Latent Variable Models for Text, Event, and Network Data

MURI Project: University of California, Irvine

Annual Review Meeting

December 8th 2009

Padhraic Smyth

(joint work with Arthur Asuncion and Chris DuBois)

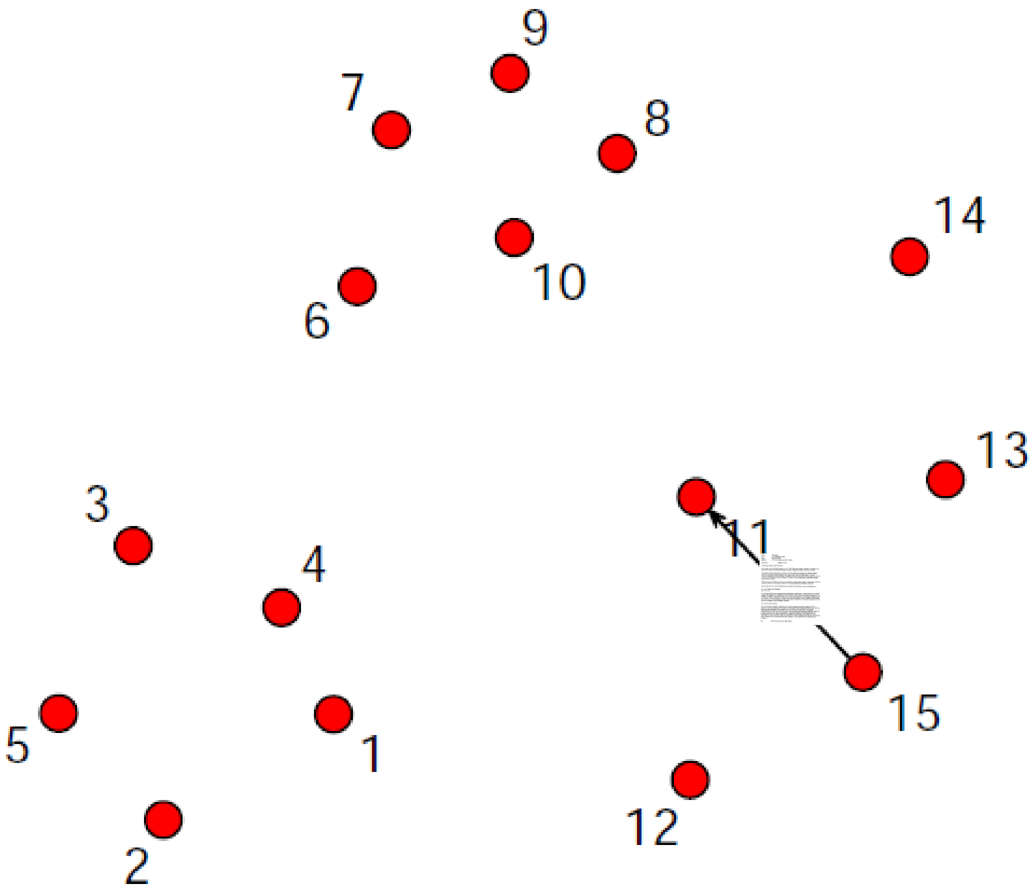
Event, Text, Network Data

- Network: N actors

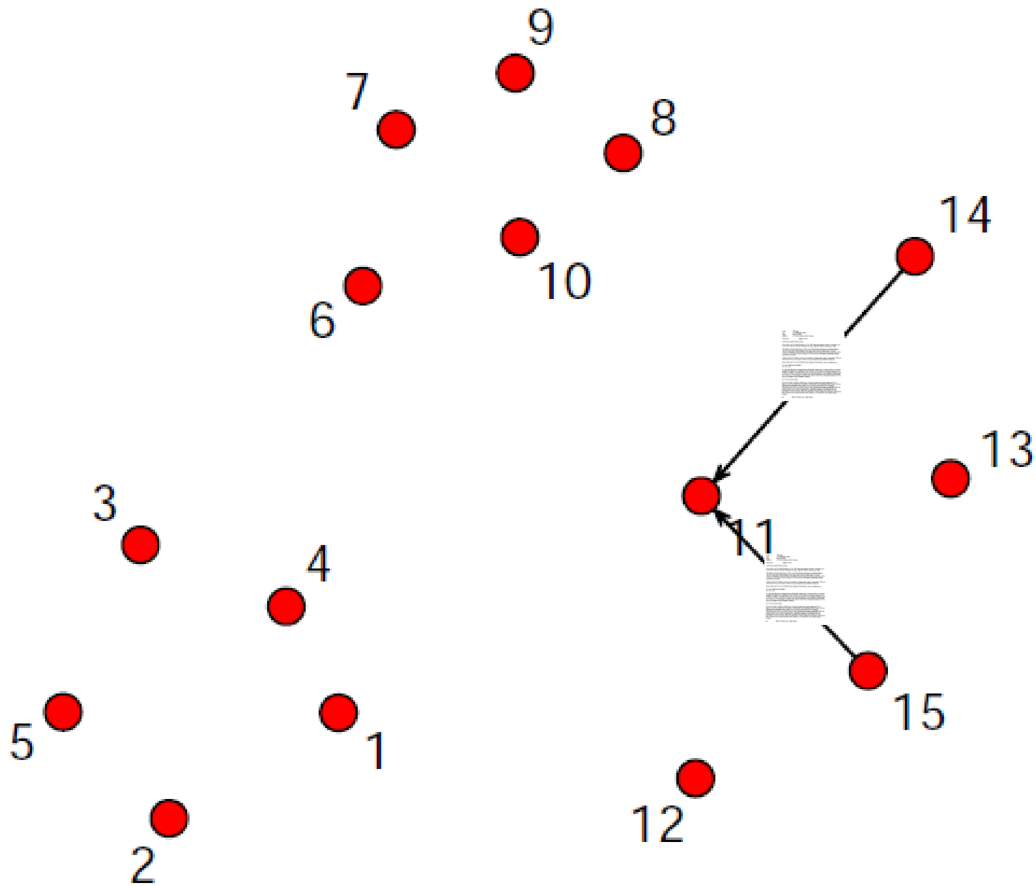
- Events:
  - Event i occurs at timestamp t with sender s and receiver r
  - Events are instantaneous
  - Note: interested in event-level data, not aggregates

- Text
  - e.g., document for each event i, e.g., email
  - e.g., text data for each actor

