

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

Bomin Kim<sup>1</sup>      Aaron Schein<sup>3</sup>  
Bruce Desmarais<sup>1</sup>      Hanna Wallach<sup>2,3</sup>

<sup>1</sup> The Pennsylvania State University

<sup>2</sup> Microsoft Research NYC

<sup>3</sup> 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

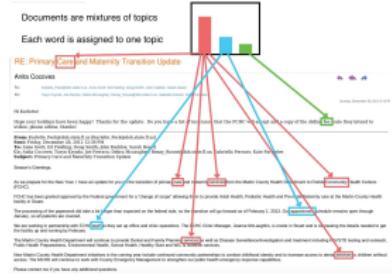
	<b>k = 1</b>	<b>k = 2</b>	<b>k = 3</b>
support	services		budget
position	care		funds
fill	child		money
staff	information		budgeted
desk	system		including
service	community		cost
customer	nurse		salary
begin	completed		amount
duties	provided		revenues
vacancy	pregnancy		debt
:	:		:
<b>IP = 1</b>	<b>IP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

- 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 <sub>i,1</sub>	2-send <sub>i,1</sub>	2-send <sub>i,1</sub>
[t-96h, t-24h)	2-send <sub>i,1</sub>	2-send <sub>i,2</sub>	2-send <sub>i,2</sub>
[t-384h, t-96h)	2-send <sub>i,1</sub>	2-send <sub>i,2</sub>	2-send <sub>i,3</sub>

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

	[t-96h, t-24h)	[t-384h, t-96h)
	2-send <sub>i,1</sub>	2-send <sub>i,1</sub>

## 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  $\delta$  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:

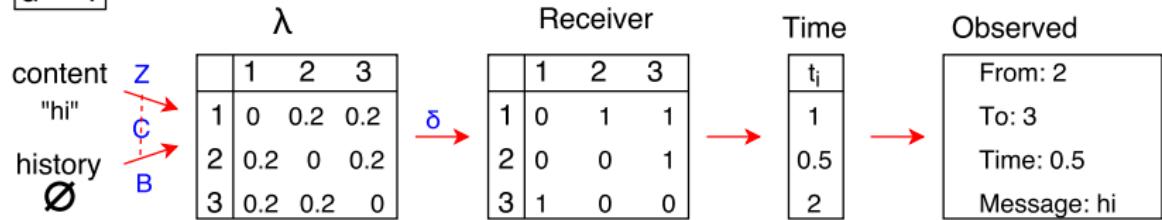
$$i^{(d)} = i_{\min(\Delta T_{iJ_i}^{(d)})}$$

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

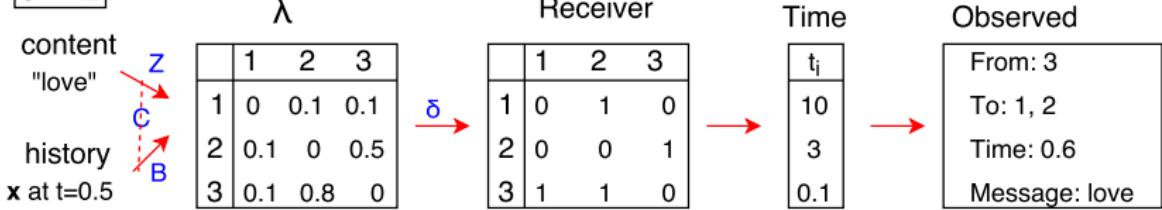
$$t^{(d)} = t^{(d-1)} + \min(\Delta T_{iJ_i}^{(d)})$$

# Joint Generating Process

$d = 1$



$d = 2$



# Inference

- Take a Bayesian approach to inference
- $\mathcal{B}$  and  $\delta$  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**  $\delta$  via Metropolis-Hastings

