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

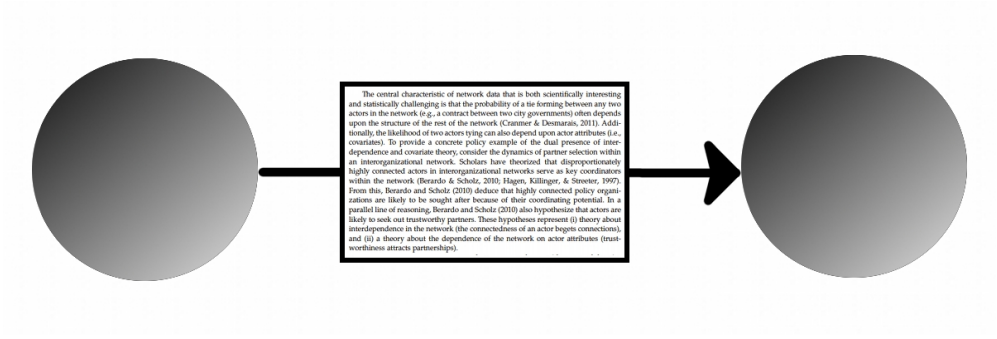
[3] University of Massachusetts Amherst

October 14, 2017

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

Motivation

- Many networks have ties that are attributed with text:
  - International treaties, legislative cosponsorship, online discussions



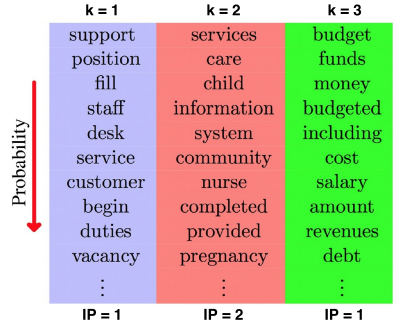
- Network models can't account for text
- Models for text either can't account for ties or only account for simplistic network structure

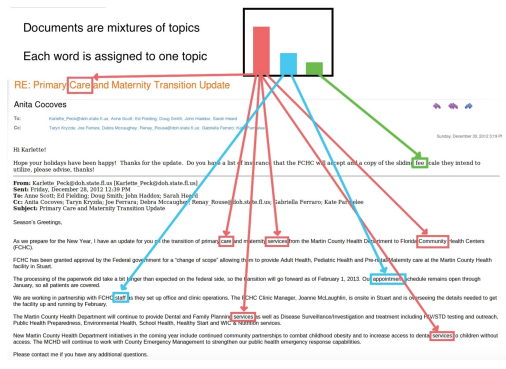
Interaction-Partitioned Topic Model (IPTM)

- Model for networks with timestamped, text-valued ties
- Draws on ideas from two existing models:
  - Latent Dirichlet allocation (LDA) for topic-based document content
  - Dynamic exponential random graph model (ERGM) for ties
- *Who communicates with whom about what, and when?*

Generating Content: LDA (Blei et al., 2003)

- For each topic $k = 1, ..., K$:
  - Draw a distribution over vocabulary: $\boldsymbol{\phi}_k \sim \text{Dir}(\beta, (\frac{1}{V}, \ldots, \frac{1}{V}))$
- For each email $d = 1, ..., D$:
  - Draw a distribution over topics: $\boldsymbol{\theta}_d \sim \text{Dir}(\alpha, (m_1, \ldots, m_K))$
  - For each token $n = 1$ to N_d:
    - Draw a topic: $z_{dn} \sim \boldsymbol{\theta}_d$
    - Draw a word type: $w_{dn} \sim \boldsymbol{\phi}_{z_{dn}}$





Key Idea: Interaction Patterns

- Different topics associated with different interaction patterns
- For each topic $k = 1, ..., K$:
  - Assign topic k to an interaction pattern: $l_k \sim \text{Unif}(1, C)$
- For each email $d = 1, ..., D$:
  - Calculate distribution over interaction patterns:

$$\pi_{dc} = \frac{\sum_{k:l_k=c} N_{dk}}{N_d}$$

Generating Ties and Timestamps

- Continuous-time process
- Ties predicted using dynamic network features:
  - Popularity, activity
  - Repetition, reciprocity
  - Transitivity
- Author determines recipients and timestamp
- Innovative approach to modeling multiple recipients

