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 18, 2017

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

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

Probability

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

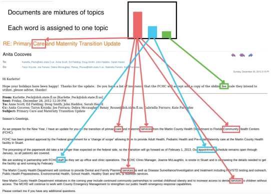
- For each document $d = 1, ..., D$:

  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:

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

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 $(l = 1, 2, 3)$

- $\boldsymbol{x}_{t,l}^{(c)}(i, j)$ 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	$(i \rightarrow \forall j)$	**send**	$(i \rightarrow j)$	**2-send**	$\sum_h (i \rightarrow h \rightarrow j)$	**sibling**	$\sum_h (h \rightarrow i, h \rightarrow j)$
indegree	$(i \leftarrow \forall j)$	**receive**	$(i \leftarrow j)$	**2-receive**	$\sum_h (i \leftarrow h \leftarrow j)$	**cosibling**	$\sum_h (h \leftarrow i, h \leftarrow j)$

Tie Generating Process: Latent Edges

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:

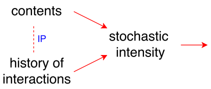
$$\lambda_{ij}^{(d)} = \sum_{c=1}^{C} p_c^{(d)} \cdot \exp\Big\{\boldsymbol{b}_0^{(c)} + \boldsymbol{b}^{(c)T}\boldsymbol{x}_{t_{(d-1)}}^{(c)}(i,j)\Big\},$$

   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)

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

   where δ is a real-valued intercept controlling recipient size parameter and $Z(\delta, \log(\lambda_i^{(d)})) = (\prod_{j \in \mathcal{A}_{\setminus i}} (\exp\{\delta + \log(\lambda_{ij}^{(d)})\} + 1)) - 1$ 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

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

$$\Delta T^{(d)}_{iJ_i} \sim \text{Exponential}(\lambda^{(d)}_{iJ_i}),$$

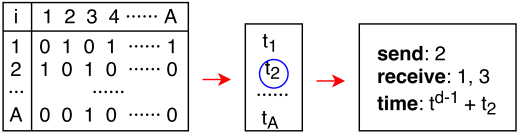
where $\lambda^{(d)}_{iJ_i} = \sum_{c=1}^{C} p^{(d)}_c \cdot \exp\Big\{\lambda^{(c)}_0 + \frac{1}{|J_i|}\sum_{j \in J_i} \boldsymbol{b}^{(c)T}\boldsymbol{x}^{(c)}_{t^{(d-1)}}(i,j)\Big\}$ is the updated sender-specific stochastic intensity given the receivers.

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

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

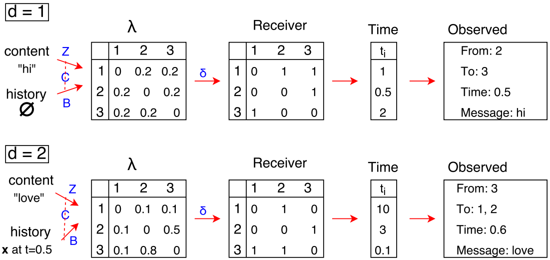
$$J^{(d)} = J_{i^{(d)}}$$

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



Joint Generating Process and Bayesian Inference

- Joint Generating Process



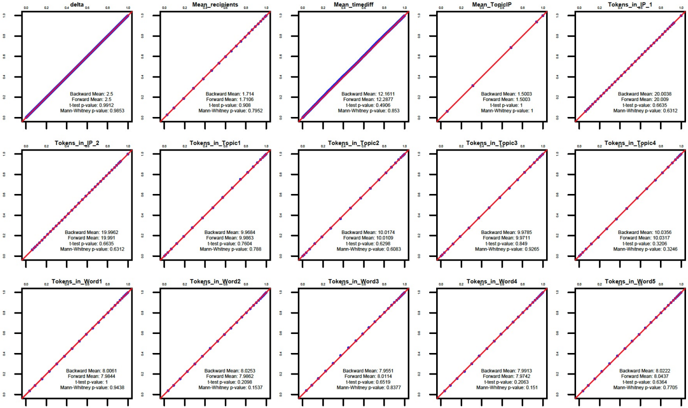
- Bayesian Inference using Markov Chain Monte Carlo (MCMC)