**end**

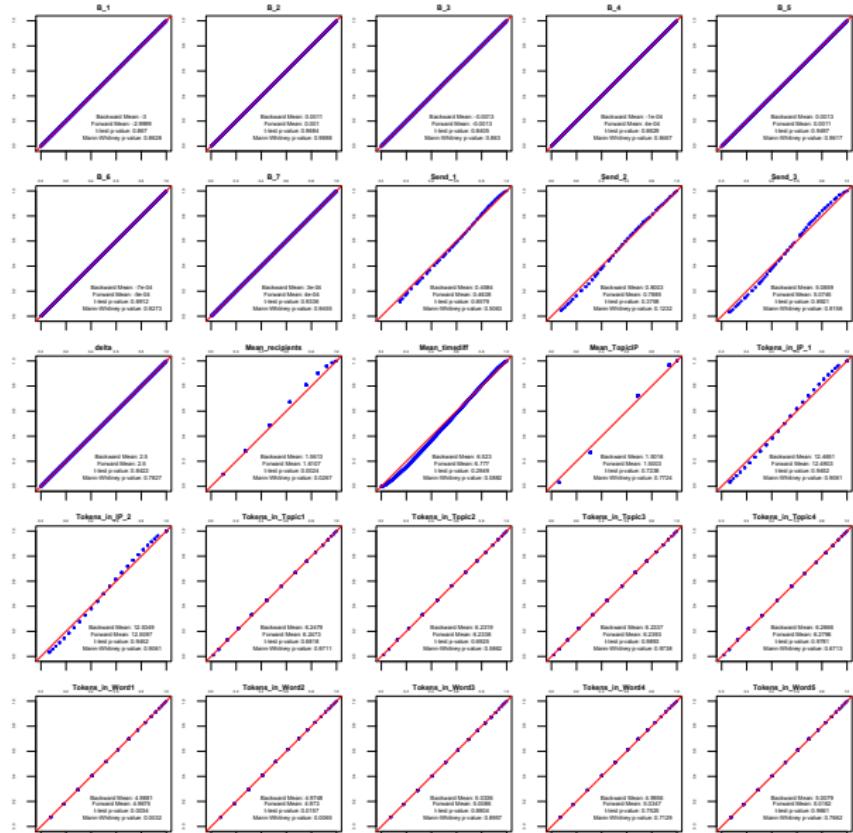
---

# Getting it Right: Jointly testing math and code

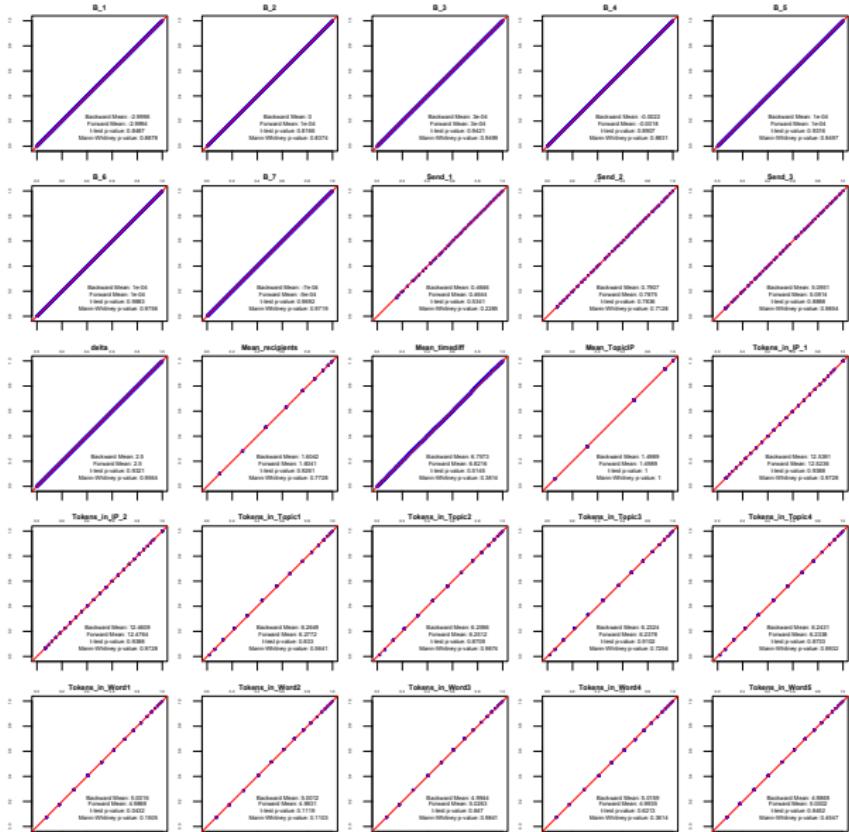
Geweke (2004) proposed a test for Bayesian posterior samplers

