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

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

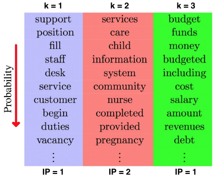
Interaction-Partitioned Topic Model (IPTM)

- 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 over the word types
  2. Choose a topic-interaction pattern assignment



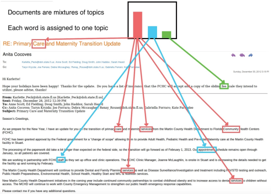
- For each document $d = 1, ..., D$:

  3-1. Choose a document-topic distribution

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

    (a) Choose a topic from document-topic distribution

    (b) Choose a word from topic-word distribution



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

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

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

Conditioning feagures on recency

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

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

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

h → j

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

Tie Generating Process: Receivers

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

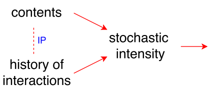
$$\lambda_{ij}^{(d)} = \sum_{c=1}^{C} p_c^{(d)} \cdot \text{exp}\Big\{\boldsymbol{b}_0^{(c)} + \boldsymbol{b}^{(c)T}\boldsymbol{x}_{t^{(d-1)}}^{(c)}(i,j)\Big\},$$

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

2. 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)

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

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



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 $i \in \{1, ..., A\}$, generate the time increments for document d

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

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

$\boxed{d = 1}$

content "hi"

history $\varnothing$

Z C B

λ

	1	2	3
1	0	0.2	0.2
2	0.2	0	0.2
3	0.2	0.2	0

$\xrightarrow{\delta}$

Receiver

	1	2	3
1	0	1	1
2	0	0	1
3	1	0	0

$\rightarrow$

Time

t_i
1
0.5
2

$\rightarrow$

Observed

From: 2
To: 3
Time: 0.5
Message: hi

$\boxed{d = 2}$

content "love"

history **x** at t=0.5

Z C B

λ

	1	2	3
1	0	0.1	0.1
2	0.1	0	0.5
3	0.1	0.8	0

$\xrightarrow{\delta}$

Receiver

	1	2	3
1	0	1	0
2	0	0	1
3	1	1	0

$\rightarrow$

Time

t_i
10
3
0.1

$\rightarrow$

Observed

From: 3
To: 1, 2
Time: 0.6
Message: love

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

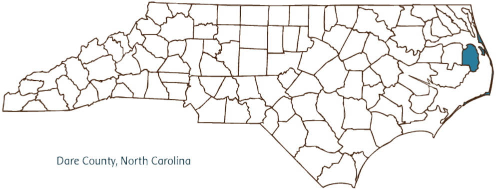
- *Forward samples*:
  1. Draw parameters from prior
  2. Draw data conditional on parameters
  3. Repeat
- *Backward samples*:
  1. Start with a forward sample of data
  2. Run inference on data
  3. Generate new data conditioned on inferred parameters
  4. Run inference on new data
  5. Repeat
- Forward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

GiR: Results with full model

GiR: Results with fixed C

Data: North Carolina Dare county email data

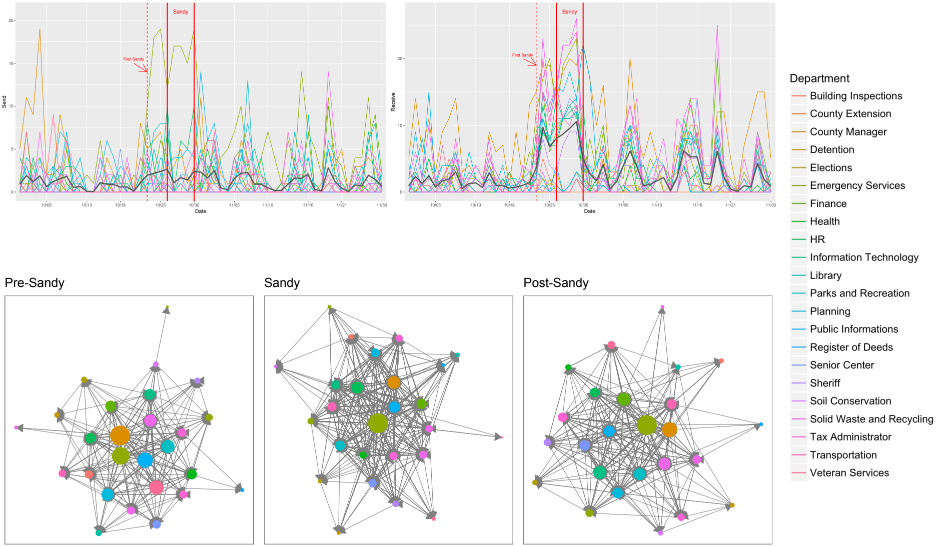
- $D = 1456$ emails between $A = 27$ county government managers, covering 2 month periods (October 1 - November 30) in 2012



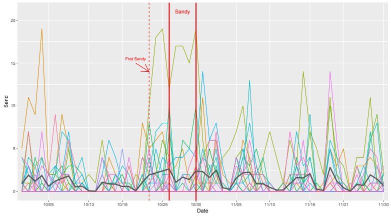
- Hurricane Sandy passed by NC: October 26 - October 30

