Analyzing Text and Social Network Data with Probabilistic Models

Padhraic Smyth
Department of Computer Science
University of California, Irvine

Acknowledgements

- **Students (past and present)**
  - Arthur Asuncion, Chaitanya Chemudugunta, America Chambers, Chris DuBois, Jimmy Foulds, Tim Rubin, Nick Navaroli

- **Collaborators**
  - Mark Steyvers, Dave Newman, Michal Rosen-Zvi, Tom Griffiths, Carter Butts, Dave Hunter, Max Welling

- **Funding**
  - NSF, ONR, Google, Yahoo!, Microsoft, Experian

ECML/PKDD-2002

13th European Conference on Machine Learning (ECML'02)

6th European Conference on Principles and Practice of Knowledge Discovery in Databases (PKDD'02)

19-23 August 2002

Helsinki, Finland

$$p(x) = \sum_{k=1}^{K} p(x|z_k)p(z_k)$$

ECML/PKDD-2002

13th European Conference on Machine Learning (ECML'02)

6th European Conference on Principles and Practice of Knowledge Discovery in Databases (PKDD'02)

19-23 August 2002

Helsinki, Finland

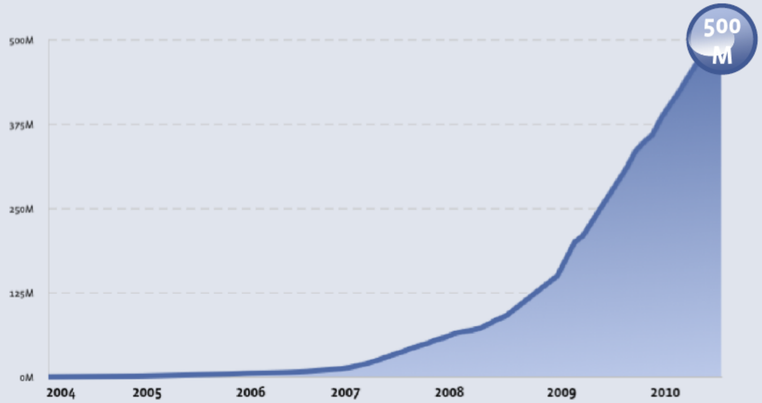
$$p(x) = \sum_{k=1}^{K} p(x|z_k)p(z_k)$$

$$p(x|d) = \sum_{k=1}^{K} p(x|z_k)p(z_k|d)$$

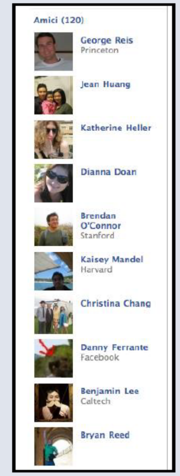
facebook

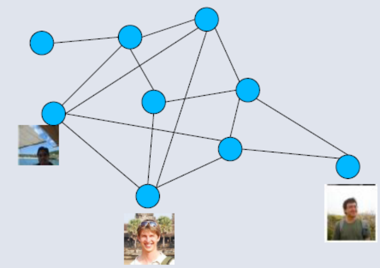
Graphics from Lars Backstrom, ESWC 2012

500 million 30-day active users



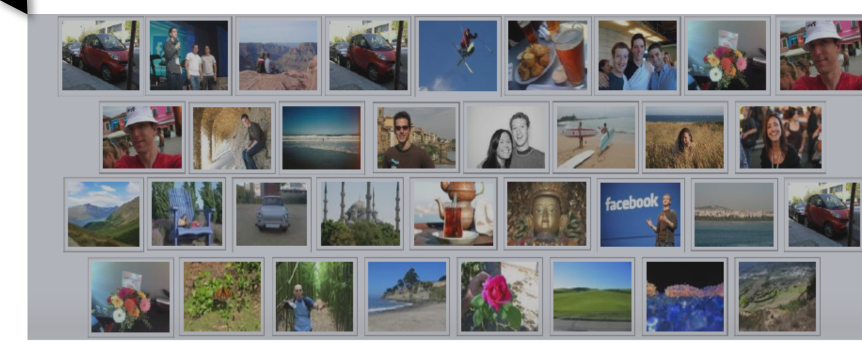
The Friendship graph





500M users each connect to an average of 130 other users = ~ 60 Billion Edges

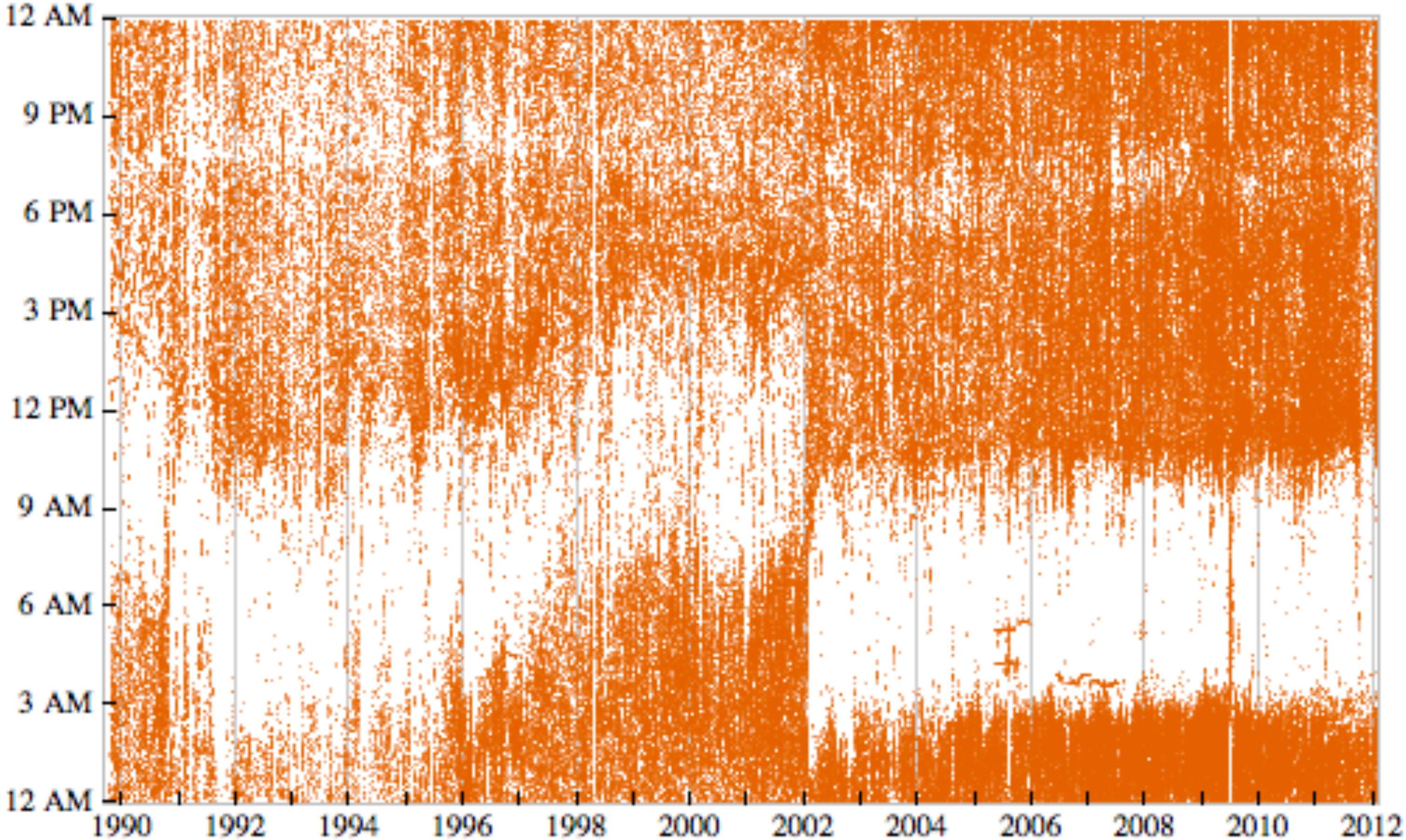
Over 30 billion pieces of content shared every month