- *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

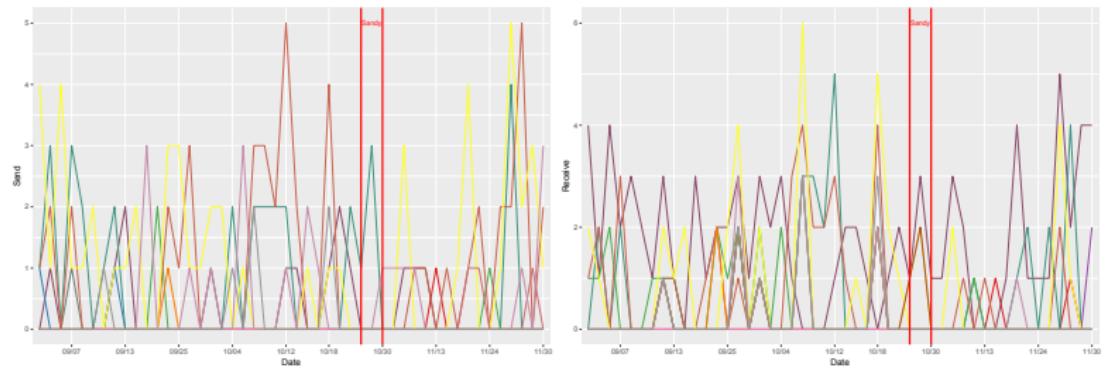
- Dare County:
  - $D = 1456$  emails
  - $A = 27$  county government managers
  - covering 2 month period (October 1 - November 30) in 2012
- Vance County:
  - $D = 183$  emails
  - $A = 17$  county government managers
  - covering 3 month period (September 4 - November 30) in 2012
- Hurricane Sandy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



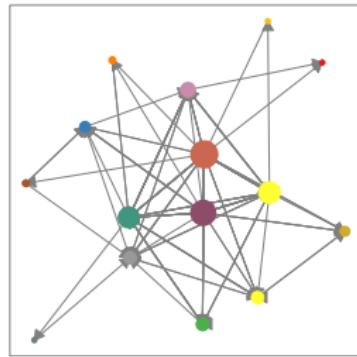
## Theoretical considerations

- Personal/friendship topics exhibit reciprocity and transitivity
- Professional communications avoid loops
- Sandy communications represent pattern breakdowns

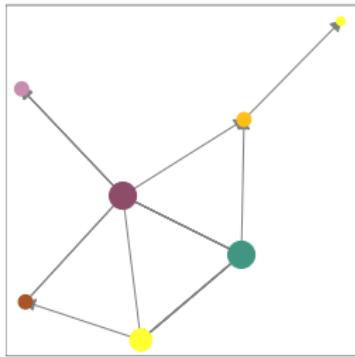
# Exploratory Data Analysis: Vance County



