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

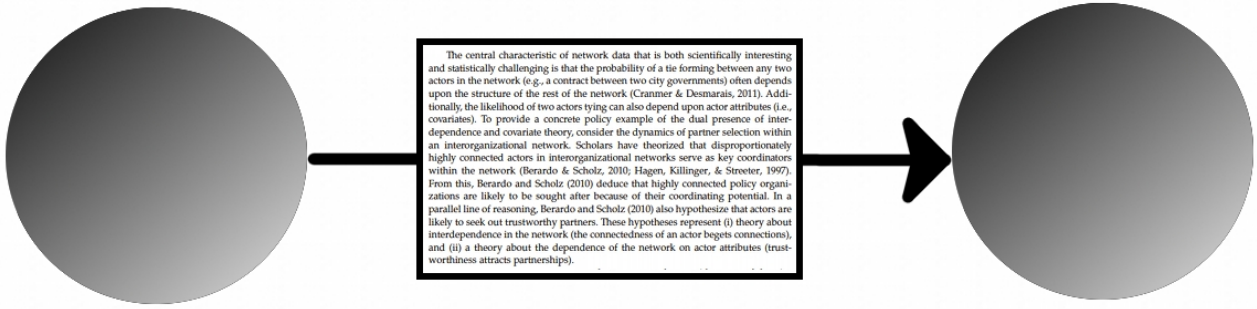
[3] University of Massachusetts Amherst

October 11, 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)

- 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

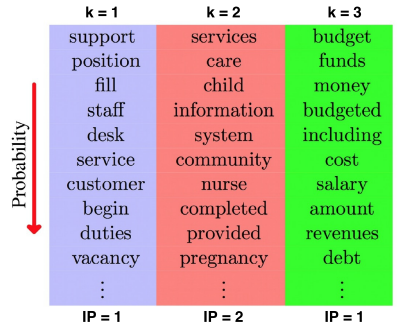
"who communicates with whom about what, and when?"

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



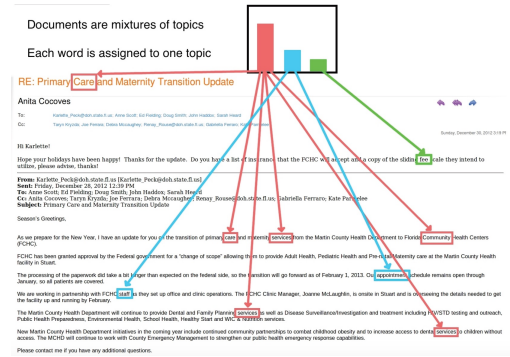
- For each document $d = 1, ..., D$:

  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:

