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

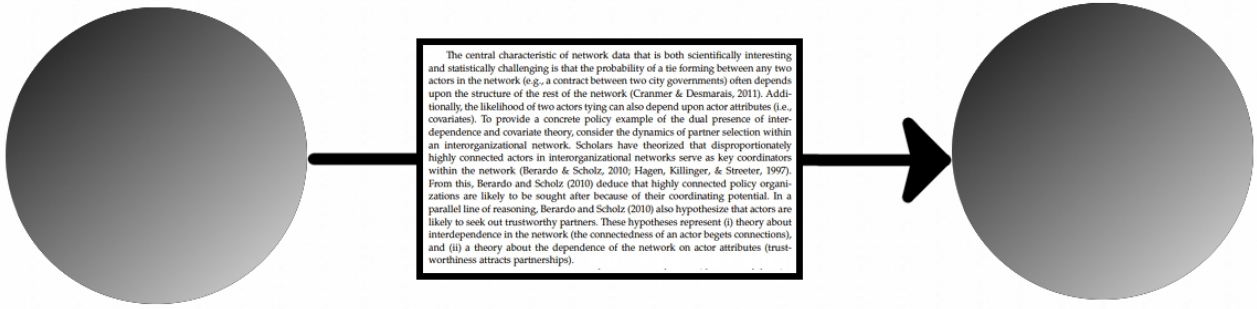
[3] University of Massachusetts Amherst

October 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)

- 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

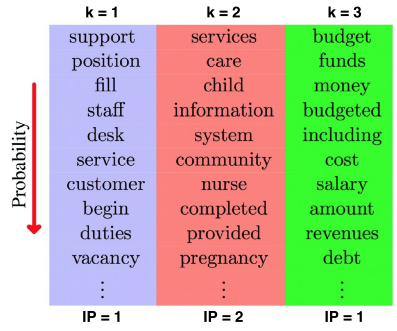
"who communicates with whom about what, and when?"

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



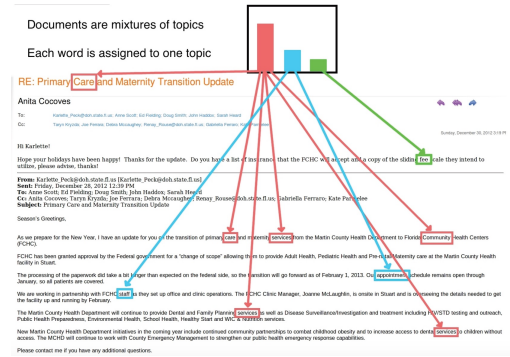
- For each document $d = 1, ..., D$:

  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:

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

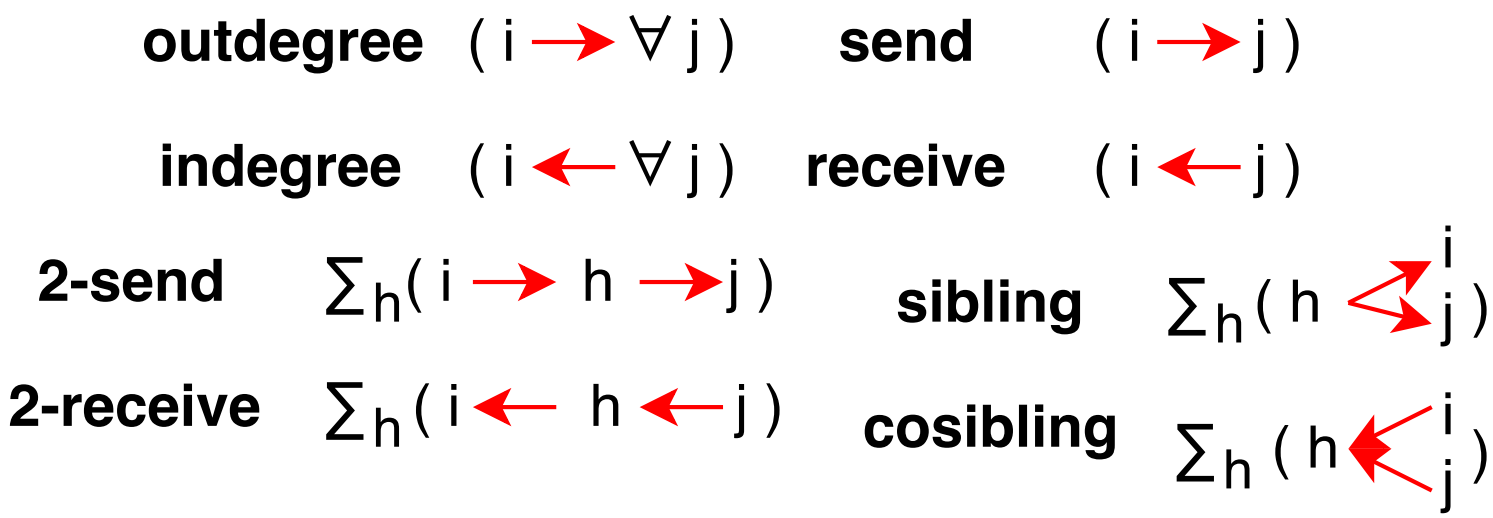
Network model components

- 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

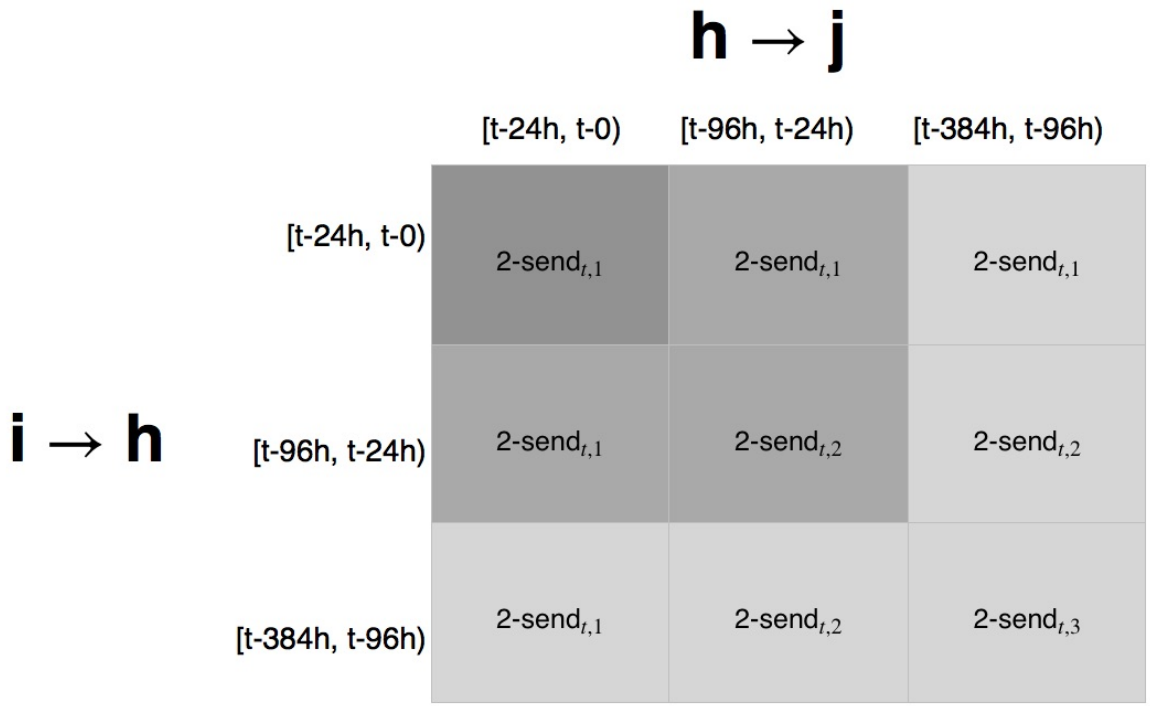


Conditioning features on recency

- 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:

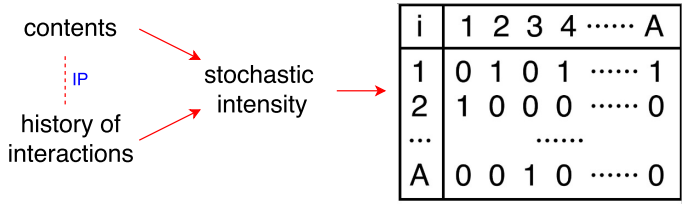
$$\lambda_{ij}^{(d)} = \sum_{c=1}^{C} p_c^{(d)} \cdot \exp\Big\{ \boldsymbol{b}^{(c)T} \boldsymbol{x}_{t^{(d-1)}}^{(c)}(i,j) \Big\},$$

which is a mixture of contents 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