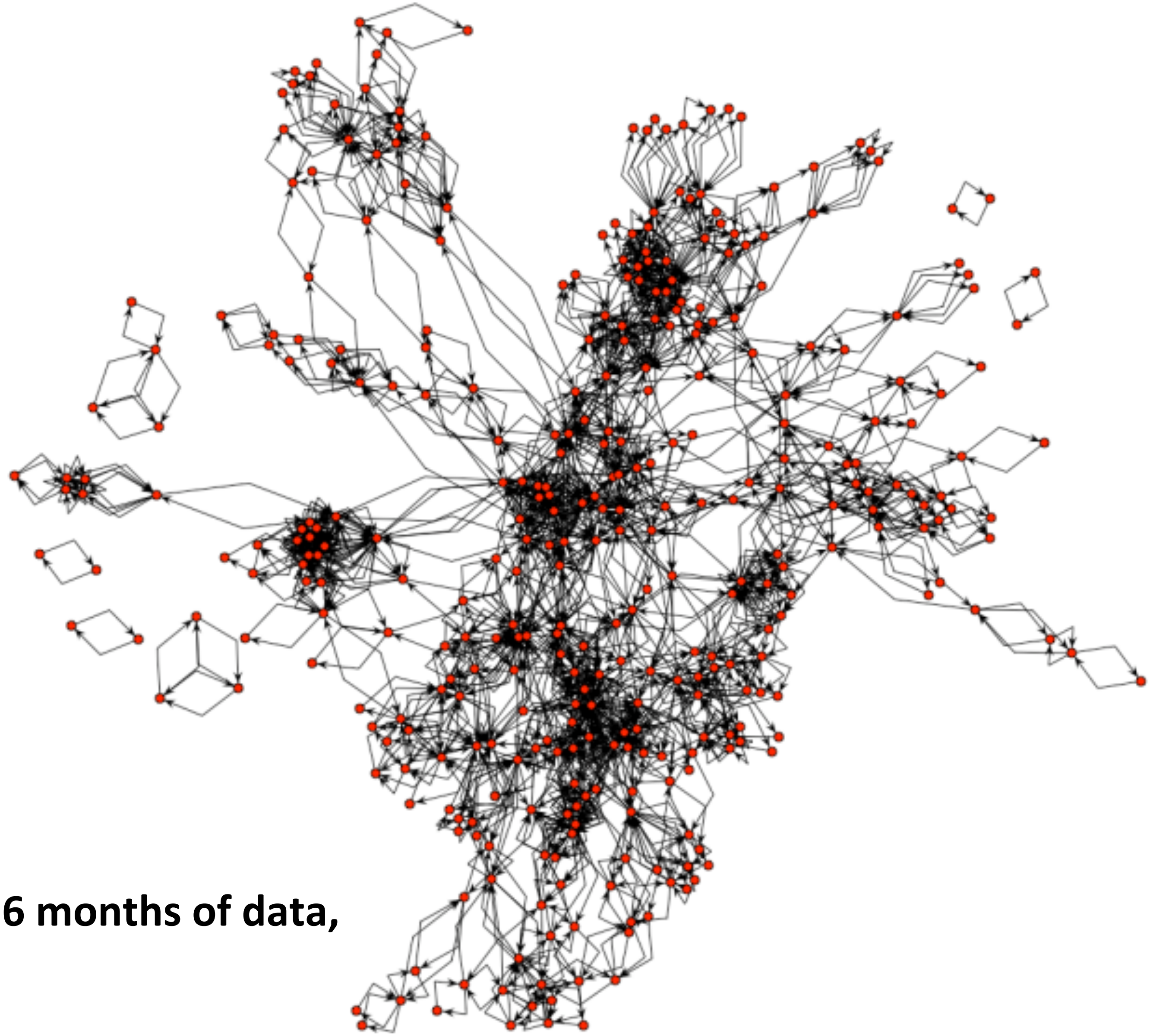


Over **3 billion** photos uploaded each month

Email Data

Record of 300,000 emails sent since 1990
From: blog.stephenwolfram.com
***The Personal Analytics of My Life,* March 8th 2012**





Email Network

aggregation of 6 months of data, HP Labs

July 20. 1749. NUMB. 1075.

The Pennsylvania GAZETTE.

Containing the freshest Advices, Foreign and Domestick.

To his Excellency the honourable GEORGE CLINTON, Captain general and governor in chief of the colony of New-York, and territories thereon depending in America, vice admiral of the same, and admiral of the white squadron of his majesty's fleet;

The humble Address of the General Assembly of the said colony.

May it please your Excellency,

WE his majesty's most dutiful and loyal subjects, the general assembly of the colony of New-York, return your excellency our thanks, for your congratulations upon the re-establishment of peace; and most sincerely and heartily wish, that this, and his majesty's other colonies on the continent, may be entirely freed from those calamities to which they have been expos'd during the war; and that your excellency may be more successful in cultivating among us, the arts of peace, as well for the future security, as present prosperity, of the people of this colony.

We must acknowledge, Sir, with all mankind, that it can be of no service, or rather that it would be in vain to make laws, if they with whom the executive powers of government are intrusted, be disabled from putting them in execution: but we are not sensible that we have been accessary to any such disability ~~in this government; the general~~ assemblies having always (as we humbly conceive) manifested their duty to the crown, in making ample and honourable provision for the support of this his majesty's government; though they may sometimes have differed with governors, as to modes and forms in the method of doing it; and yet have had a proper regard to the royal intention, in the commission and instructions to governors, by taking care that the money given by them, should be duly applied to the respective uses for which they gave it.

We hope your excellency will excuse us, if in our considerations on the extracts of his majesty's instructions, we happen to differ in sentiments on the construction; though your excellency has not indulg'd us with communicating yours: We presume your excellency has not just now receiv'd them (tho' it should seem the speech insinuates as much) but that they have been standing instructions to all your excellency's predecessors (some alterations in the style only of your excellency's copy excepted:) And though there be no royal injunction apparent in them, for the practice so much exploded for many years past, for the mischiefs of which the grant of a five years support was productive, and of which your excellency seem'd at first to be made sensible upon ample considerations; yet we find governors, upon more mature deliberation, would willingly engross the sweets of that method to themselves, of which the people neither ever had, or ever can expect any benefit.

We must beg leave to observe to your excellency, that the general assembly do not perceive any essential difference with respect to his majesty's service, and the royal intention concerning the support of the government of this colony, whether the governor, and other officers in the government, have their salaries provided annually, or for the term of five years; which last your excellency demanded by your speech on the fourteenth of October last; and the same claim (as we take it) your excellency now renews, by your reference to it.

It must be obvious to every one, that the time of our present meeting is most inconvenient to much the greater part of the members, for the very reason mentioned by your excellency: And though we shall at all times chearfully postpone our private interest to the attendance on the publick exigency; yet we are constrained to declare, that from the experience we have had of the unhappy influence which has for some time past, and still does direct your excellency's administration, we have but little hopes of indulgence from your excellency at any time, however consistent it may be with the publick welfare.

And as your excellency is pleas'd to tell us, you cou'd not think it proper to meet us, till you knew the sentiments of his majesty's ministers in respect to our refusing to grant the support of government in the manner you asked it last fall; so as we are not further enlightened in the matter at this time, we cannot but continue in the same opinion we then were.

Whatever just demands there may be due for services done the publick, besides the ordinary occurrences; that they have remain'd so long unpaid, cannot with justice be ascrib'd to any default in the general assembly, who have ever been careful to provide for such, and did pass a bill wherein provision was made, as may be seen in the printed journal of our proceedings; and therefore, their remaining still undischarg'd, is solely to be imputed to the unseasonable prorogation of the general assembly on the twelfth of November last, whereby that bill, with several others, was defeated.

We sincerely lament the hardships which some of our poor brethren, prisoners in Canada, have, and still do suffer, by their long confinement there: and are fully persuaded they might long since have been released, had proper measures been taken by those whose province it is: And here we beg leave to remind your excellency, of the assurances we gave you on that head, in our address of the nineteenth of October last, that we would chearfully provide for all reasonable expence attending such service.

We must conclude upon the whole, with declaring to your excellency, that we can give no other or better answer to the present speech, than we did towards the conclusion of our address, in answer to your excellency's speech of the fourteenth of October last, which was (among other things) in substance, that we were confirm'd in our opinion, that the faithful representatives of the people, could never recede from the method of an annual support.

By order of the general assembly,

City of New-York, DAVID JONES, Speaker.
Assembly chamber, 5th of July, 1749.

HAMBURGH, April 5.

WE are told, that all military preparations in his Prussian majesty's dominions are very much slackened of late, from whence it is hoped, that some method will be found for calming the troubles in the North, before they rise so high as an open rupture; and it is still rendered the more likely by the handing about the following paper, which is said to be the copy of a letter to his Britannick majesty, from his nephew the king of Prussia.

Sir, and Brother,

YOUR majesty's interest and mine are the same with regard to the tranquility of the North. Reports are spread all over Europe; that this tranquility may be disturbed. For my part, I see no likelihood of it in the main; and it seems, that nothing but reciprocal distrust and ill-grounded suspicions can hitherto have gained those rumours any credit.

But as the smallest objects may, by encreasing, become material, as one ought to neglect nothing for the maintenance of peace, and that every thing becomes important to those who are fond of preserving it, I apply to your majesty, whom I know to be in the same sentiments, to the end that, by our joint endeavours, we may so much the more effectually contribute thereto. The suspicions which Sweden's neighbours entertain of her, can rest only on two objects.

The first, which is manifestly frivolous, regards the dangerous projects which they seem resolved to impute to that power against her neighbours. Your majesty's discernment is too quick not to perceive the falsity of it at the first glance. The other falls on the change of the present form of government in Sweden; a project which they father on the prince successor. Methinks the declaration which the prince and the senate made very lately to the court of Russia upon this subject, is so perspicuous, so positive, and so prudent, that it leaves nothing farther to be wished by such powers as interest themselves in maintaining the present government of that kingdom.

The defensive alliance that I made with Sweden, to which France acceded, and the original of which was shewn to the count de Keyserling, the Russian minister at my court, and a copy of which was communicated in due time to your majesty's ministry at London, hath no relation to any new measures; but is, nevertheless, binding on France and myself to maintain the succession, actually established in Sweden, and mutually to defend each other, in case any should attack us.

God forbid that I should suppose any powers in friendship with us capable of so black designs, or that I should so much as suspect them of such dangerous projects! But I entreat your majesty to join your endeavours with mine, to engage both parties to proper explications, which will be found equally salutary for them. I must intreat your attention to all the points that I have been explaining, and that your majesty would employ your credit and good offices to extinguish that fire which glows at present under the embers, and which if it once break out, will spread into flames thro' all Europe.

I am very ready, and offer, with great pleasure, to enter into all the measures which your majesty shall think requisite for the preserving of peace, perswaded that his most christian majesty, who has no less at heart than we, the maintenance of the peace in Europe, and the tranquility of the North, will join his efforts to ours, to contribute the more powerfully thereto.