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

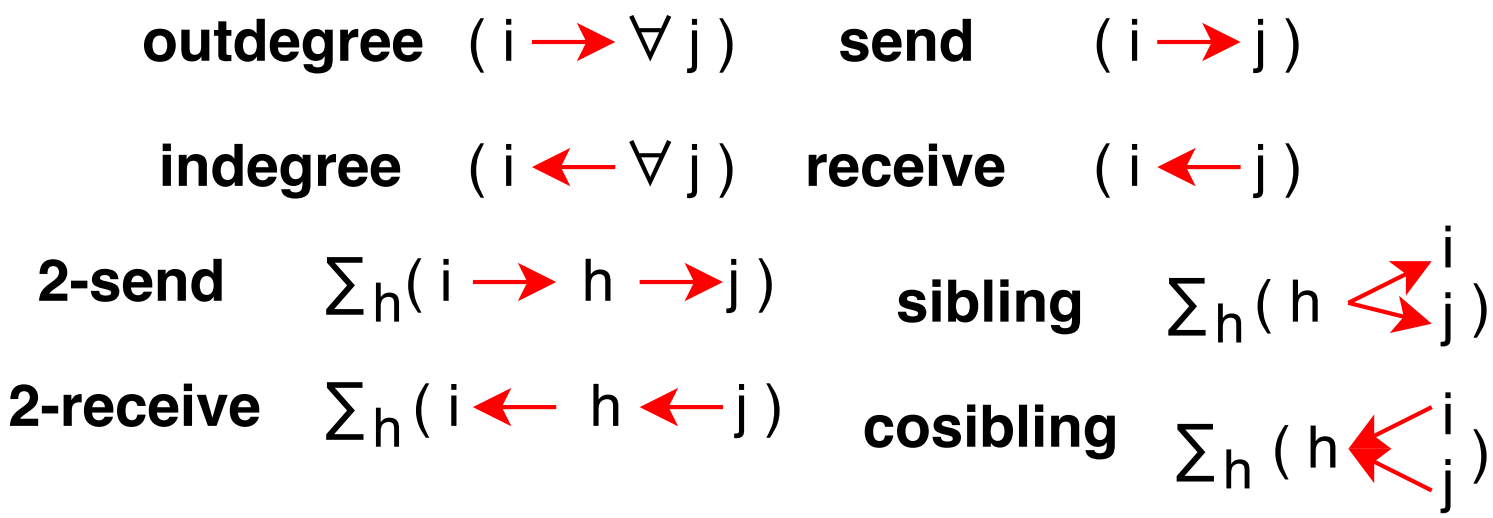
Network Model Components

- 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

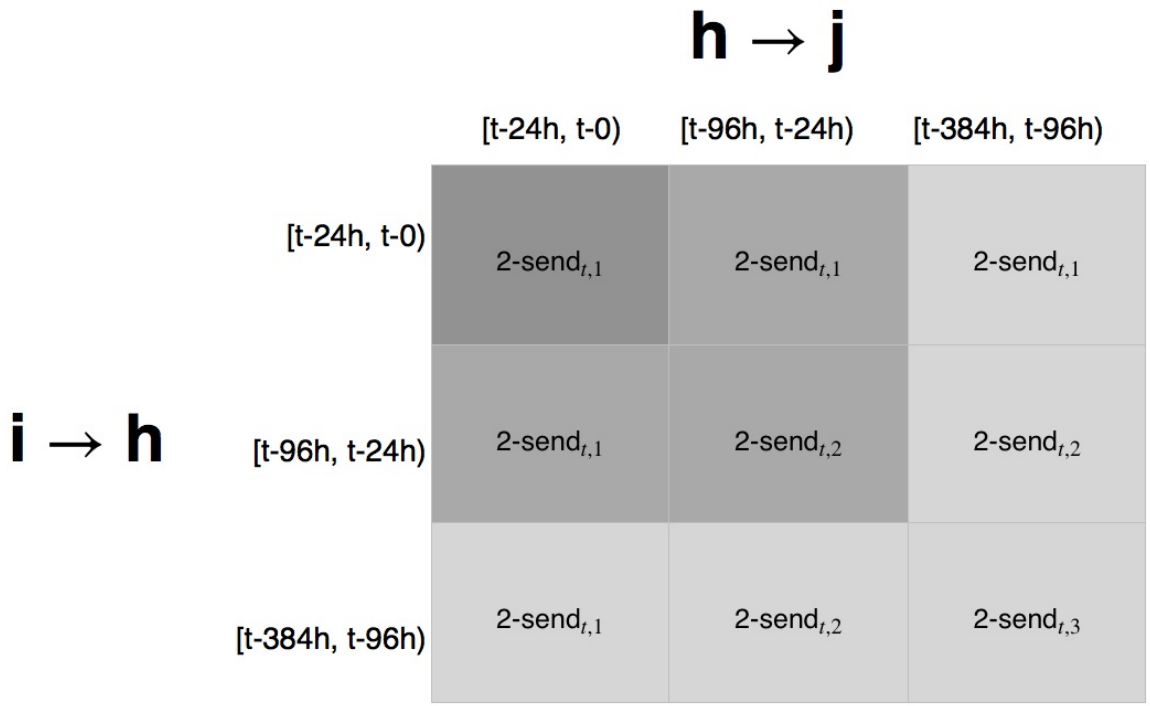


Conditioning features on recency

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

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

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



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:

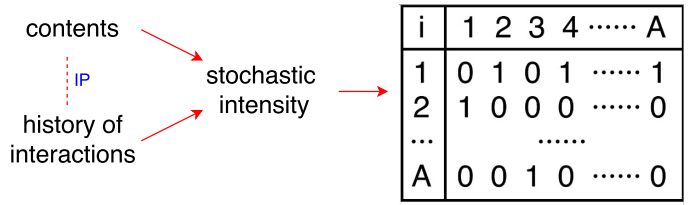
$$\lambda_{ij}^{(d)} = \sum_{c=1}^{C} p_c^{(d)} \cdot \exp\Big\{ \boldsymbol{b}_0^{(c)} + \boldsymbol{b}^{(c)T} \boldsymbol{x}_{t^{(d-1)}}^{(c)}(i,j) \Big\},$$

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)

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

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



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.

