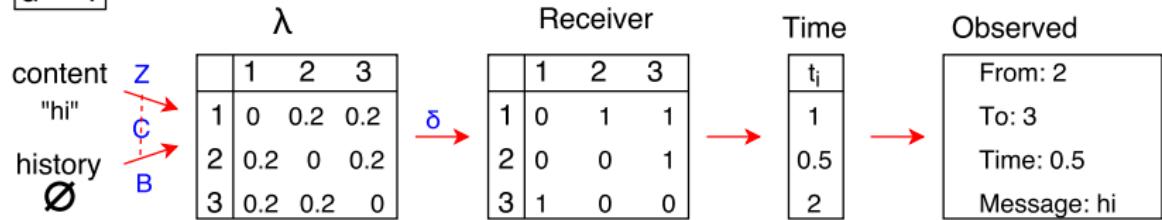
$$i^{(d)} = i_{\min(\Delta T_{iJ_i}^{(d)})}$$

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

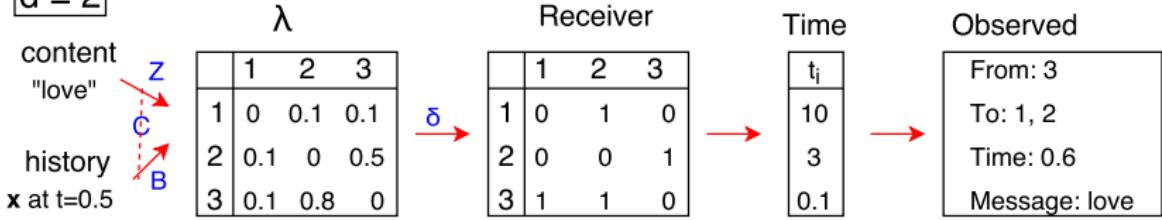
$$t^{(d)} = t^{(d-1)} + \min(\Delta T_{iJ_i}^{(d)})$$

Joint Generating Process

$d = 1$



$d = 2$



Inference

- Take a Bayesian approach to inference
- \mathcal{B} and δ interpreted at fixed \mathcal{Z} and \mathcal{C}

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 receivers** $J_{ij}^{(d)}$ via Gibbs sampling