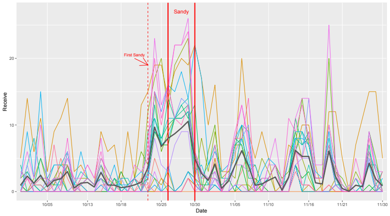
Theoretical expectations

Exploratory Data Analysis: SMALL COUNTY

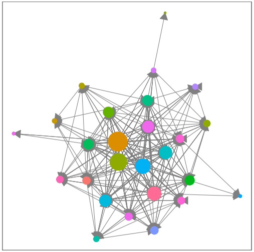


Exploratory Data Analysis: DARE COUNTY

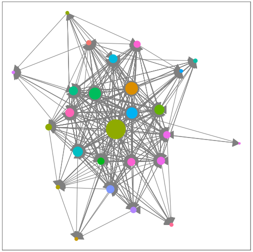




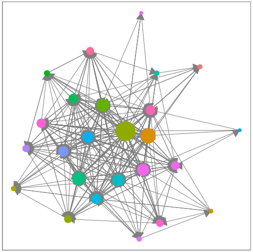
Pre-Sandy



Sandy



Post-Sandy



IPTM Result: Contents

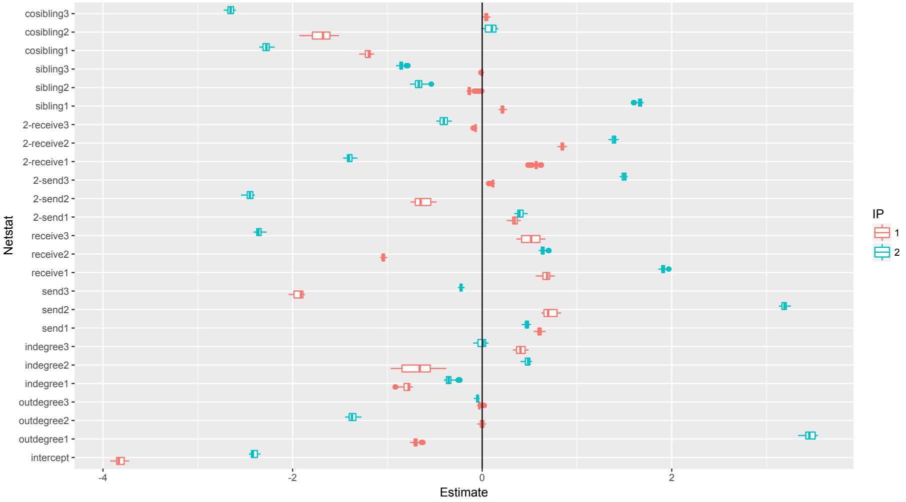
- IPTM result with $C = 2$, $K = 20$ and $O = 20^*$:

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

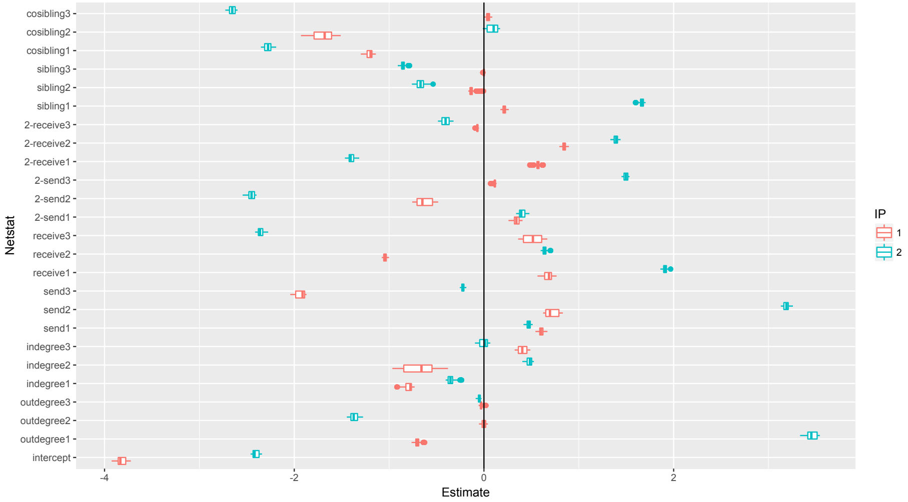
- IPTM result with $C = 2$, $K = 20$ and $O = 20^{\ddagger}$:

IP	1	1	1	2	2	2
Topic	**2**	**13**	**7**	**10**	**9**	**12**
Word	winds	track	offices	sanitation	marshall	morning
	flooding	offices	hurricane	billed	human	fema
	policy	obx	sandy	long	collins	weather
	mph	shore	update	bill	phone	ems
	moving	winds	force	question	resources	risks
	outer	exam	reading	staff	phr	sure
	banks	area	contact	vehicles	drive	tomorrow
	rain	change	updates	additional	box	opening
	will	continues	amount	form	fax	address
	duration	expect	northwest	estimate	bridge	elections
	monday	curves	tuesday	total	director	thought
	ocean	side	expected	doors	monday	minutes
	open	east	good	services	manteo	starting
	heads	better	well	tomorrow	summary	wrote
	late	mile	night	haterras	october	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 $O = 20^{\S}$:



§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'