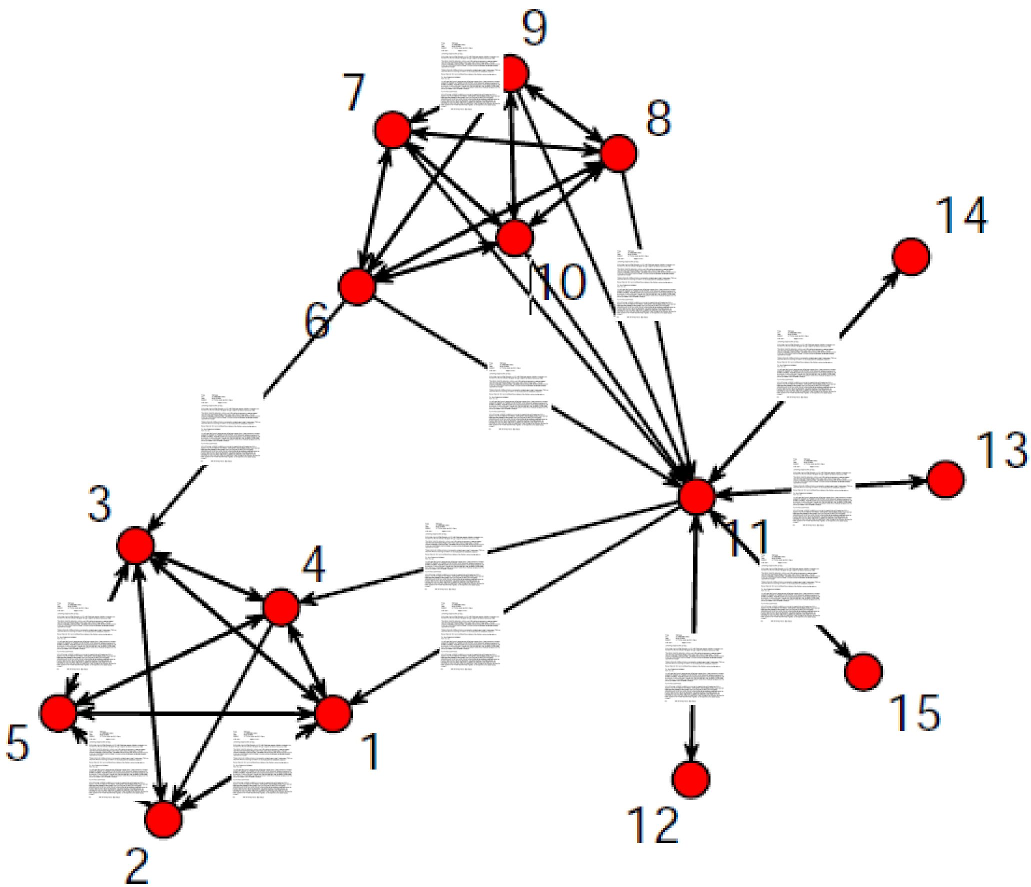
Time 1



Time 2



Time 50



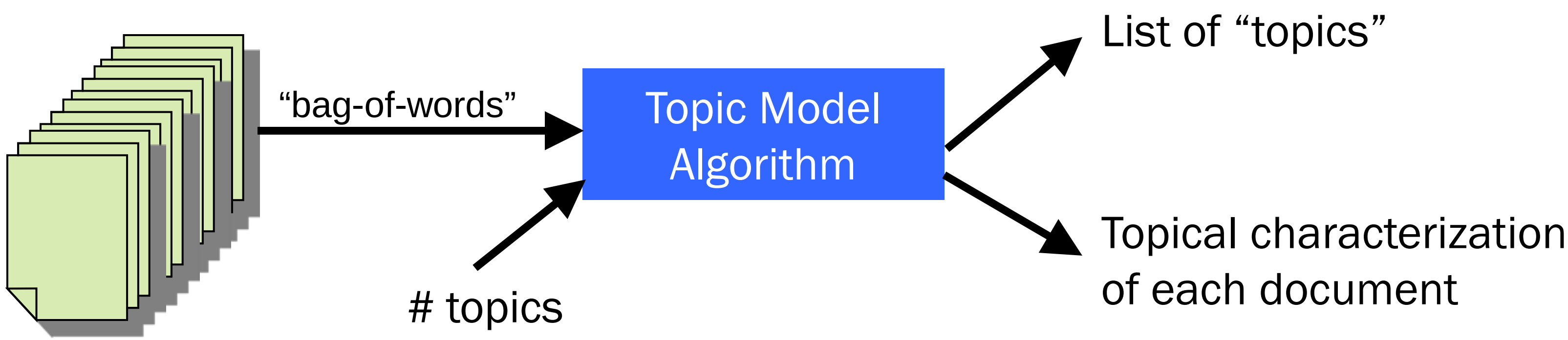
Motivation

- Real-world social networks often involve events and text
  - Email communications
  - Facebook postings
  - Blogs
  - Etc
- Want to build statistical models that
  - Provide insight into underlying processes
  - Allow us to make predictions
- Focus on “semi-parametric” models
  - Hidden/latent variables
  - Provides dimensionality reduction (and insight)

Outline

- Statistical topic models
  - "building block" for text modeling
- Relational topic models
  - Extending topic models to documents with links
- Scalable parallel algorithms for large data sets
- Event data
  - Learning "modes" of behavior for relational events
- Putting it together….
  - Current and future directions

Statistical Topic Modeling



- Original work by Blei, Ng, Jordan (2003)

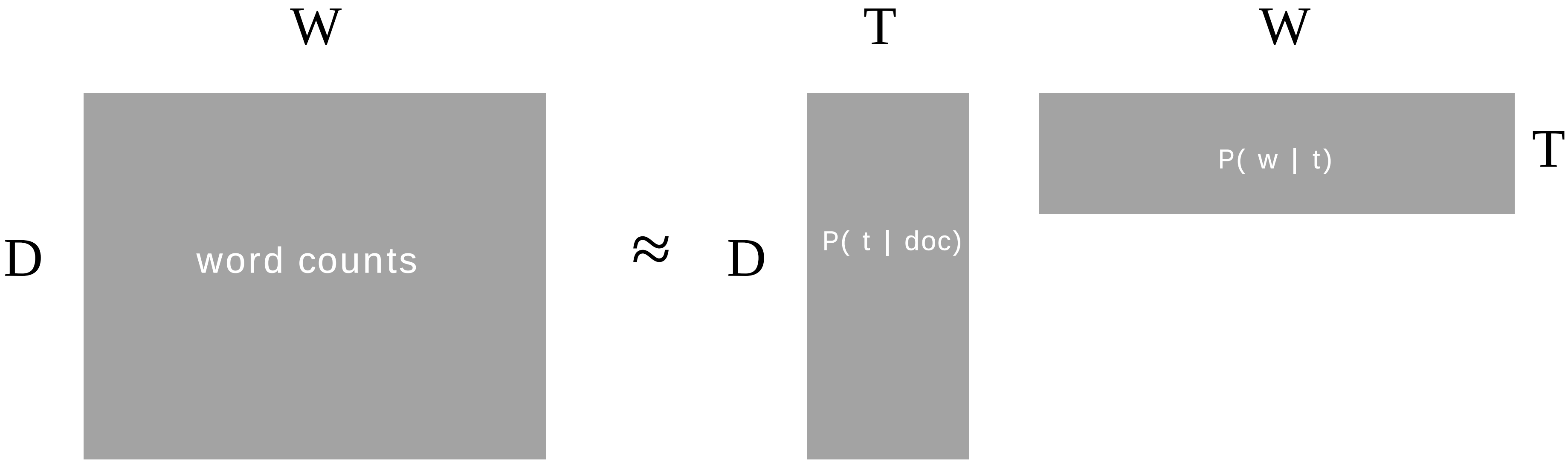
- Multiple applications:
  - Improved web searching
  - Automatic indexing of digital historical archives
  - Specialized search browsers (e.g. medical applications)
  - Legal applications (e.g. email forensics)

Statistical Topic Modeling

- Document = vector of word counts $\underline{w}$
- Topic = multinomial distribution over $\underline{w}$

  $= P(w_1, w_2, \ldots\ldots\ldots, w_W \mid t)$
- Assume T latent topics –> act as "basis functions"
- Words are generated by
  - Selecting a topic given a document from p(t | doc)
  - Selecting a word given a topic from P(w | t)
- Estimation:
  - Find P(w | t) by maximizing likelihood of observed words
  - Use collapsed Gibbs sampling: linear per iteration

Topics as Matrix Factorization



$$p(w_i|d) = \sum_{j=1}^{T} p(w_i|z_j)p(z_j|d)$$

Examples of Word-Topic Distributions

word	prob.
oxygen	0.136
carbon	0.097
dioxide	0.050
air	0.046
ramona	0.037
gas	0.036
nitrogen	0.030
gases	0.026
atmosphere	0.020
hydrogen	0.020
water	0.016
respiraion	0.014
process	0.014
beezus	0.012
breathe	0.011

word	prob.
president	0.129
roosevelt	0.032
congress	0.030
johnson	0.026
office	0.021
wilson	0.021
nixon	0.020
reagan	0.018
kennedy	0.018
carter	0.017
presidents	0.012
administration	0.012
presidential	0.011
white	0.011
budget	0.010

word	prob.
france	0.071
french	0.069
europe	0.051
germany	0.043
german	0.041
countries	0.030
britain	0.024
italy	0.019
western	0.019
european	0.019
british	0.016
war	0.015
germans	0.013
country	0.012
nations	0.012

From: PGE News
To: ALL PGE EMPLOYEES
Date: 8/14/01 2:54PM
Subject: Jeff Skilling resigns as CEO of Enron

PGE News .... August 14, 2001

Jeff Skilling resigns as CEO of Enron

Enron today announced that President and CEO Jeff Skilling has resigned, effective immediately, and that the Enron Board of Directors has asked Ken Lay to resume his role as Chairman and CEO.

"Stan Horton called this afternoon to inform me of Jeff's decision to step down for personal reasons," says PGE CEO and President Peggy Fowler. Horton, CEO of Enron Transportation, is Fowler's executive connection to the Enron team. "He wanted to let me know that Mr. Skilling's departure will not in any way impact Enron's ongoing strategy for success and we should expect no near-term dramatic organizational changes."

"Clearly, Enron will continue to focus on increasing the company's stock value," Fowler added. "PGE can help in this effort by remaining committed to our Scorecard goals and operational excellence."

Below is the letter Ken Lay is sending to Enron employees this afternoon announcing the decision:

To: Enron Employees Worldwide
From: Ken Lay

It is with regret that I have to announce that Jeff Skilling is leaving Enron. Today, the Board of Directors accepted his resignation as President and CEO of Enron. Jeff is resigning for personal reasons and his decision is voluntary. I regret his decision, but I accept and understand it. I have worked closely with Jeff for more than 15 years, including 11 here at Enron, and have had few, if any, professional relationships that I value more. I am pleased to say that he has agreed to enter into a consulting arrangement with the company to advise me and the Board of Directors.

Now it's time to look forward.

With Jeff leaving, the Board has asked me to resume the responsibilities of President and CEO in addition to my role as Chairman of the Board. I have agreed. I want to assure you that I have never felt better about the prospects for the company. All of you know that our stock price has suffered substantially over the last few months. One of my top priorities will be to restore a significant amount of the stock value we have lost as soon as possible. Our performance has never been stronger; our business model has never been more robust; our growth has never been more certain; and most importantly, we have never had a better nor deeper pool of talent throughout the company. We have the finest organization in American business today. Together, we will make Enron the world's leading company.