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

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

 Sample the **network effect parameters** \mathcal{B} via Metropolis-Hastings

 Sample the **receiver size parameter** δ via Metropolis-Hastings

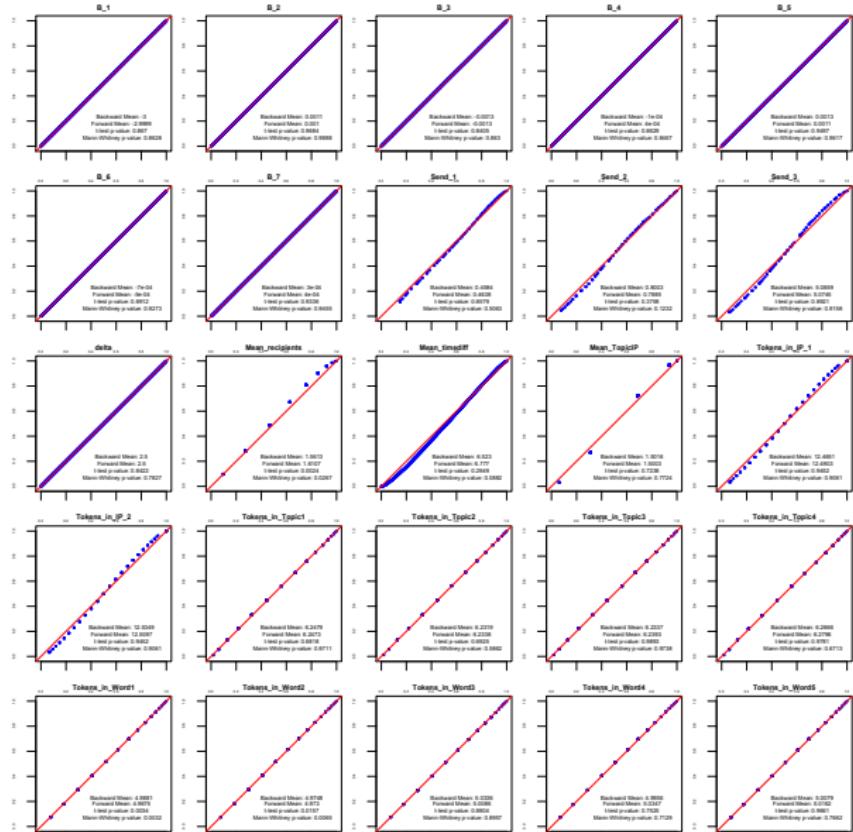
end

Getting it Right: Jointly testing math and code

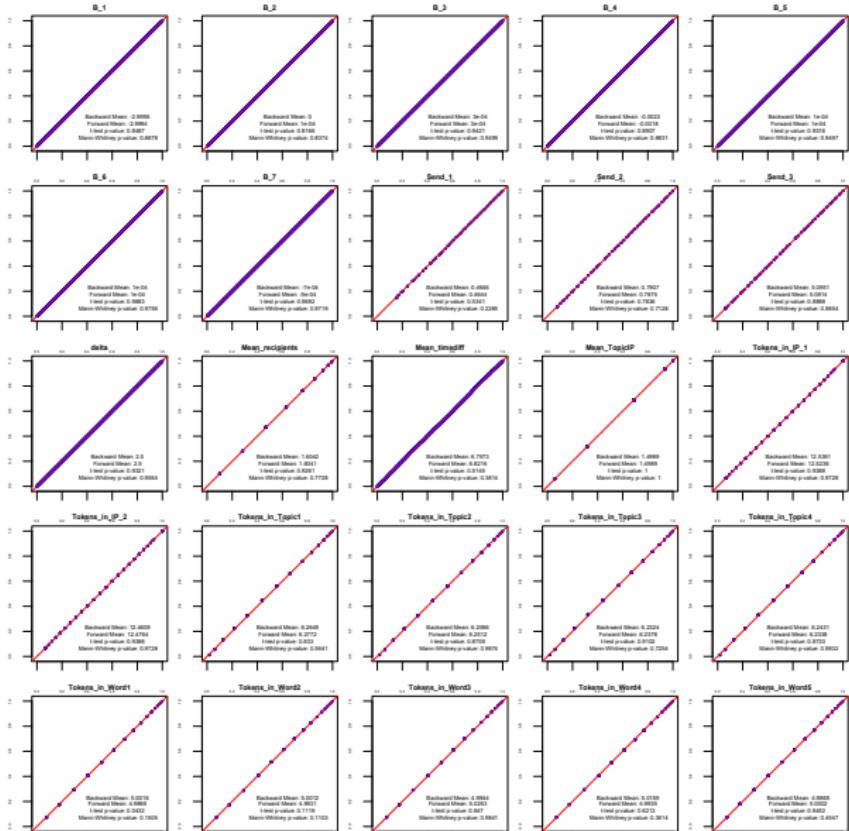
Geweke (2004) proposed a test for Bayesian posterior samplers

- *Forward samples:*
 - ① Draw parameters from prior
 - ② Draw data conditional on parameters
 - ③ Repeat
- *Backward samples:*
 - ① Start with a forward sample of data
 - ② Run inference on data
 - ③ Generate new data conditioned on inferred parameters
 - ④ Run inference on new data
 - ⑤ Repeat
- Forward samples and backward samples should match