Generating Hypothetical Authors and Recipients

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

$$\nu_{idjc} = \exp(\boldsymbol{b}_c{}^{\top} \boldsymbol{x}_{idjc})$$

$$\lambda_{idj} = \sum_{c=1}^{C} \pi_{dc}\, \nu_{idjc}$$

   which combines information about content and network structure

2. For each author $i \in \{1, ..., A\}$, draw a binary vector $\boldsymbol{u}_{id}$ of length A from a non-empty Gibbs measure (Fellows and Handcock, 2017)

$$\boldsymbol{u}_{id} = (u_{id1}, \ldots, u_{idA}) \sim \text{Gibbs}(\delta, (\lambda_{id1}, \ldots, \lambda_{idA}))$$

   where δ is a real-valued value that controls the number of recipients

Generating Hypothetical Timestamps and Actual Data

3. For each author $i \in \{1, ..., A\}$, generate a timestamp:

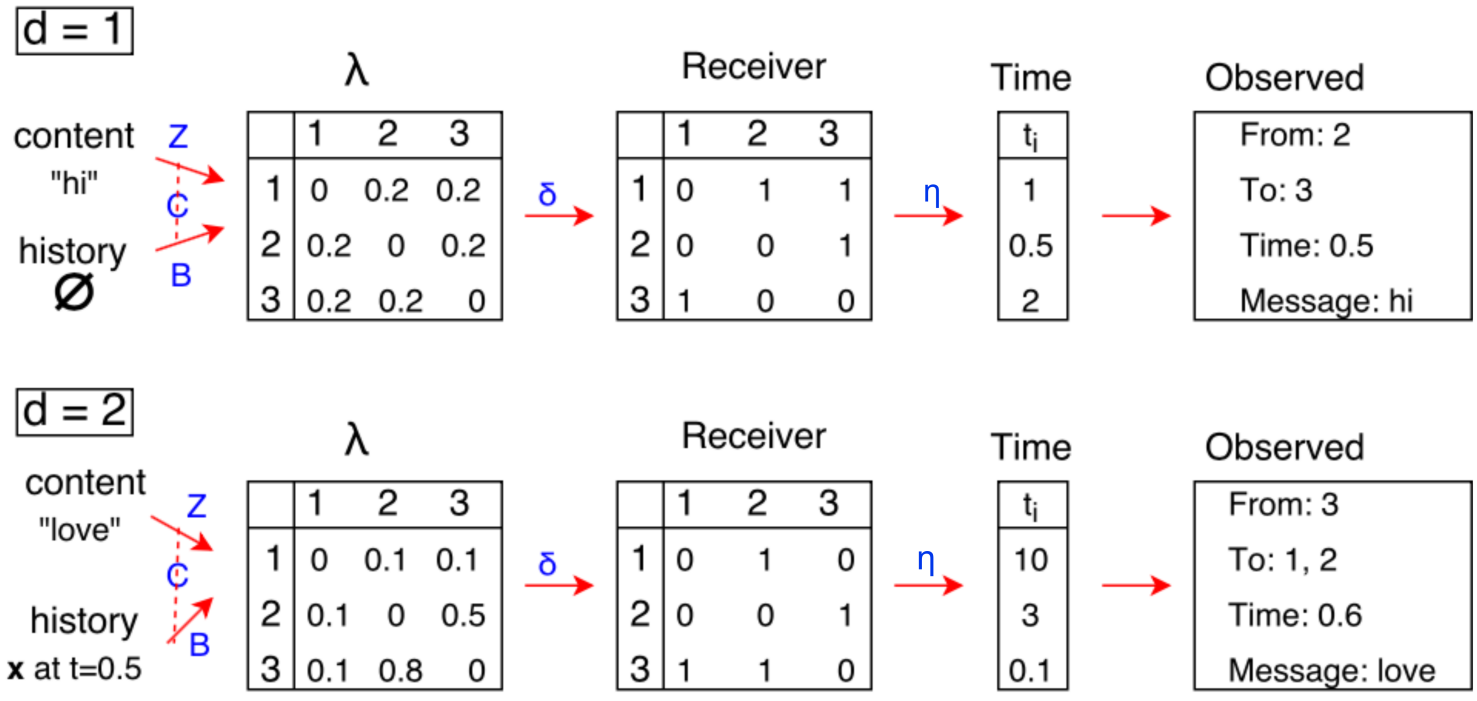
$$\mu_{id} = \sum_{c=1}^{C} \pi_{dc} \, \text{GeomMean}(\{\nu_{idjc}\}_{j:u_{idj}=1})$$
$$\tau_{id} \sim \text{Exp}(\eta \, \mu_{id})$$
$$t_{id} = t_{d-1} + \tau_{id}$$