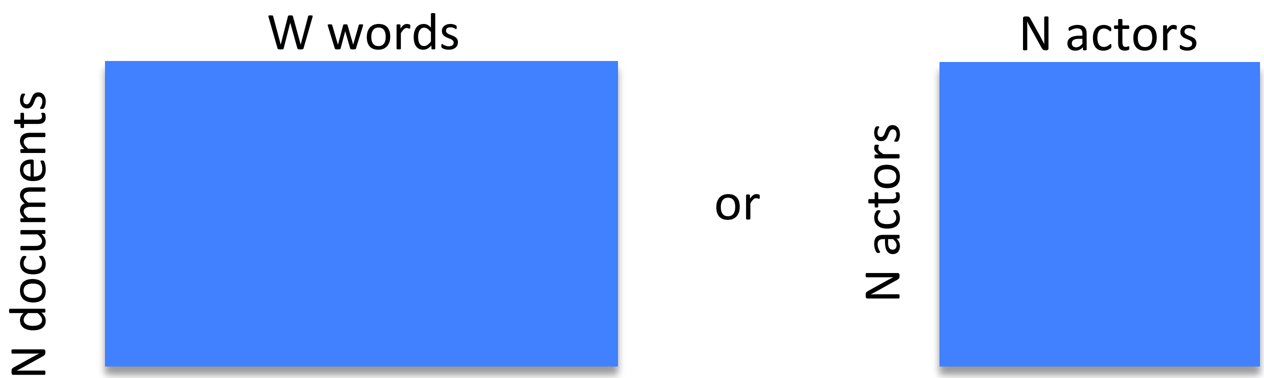
The present occasion which offers itself to your majesty, is one of the most favourable for augmenting the glory of your reign, for supporting

Pennsylvania Gazette

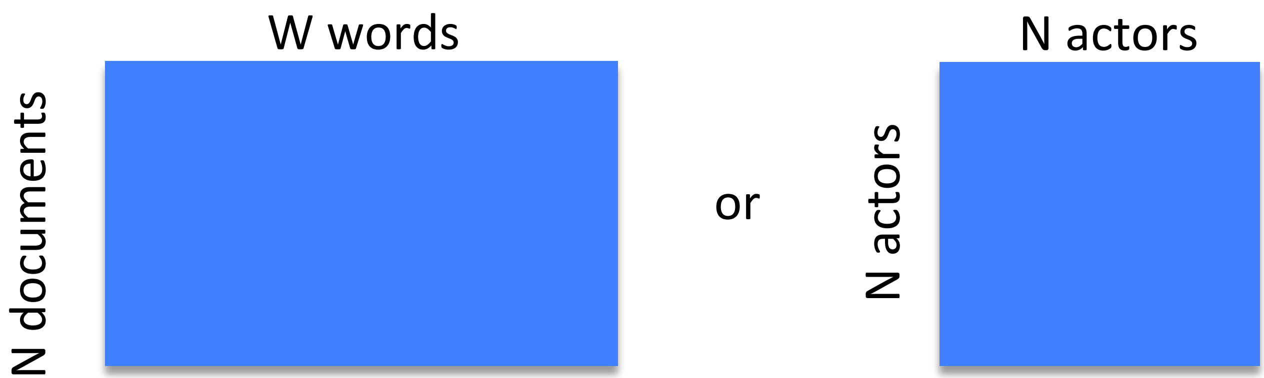
80,000 articles

1728-1800

Common Themes



Common Themes



Additional "side information"

Time

Covariates/Attributes

Supervised labels

Modeling with Latent Variables

- **Modeling high-dimensional data is difficult**

- **Latent variables provide a simplification….**
  - Usually assume conditional independence given the latent variables
  - Representational form of the model is simple
  - Learning the latent variables is not so simple

- **Different roles for latent variable models**
  - Measurable, e.g., distance of an object from a camera in computer vision
  - Not measurable and interpretation is not important, e.g., HMMs in speech
  - Not measurable and interpretation is important, e.g.,. topics, PCA

Statistical Latent Variable Models

- **Rich literature on latent variable models in statistics and probability**
  - Principal component analysis (Pearson, 1901; Hotelling, 1936)
  - Finite mixtures (Pearson, 1894)
  - Hidden Markov models (Baum and Petrie, 1966)

- **Statistical framework provides a useful foundation**
  - Can incorporate metadata such as attributes, times, relational information
  - Tools and techniques for learning

- **However….**
  - Additional complexity (computational and engineering) may cancel out the benefits (e.g., compared to clustering)

Outline

- **Latent variable models for text data**
  - Basic concepts
  - Semi-supervised learning

- **Latent variable models for network data**
  - Static networks
  - Dynamic networks

- **Concluding comments**

Multinomial Distributions on Words

Word	Probability
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

Multinomial Model for Words

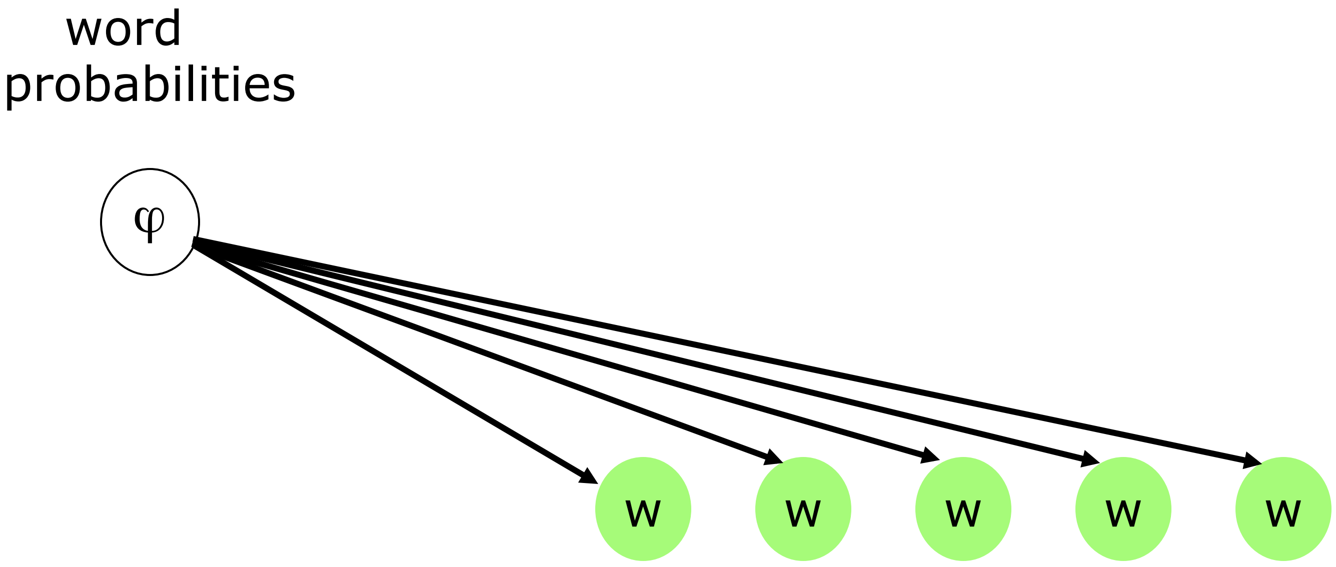
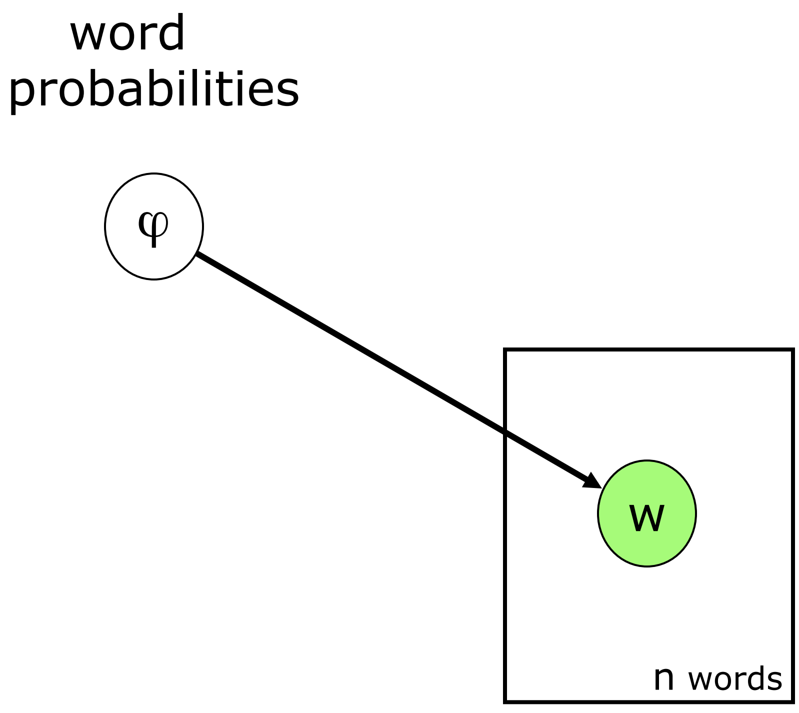