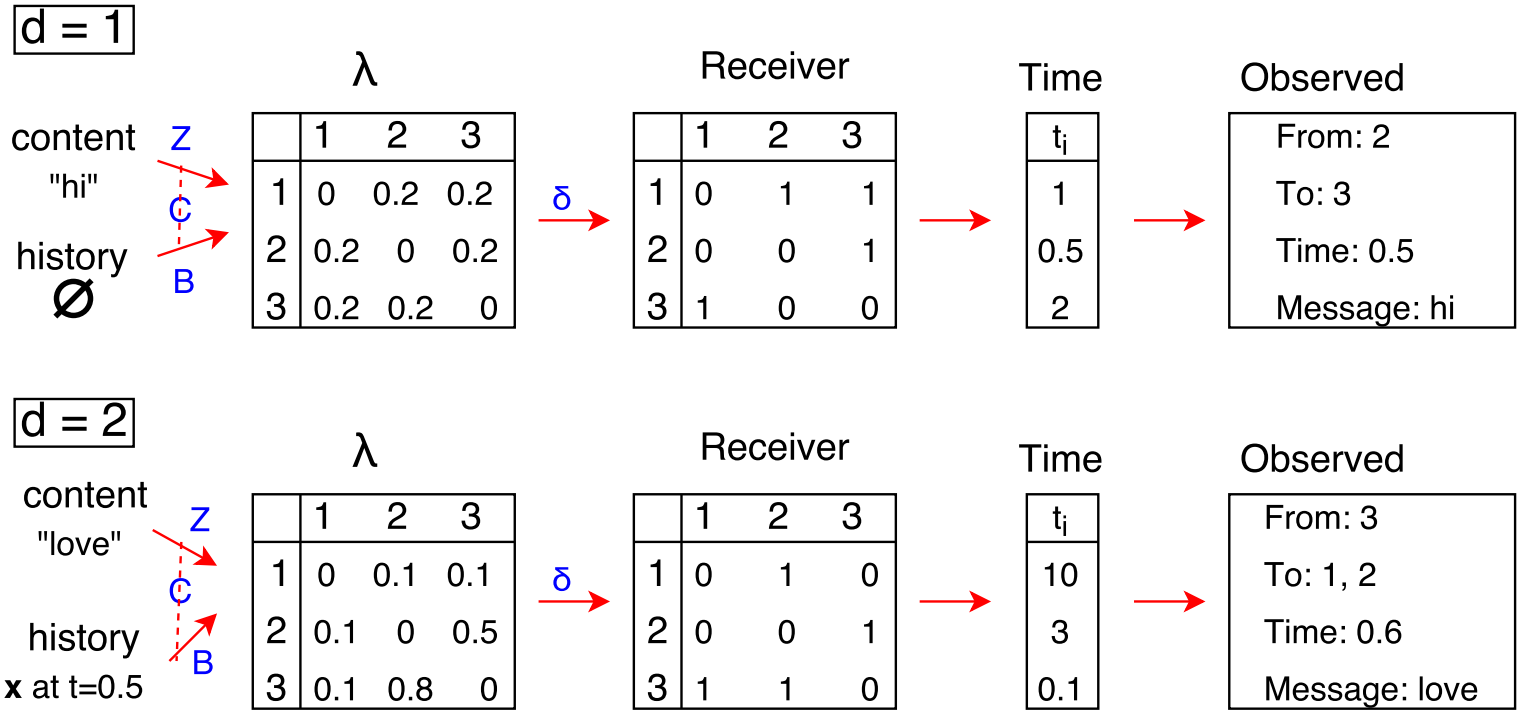
4. Set the observed sender, receivers and timestamp simultaneously:

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

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

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

Joint Generating Process



Inference

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

Algorithm 1 MCMC

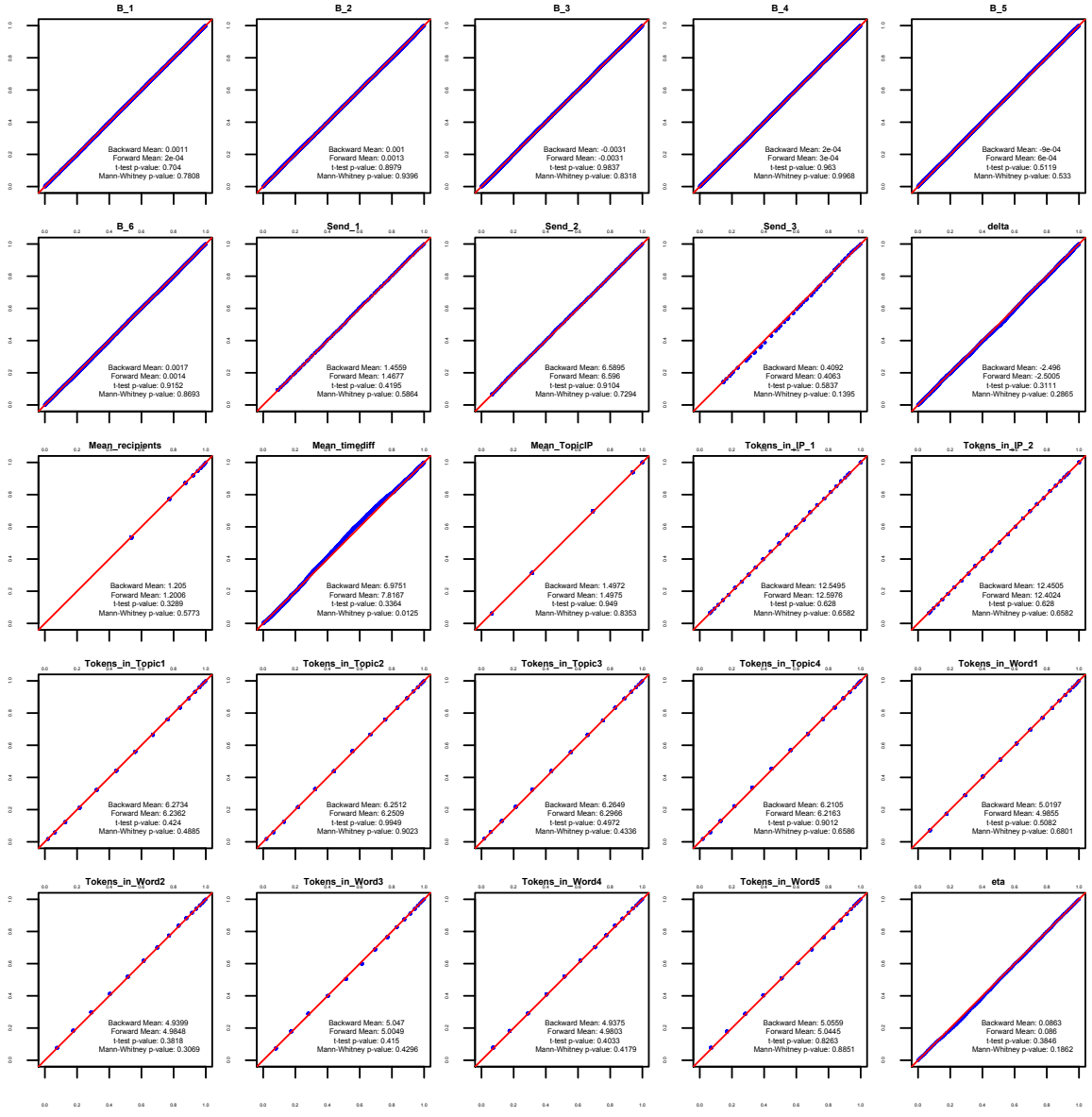
Set initial values $\mathcal{Z}^{(0)}, \mathcal{C}^{(0)}$, and $(\mathcal{B}^{(0)}, \delta^{(0)})$
for *o=1 to O* **do**
  Sample the latent receivers $J_{ij}^{(d)}$ via Gibbs sampling
  Sample the topic assignments $\mathcal{Z}$ via Gibbs sampling
  Sample the interaction pattern assignments $\mathcal{C}$ via Gibbs sampling
  Sample the network effect parameters $\mathcal{B}$ via Metropolis-Hastings
  Sample the receiver size parameter δ via Metropolis-Hastings
end

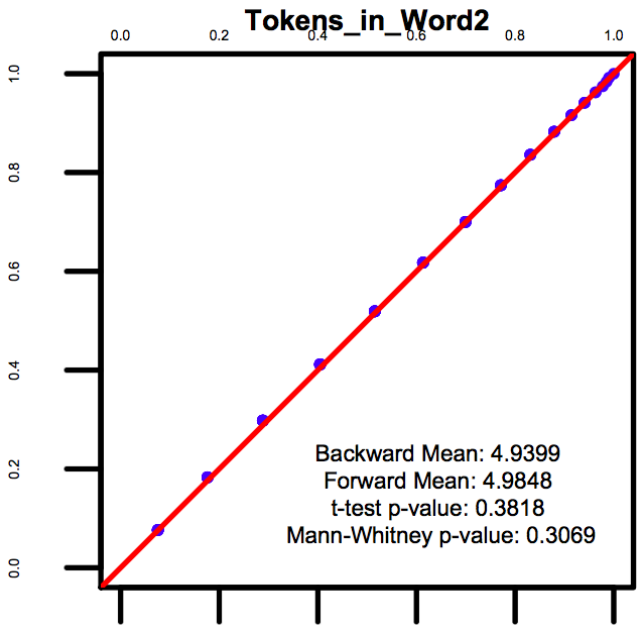
Getting it Right: Jointly testing math and code