   where η is a positive real-valued value

4. Select the actual author, recipients, and timestamp:

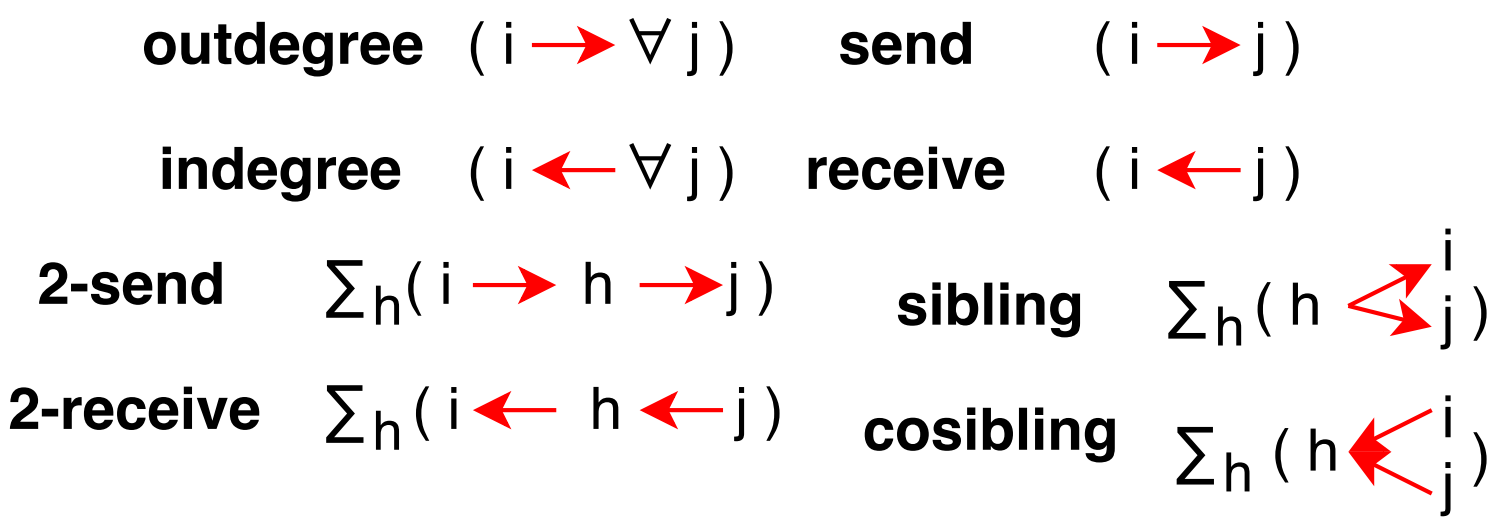
$$a_d = \text{argmin}_i(t_{id})$$
$$\boldsymbol{y}_d = \boldsymbol{u}_{a_d d}$$
$$t_d = t_{a_d d}$$

Generating Ties and Timestamps



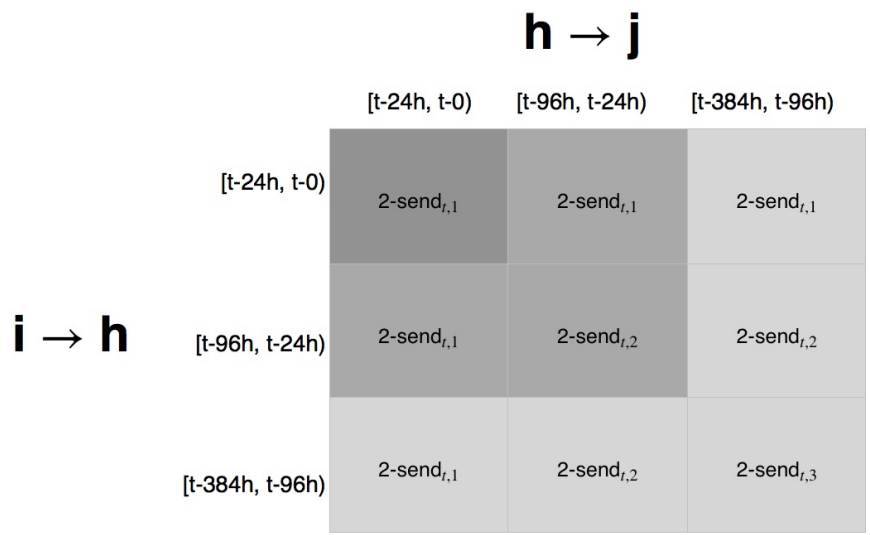
Network Features (Perry and Wolfe, 2012)