CC: Kathy & George Wyatt; Kathy Wyatt

Enron email data set:

250,000 emails

1999-2002

Enron email topics

TOPIC 36	
WORD	**PROB.**
FEEDBACK	0.0781
PERFORMANCE	0.0462
PROCESS	0.0455
PEP	0.0446
MANAGEMENT	0.03
COMPLETE	0.0205
QUESTIONS	0.0203
SELECTED	0.0187
COMPLETED	0.0146
SYSTEM	0.0146
SENDER	**PROB.**
perfmgmt	0.2195
perf eval process	0.0784
enron announcements	0.0489
***	0.0089
***	0.0048

TOPIC 72	
WORD	**PROB.**
PROJECT	0.0514
PLANT	0.028
COST	0.0182
CONSTRUCTION	0.0169
UNIT	0.0166
FACILITY	0.0165
SITE	0.0136
PROJECTS	0.0117
CONTRACT	0.011
UNITS	0.0106
SENDER	**PROB.**
***	0.0288
***	0.022
***	0.0123
***	0.0111
***	0.0108

TOPIC 54	
WORD	**PROB.**
FERC	0.0554
MARKET	0.0328
ISO	0.0226
COMMISSION	0.0215
ORDER	0.0212
FILING	0.0149
COMMENTS	0.0116
PRICE	0.0116
CALIFORNIA	0.0110
FILED	0.0110
SENDER	**PROB.**
***	0.0532
***	0.0454
***	0.0384
***	0.0334
***	0.0317

TOPIC 23	
WORD	**PROB.**
ENVIRONMENTAL	0.0291
AIR	0.0232
MTBE	0.019
EMISSIONS	0.017
CLEAN	0.0143
EPA	0.0133
PENDING	0.0129
SAFETY	0.0104
WATER	0.0092
GASOLINE	0.0086
SENDER	**PROB.**
***	0.1339
***	0.0275
***	0.0205
***	0.0166
***	0.0129

Non-work Topics...

TOPIC 66	
WORD	**PROB.**
HOLIDAY	0.0857
PARTY	0.0368
YEAR	0.0316
SEASON	0.0305
COMPANY	0.0255
CELEBRATION	0.0199
ENRON	0.0198
TIME	0.0194
RECOGNIZE	0.019
MONTH	0.018
SENDER	**PROB.**
chairman & ceo	0.131
***	0.0102
***	0.0046
***	0.0022
general announcement	0.0017

TOPIC 182	
WORD	**PROB.**
TEXANS	0.0145
WIN	0.0143
FOOTBALL	0.0137
FANTASY	0.0129
SPORTSLINE	0.0129
PLAY	0.0123
TEAM	0.0114
GAME	0.0112
SPORTS	0.011
GAMES	0.0109
SENDER	**PROB.**
cbs sportsline com	0.0866
houston texans	0.0267
houstontexans	0.0203
sportsline rewards	0.0175
pro football	0.0136

TOPIC 113	
WORD	**PROB.**
GOD	0.0357
LIFE	0.0272
MAN	0.0116
PEOPLE	0.0103
CHRIST	0.0092
FAITH	0.0083
LORD	0.0079
JESUS	0.0075
SPIRITUAL	0.0066
VISIT	0.0065
SENDER	**PROB.**
crosswalk com	0.2358
wordsmith	0.0208
***	0.0107
doctor dictionary	0.0101
***	0.0061

TOPIC 109	
WORD	**PROB.**
AMAZON	0.0312
GIFT	0.0226
CLICK	0.0193
SAVE	0.0147
SHOPPING	0.0140
OFFER	0.0124
HOLIDAY	0.0122
RECEIVE	0.0102
SHIPPING	0.0100
FLOWERS	0.0099
SENDER	**PROB.**
amazon com	0.1344
jos a bank	0.0266
sharperimageoffers	0.0136
travelocity com	0.0094
barnes & noble com	0.0089

Topical Topics

TOPIC 18	
WORD	**PROB.**
POWER	0.0915
CALIFORNIA	0.0756
ELECTRICITY	0.0331
UTILITIES	0.0253
PRICES	0.0249
MARKET	0.0244
PRICE	0.0207
UTILITY	0.0140
CUSTOMERS	0.0134
ELECTRIC	0.0120
SENDER	**PROB.**
***	0.1160
***	0.0518
***	0.0284
***	0.0272
***	0.0266

TOPIC 22	
WORD	**PROB.**
STATE	0.0253
PLAN	0.0245
CALIFORNIA	0.0137
POLITICIAN Y	0.0137
RATE	0.0131
BANKRUPTCY	0.0126
SOCAL	0.0119
POWER	0.0114
BONDS	0.0109
MOU	0.0107
SENDER	**PROB.**
***	0.0395
***	0.0337
***	0.0295
***	0.0251
***	0.0202

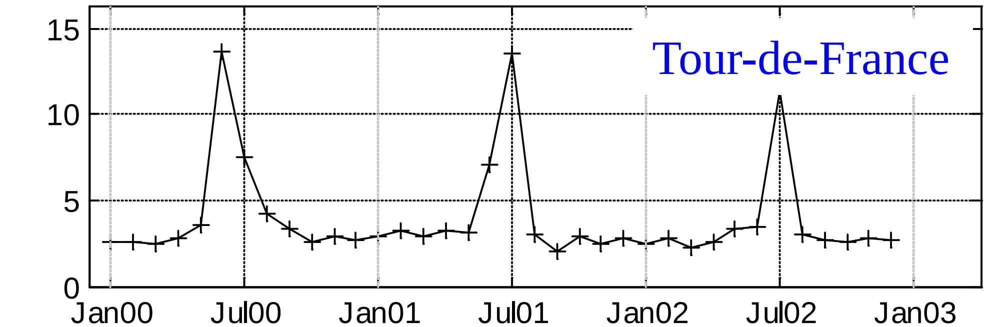
TOPIC 114	
WORD	**PROB.**
COMMITTEE	0.0197
BILL	0.0189
HOUSE	0.0169
WASHINGTON	0.0140
SENATE	0.0135
POLITICIAN X	0.0114
CONGRESS	0.0112
PRESIDENT	0.0105
LEGISLATION	0.0099
DC	0.0093
SENDER	**PROB.**
***	0.0696
***	0.0453
***	0.0255
***	0.0173
***	0.0317

TOPIC 194	
WORD	**PROB.**
LAW	0.0380
TESTIMONY	0.0201
ATTORNEY	0.0164
SETTLEMENT	0.0131
LEGAL	0.0100
EXHIBIT	0.0098
CLE	0.0093
SOCALGAS	0.0093
METALS	0.0091
PERSON Z	0.0083
SENDER	**PROB.**
***	0.0696
***	0.0453
***	0.0255
***	0.0173
***	0.0317