Plate Diagram

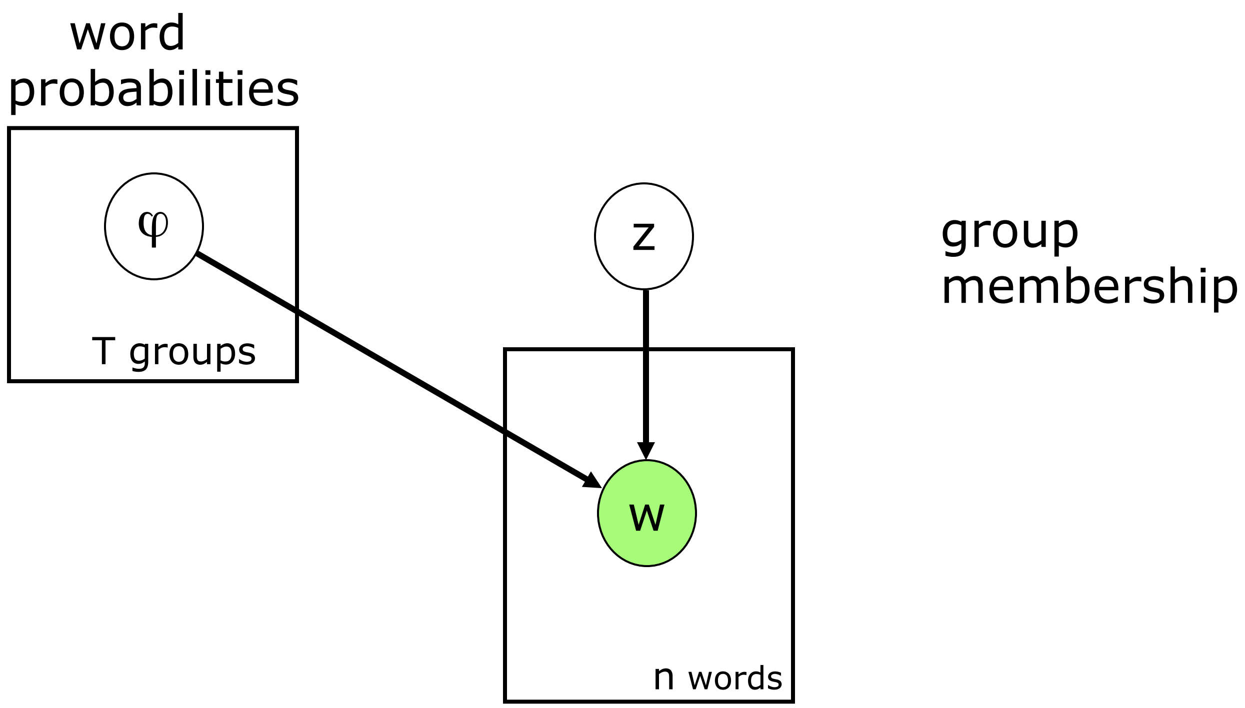


Multinomials = Topics

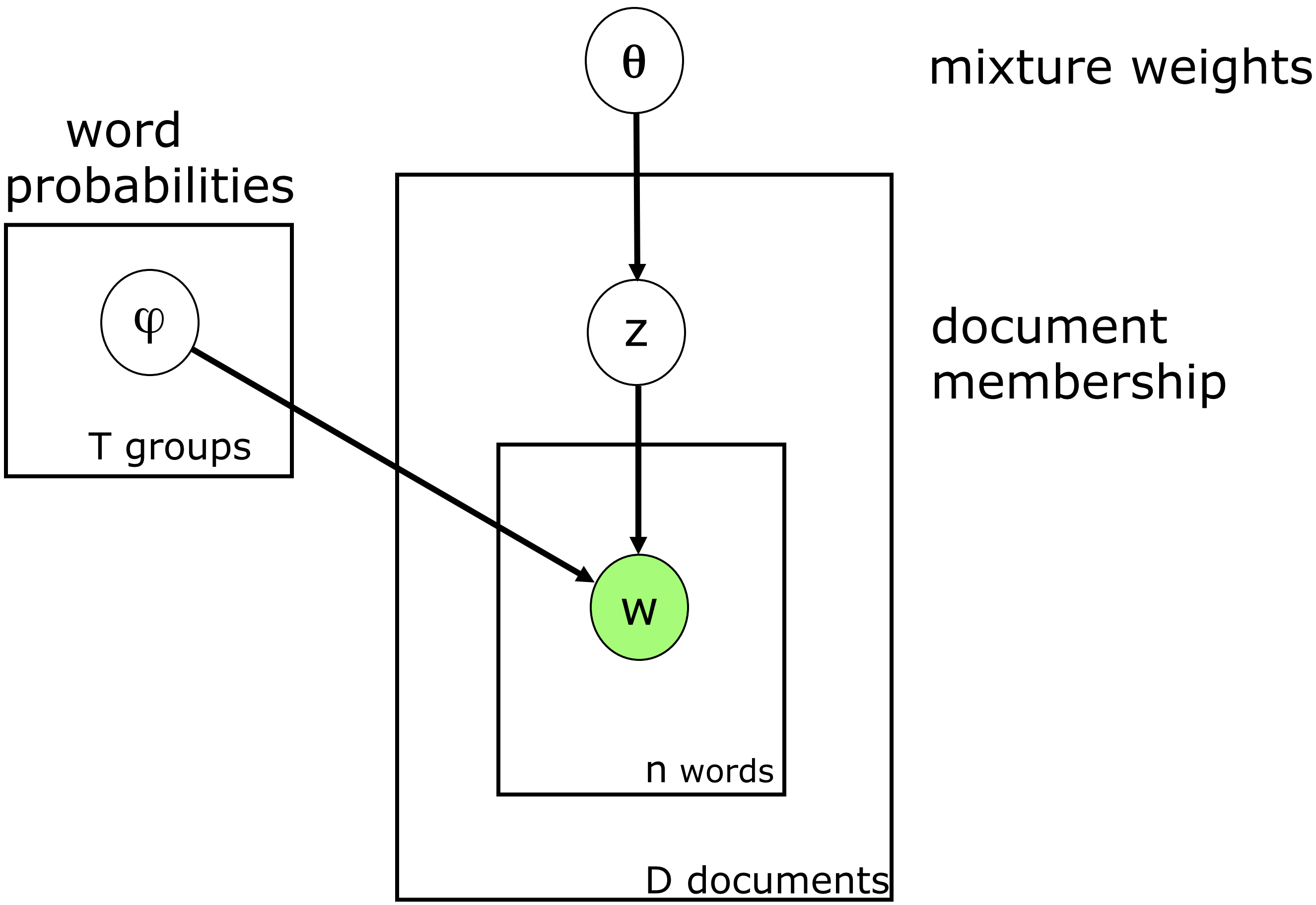
Word	Probability
red	0.202
blue	0.099
green	0.096
yellow	0.073
white	0.048
color	0.030
bright	0.029
colors	0.027
brown	0.027
pink	0.017

Word	Probability
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

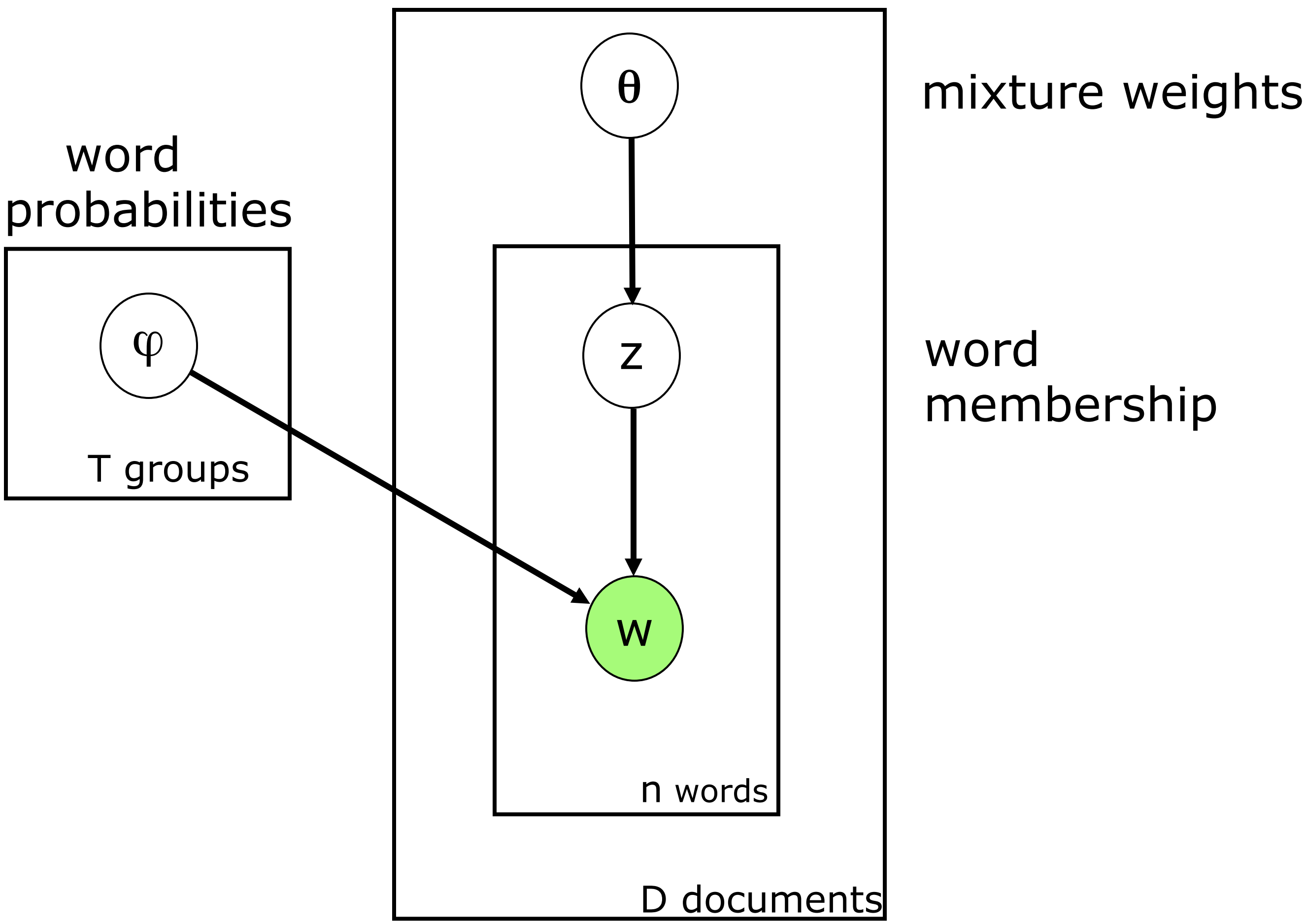
Multiple Word Distributions



Mixture Model Clustering



Topic Model



θ
mixture weights
word
probabilities
φ
T groups
z
word
membership
w
n words
D documents

Clusters v. Topics

Original Document

Hidden Markov Models in Molecular Biology: New Algorithms and Applications

Pierre Baldi, Yves C Hauvin, Tim Hunkapiller, Marcella A. McClure

Hidden Markov Models (HMMs) can be applied to several important problems in molecular biology. We introduce a new convergent learning algorithm for HMMs that, unlike the classical Baum-Welch algorithm is smooth and can be applied on-line or in batch mode, with or without the usual Viterbi most likely path approximation. Left-right HMMs with insertion and deletion states are then trained to represent several protein families including immunoglobulins and kinases. In all cases, the models derived capture all the important statistical properties of the families and can be used efficiently in a number of important tasks such as multiple alignment, motif detection, and classification.