Geweke (2004) proposed a test for Bayesian posterior samplers

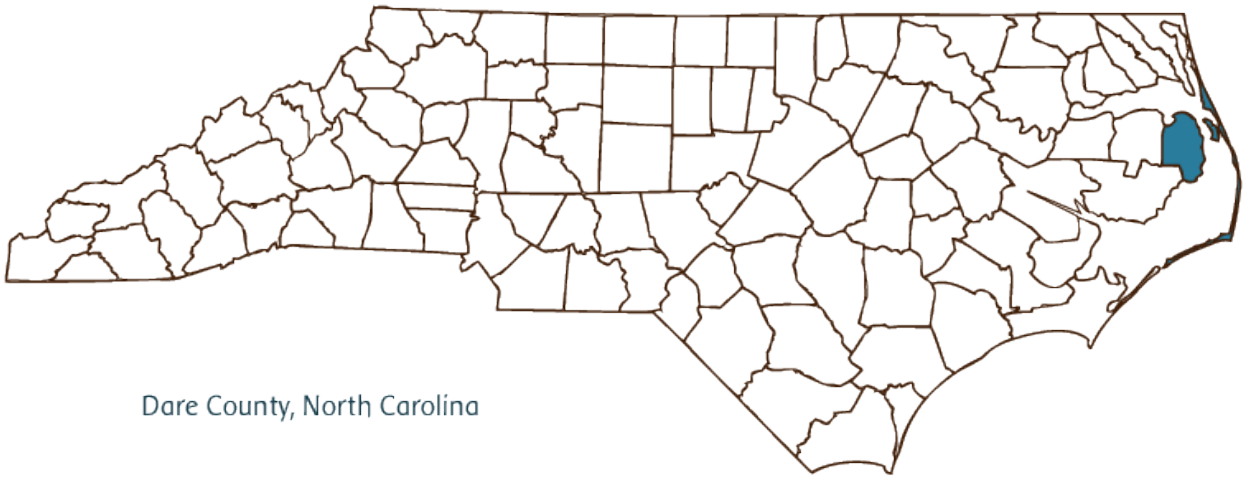
- *Forward samples*:
  1. Draw parameters from prior
  2. Draw data conditional on parameters
  3. Repeat
- *Backward samples*:
  1. Start with a forward sample of data
  2. Run inference on data
  3. Generate new data conditioned on inferred parameters
  4. Run inference on new data
  5. Repeat
- Forward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