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

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

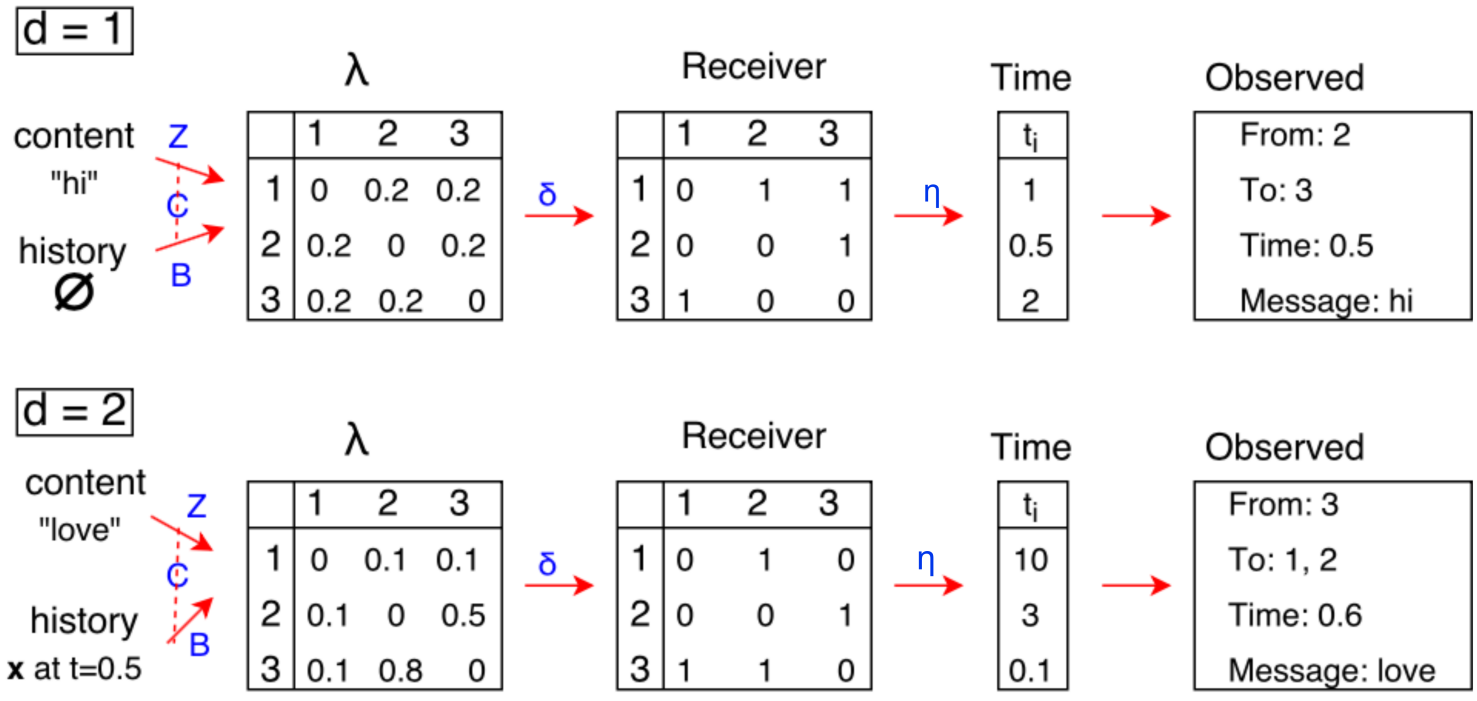
4. Set the observed sender, receivers and timestamp simultaneously:

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

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

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

Joint generating process



Inference

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

Algorithm 1 MCMC

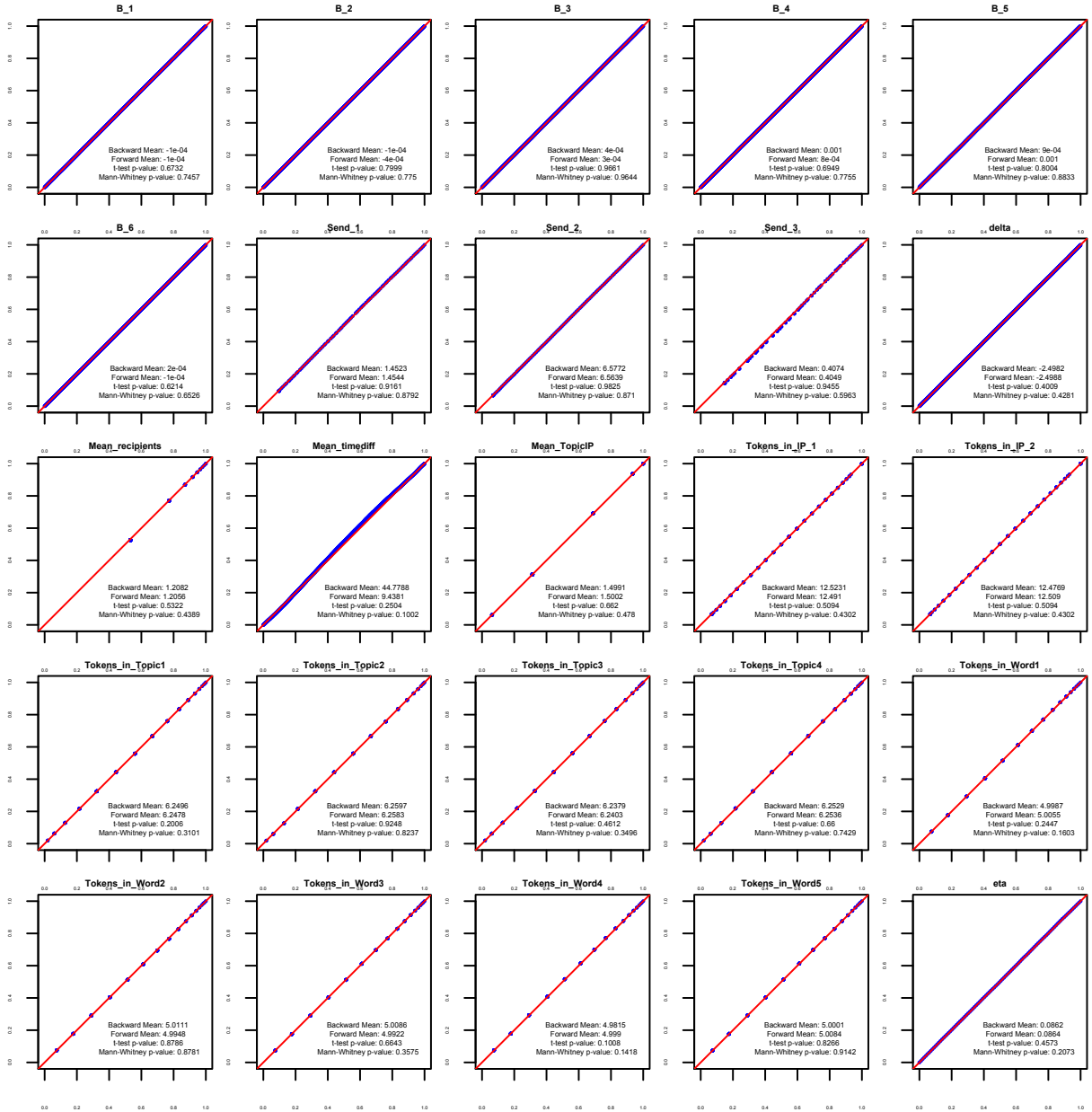
Set initial values $\mathcal{Z}^{(0)}, \mathcal{C}^{(0)}$, and $(\mathcal{B}^{(0)}, \delta^{(0)}, \eta^{(0)})$
for *o=1 to O* **do**
  Sample the latent receivers $J_{ij}^{(d)}$ via Gibbs sampling
  Sample the topic assignments $\mathcal{Z}$ via Gibbs sampling
  Sample the interaction pattern assignments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

Getting it Right: jointly testing math and code

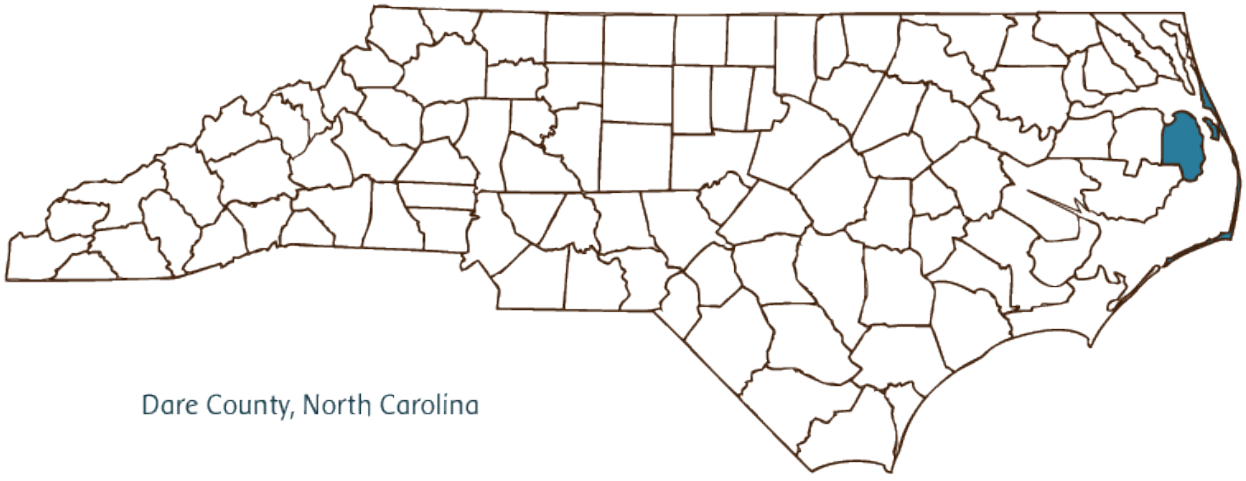
Geweke (2004) proposed a test for Bayesian posterior samplers