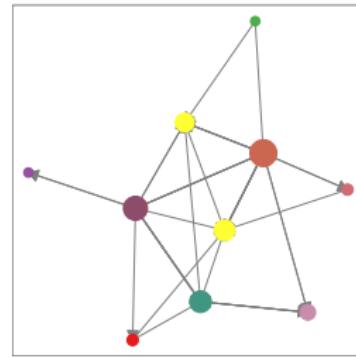
Pre-Sandy



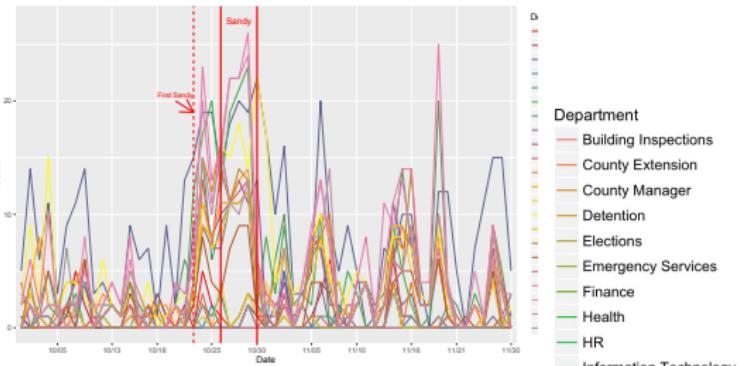
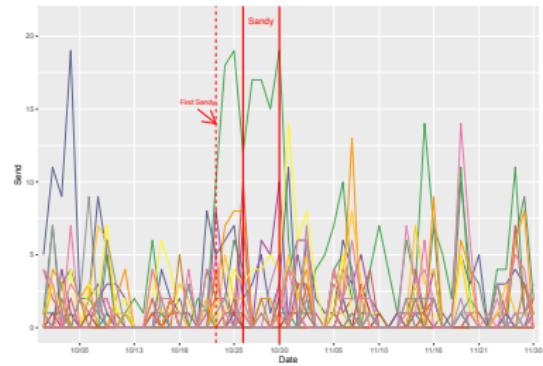
Sandy



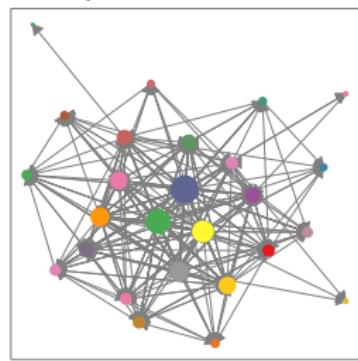
Post-Sandy



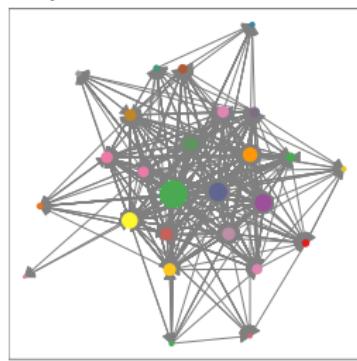
# Exploratory Data Analysis: Dare County



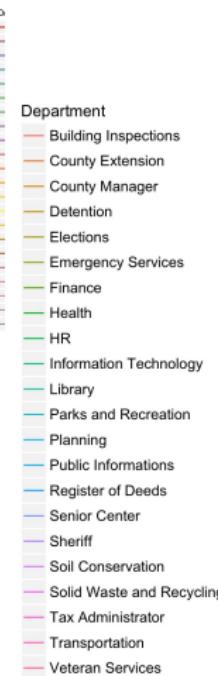
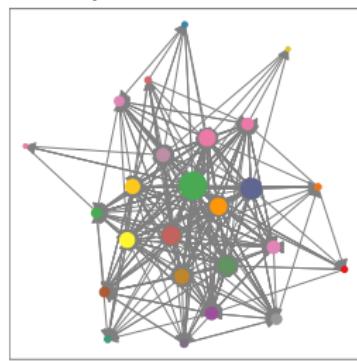
Pre-Sandy