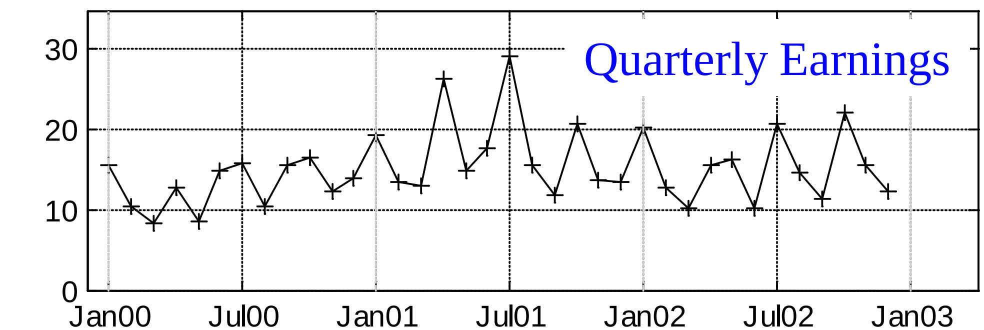
Topic trends from New York Times

330,000 articles

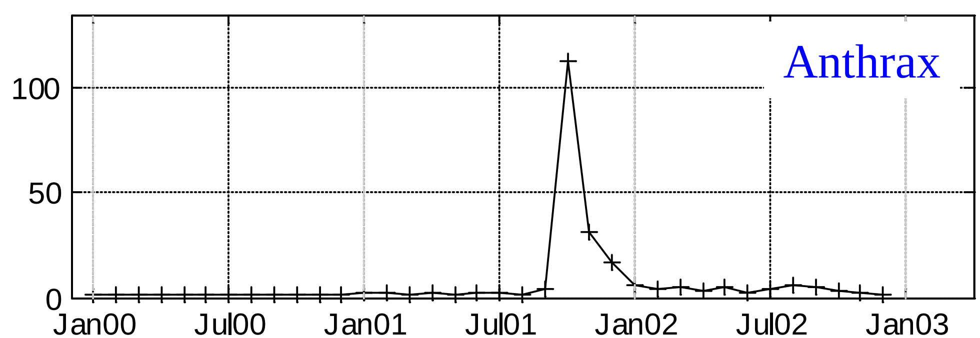
2000-2002



TOUR
RIDER
LANCE_ARMSTRONG
TEAM
BIKE
RACE
FRANCE



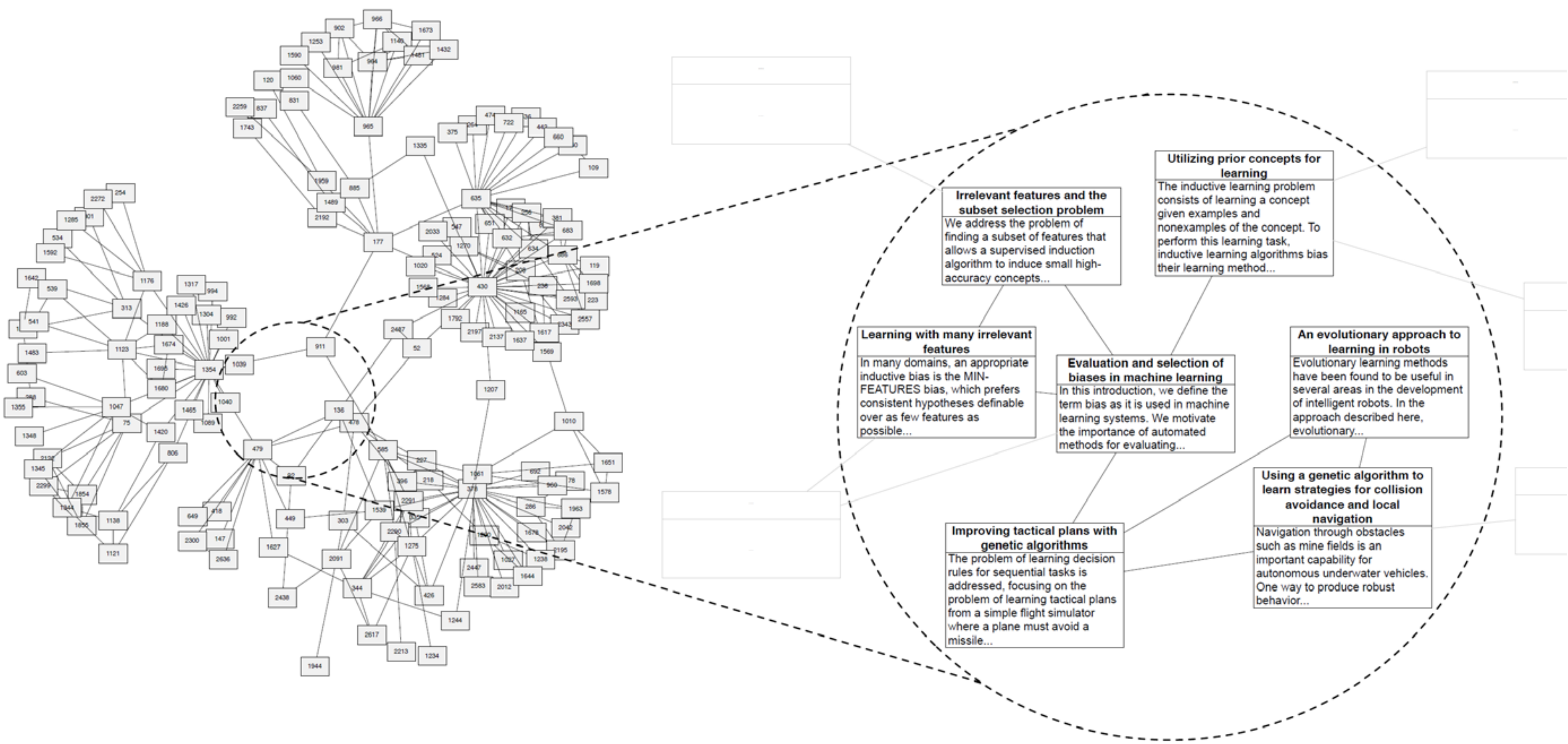
COMPANY
QUARTER
PERCENT
ANALYST
SHARE
SALES
EARNING



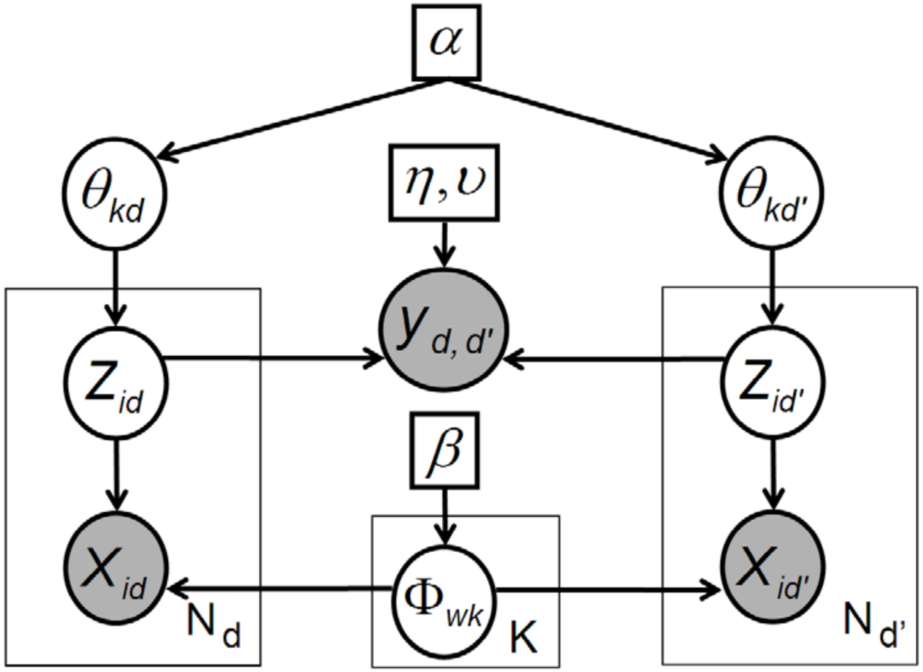
ANTHRAX
LETTER
MAIL
WORKER
OFFICE
SPORES
POSTAL
BUILDING

Relational Topic Models

[Chang, Blei, 2009]



Relational Topic Models



$$y_{d,d'} \sim \psi(y_{d,d'} | \mathbf{z}_d, \mathbf{z}_{d'}, \eta, \nu)$$

"Link probability function"

Where, for example

$$\psi(y_{d,d'} = 1) = \exp(\eta^T(\bar{\mathbf{z}}_d \circ \bar{\mathbf{z}}_{d'}) + \nu)$$

(similar to latent-space model)

Collapsed Gibbs sampling for RTM

- Conditional distribution of each z:

$$p(z_{id} = k \mid \mathbf{z}^{\neg id}, -) \propto (N_{dk}^{\neg id} + \alpha) \frac{(N_{kw}^{\neg id} + \beta)}{(N_k^{\neg id} + W\beta)}$$ ⟵ LDA term

$$\prod_{d' \neq d : y_{d,d'} = 1} \psi_e(y_{d,d'} = 1 | \mathbf{z}_d, \mathbf{z}_{d'}, \eta, \nu)$$ ⟵ "Edge" term

$$\prod_{d' \neq d : y_{d,d'} = 0} \psi_e(y_{d,d'} = 0 | \mathbf{z}_d, \mathbf{z}_{d'}, \eta, \nu)$$ ⟵ "Non-edge" term

- Using the exponential link probability function, it is computationally efficient to calculate the "edge" term.

- It is **very costly** to compute the "non-edge" term exactly
  -> can explore various efficient ways to approximate this term

Results on Movie Data

Wikipedia pages of 10,000 movies

Movies are linked if they have a common director or common actor

Model trained on subgraph and tested on different subgraph

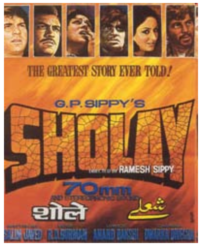
ALGORITHM	MEAN LINK RANK OF PREDICTIONS
Random Guessing	5000
LDA + Regression	2321
Ignoring Non-Edges	1955
Fast Approximation	2089
Subsampling 5% + Caching	1739

Examples of Movie Data Topics

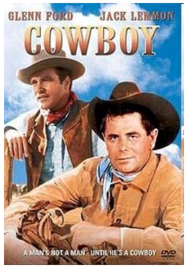
POLICE:	[t2] police agent kill gun action escape car film
DISNEY:	[t4] disney film animated movie christmas cat animation story
AMERICAN:	[t5] president war american political united states government against
CHINESE:	[t6] film kong hong chinese chan wong china link
WESTERN:	[t7] western town texas sheriff eastwood west clint genre
SCI-FI:	[t8] earth science space fiction alien bond planet ship
AWARDS:	[t9] award film academy nominated won actor actress picture
WAR:	[t20] war soldier army officer captain air military general
FRENCH:	[t21] french film jean france paris fran les link
HINDI:	[t24] film hindi award link india khan indian music
MUSIC:	[t28] album song band music rock live soundtrack record
JAPANESE:	[t30] anime japanese manga series english japan retrieved character
BRITISH:	[t31] british play london john shakespeare film production sir
FAMILY:	[t32] love girl mother family father friend school sister
SERIES:	[t35] series television show episode season character episodes original
SPIELBERG:	[t36] spielberg steven park joe future marty gremlin jurassic
MEDIEVAL	[t37] king island robin treasure princess lost adventure castle
GERMAN:	[t38] film german russian von germany language anna soviet
GIBSON:	[t41] max ben danny gibson johnny mad ice mel
MUSICAL:	[t42] musical phantom opera song music broadway stage judy
BATTLE:	[t43] power human world attack character battle earth game
MURDER:	[t46] death murder kill police killed wife later killer
SPORTS:	[t47] team game player rocky baseball play charlie ruth
KING:	[t48] king henry arthur queen knight anne prince elizabeth
HORROR:	[t49] horror film dracula scooby doo vampire blood ghost