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 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

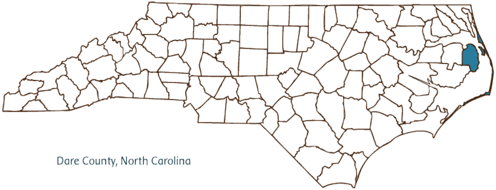
Joint Distribution Tests: Getting it Right (Geweke, 2004)

- Forward sampling: draws parameters from the prior and then generate data conditional on the parameters
- Backward sampling: estimate the parameters from inference algorithm and then generate data conditional on the posterior estimates



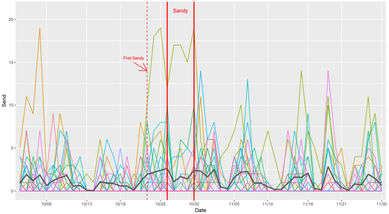
Data: North Carolina Dare county email data

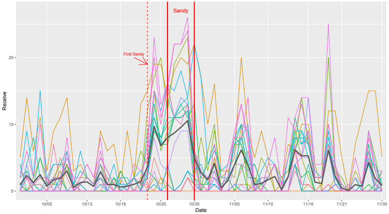
- $D = 1456$ emails between $A = 27$ county government managers, covering 2 month periods (October 1 - November 30) in 2012



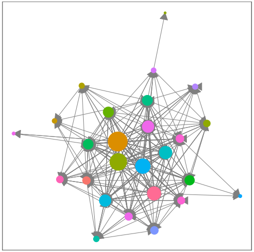
- Hurricane Sandy passed by NC: October 26 - October 30

Exploratory Data Analysis: Effect of Sandy

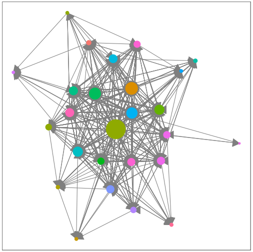




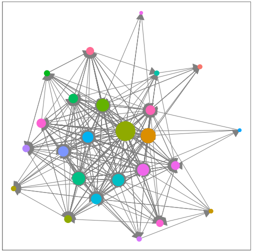
Pre-Sandy



Sandy



Post-Sandy



IPTM Result: Contents

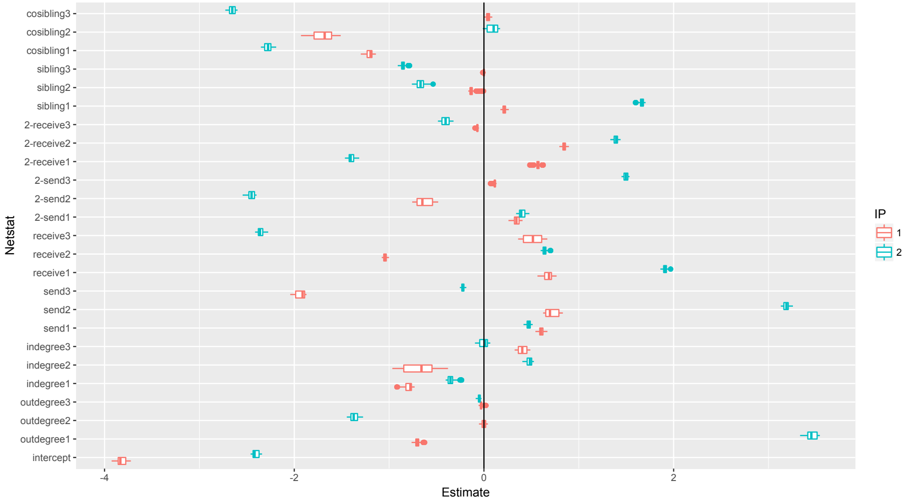
- IPTM result with $C = 2$, $K = 20$ and $O = 20^*$:

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.

Conclusion

- Interaction Pattern (IP): cluster of topics that share network properties
- 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'