- *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


B_1
Backward Mean: -1e-04
Forward Mean: -1e-04
t-test p-value: 0.6732
Mann-Whitney p-value: 0.7457
B_2
Backward Mean: -1e-04
Forward Mean: -4e-04
t-test p-value: 0.7999
Mann-Whitney p-value: 0.775
B_3
Backward Mean: 4e-04
Forward Mean: 3e-04
t-test p-value: 0.9661
Mann-Whitney p-value: 0.9644
B_4
Backward Mean: 0.001
Forward Mean: 8e-04
t-test p-value: 0.6949
Mann-Whitney p-value: 0.7755
B_5
Backward Mean: 9e-04
Forward Mean: 0.001
t-test p-value: 0.8004
Mann-Whitney p-value: 0.8833
B_6
Backward Mean: 2e-04
Forward Mean: -1e-04
t-test p-value: 0.6214
Mann-Whitney p-value: 0.6526
Send_1
Backward Mean: 1.4523
Forward Mean: 1.4544
t-test p-value: 0.9161
Mann-Whitney p-value: 0.8792
Send_2
Backward Mean: 6.5772
Forward Mean: 6.5639
t-test p-value: 0.9825
Mann-Whitney p-value: 0.871
Send_3
Backward Mean: 0.4074
Forward Mean: 0.4049
t-test p-value: 0.9455
Mann-Whitney p-value: 0.5963
delta
Backward Mean: -2.4982
Forward Mean: -2.4988
t-test p-value: 0.4009
Mann-Whitney p-value: 0.4281
Mean_recipients
Backward Mean: 1.2082
Forward Mean: 1.2056
t-test p-value: 0.5322
Mann-Whitney p-value: 0.4389
Mean_timediff
Backward Mean: 44.7788
Forward Mean: 9.4381
t-test p-value: 0.2504
Mann-Whitney p-value: 0.1002
Mean_TopicIP
Backward Mean: 1.4991
Forward Mean: 1.5002
t-test p-value: 0.662
Mann-Whitney p-value: 0.478
Tokens_in_IP_1
Backward Mean: 12.5231
Forward Mean: 12.491
t-test p-value: 0.5094
Mann-Whitney p-value: 0.4302
Tokens_in_IP_2
Backward Mean: 12.4769
Forward Mean: 12.509
t-test p-value: 0.5094
Mann-Whitney p-value: 0.4302
Tokens_in_Topic1
Backward Mean: 6.2496
Forward Mean: 6.2478
t-test p-value: 0.2006
Mann-Whitney p-value: 0.3101
Tokens_in_Topic2
Backward Mean: 6.2597
Forward Mean: 6.2583
t-test p-value: 0.9248
Mann-Whitney p-value: 0.8237
Tokens_in_Topic3
Backward Mean: 6.2379
Forward Mean: 6.2403
t-test p-value: 0.4612
Mann-Whitney p-value: 0.3496
Tokens_in_Topic4
Backward Mean: 6.2529
Forward Mean: 6.2536
t-test p-value: 0.66
Mann-Whitney p-value: 0.7429
Tokens_in_Word1
Backward Mean: 4.9987
Forward Mean: 5.0055
t-test p-value: 0.2447
Mann-Whitney p-value: 0.1603
Tokens_in_Word2
Backward Mean: 5.0111
Forward Mean: 4.9948
t-test p-value: 0.8786
Mann-Whitney p-value: 0.8781
Tokens_in_Word3
Backward Mean: 5.0086
Forward Mean: 4.9922
t-test p-value: 0.6643
Mann-Whitney p-value: 0.3575
Tokens_in_Word4
Backward Mean: 4.9815
Forward Mean: 4.999
t-test p-value: 0.1008
Mann-Whitney p-value: 0.1418
Tokens_in_Word5
Backward Mean: 5.0001
Forward Mean: 5.0084
t-test p-value: 0.8266
Mann-Whitney p-value: 0.9142
eta
Backward Mean: 0.0862
Forward Mean: 0.0864
t-test p-value: 0.4573
Mann-Whitney p-value: 0.2073