Predictions on Movie Data

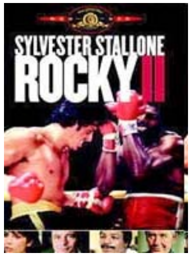
- **'Sholay'**
  - Indian film, 45% of words belong to topic 24 (Hindi topic)
  - Top 5 most probable movie links in training set:
    - 'Laawaris'
    - 'Hote Hote Pyaar Ho Gaya'
    - 'Trishul'
    - 'Mr. Natwarlal'
    - 'Rangeela'



- **'Cowboy'**
  - Western film, 25% of words belong to topic 7 (western topic)
  - Top 5 most probable movie links in training set:
    - 'Tall in the Saddle'
    - 'The Indian Fighter'
    - 'Dakota'
    - 'The Train Robbers'
    - 'A Lady Takes a Chance'



- **'Rocky II'**
  - Boxing film, 40% of words belong to topic 47 (sports topic)
  - Top 5 most probable movie links in training set:
    - 'Bull Durham'
    - '2003 World Series'
    - 'Bowfinger'
    - 'Rocky V'
    - 'Rocky IV'



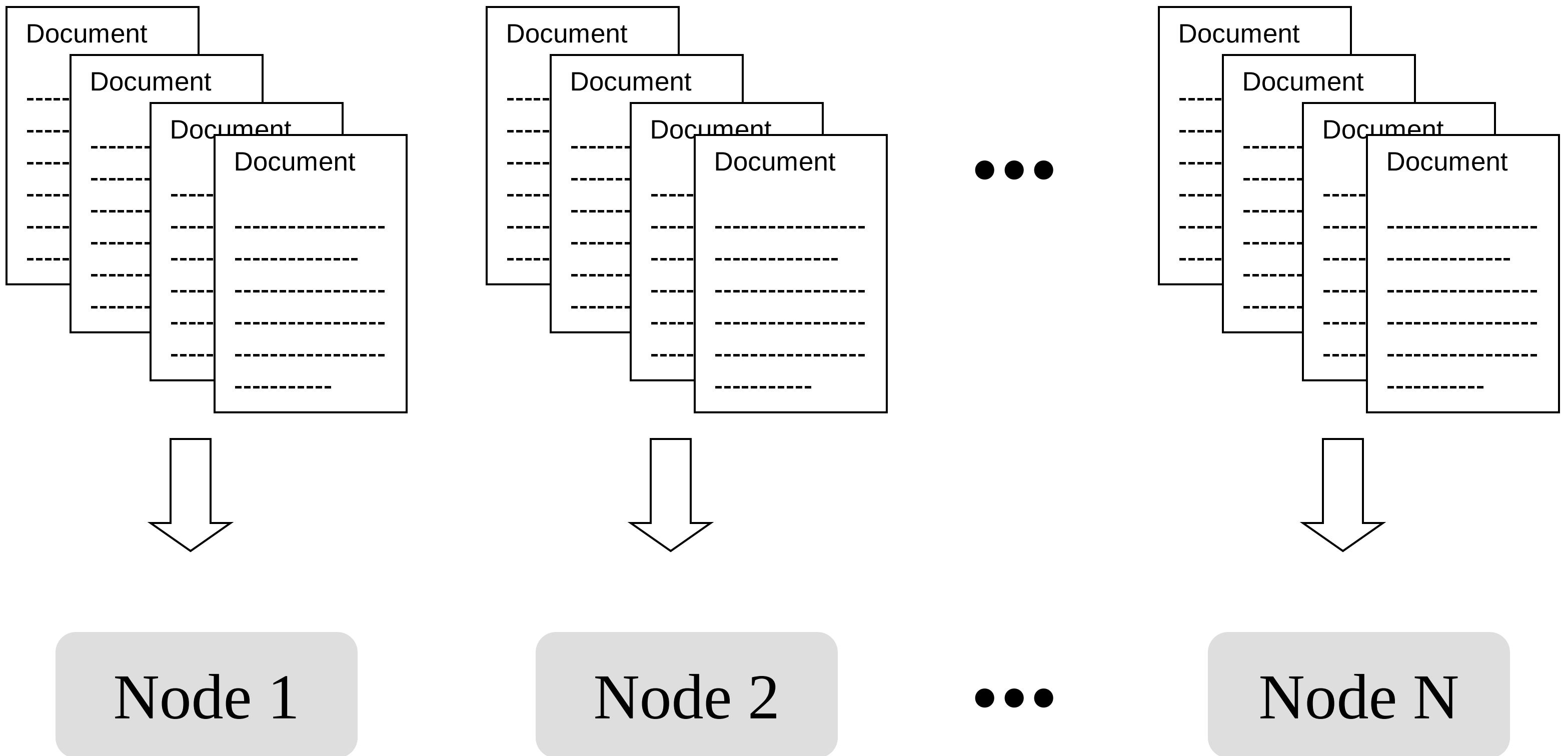
Scalability

- Two Problems:
  - – Very large data sets will not fit in main memory
  - – Topic model learning is not real-time
    - Algorithm is linear time, but constant can be large

- Solutions:
  - – Distributed topic learning (Newman et al, NIPS 2007; JMLR in press)
    - Factor of P speedup, with P processors, 70% efficiency
  - – Fast sampling algorithms (Porteous et al, ACM SIGKDD, 2008)
  - – More general extensions
    - – Asuncion, Welling, Smyth, NIPS 2008
    - – Asuncion, Welling, Smyth, UAI 2009