Sandy



Post-Sandy



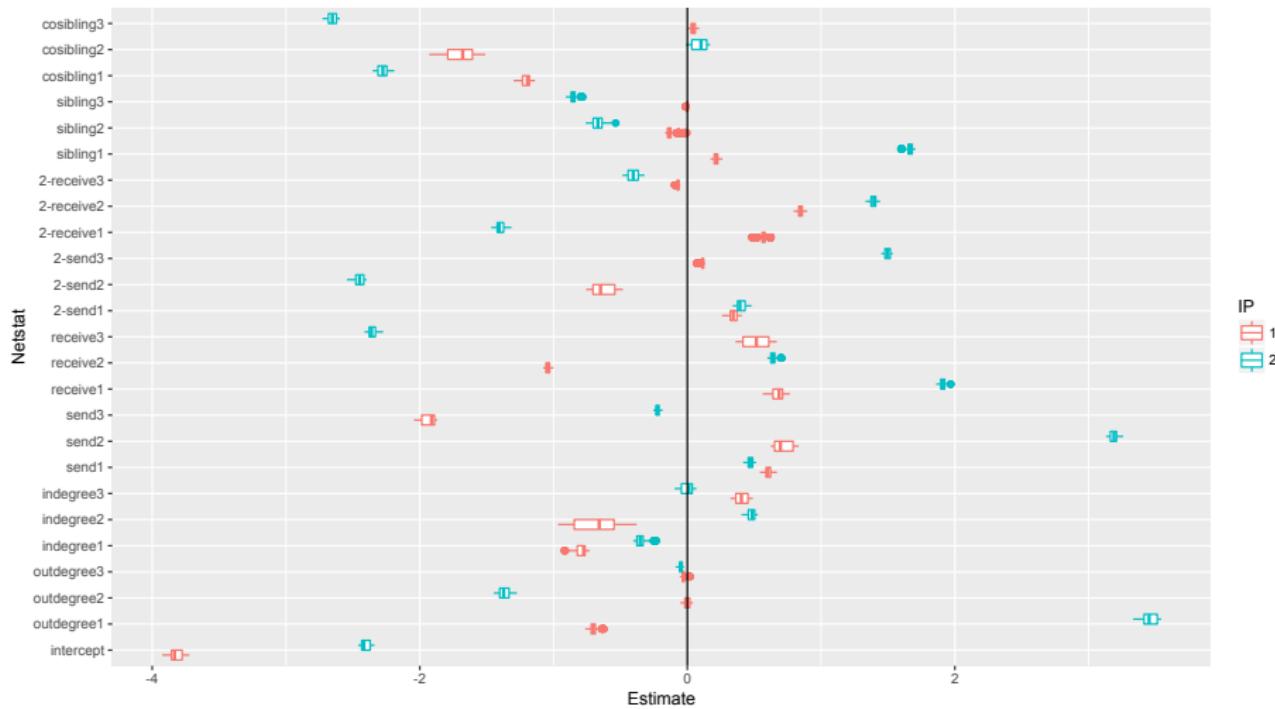
# IPTM Result: Topics, Vance County

$C = 2$ ,  $K = 20$  and  $O = 500$

IP	1	1	1	2	2	2
Topic	1 (0.078)	3 (0.224)	5 (0.141)	2 (0.139)	4 (0.369)	6 (0.038)
Word	directory switch network address extension tax latest department henderson january young wireless installation cutting rest	message electronic ncgs chapter response public manager attachments siemens pursuant subject review records jail hereto	operations emergency office communications center lines fax enp cem suite good asap henderson street church	phase description planning board water taps keep phone compliance signups meter suite fax meeting tuesday	dropbox cecd henderson-vance box commission economic unemployment licensed rural development reduced financial private labor force	phones october will polycom intstructions training conference three finalized room cutover contact thursday folks finishing

# IPTM Result: Dynamic Network Effects

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



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

# IPTM Result: Topics, Dare County (IP 1)

$$C = 2, K = 20 \text{ and } O = 100$$

IP	1	1	1	1	1
Topic	3 (0.078)	5 (0.065)	19 (0.064)	13 (0.057)	15 (0.053)
Word	water relocation location hills utilities mustian hydrant department skyco kill devil road lane tank map	planning meter room asked needed sure afternoon cheryl johnson issues case letter antennas inspection keep	phone collins drive marshall director human resources manteo phr fax box timesheets -lsb- wanted touch	questions board december call sheets agenda nov hope item weekly management internet told care comp	contact info problem release check weather priority readings rodanthe top collection located health heads ahead