Dare County, NC email data

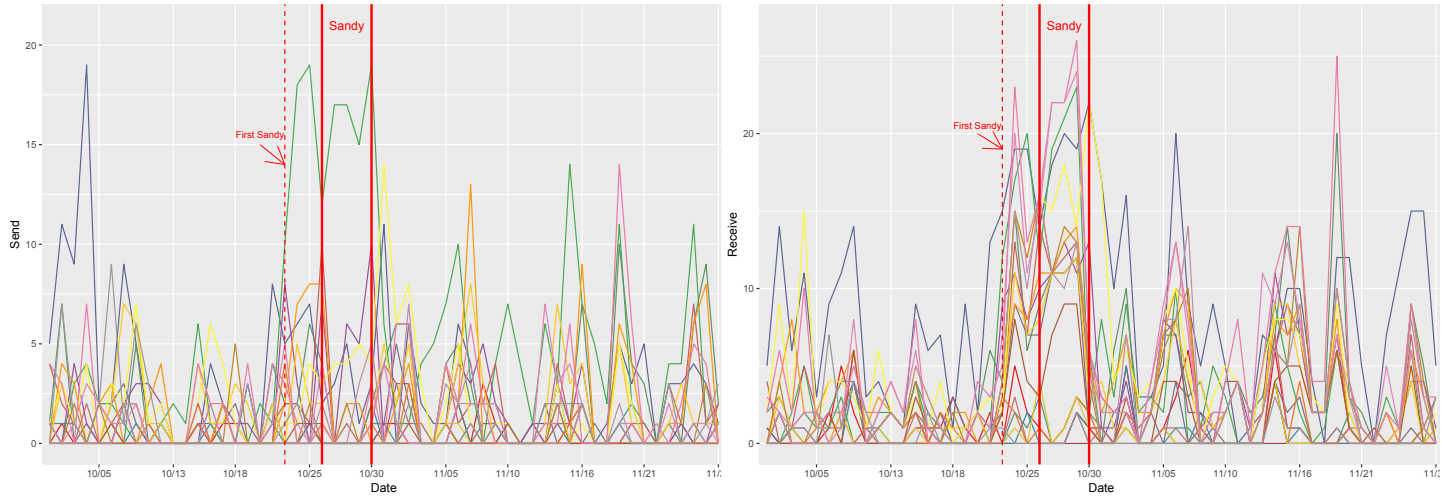


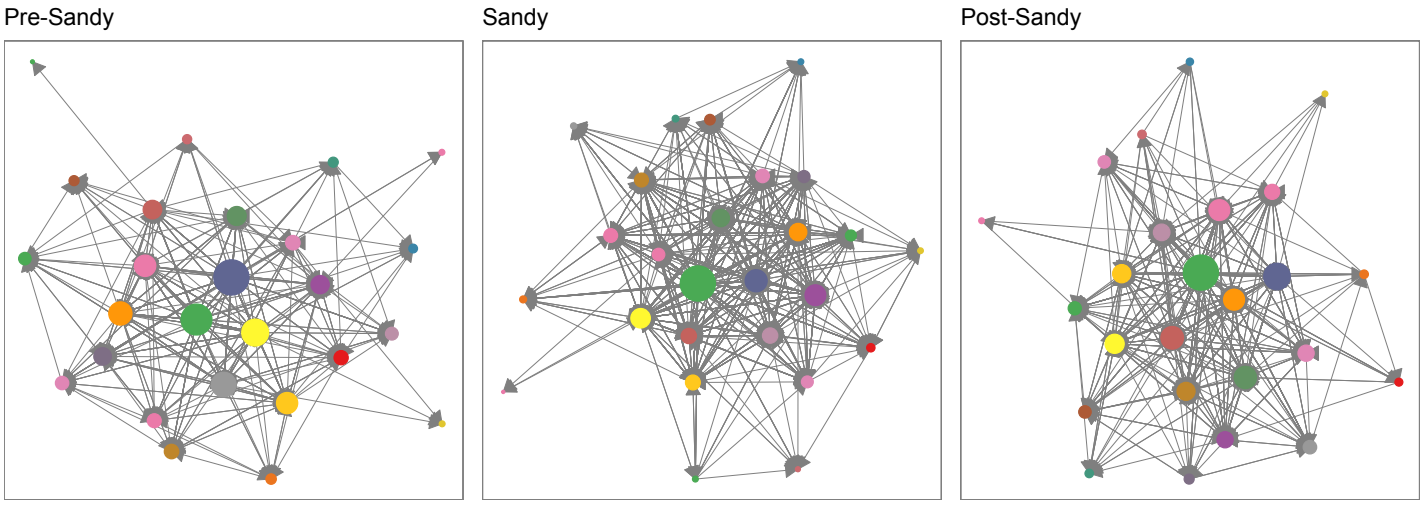
- $D = 2210$ emails
- $A = 27$ county government managers
- $W = 2907$ unique words
- covering 3 month period (September 1 - November 30) in 2012
- Hurricane Sandy passed by NC: October 26 - October 30