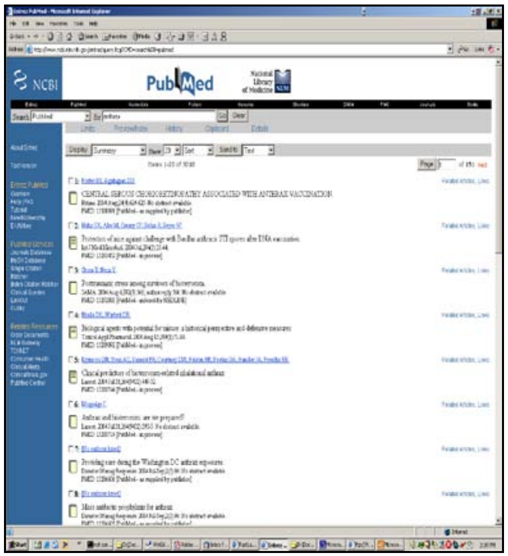
Distributed Topic Modeling

Newman, Asuncion, Smyth, Welling, NIPS 2007, NIPS 2008

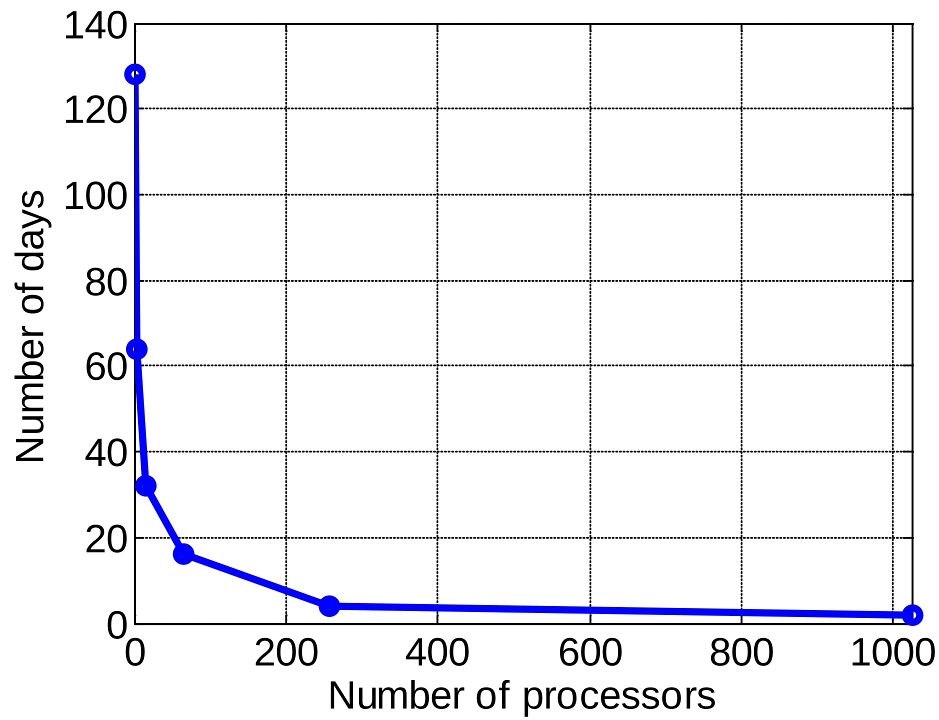


Global synchronization of statistics after each local sampling pass

Large Scale Experiments



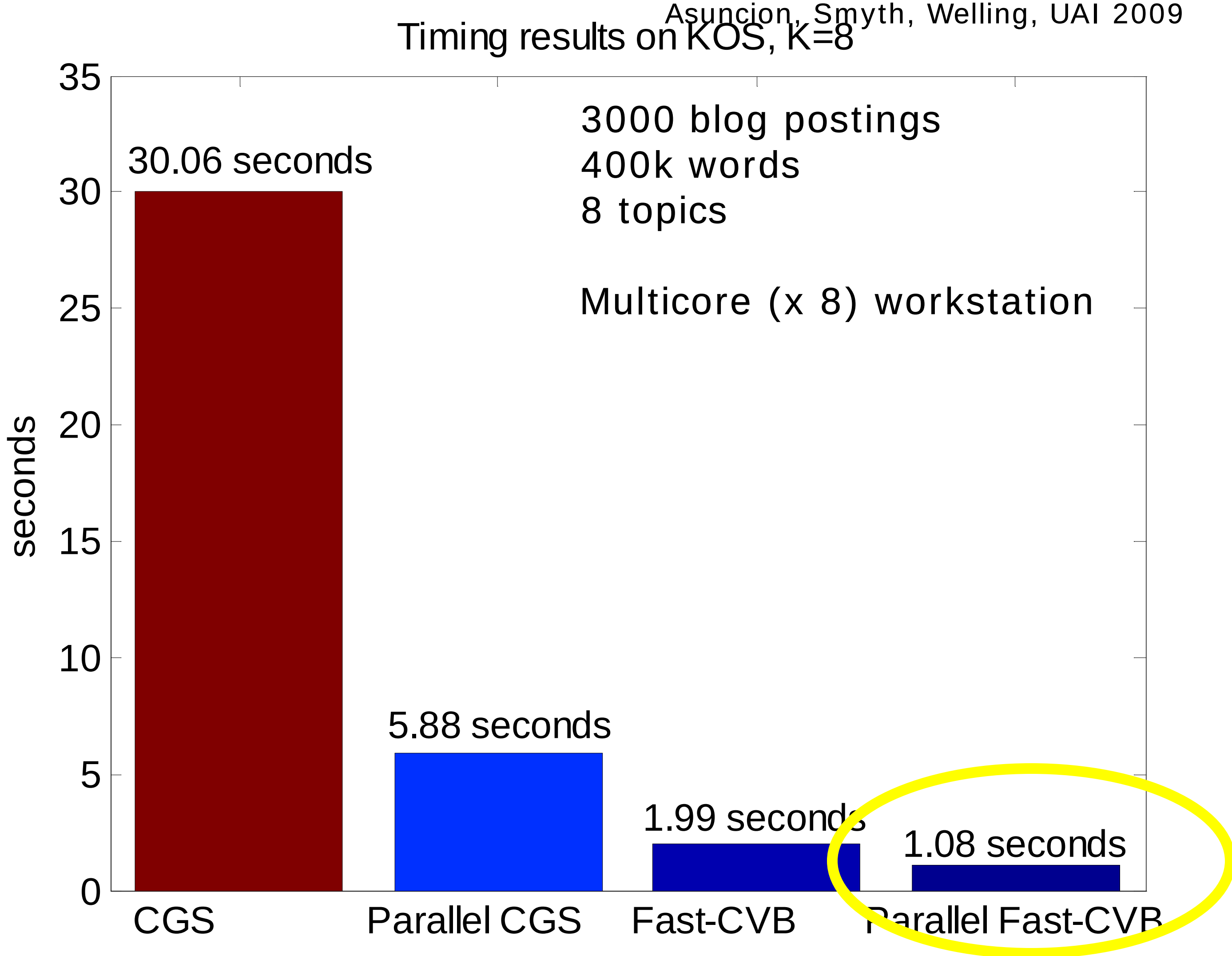
MEDLINE
8 million abstracts
1 billion words
2000 topics



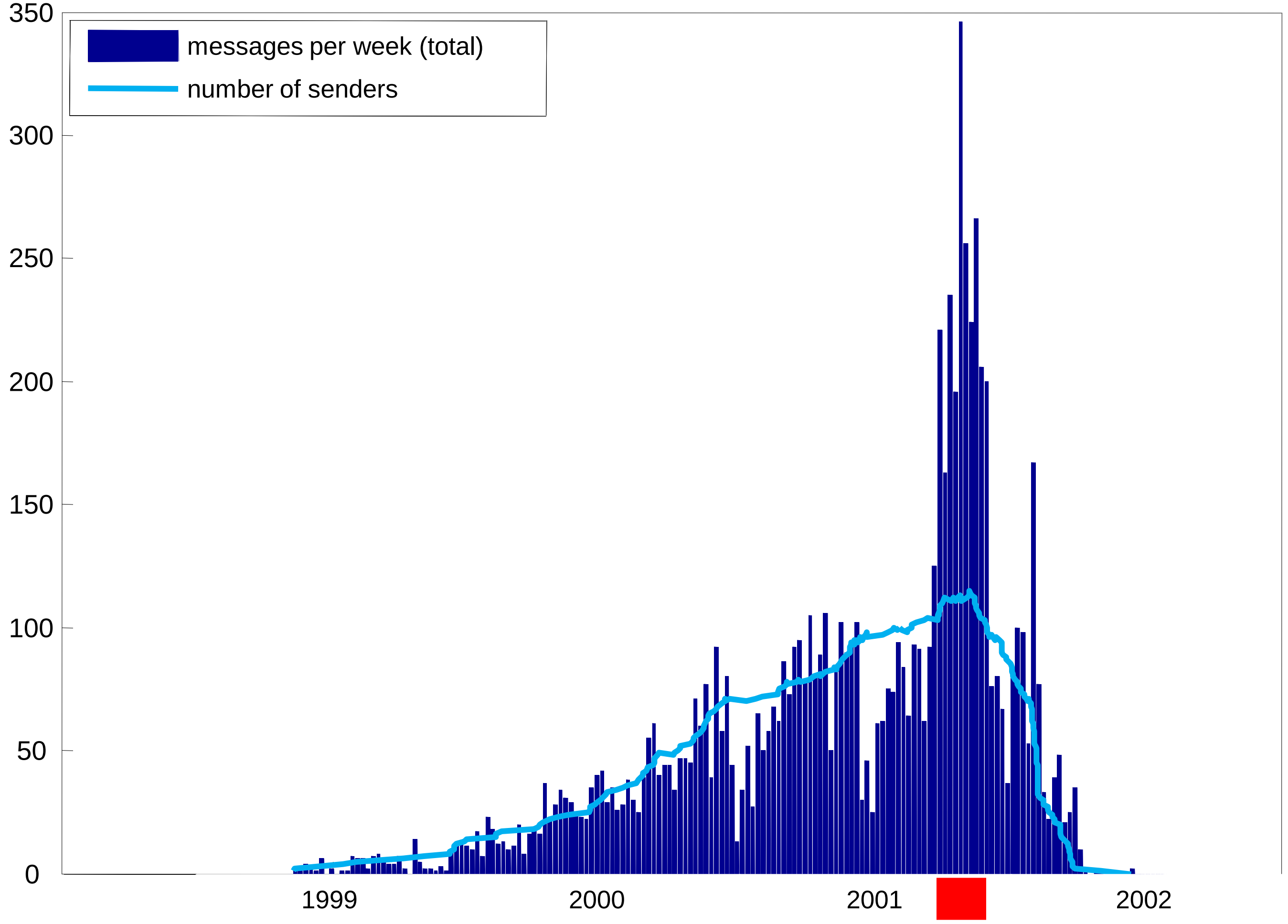
Experiments with 1000 processors
at the San Diego Supercomputing
Center (SDSC)