Clusters v. Topics

Original Document

Hidden Markov Models in Molecular Biology: New Algorithms and Applications

Pierre Baldi, Yves C Hauvin, Tim Hunkapiller, Marcella A. McClure

Hidden Markov Models (HMMs) can be applied to several important problems in molecular biology. We introduce a new convergent learning algorithm for HMMs that, unlike the classical Baum-Welch algorithm is smooth and can be applied on-line or in batch mode, with or without the usual Viterbi most likely path approximation. Left-right HMMs with insertion and deletion states are then trained to represent several protein families including immunoglobulins and kinases. In all cases, the models derived capture all the important statistical properties of the families and can be used efficiently in a number of important tasks such as multiple alignment, motif detection, and classification.

One Cluster

[cluster 88]

model data
models time
neural figure state
learning set
parameters
network
probability
number networks
training function
system algorithm
hidden markov

Clusters v. Topics

Original Document

Hidden Markov Models in Molecular Biology: New Algorithms and Applications

Pierre Baldi, Yves C Hauvin, Tim Hunkapiller, Marcella A. McClure

Hidden Markov Models (HMMs) can be applied to several important problems in molecular biology. We introduce a new convergent learning algorithm for HMMs that, unlike the classical Baum-Welch algorithm is smooth and can be applied on-line or in batch mode, with or without the usual Viterbi most likely path approximation. Left-right HMMs with insertion and deletion states are then trained to represent several protein families including immunoglobulins and kinases. In all cases, the models derived capture all the important statistical properties of the families and can be used efficiently in a number of important tasks such as multiple alignment, motif detection, and classification.

One Cluster

[cluster 88]
model data models time neural figure state learning set parameters network probability number networks training function system algorithm hidden markov

Multiple Topics

[topic 10] state hmm markov sequence models hidden states probabilities sequences parameters transition probability training hmms hybrid model likelihood modeling

[topic 37] genetic structure chain protein population region algorithms human mouse selection fitness proteins search evolution generation function sequence sequences genes

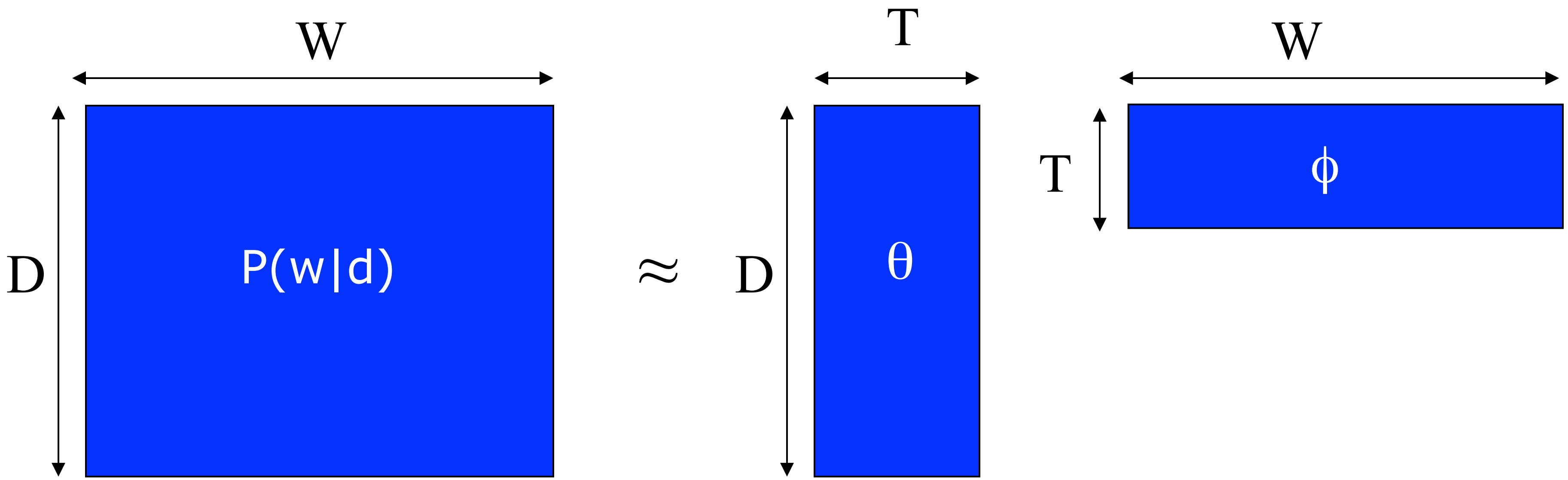
Topic Models

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

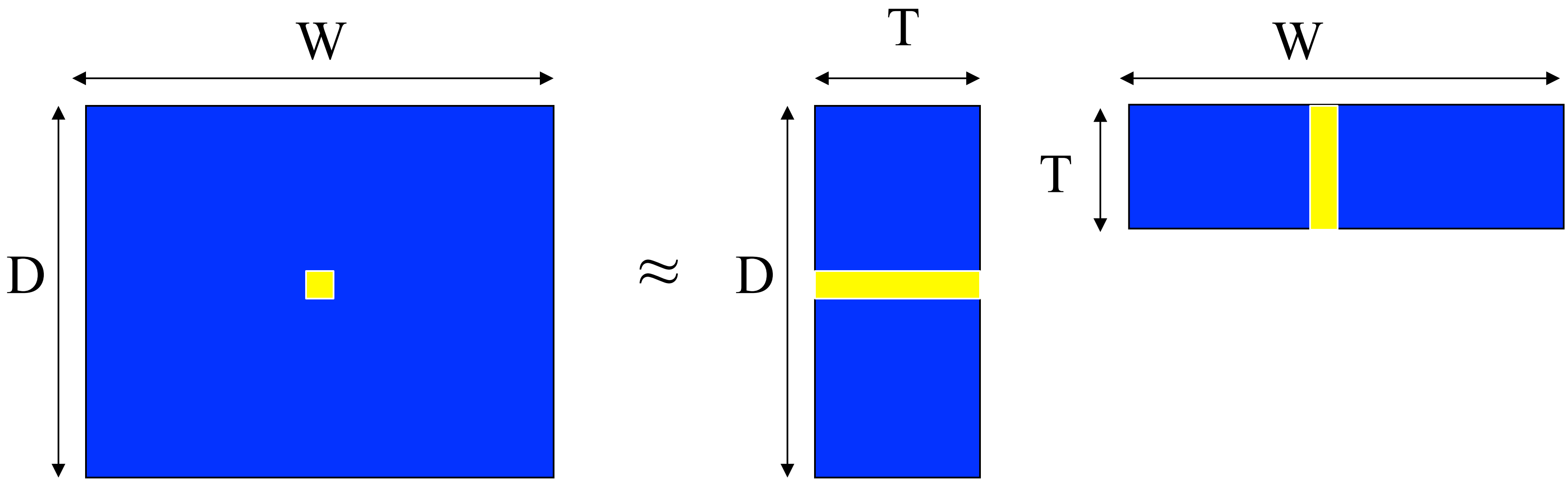
Probability of word i under different topics (the ϕ's)

Multinomial over topics for document d (the θ's)