GiR: Results with full model



GiR: Results with fixed C



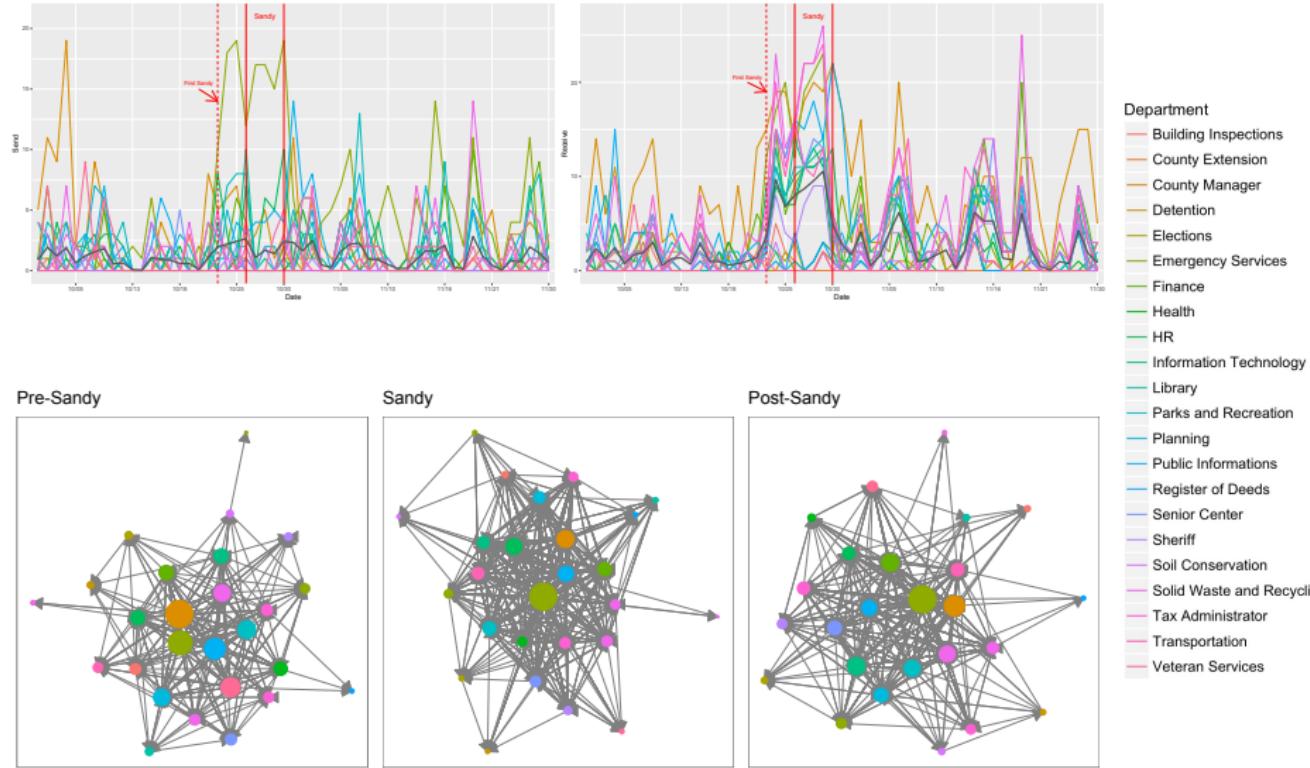
North Carolina county email data

- Dare County:
 - $D = 1456$ emails
 - $A = 27$ county government managers
 - covering 2 month period (October 1 - November 30) in 2012
- Vance County:
 - $D = 183$ emails
 - $A = 17$ county government managers
 - covering 3 month period (September 4 - November 30) in 2012
- Hurricane Sandy passed by NC: October 26 - October 30