Real-Time Topic Modeling

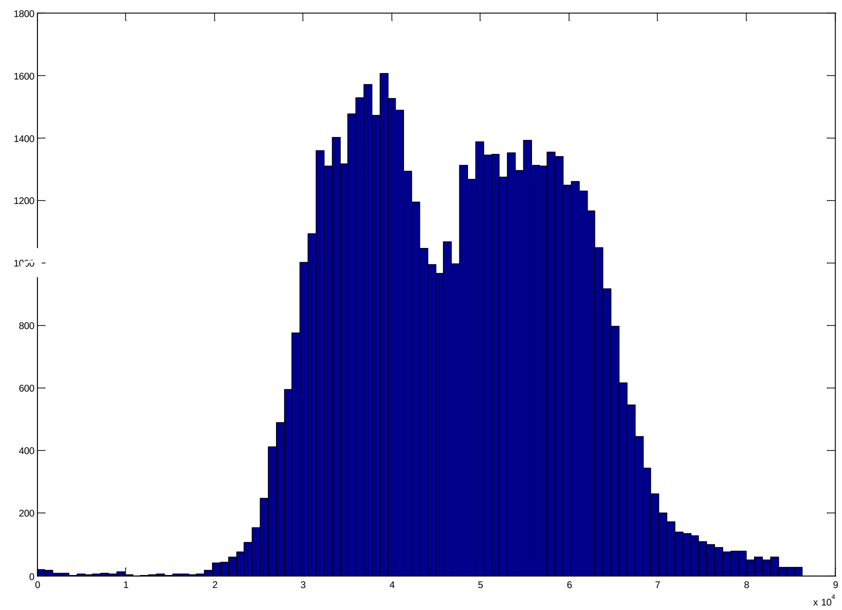
Asuncion, Smyth, Welling, UAI 2009

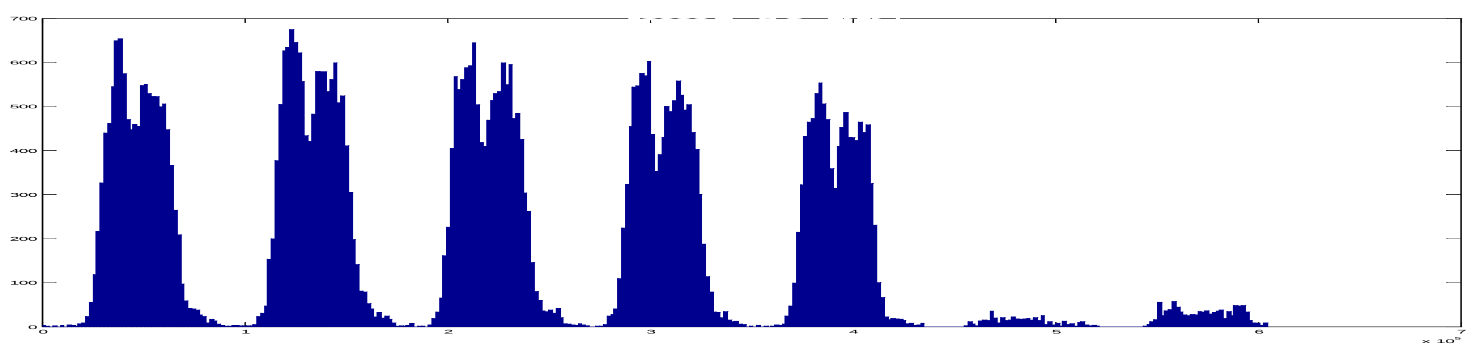


Enron email dataset



Daily and weekly variation





Latent Model for Event Data

Poster by Chris DuBois

- Data
  - Events = { <sender, receiver, timestamp> }

- Notation
  - Sender s, receiver r
  - K latent modes, m_k

- Generative model

$$m_k \sim P(m_k \mid \text{time } t)$$
$$s_i \sim P(s \mid m_k)$$
$$r_i \sim P(r \mid m_k)$$

- The m_k represent latent "modes" of network behavior
  - can be learned from the data
  - low-dimensional "space" for large network

Similarities to Topic Model

Topics for Text

Topic: $P(z_k \mid doc)$

Word: $P(w \mid z_k)$

$P(w \mid doc)$

$= \sum P(w \mid z_k) \, P(z_k \mid doc)$

Similarities to Topic Model

Topics for Text

Topic: $P(z_k \mid doc)$

Word: $P(w \mid z_k)$

$P(w \mid doc)$

$= \sum P(w \mid z_k) P(z_k \mid doc)$

Modes for Events

Mode: $P(m_k \mid time)$

Event: $P(s, r \mid m_k)$

$P(s, r \mid time)$

$= \sum P(s, r \mid m_k) P(m_k \mid time)$

Similarities to Topic Model

Topics for Text	**Modes for Events**
Topic: $P(z_k \mid doc)$	Mode: $P(m_k \mid time)$
Word: $P(w \mid z_k)$	Event: $P(s, r \mid m_k)$
$P(w \mid doc)$	$P(s, r \mid time)$
$= \Sigma P(w \mid z_k) P(z_k \mid doc)$	$= \Sigma P(s, r \mid m_k) P(m_k \mid time)$

Similarities to Topic Model

Topics for Text	**Modes for Events**
Topic: $P(z_k \mid \text{doc})$	Mode: $P(m_k \mid \text{time})$
Word: $P(w \mid z_k)$	Event: $P(s, r \mid m_k)$
$P(w \mid \text{doc}) = \sum P(w \mid z_k)\, P(z_k \mid \text{doc})$	$P(s, r \mid \text{time}) = \sum P(s, r \mid m_k)\, P(m_k \mid \text{time})$