Topics as Matrix Factorization



Topics as Matrix Factorization

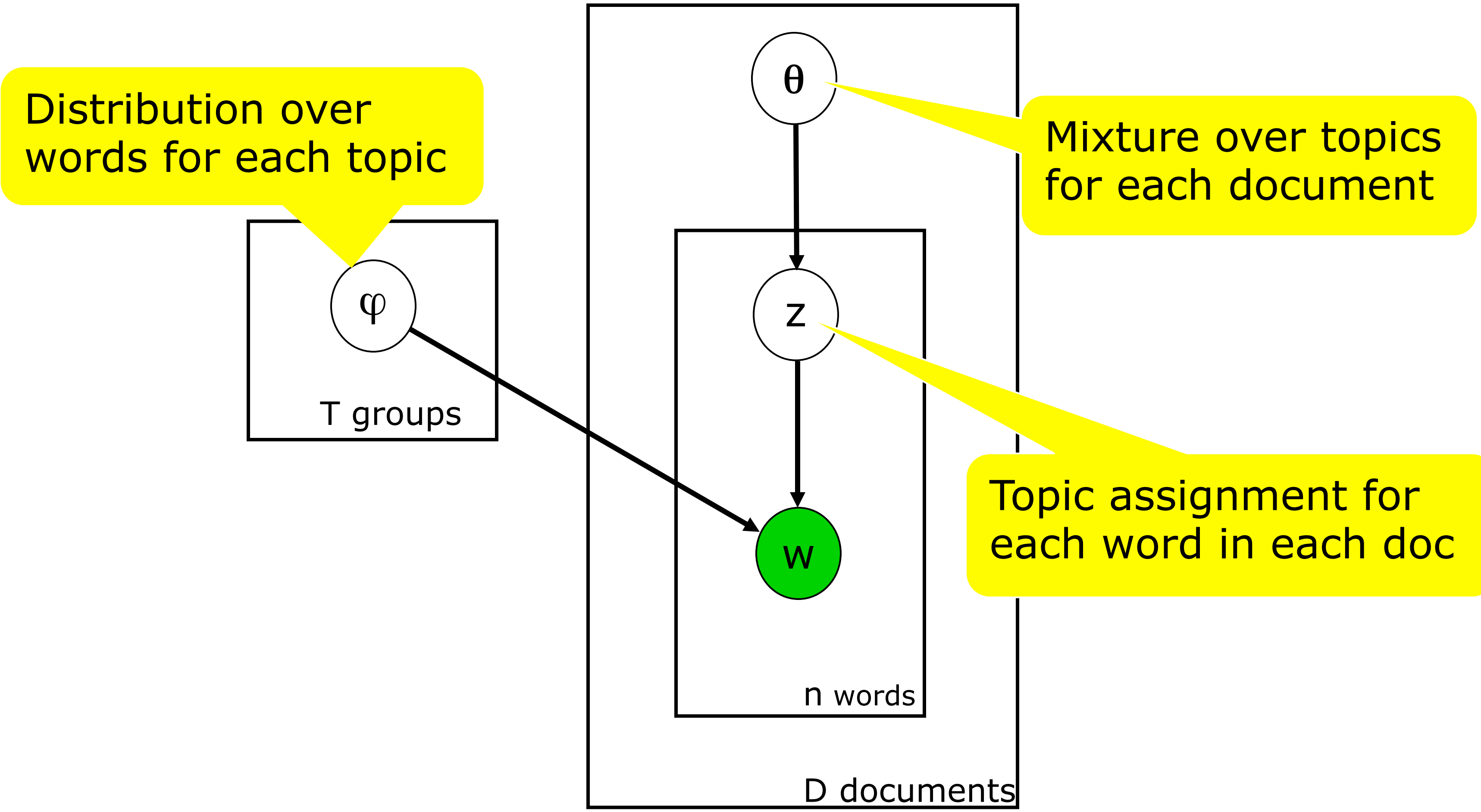


$$p(w_i|d) = \sum_{j=1}^{T} p(w_i|z_j)p(z_j|d)$$

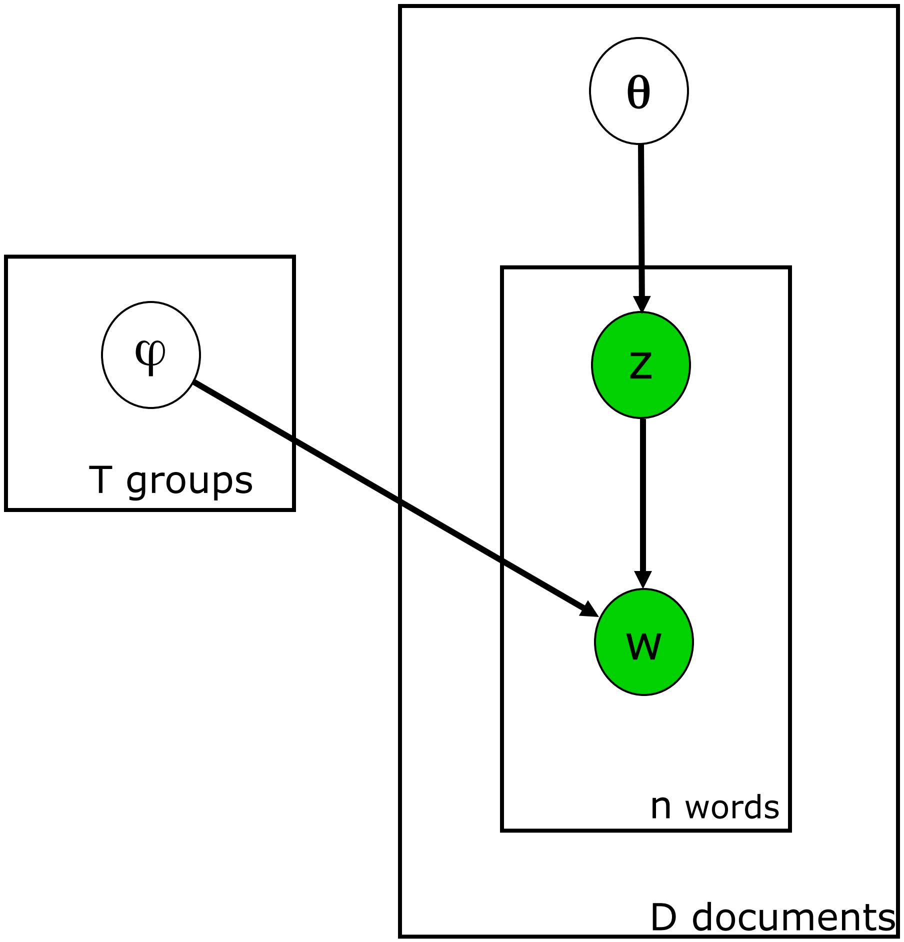
History

1990	Latent Semantic Analysis (Deerwester et al)
1999	Probabilistic Latent Semantic Analysis (Hoffman)
2003	Latent Dirichlet Allocation (Blei, Ng, Jordan)
2004	Gibbs sampling (Griffiths & Steyvers)
2004+	Many extensions and applications……

What do we need to learn ?



Imagine if the z's were known…



Learning Algorithm

- **Inputs**
  - N documents, each as "bag of words"
  - Number of topics T
  - No labels (completely unsupervised)

Learning Algorithm

- **Inputs**
  - N documents, each as "bag of words"
  - Number of topics T
  - No labels (completely unsupervised)

- **Learning via Collapsed Gibbs sampling**
  - Sample z's, marginalize over θ and ϕ
  - Linear time per iteration in number of word tokens and T
  - Given z's, easy to estimate θ and ϕ

Learning Algorithm

- **Inputs**
  - N documents, each as "bag of words"
  - Number of topics T
  - No labels (completely unsupervised)

- **Learning via Collapsed Gibbs sampling**
  - Sample z's, marginalize over θ and ϕ
  - Linear time per iteration in number of word tokens and T
  - Given z's, easy to estimate θ and ϕ

- **Outputs**
  - ϕ: Topic-word probability distributions for each topic
  - θ: Document-topic probability distributions
  - z's: Assignment of each word in each doc to a topic

Sampling Equations

count of topic ***t*** assigned to doc ***d***

count of word ***w*** assigned to topic ***t***

$$p(z_i = t \mid z_{-i}) = \frac{n_{td}^{-i} + \alpha}{\sum_{t'} n_{t'd}^{-i} + T\alpha} \times \frac{n_{wt}^{-i} + \beta}{\sum_{w'} n_{w't}^{-i} + W\beta}$$

probability that word ***i*** is assigned to topic ***t***

Example: Topics from DNA Microarray Literature

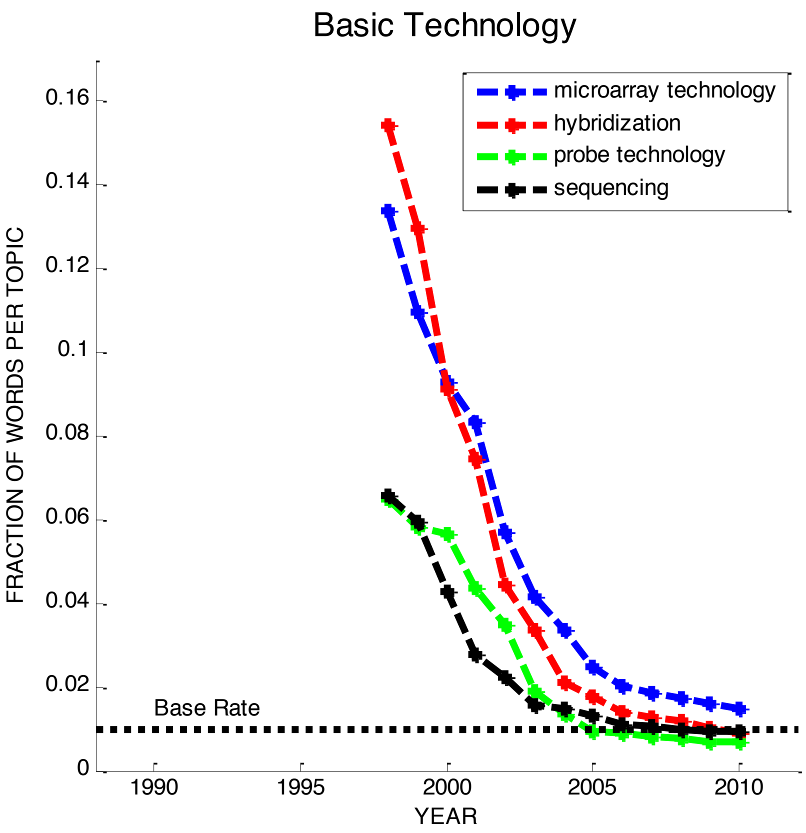
49,000 PubMed abstracts related to DNA Microarrays