GiR Results


B_1
Backward Mean: 0.0011
Forward Mean: 2e-04
t-test p-value: 0.704
Mann-Whitney p-value: 0.7808
B_2
Backward Mean: 0.001
Forward Mean: 0.0013
t-test p-value: 0.8979
Mann-Whitney p-value: 0.9396
B_3
Backward Mean: -0.0031
Forward Mean: -0.0031
t-test p-value: 0.9837
Mann-Whitney p-value: 0.8318
B_4
Backward Mean: 2e-04
Forward Mean: 3e-04
t-test p-value: 0.963
Mann-Whitney p-value: 0.9968
B_5
Backward Mean: -9e-04
Forward Mean: 6e-04
t-test p-value: 0.5119
Mann-Whitney p-value: 0.533
B_6
Backward Mean: 0.0017
Forward Mean: 0.0014
t-test p-value: 0.9152
Mann-Whitney p-value: 0.8693
Send_1
Backward Mean: 1.4559
Forward Mean: 1.4677
t-test p-value: 0.4195
Mann-Whitney p-value: 0.5864
Send_2
Backward Mean: 6.5895
Forward Mean: 6.596
t-test p-value: 0.9104
Mann-Whitney p-value: 0.7294
Send_3
Backward Mean: 0.4092
Forward Mean: 0.4063
t-test p-value: 0.5837
Mann-Whitney p-value: 0.1395
delta
Backward Mean: -2.496
Forward Mean: -2.5005
t-test p-value: 0.3111
Mann-Whitney p-value: 0.2865
Mean_recipients
Backward Mean: 1.205
Forward Mean: 1.2006
t-test p-value: 0.3289
Mann-Whitney p-value: 0.5773
Mean_timediff
Backward Mean: 6.9751
Forward Mean: 7.8167
t-test p-value: 0.3364
Mann-Whitney p-value: 0.0125
Mean_TopicIP
Backward Mean: 1.4972
Forward Mean: 1.4975
t-test p-value: 0.949
Mann-Whitney p-value: 0.8353
Tokens_in_IP_1
Backward Mean: 12.5495
Forward Mean: 12.5976
t-test p-value: 0.628
Mann-Whitney p-value: 0.6582
Tokens_in_IP_2
Backward Mean: 12.4505
Forward Mean: 12.4024
t-test p-value: 0.628
Mann-Whitney p-value: 0.6582
Tokens_in_Topic1
Backward Mean: 6.2734
Forward Mean: 6.2362
t-test p-value: 0.424
Mann-Whitney p-value: 0.4885
Tokens_in_Topic2
Backward Mean: 6.2512
Forward Mean: 6.2509
t-test p-value: 0.9949
Mann-Whitney p-value: 0.9023
Tokens_in_Topic3
Backward Mean: 6.2649
Forward Mean: 6.2966
t-test p-value: 0.4972
Mann-Whitney p-value: 0.4336
Tokens_in_Topic4
Backward Mean: 6.2105
Forward Mean: 6.2163
t-test p-value: 0.9012
Mann-Whitney p-value: 0.6586
Tokens_in_Word1
Backward Mean: 5.0197
Forward Mean: 4.9855
t-test p-value: 0.5082
Mann-Whitney p-value: 0.6801
Tokens_in_Word2
Backward Mean: 4.9399
Forward Mean: 4.9848
t-test p-value: 0.3818
Mann-Whitney p-value: 0.3069
Tokens_in_Word3
Backward Mean: 5.047
Forward Mean: 5.0049
t-test p-value: 0.415
Mann-Whitney p-value: 0.4296
Tokens_in_Word4
Backward Mean: 4.9375
Forward Mean: 4.9803
t-test p-value: 0.4033
Mann-Whitney p-value: 0.4179
Tokens_in_Word5
Backward Mean: 5.0559
Forward Mean: 5.0445
t-test p-value: 0.8263
Mann-Whitney p-value: 0.8851
eta
Backward Mean: 0.0863
Forward Mean: 0.086
t-test p-value: 0.3846
Mann-Whitney p-value: 0.1862


Tokens_in_Word2
Backward Mean: 4.9399
Forward Mean: 4.9848
t-test p-value: 0.3818
Mann-Whitney p-value: 0.3069

Dare County, NC email data

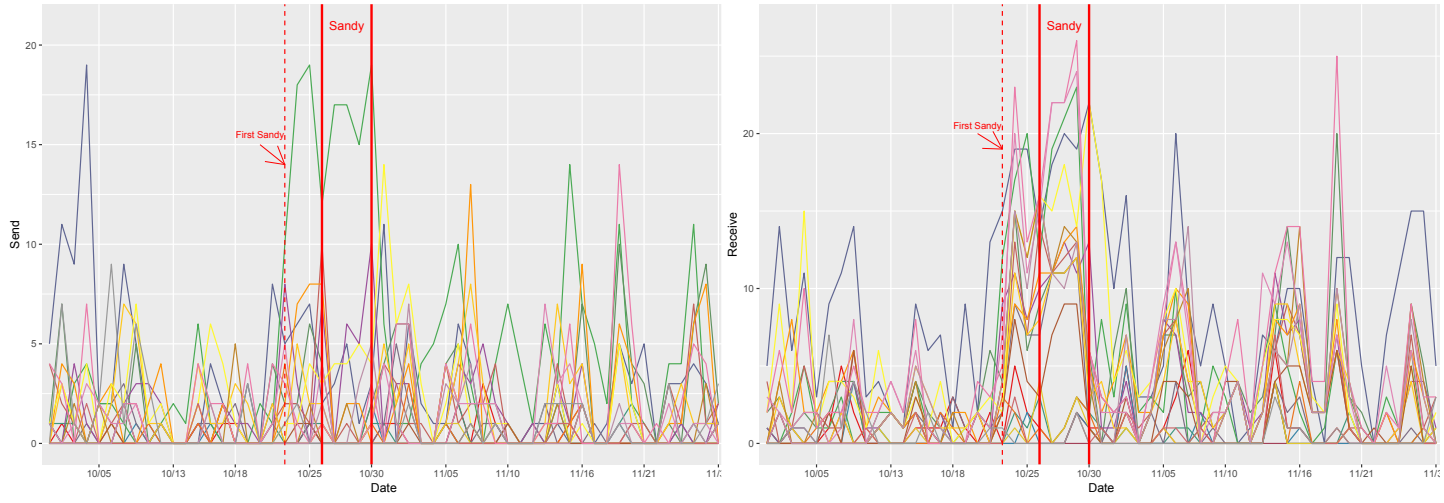


- $D = 1456$ emails
- $A = 27$ county government managers
- covering 2 month period (October 1 - November 30) in 2012
- Hurricane Sandy passed by NC: October 26 - October 30

