

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

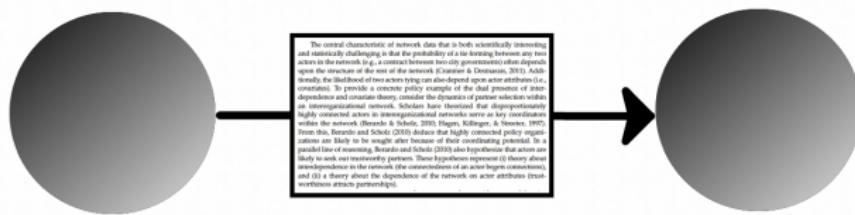
October 12, 2017

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



Motivation

- ▶ Ties attributed with text
 - ▶ International treaties
 - ▶ Legislative cosponsorship
 - ▶ Discussion networks on social media



- ▶ Network models can't model text
- ▶ Models for text...
 - ▶ not designed for networks
 - ▶ 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

- ▶ 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 $N^{(d)}$:

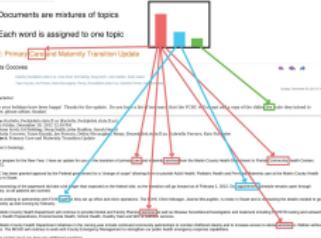
- (a) Choose a topic from document-topic distribution

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

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

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

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



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

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 :

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

which is a mixture of contents 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' (indicated by a red dotted line), 'stochastic intensity' (indicated by a red arrow), and 'history of interactions' (indicated by a red arrow). These three components point to a binary matrix on the right. The matrix has a header row 'i | 1 2 3 4 A' and a header column 'i'. The rows are labeled 1, 2, ..., A. The matrix contains binary values (0 or 1) representing the presence or absence of ties between senders and receivers. For example, row 1 has values [0, 1, 0, 1, ..., 1], row 2 has [1, 0, 0, 0, ..., 0], and so on.

| 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

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

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

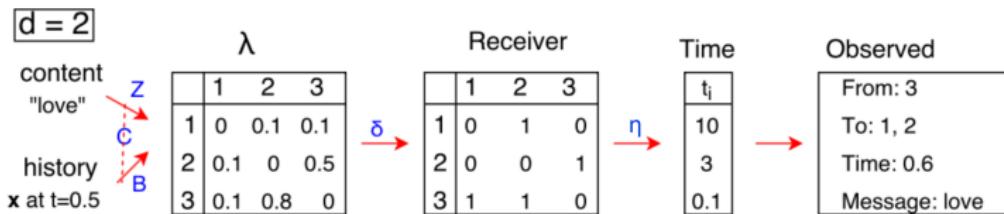
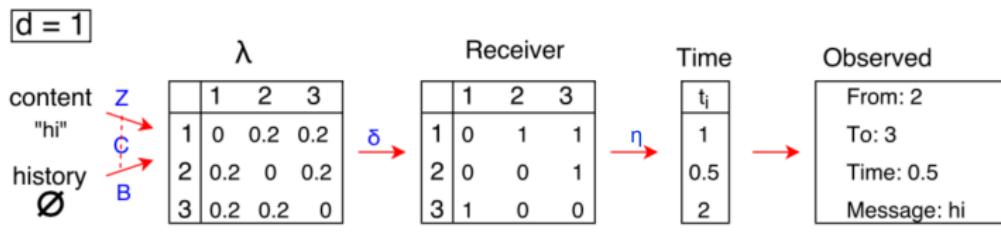
4. Set the observed sender, receivers and timestamp simultaneously:

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

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

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

Joint generating process



Inference

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

Algorithm 1 MCMC

Set initial values $\mathcal{Z}^{(0)}, \mathcal{C}^{(0)}$, and $(\mathcal{B}^{(0)}, \delta^{(0)}, \eta^{(0)})$

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
Sample the **time-increment parameter** η via Metropolis-Hastings