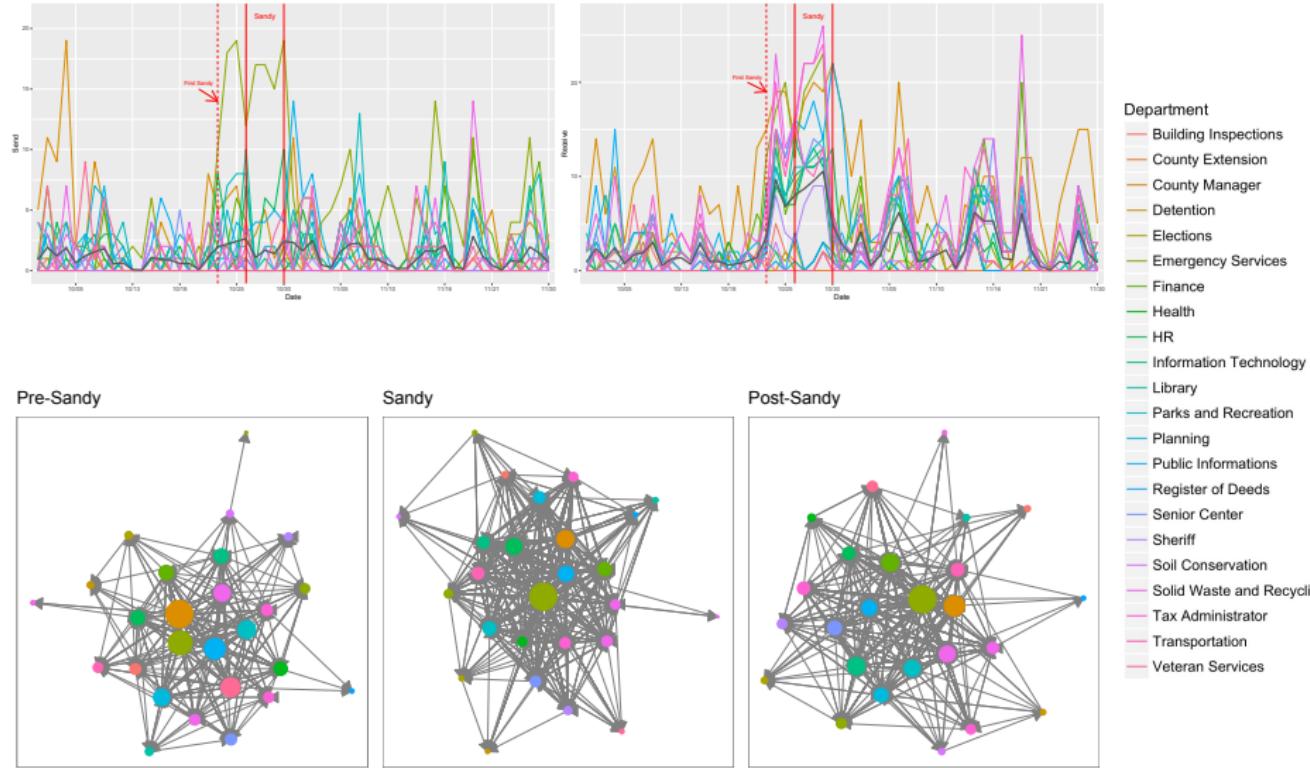


Theoretical expectations

Exploratory Data Analysis: SMALL COUNTY



Exploratory Data Analysis: DARE COUNTY



IPTM Result: Contents

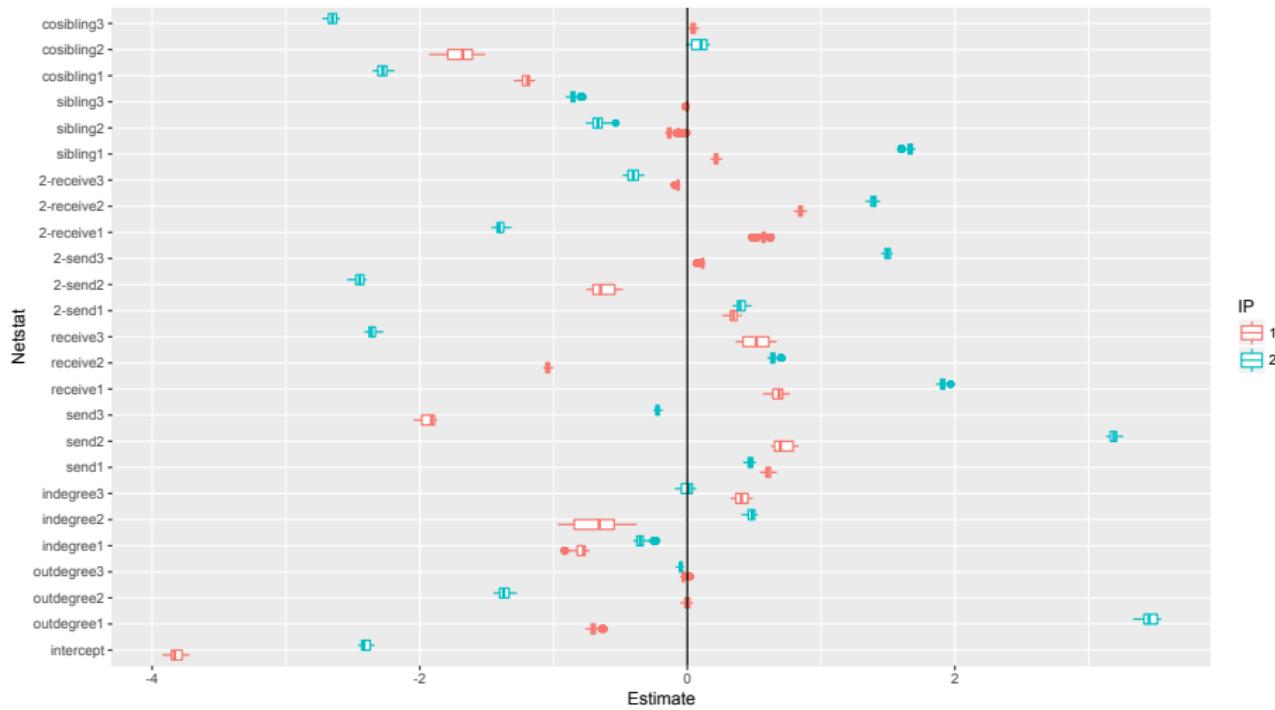
- IPTM result with $C = 2$, $K = 20$ and $O = 20^*$:

| IP | 1 | 1 | 1 | 2 | 2 | 2 |
|-------|--|--|--|---|--|---|
| Topic | 2 | 13 | 7 | 10 | 9 | 12 |
| Word | winds flooding policy mph moving outer banks rain will duration monday ocean open heads late | track offices obx shore winds exam area change continues expect curves side east better mile | offices hurricane sandy update force reading contact updates amount northwest tuesday expected good well night | sanitation billed long bill question staff vehicles additional form estimate total doors services tomorrow haterras | marshall human collins phone resources phr drive box fax bridge director monday manteo summary october | morning fema weather ems risks sure tomorrow opening address elections thought minutes starting wrote operation |