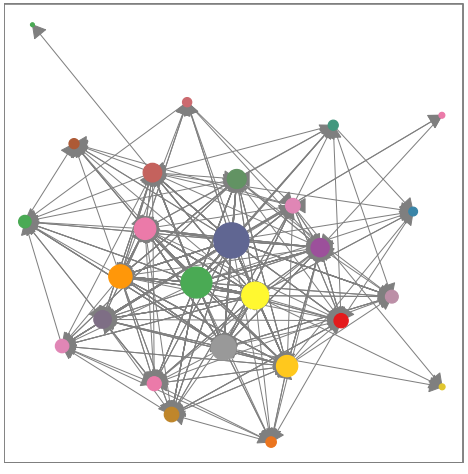
Theoretical considerations

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

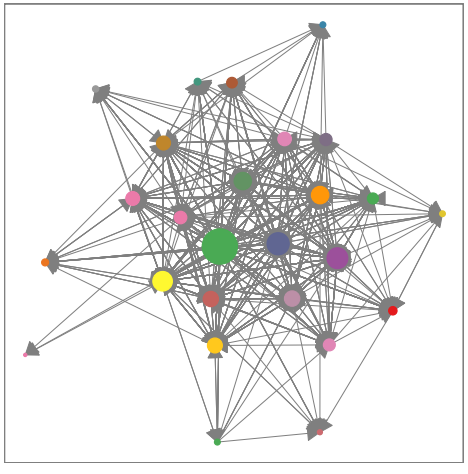
Exploratory Data Analysis: Dare County



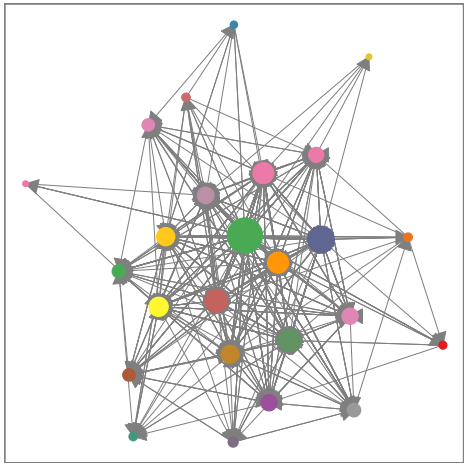
Pre-Sandy



Sandy



Post-Sandy



IPTM Result: Topics, Dare County (IP 1)