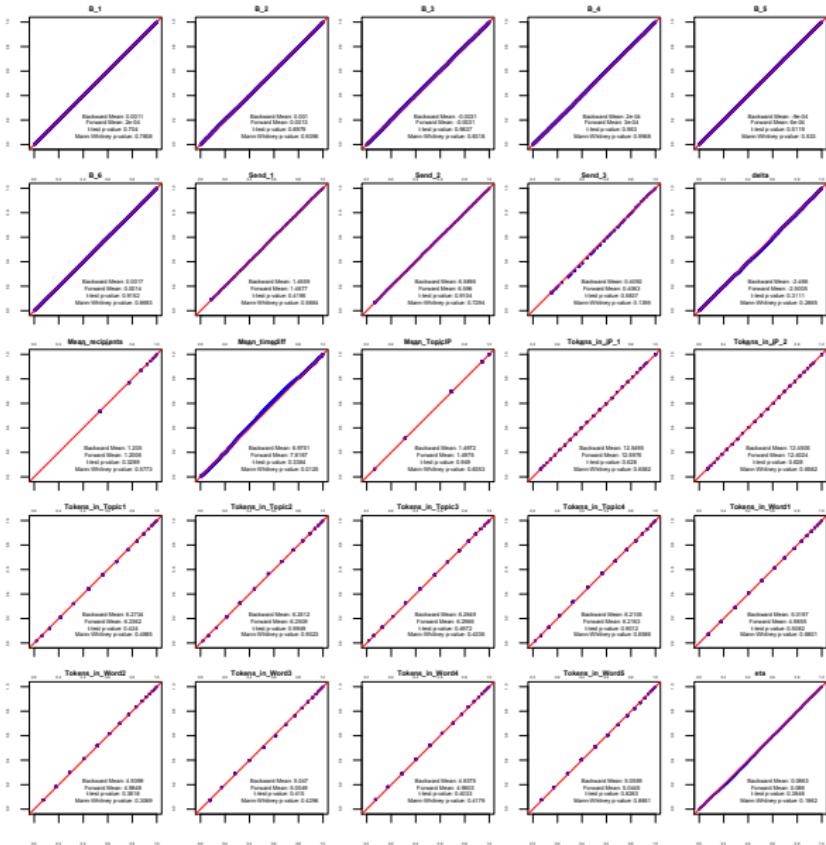
end

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



Dare County, NC email data

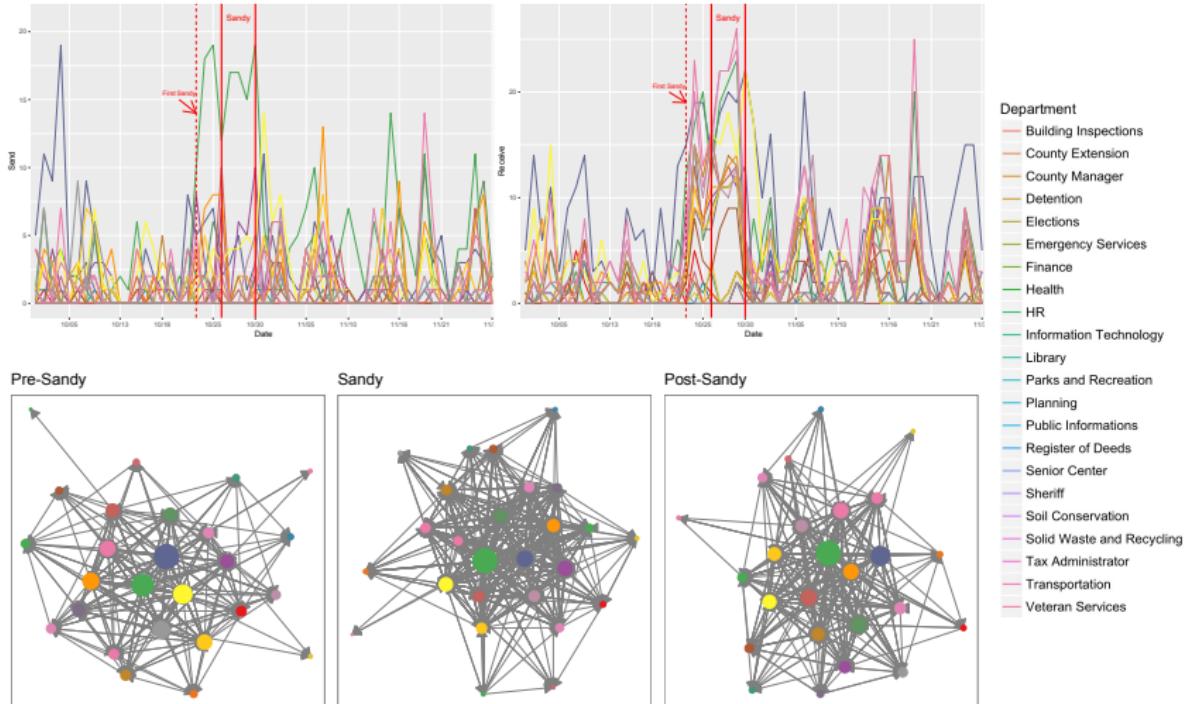


- ▶ $D = 2210$ emails
- ▶ $A = 27$ county government managers
- ▶ $W = 2907$ unique words
- ▶ covering 3 month period (September 1 - November 30) in 2012
- ▶ Hurricane Sandy passed by NC: October 26 - October 30

Theoretical considerations

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

Exploratory data analysis: Dare County



IPTM result: topics (IP=1)

$C = 2$, $K = 25$ and $O = 100$

| IP | 1 | 1 | 1 | 1 | 1 |
|-------|--|---|---|--|---|
| Topic | 23 (0.055) | 18 (0.054) | 11 (0.048) | 20 (0.033) | 8 (0.032) |