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

IPTM Result: Dynamic Network Effects

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



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

IPTM Result: Contents DARE

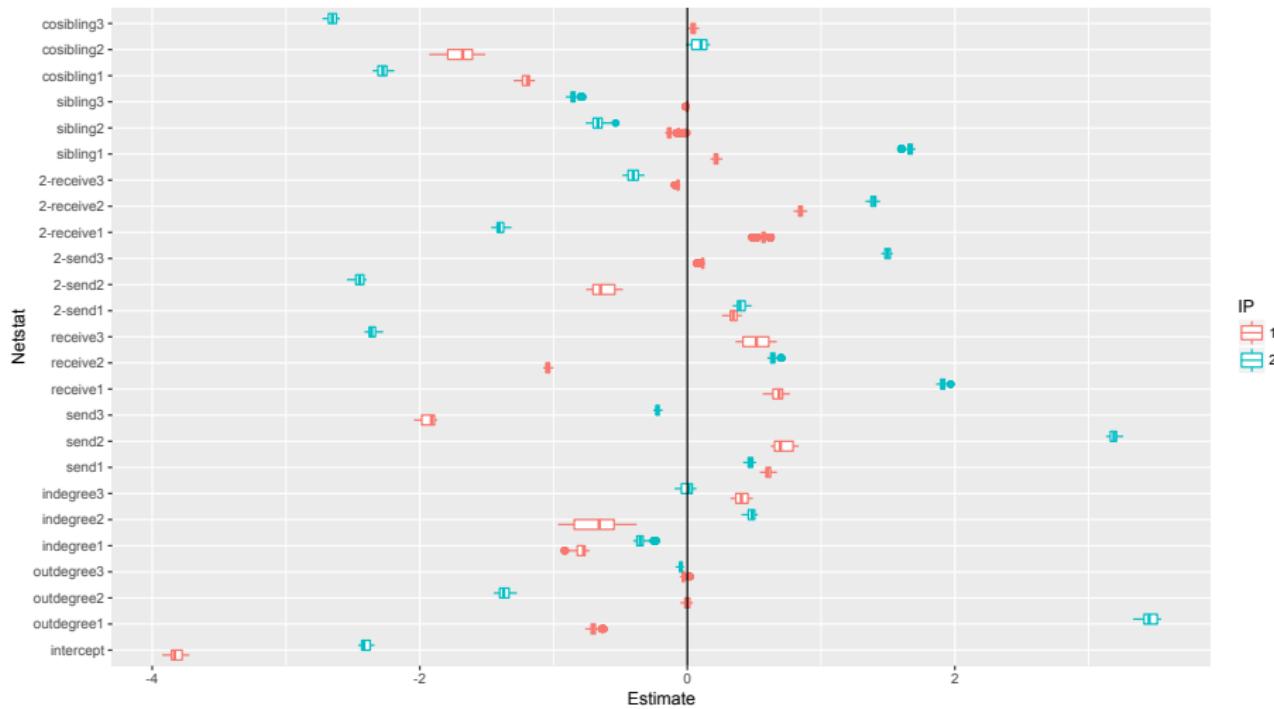
- IPTM result with $C = 2$, $K = 20$ and $O = 20^{\ddagger}$:

| IP | 1 | 1 | 1 | 2 | 2 | 2 |
|-------|--|--|--|---|--|---|
| Topic | 2 | 13 | 7 | 10 | 9 | 12 |
| Word | winds flooding policy mph moving outer banks rain will duration monday ocean open heads late | track offices obx shore winds exam area change continues expect curves side east better mile | offices hurricane sandy update force reading contact updates amount northwest tuesday expected good well night | sanitation billed long bill question staff vehicles additional form estimate total doors services tomorrow haterras | marshall human collins phone resources phr drive box fax bridge director monday manteo summary october | morning fema weather ems risks sure tomorrow opening address elections thought minutes starting wrote operation |

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

IPTM Result: Dynamic Network Effects DARE

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



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

Showing MCMC convergence

Predictive experiment design

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
- Developement of R package 'IPTM'