- Features capture different types of interaction: popularity, activity, repetition, reciprocity, transitivity



Dynamic Network Features

- To form $\boldsymbol{x}_{idjc}$, consider 3 time intervals prior to t^{+}_{d-1}:
  - 3–16 days, 1–3 days, 0–1 days
- For each interval, compute each network feature focusing on emails in that interval: 24 dynamic network features in $\boldsymbol{x}_{idjc}$



Inference

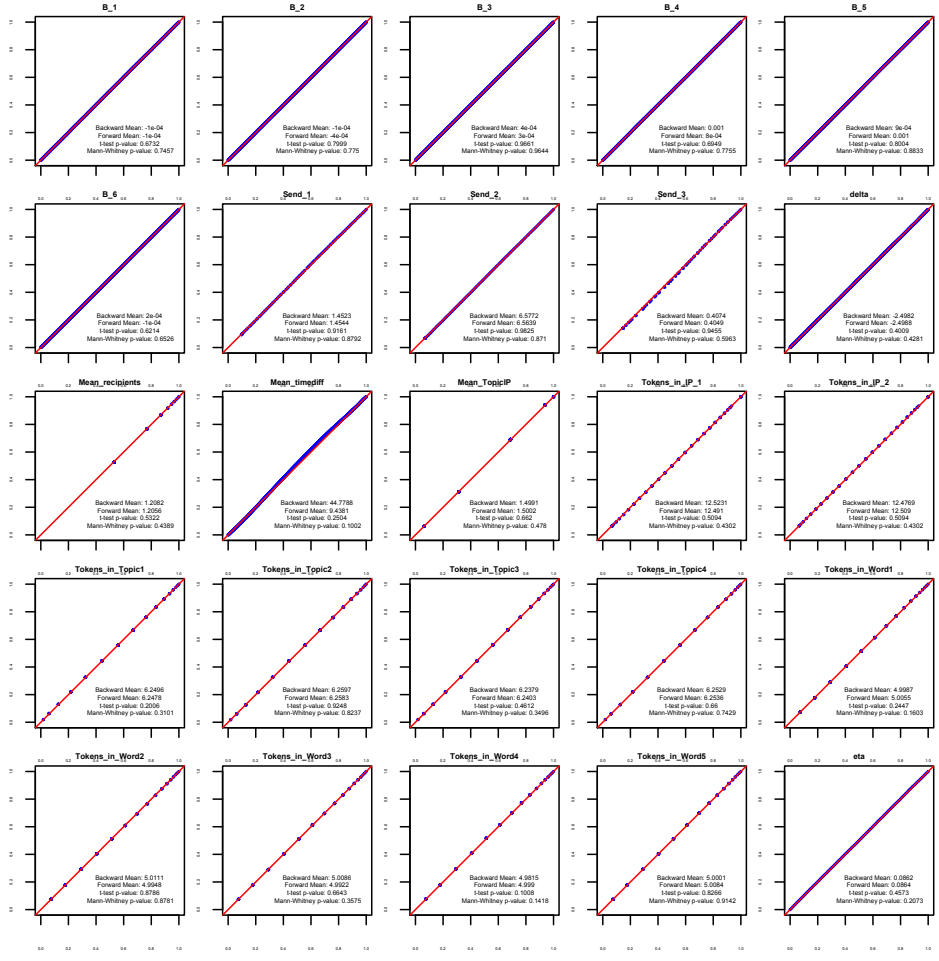
Algorithm 1 MCMC: Metropolis-Hastings within Gibbs

Initialize latent variables
for *o=1 to O* **do**
  Sample token–topic assignments z_{dn}
  Sample topic–interaction pattern assignments l_k
  Sample hypothetical author–recipient set pairs $\boldsymbol{u}_{ad}$
  Sample coefficients $\boldsymbol{b}_c$ using Metropolis-Hastings
  Sample recipient parameter δ using Metropolis-Hastings
  Sample timestamp parameter η using Metropolis-Hastings
end