| Word | change order manager storm emergency coastal statute evacuation track couple changes well concerns things consistent boat misdemeanor program system powerpoint | will winds location beach hydrant water relocation mirlo road high moving gas forecast saturday project outer map airport called banks | sandy munis hurricane position monday point power update storm hey release weeks weekend working month problems strong impacts three ncdot | time hours monday leave employees timesheets storm employee tomorrow work regular period comp sheets vacation administrative operation timesheet personnel will | services public white director fyi tim update status board approval wanted rec adult older today storm reminder called sep charlotte |

IPTM result: topics (IP=2)

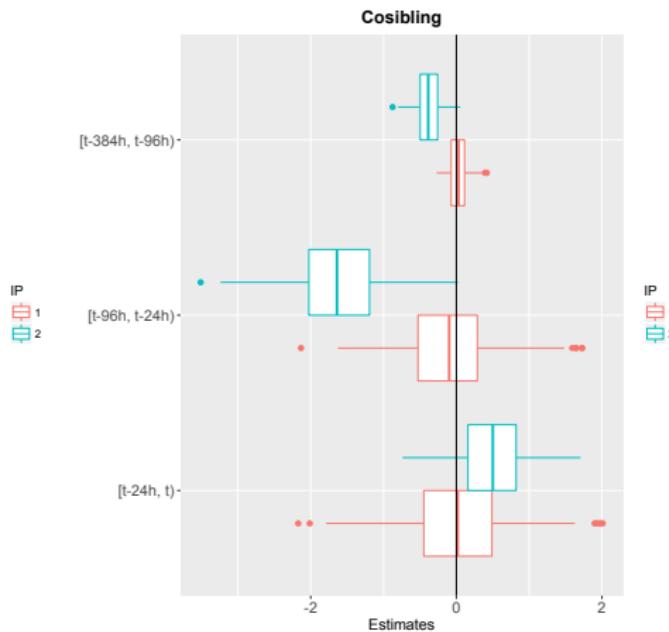
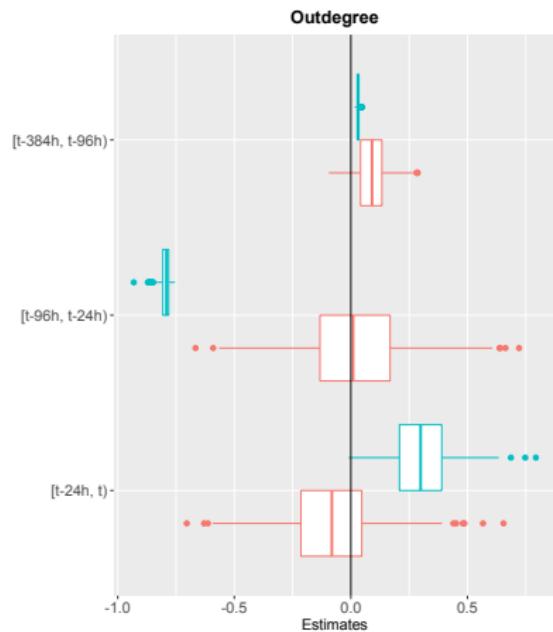
$$C = 2, K = 25 \text{ and } O = 100$$