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

IPTM result: topics (IP=2)

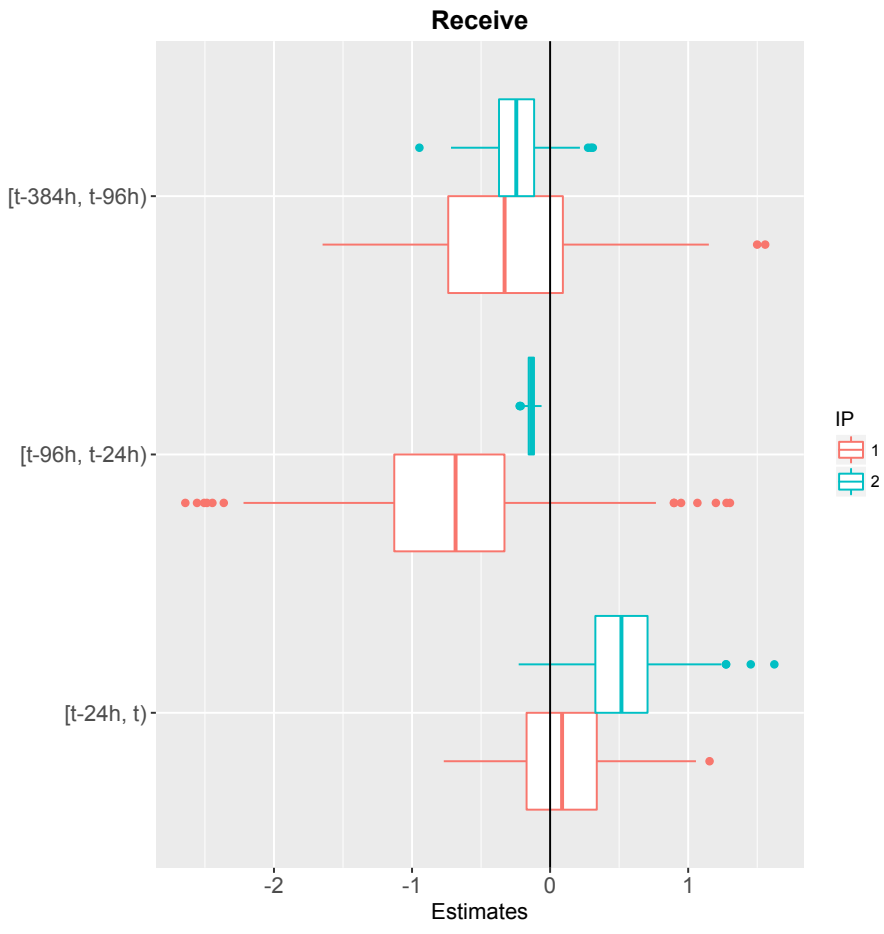
$C = 2$, $K = 25$ and $O = 100$

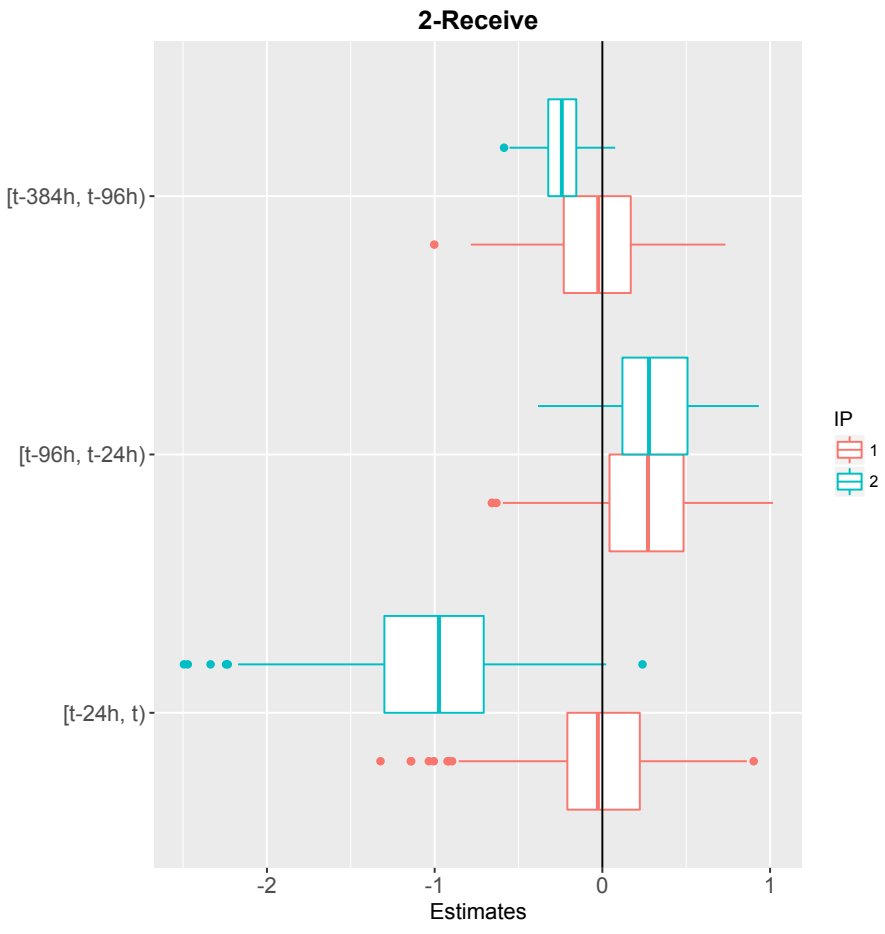
Topic	**21** (0.057)	**12** (0.037)	**14** (0.037)	**4** (0.035)	**10** (0.026)
Word	library	marshall	will	board	survey
	best	collins	center	meeting	request
	start	drive	day	property	call
	web	manteo	office	planning	sure
	place	phone	great	will	people
	visit	box	manager	notice	emergency
	albermarle	fax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