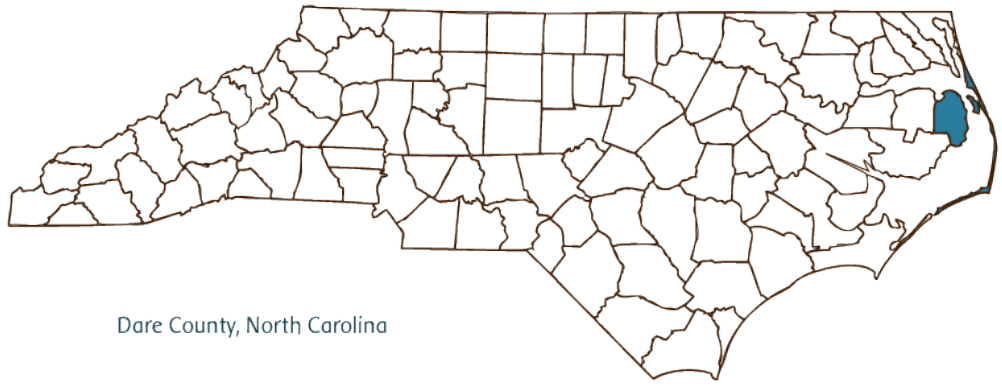
Getting it Right: Testing Math and Code (Geweke, 2004)

- *Generate forward samples*:
  1. Draw latent variables from priors
  2. Draw synthetic data conditioned on latent variables
  3. Repeat 1 and 2 many times
- *Backward samples*:
  1. Start with a forward sample
  2. Draw latent variables from posterior using inference code
  3. Draw synthetic data conditioned on latent variables
  4. Repeat 2 and 3 many times
- Forward samples and backward samples should match

Getting it Right Results



NC Email Corpus: Dare County



- $D = 2,210$ emails
- $A = 27$ department managers
- $V = 2,907$ words in vocabulary
- Covers 3 months (September 1 to November 30) in 2012:
  - Hurricane Sandy: October 26 to October 30

Example Email

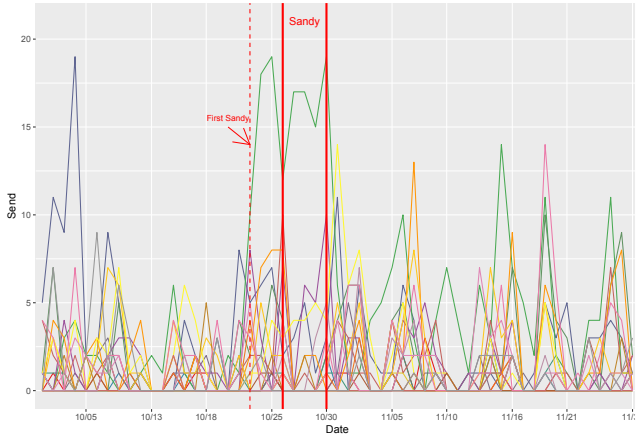
- From: Health
- To: County Manager
- At: 8 April 2012 8:46:00
- Content: ... A few months ago Jennifer was on call and Friday night she was out till 2 am and started the calls before 5pm, there were 2 bites at one time and then a goat in the highway, she took the goat to her barn because we did not have anywhere to keep it...

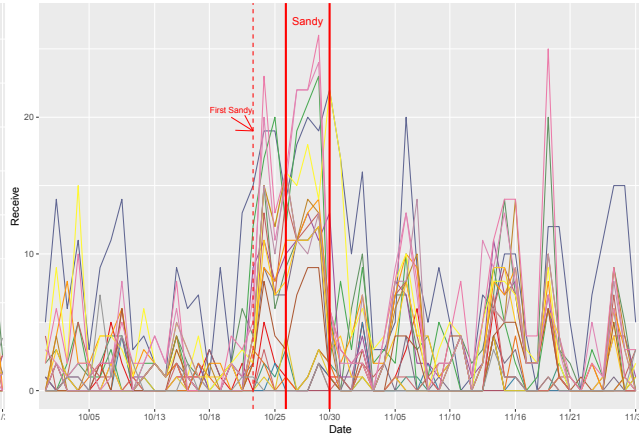
(Not the actual goat.)

Hypotheses

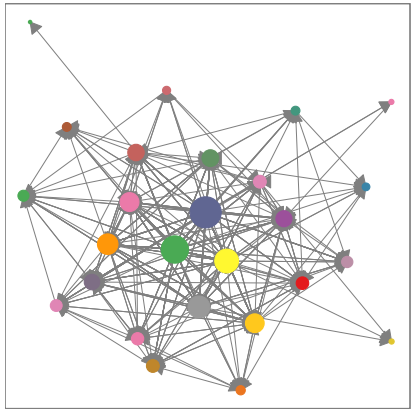
- Personal/social topics exhibit reciprocity and transitivity
- Dissemination topics don't exhibit reciprocity
- Topics about Hurricane Sandy look different