Displayed below are the top 5 highest probability words for 5 selected topics

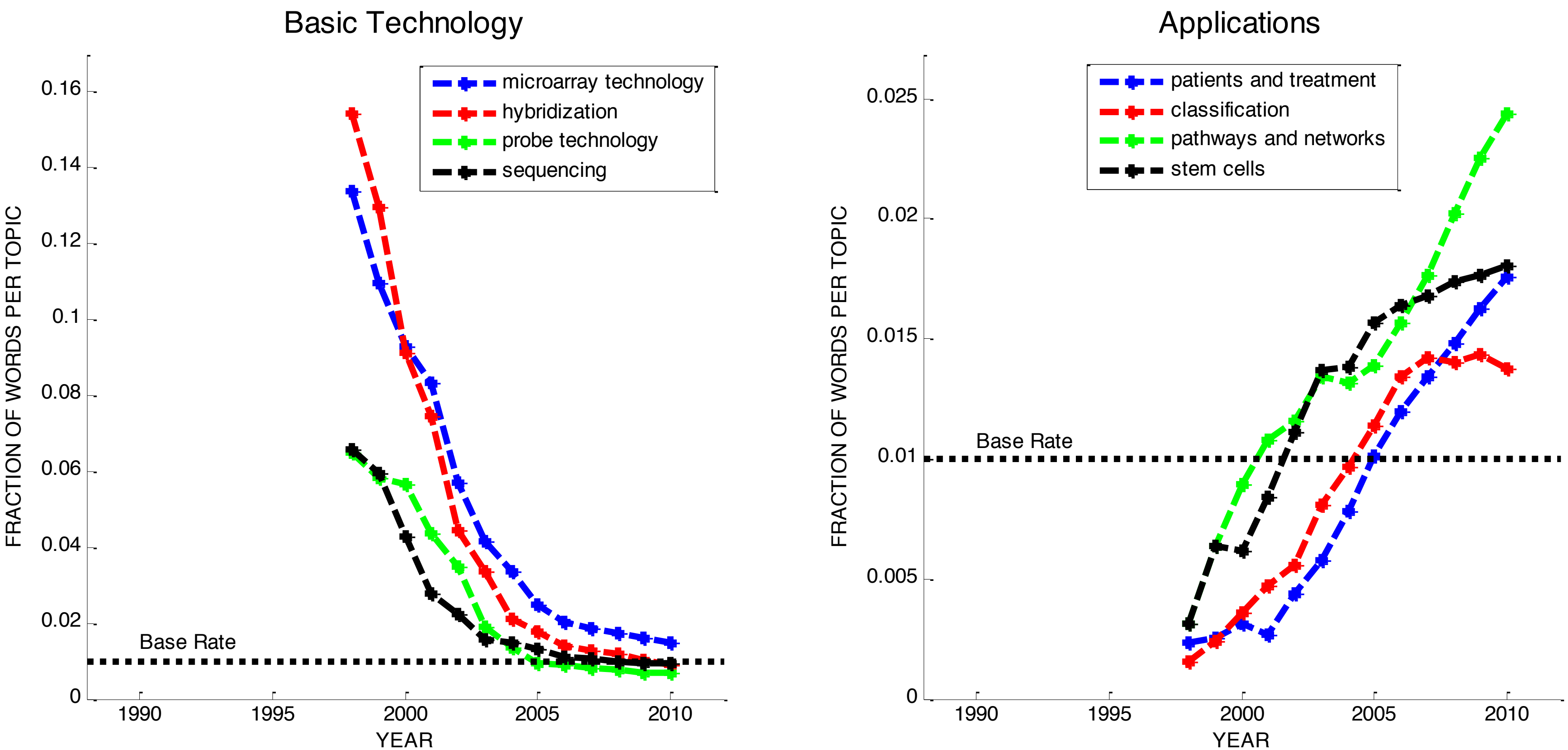
Microarray Chip Technology	Classification Methods	Databases and Annotation	Regulatory Networks	Cancer
detection	classification	databases	network	patient
surface	selection	tool	regulatory	tumor
fluorescence	cancer	annotation	pathway	cancer
hybridization	algorithm	data set	interaction	survival
array	feature	web	transcriptional	prognostic

From basic technology to applications

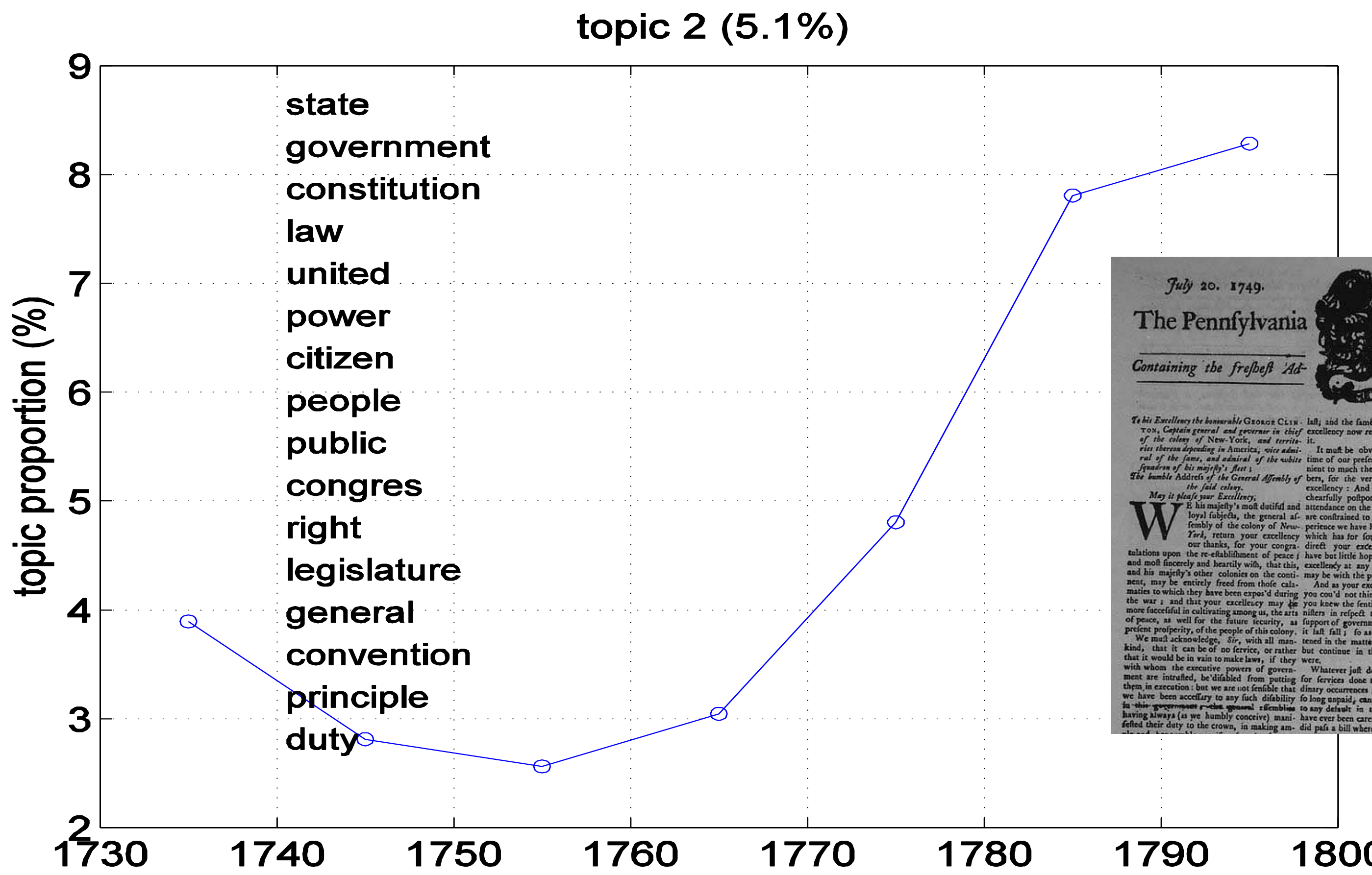
Technology v. Application Topics



Technology v. Application Topics



Pennsylvania Gazette Data



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

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

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

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

"Personal" 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

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

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

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

Political 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

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

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

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

Extensions

Topic models over time (Blei et al)

Correlated topic models (McCallum et al)

Non-parametric Dirichlet processes (Teh et al)

Author-topic models (Steyvers et al)

Dirichlet multinomial regression for metadata (Mimno and McCallum)

… and many more

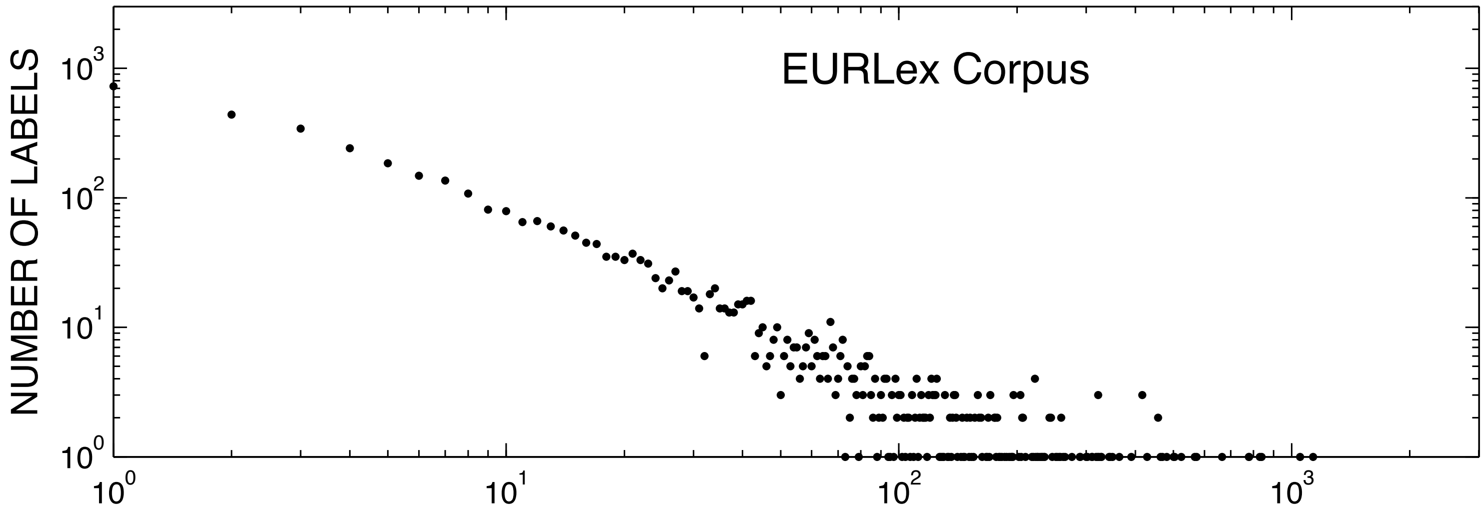
Combining Prior Knowledge and Learned Topics