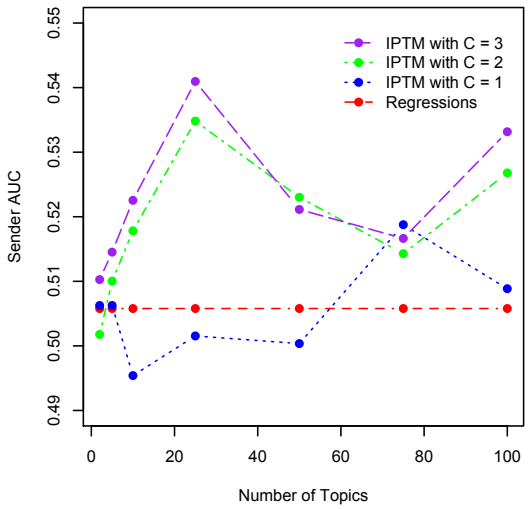
Algorithm 2 Predicting tie data for the next document

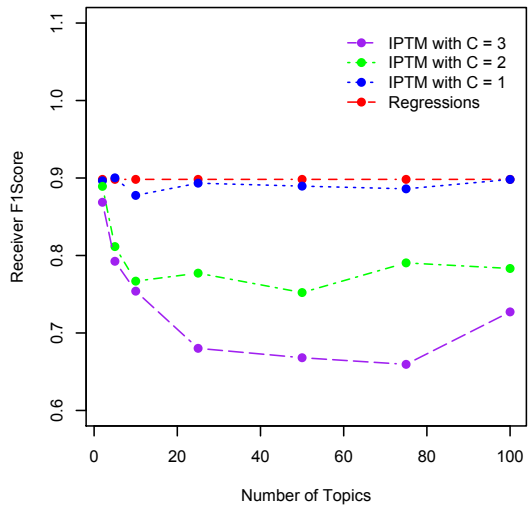
Input

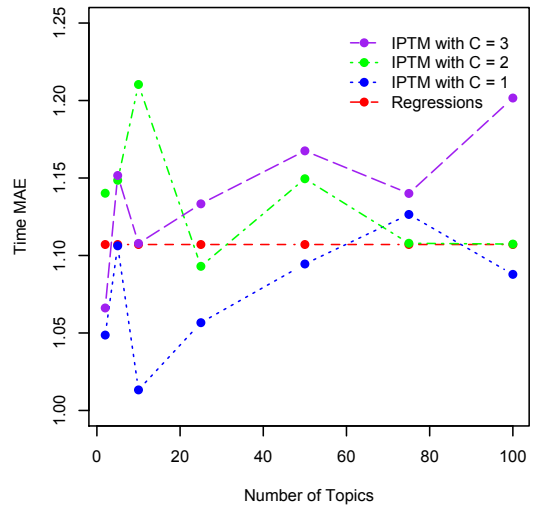
1. O, number of outer iterations of inference from which to generate predictions
2. d, the last document to use in inference
3. R, the number of iterations to sample predicted data within each outer iteration

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







- Area under the ROC curve (AUC) for sender, F1-score for Receivers, and Mean Absolute Error (MAE) for timing
- $C = 1$ loses the connection between ties and text
- Lack of fit in predicting receivers and timing
- Further modfications 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'