Data Exploration

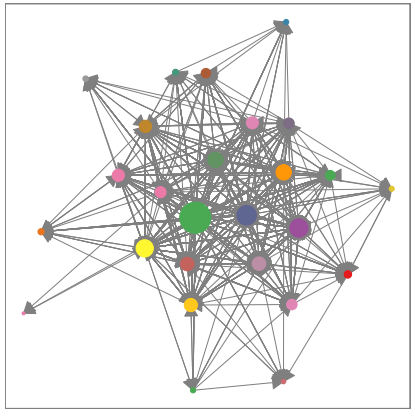




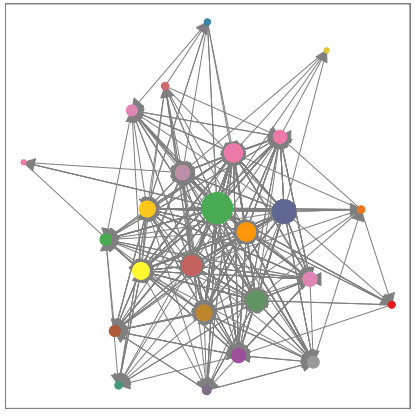
Pre-Sandy



Sandy



Post-Sandy



Department

- Building Inspections
- County Extension
- County Manager
- Detention
- Elections
- Emergency Services
- Finance
- Health
- HR
- Information Technology
- Library
- Parks and Recreation
- Planning
- Public Informations
- Register of Deeds
- Senior Center
- Sheriff
- Soil Conservation
- Solid Waste and Recycling
- Tax Administrator
- Transportation
- Veteran Services