| Topic | 21 (0.057) | 12 (0.037) | 14 (0.037) | 4 (0.035) | 10 (0.026) |
|--------------|---|--|---|---|---|
| Word | library best start web place visit albermarle east regional system voice librarian learning discovery manteo expressed views box conversion director | marshall collins drive manteo phone box fax resources director human phr policy time october hire will offices approved hour told | will center day office great manager assistance problem lot call currently september early parking received monday baum questions year today | board meeting property planning will notice january review commissioners list thoughts amendment members dot budget building inspections night text copy | survey request call sure people emergency thought seafood mail grant ncacc community check rodanthe district complete option best residents talk |

IPTM result: example email

- ▶ From: 11 (Public Infromations)
- ▶ To: 4 (Emergency Services) and 18 (Detention)
- ▶ At: 25 Oct 2012 19:14:47
- ▶ Content: will will storm good sure sandy morning morning morning
morning send change plan hurricane wanted copy base afternoon
keep description description saturday saturday saturday saturday
touch touch release signature plans prior ready duration submitted
eoc running exactly activation activation released mind joint jis
pasted activate

IPTM result: dynamic network effects, Dare County



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

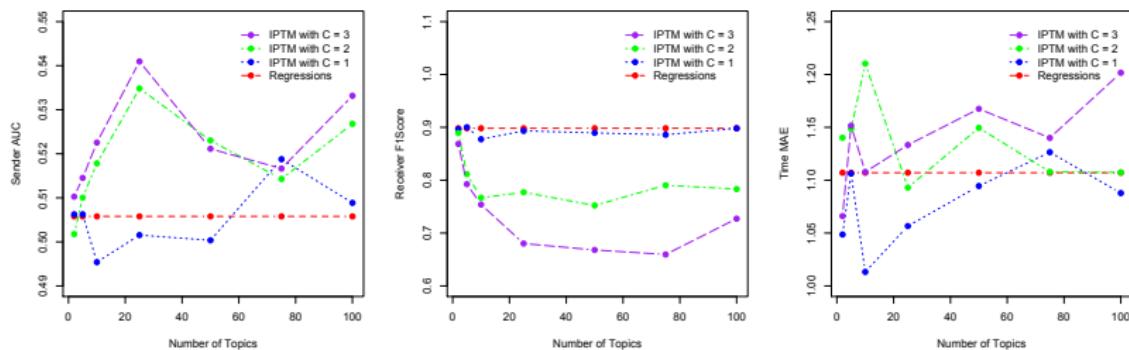
end

 store $i^{(d+1)}$, $J^{(d+1)}$, $t^{(d+1)}$, and \mathcal{Z}^{d+1}

end

Model fit evaluation

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



- ▶ Further modifications to improve model fits

Conclusion

- ▶ Joint modeling of ties (sender, receiver, time) and contents
- ▶ Contribution in distribution for non-empty multicast
- ▶ Many potential applications in political science
- ▶ Developement of R package 'IPTM'