$C = 2$, $K = 20$ and $O = 100$

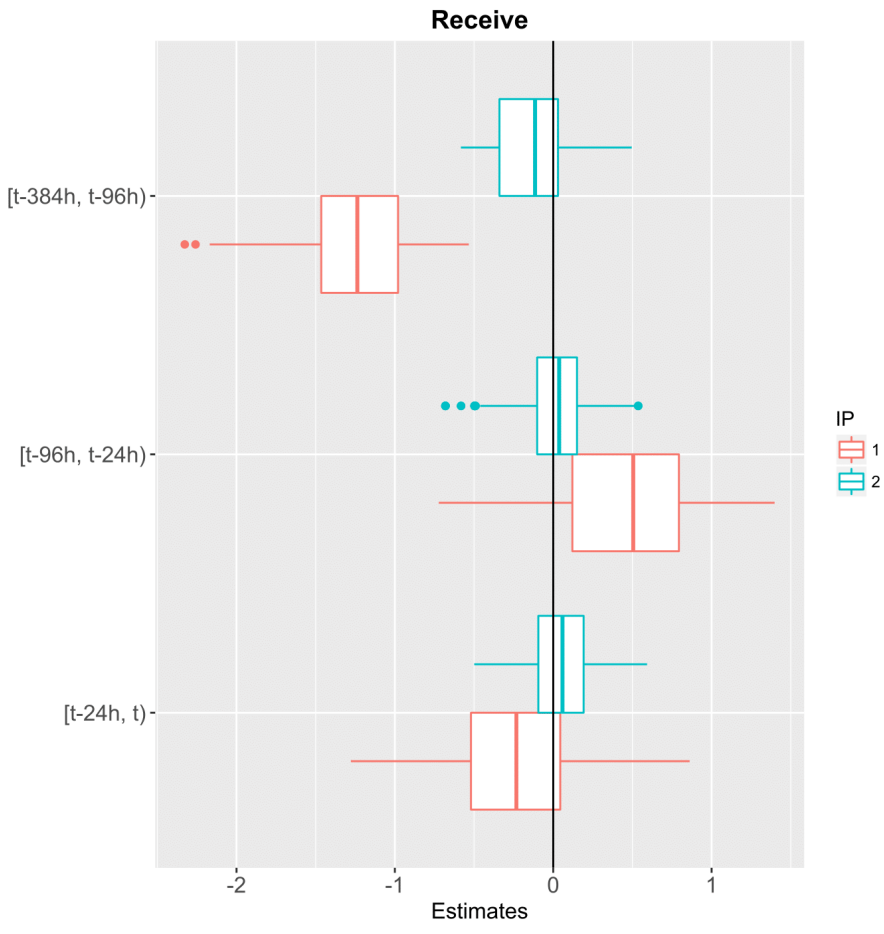
IP	**1**	**1**	**1**	**1**	**1**
Topic	**3** (0.078)	**5** (0.065)	**19** (0.064)	**13** (0.057)	**15** (0.053)
Word	water	planning	phone	questions	contact
	relocation	meter	collins	board	info
	location	room	drive	december	problem
	hills	asked	marshall	call	release
	utilities	needed	director	sheets	check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

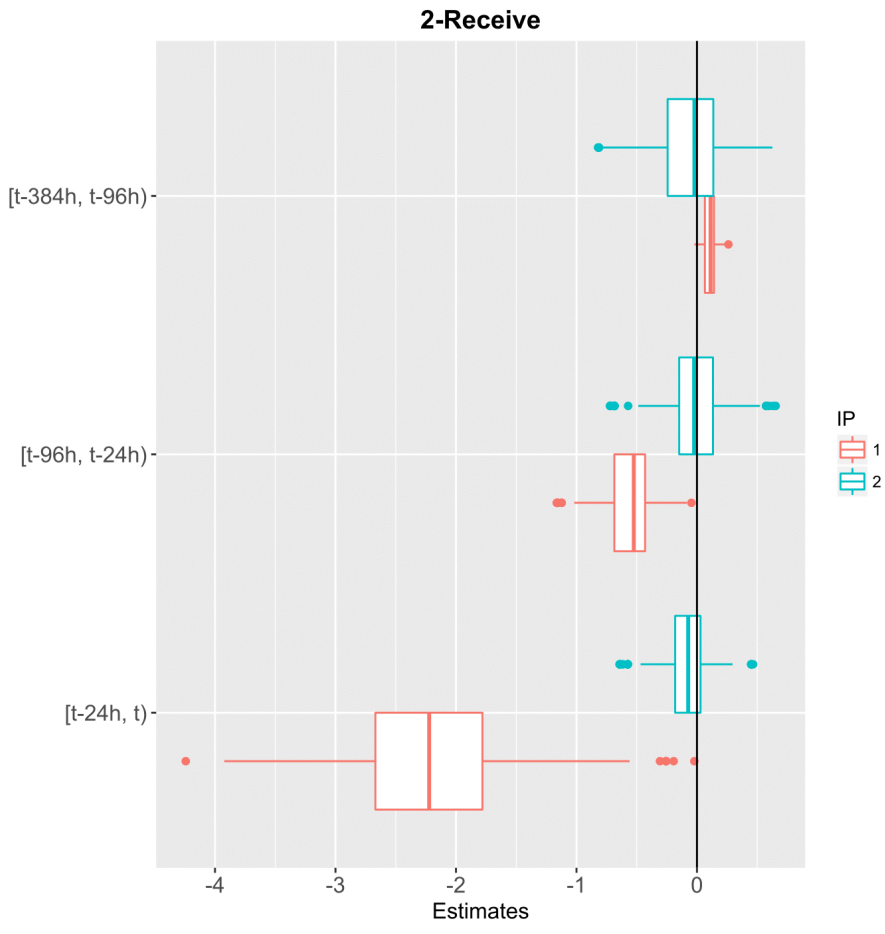
IPTM Result: Topics, Dare County (IP 2)

$C = 2$, $K = 20$ and $O = 100$

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

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

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'