Interaction Pattern 1: Top 5 Topics

23	18	11	20	8
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

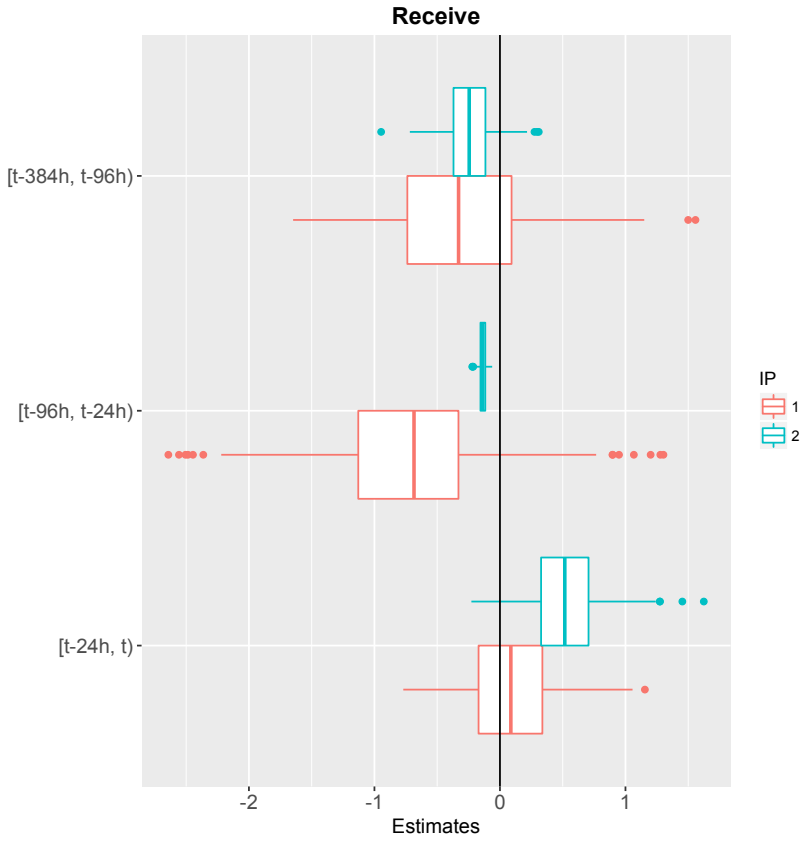
Example Email

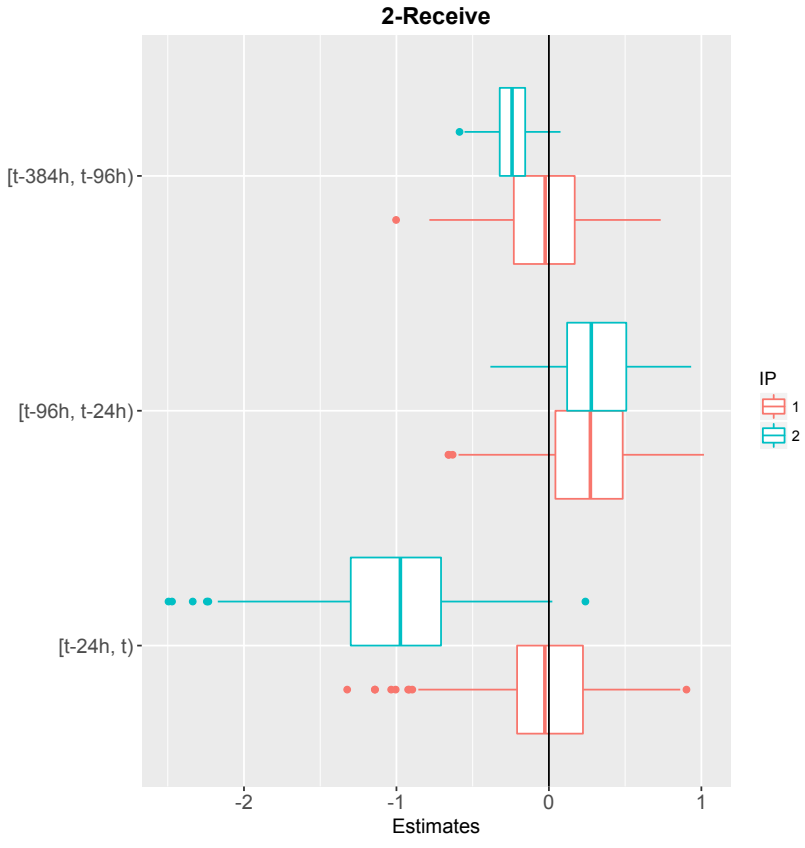
- From: Public Information
- To: Emergency Services, Detention
- At: 25 Oct 2012 19:14:47
- Content: will will storm good sure sandy morning morning morning morning send change plan hurricane wanted copy base afternoon keep description description saturday saturday saturday saturday touch touch release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