Similarities to Topic Model

Topics for Text

Topic: $P(z_k \mid doc)$

Word: $P(w \mid z_k)$

$P(w \mid doc)$

$= \sum P(w \mid z_k)\, P(z_k \mid doc)$

Modes for Events

Mode: $P(m_k \mid time)$

Event: $P(s, r \mid m_k)$

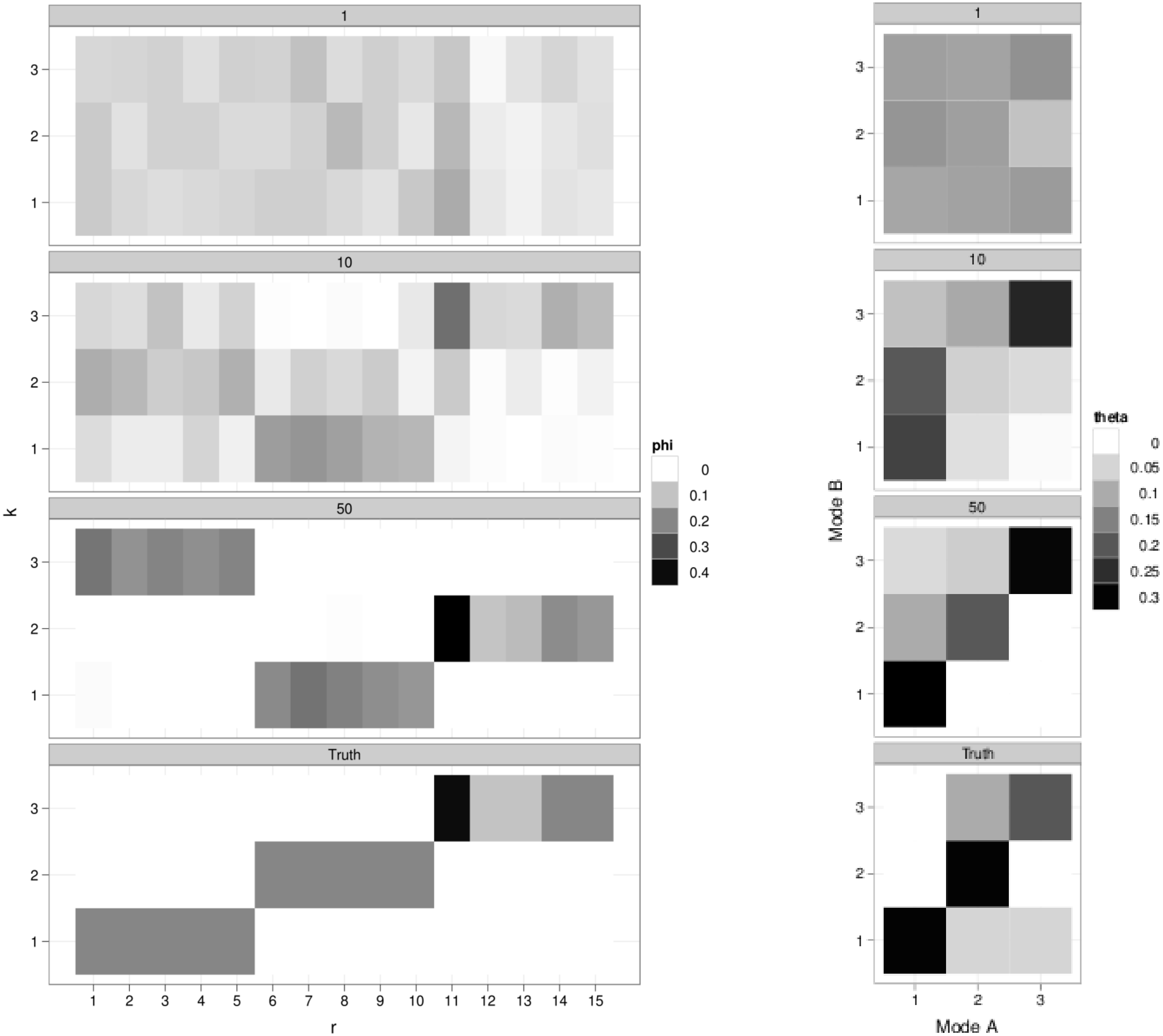
$P(s, r \mid time)$

$= \sum P(s, r \mid m_k)\, P(m_k \mid time)$

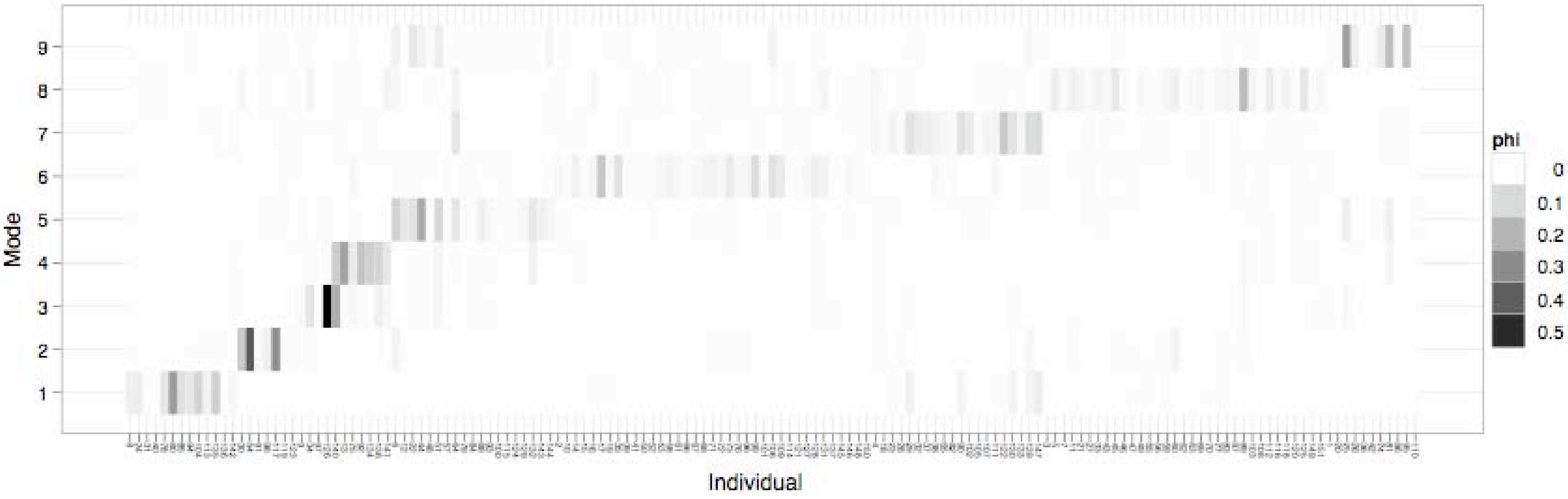
Similarities to Topic Model

Topics for Text	**Modes for Events**
Topic: $P(z_k \mid doc)$	Mode: $P(m_k \mid time)$
Word: $P(w \mid z_k)$	Event: $P(s, r \mid m_k)$
$P(w \mid doc) = \sum P(w \mid z_k) P(z_k \mid doc)$	$P(s, r \mid time) = \sum P(s, r \mid m_k) P(m_k \mid time)$

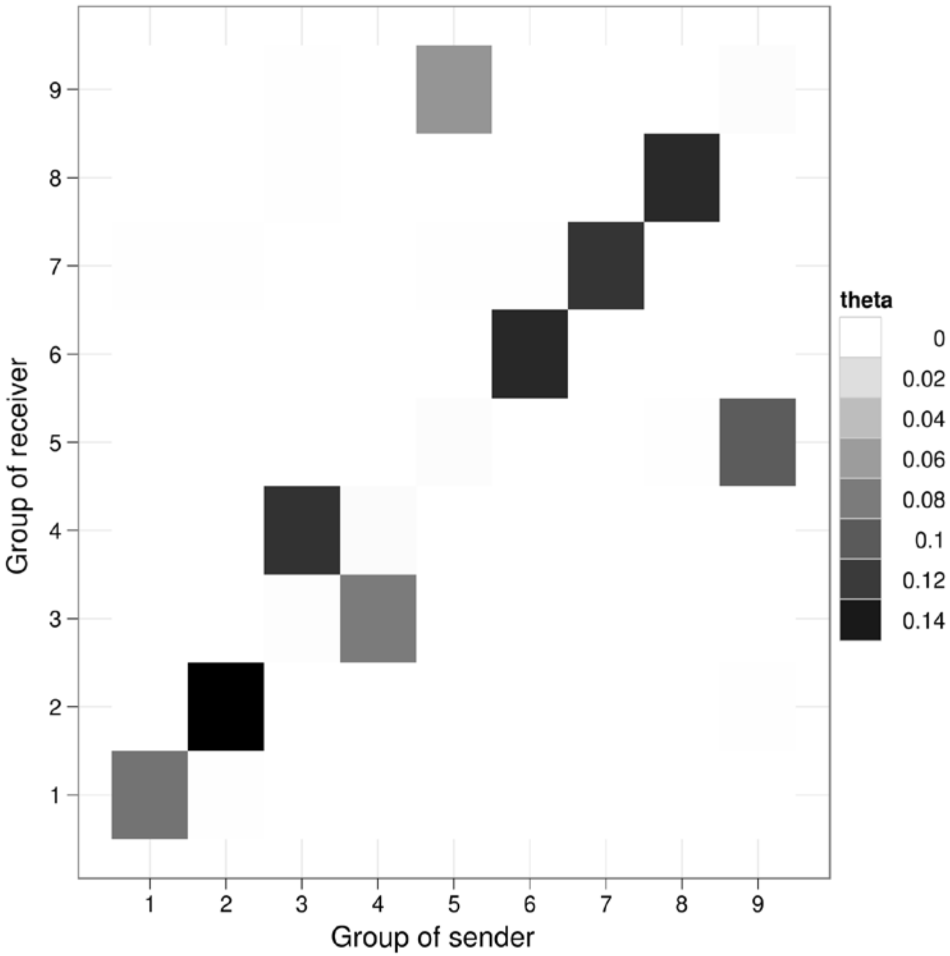
Can use same estimation techniques, e.g., collapsed Gibbs sampling


1
10
50
Truth
k
r
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15
phi
0
0.1
0.2
0.3
0.4
Mode B
Mode A
theta
0
0.05
0.1
0.15
0.2
0.25
0.3

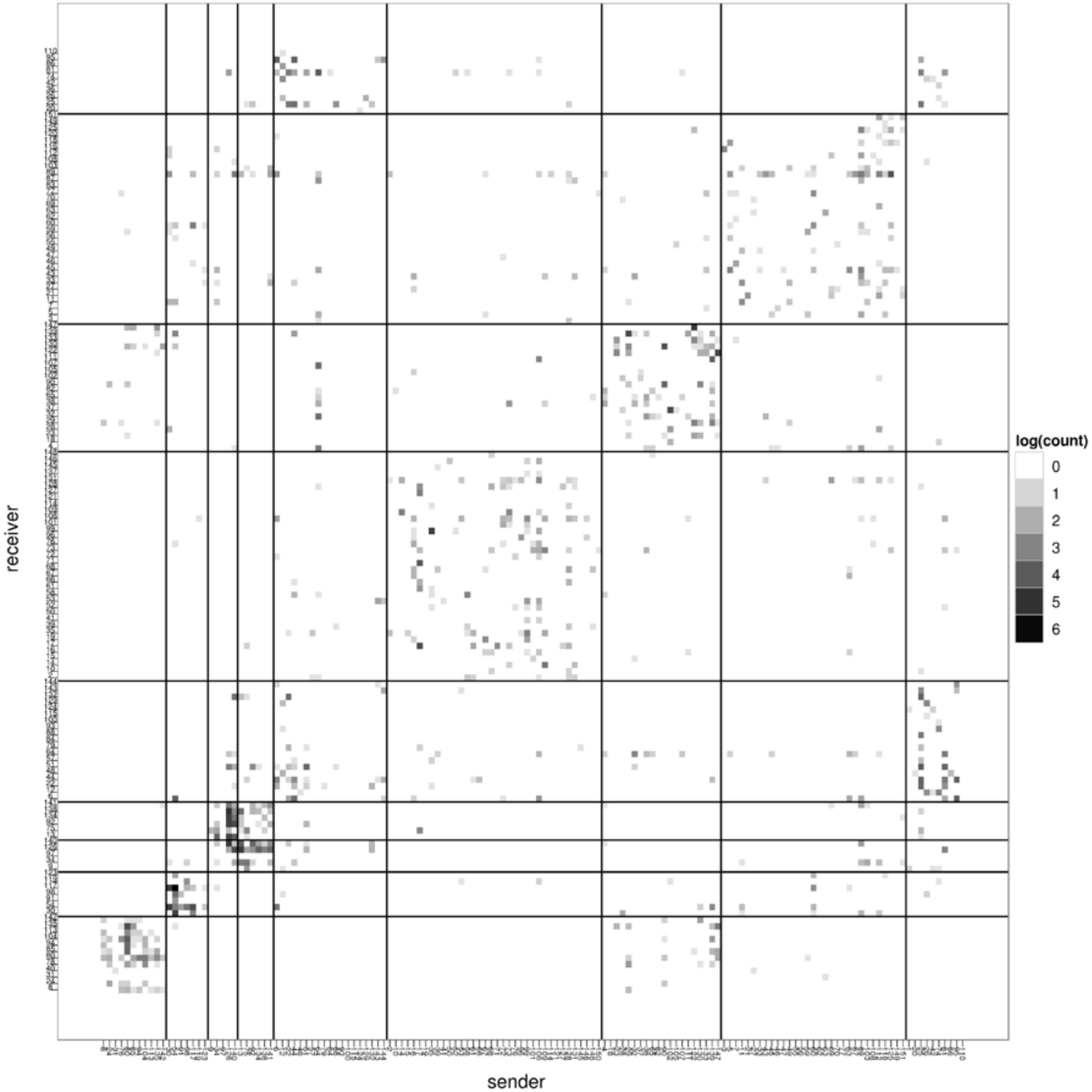
Enron: Mode Probabilities for Senders and Receivers



Enron: Joint Sender-Receiver Mode Probabilities



Number of emails sent between individuals, grouped by modes.



Ongoing and Future Work

- Add Markov dependence to the modes
  - $P(m_k \mid m_{k-1})$, e.g., model persistence
  - Results in hidden Markov model
  - Collapsed Gibbs sampling again applicable...

- Add richer structure
  - Dependence on time of day, day of week
  - Dependence on covariates
  - Extend to relational events

- Integrate events with text
  - Joint models over events and text associated with events