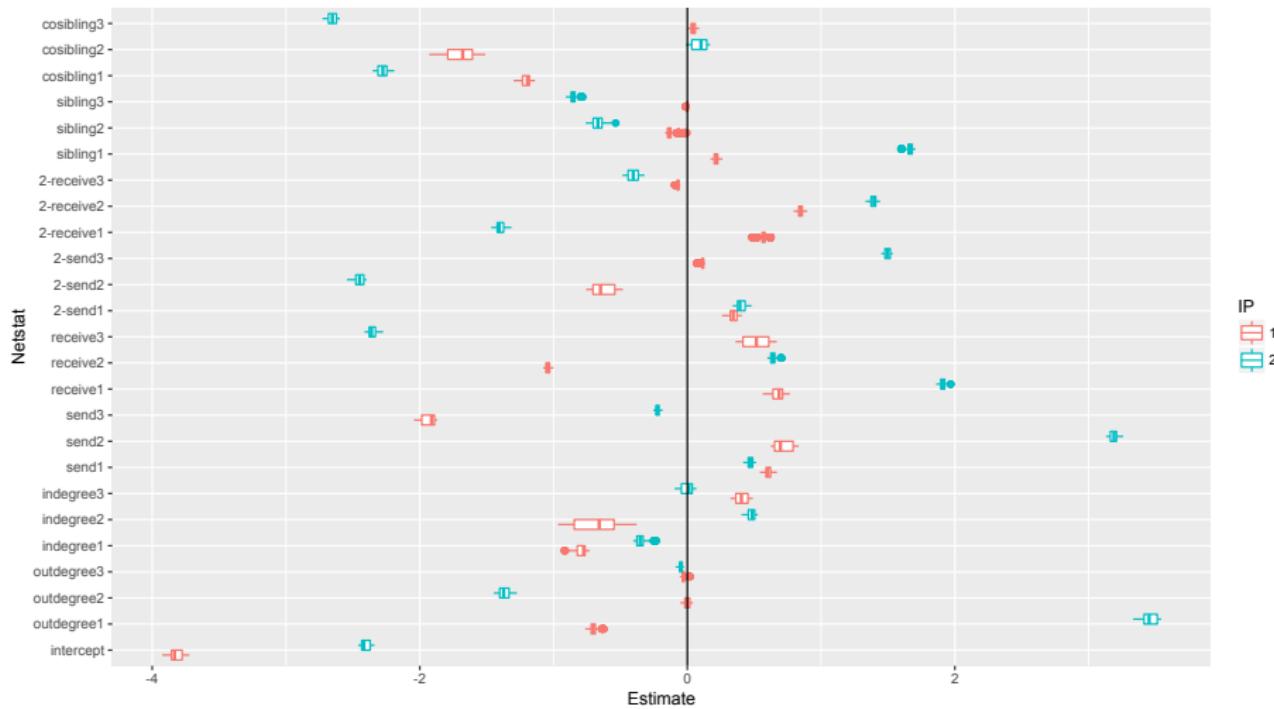
# IPTM Result: Topics, Dare County (IP 2)

$$C = 2, K = 20 \text{ and } O = 100$$

IP	2	2	2	2	2
Topic	14 (0.058)	12 (0.047)	2 (0.045)	6 (0.044)	18 (0.036)
Word	time hours leave monday administrative employees employee work day friday october storm tomorrow hour question	survey voice copy discovery regional parties disclosed elections pin sending prior editor students cost residents	road mirlo storm beach high coastal impacts saturday dot night winds hold bridge expressed normal	library week working place best start visit year albemarle librarian web learning east holiday system	status system area south forecast track pay move assessment opens damage well ocean operation addition

# IPTM Result: Dynamic Network Effects DARE

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



$\dagger$ Preliminary results with small outer iterations. Model results subject to change.

# Model fit evaluation

- Forecast topics, ties, and timing of next document
- Compare to one or more models that can generate same predictions

---

**Algorithm 2** Predicting tie data for the next document

---

Input

- ①  $O$ , number of outer iterations of inference from which to generate predictions
- ②  $d$ , the last document to use in inference
- ③  $R$ , the number of iterations to sample predicted data within each outer iteration

Run burnin iterations

```
for o=1 to O do
    run an outer iteration of inference on documents 1 through d
    initialize values for  $i^{(d+1)}$ ,  $J^{(d+1)}$ ,  $t^{(d+1)}$ , and  $\mathcal{Z}^{d+1}$ 
    for r=1 to R do
        sample  $i^{(d+1)}$ ,  $J^{(d+1)}$ , and  $t^{(d+1)}$  conditional on  $\mathcal{Z}^{d+1}$ , via the generative
        process
        sample  $\mathcal{Z}^{d+1}$  via Equation 24
    end
    store  $i^{(d+1)}$ ,  $J^{(d+1)}$ ,  $t^{(d+1)}$ , and  $\mathcal{Z}^{d+1}$ 
end
```

---

# 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'

# Convergence of Markov Chains