Interaction Pattern 2: Top 5 Topics

21	12	14	4	10
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

Coefficients



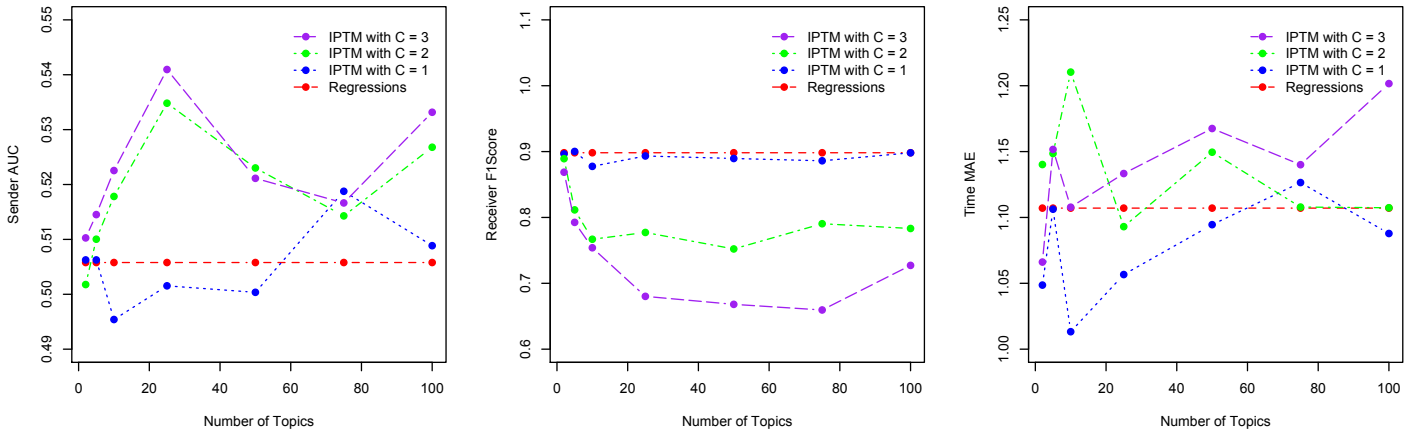


Predicting Ties and Timestamps

Algorithm 2 Predict ties and timestamp for next email

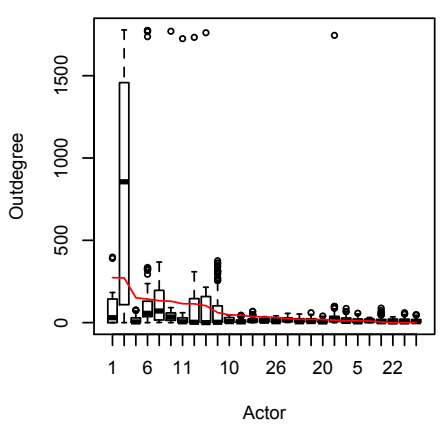
Run burn-in iterations of MCMC on emails 1 through $d-1$
for *o=1 to O* **do**
    Run single iteration of MCMC on emails 1 through $d-1$
    Initialize a_d, $\boldsymbol{y}_d$, t_d, and $\boldsymbol{z}_d$
    for *r=1 to R* **do**
        Sample a_d, $\boldsymbol{y}_d$, t_d conditioned on $\boldsymbol{z}_d$ using generative process
        Sample $\boldsymbol{z}_d$ from conditional posterior
    end
    Save a_d, $\boldsymbol{y}_d$, t_d, and $\boldsymbol{z}_d$
end

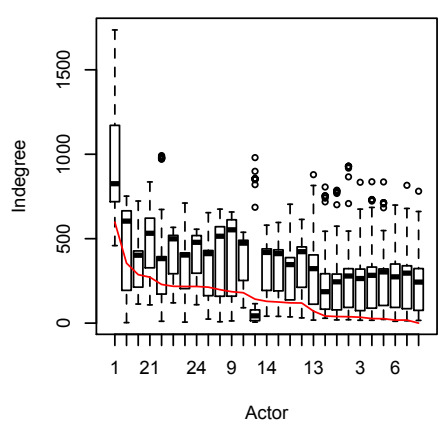
Example Results

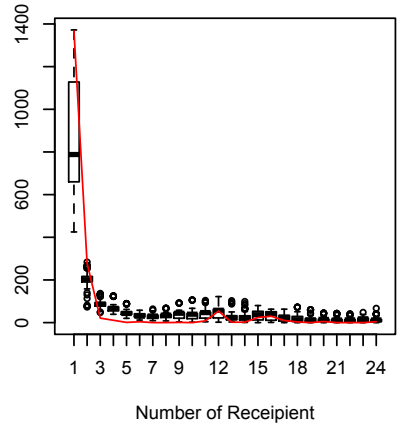


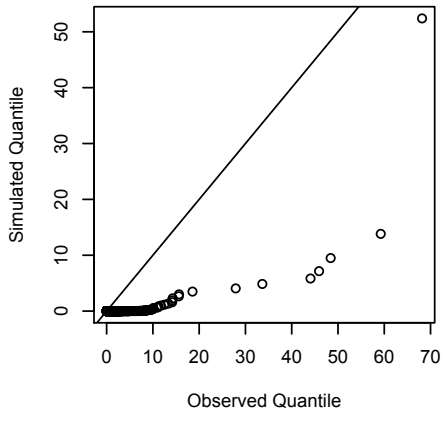
- Model does poorly at predicting ties and timestamps

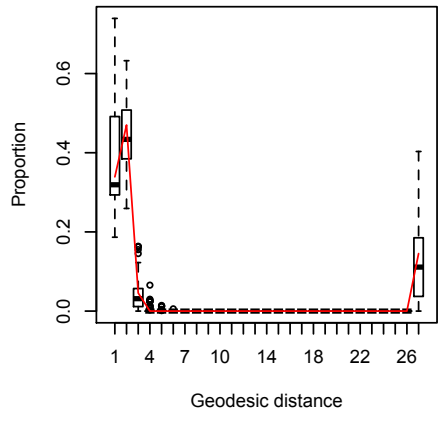
Posterior Predictive Checks

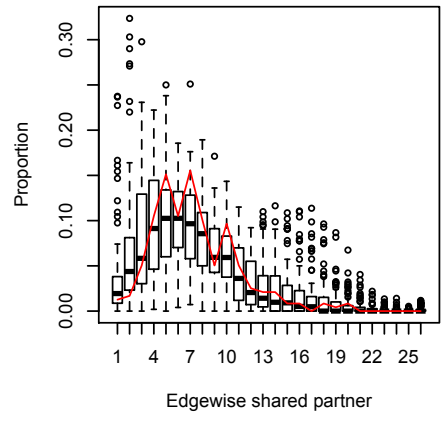












To Conclude...

- Model for networks with timestamped, text-valued ties
- Many potential applications in political science
- Developement of R package 'IPTM'

I otter know the answer...

I don't have the necessary koalafications...

If they'd gibbon me more time...

If I'd implemented the IPTM with my bear hands...

I'd be lion if I said yes...

Let's table bat discussion...