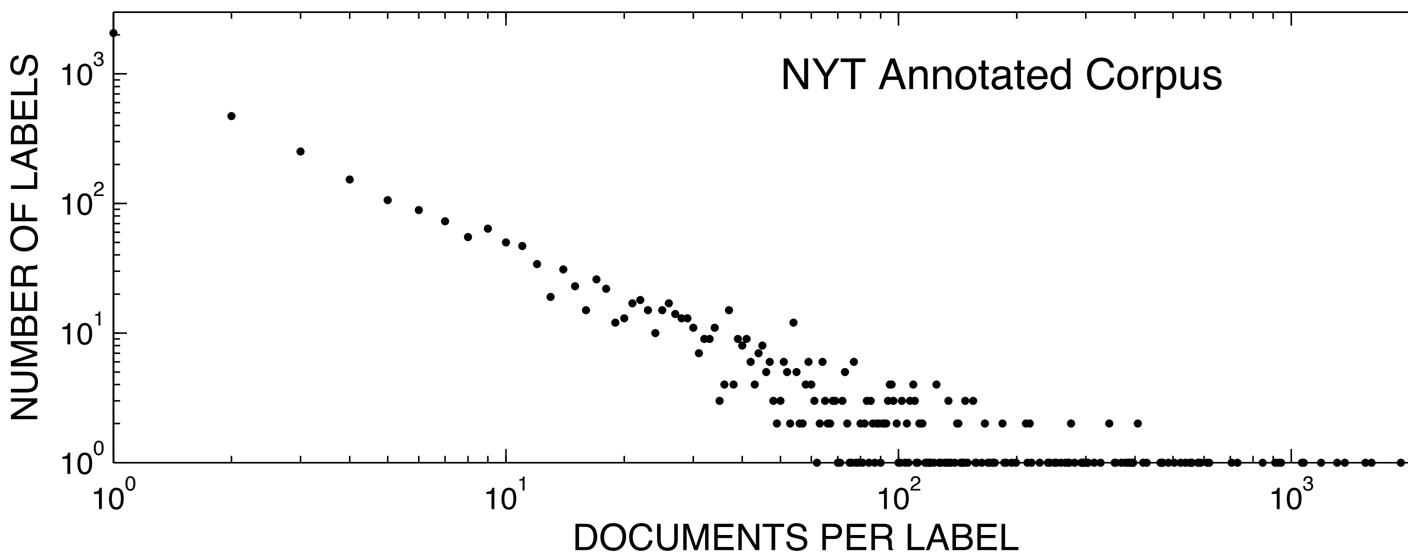
MultiLabel Document Data Sets

Data Set	Number of Unique Labels	Median Number of Documents per Label
RCV1-V2	103	7410
Yahoo! Arts	14	530
Yahoo! Health	19	500

MultiLabel Document Data Sets

Data Set	Number of Unique Labels	Median Number of Documents per Label
RCV1-V2	103	7410
Yahoo! Arts	14	530
Yahoo! Health	19	500
EUR-Lex	3993	6
New York Times	4185	3


EURLex Corpus
NUMBER OF LABELS
10^0
10^1
10^2
10^3
10^0
10^1
10^2
10^3


NYT Annotated Corpus
NUMBER OF LABELS
DOCUMENTS PER LABEL
10^0
10^1
10^2
10^3
10^0
10^1
10^2
10^3


The New York Times
Japanese Commuter Train Kills at Least 71

Applying Topic Models to Multilabel Classification

Rubin, Chambers, Smyth, Steyvers, MLJ 2012

- **Simple idea:**
  - Associate each label with a topic (see Ramage et al, EMNLP, 2009)
  - During learning, restrict the sampler to the known labels for the document
  - Algorithm learns a distribution over words for each label
  - Key difference with discriminative methods: labels are assigned per word, not per document
- **Modeling label dependencies**
  - Extend standard LDA to allow for label (topic) dependencies – significantly improves performance

NY Times Article

Document Labels	Label Freq.
ANTITRUST ACTIONS AND LAWS	19
SUITS AND LITIGATION	67
VIDEO GAMES	1

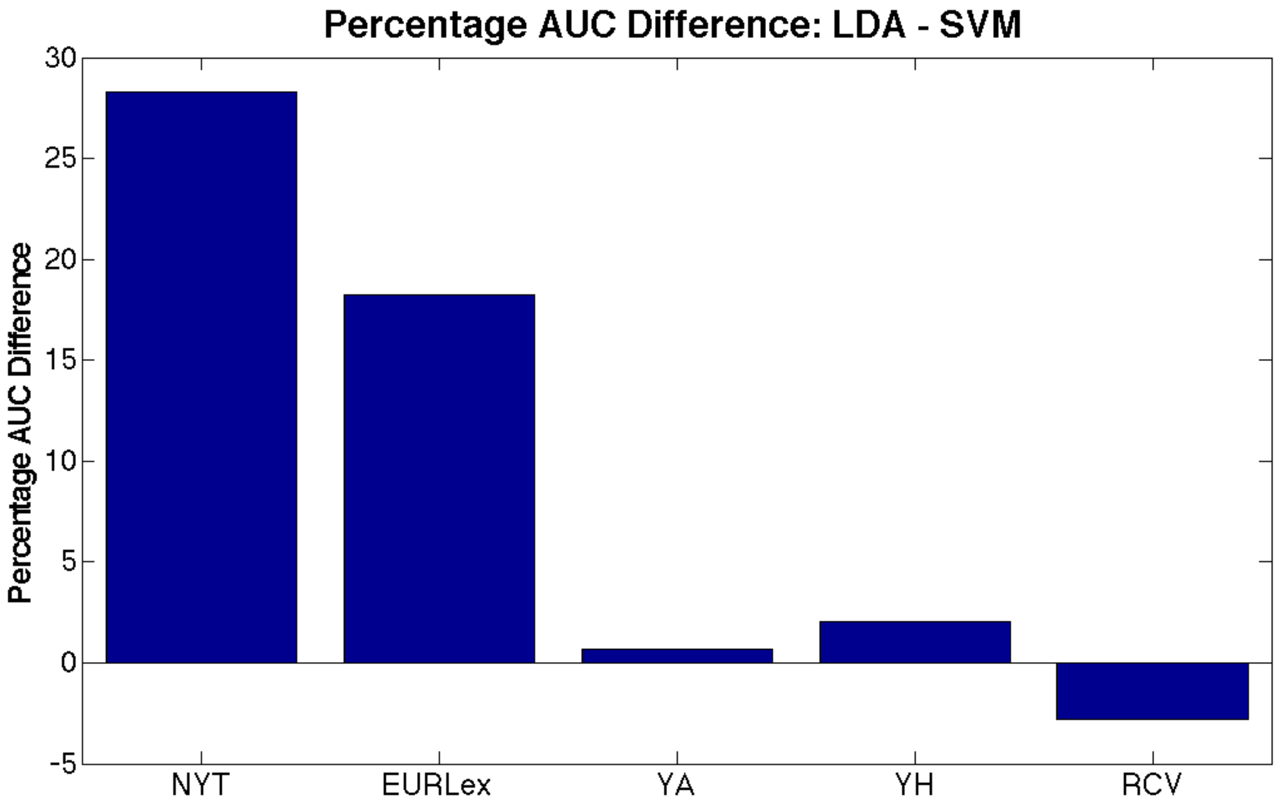
Document Excerpt

A flurry of lawsuits, started by a small American software developer, now surrounds the Nintendo Entertainment System, the best-selling toy in the United States last year...Atari Games argues that Nintendo's high degree of control is tantamount to monopoly, and is suing Nintendo for antitrust violations...

NY Times Article		Models for VIDEO GAMES	
Document Labels	**Label Freq.**	**SVM (weight)**	**LDA (prob.)**
ANTITRUST ACTIONS AND LAWS	19	nintendo	nintendo
SUITS AND LITIGATION	67	mcgowan	games
VIDEO GAMES	1	futuristic	software
		compatible	video
		illusion	system
		shrewd	game
		inception	chip
		truthful	control
		profiles	market
		billionayear	home
		suing	computer
		infringement	shortage
		architecture	say
		handheld	buy
		tantamount	demand
		payoff	developer

Document Excerpt

```
A flurry of lawsuits, started by a small American software developer, now surrounds the Nintendo Entertainment System, the best-selling toy in the United States last year...Atari Games argues that Nintendo's high degree of control is tantamount to monopoly, and is suing Nintendo for antitrust violations...
```



From Rubin, Chambers, Smyth, Steyvers, MLJ 2012

Data Set	Median Number of Documents per Label	Metrics where Topics were better	Metrics where SVMs were better
RCV1-V2	7410	1	24
Yahoo! Arts	530	11	13
Yahoo! Health	500	13	12
EUR-Lex	6	18	6
New York Times	3	22	1

From Rubin, Chambers, Smyth, Steyvers, MLJ 2012

High Probability Words in Topic about Families Learned from Data

child (0.171), parent (0.073), young (0.040),
boy (0.028), mother (0.027), father (0.021),
school (0.020), etc..

High Probability Words in Topic about Families Learned from Data

child (0.171), parent (0.073), young (0.040), boy (0.028), mother (0.027), father (0.021), school (0.020), etc..

Words under Family Concept in Thesaurus

beget, birthright, brood, brother, child, children, circle, close, closer, closest, distantly, dynastic, dynasties, dynasty, elder, eldest, estranged, families, family, father, fatherless, first-born, generation, guardian, half-brother, heirloom, etc.

From Cambridge International Dictionary of English (CIDE)

Combining Human-Defined Concepts and Topics

Chemudugunta, Holloway, Smyth, Steyvers *ISWC*, 2008

- **Treat thesauri as prior knowledge**
  - Each concept in a thesaurus associated with a topic
  - Words associated with thesaurus concept act as prior for a topic

- **From topic-learning viewpoint**
  - Provides rich and useful source of prior knowledge (e.g., rare words)

- **From the thesaurus viewpoint**
  - Overlays a probabilistic model: can tag documents with concepts

Tagging Documents with Concepts

From Steyvers, Smyth, Chemudugunta, 2011

tag	P(*c*)	Concept	P(*w*|*c*)
a	0.1702	PHYSICS	electrons (0.2767) electron (0.1367) radiation (0.0899) protons (0.0723) ions (0.0532) radioactive (0.0476) proton (0.0282)
b	0.1325	CHEMICAL ELEMENTS	oxygen (0.3023) hydrogen (0.1871) carbon (0.0710) nitrogen (0.0670) sodium (0.0562) sulfur (0.0414) chlorine (0.0398)
c	0.0959	ATOMS, MOLECULES, AND SUB-ATOMIC PARTICLES	atoms (0.3009) molecules (0.2965) atom (0.2291) molecule (0.1085) ions (0.0262) isotopes (0.0135) ion (0.0105) isotope (0.0069)
d	0.0924	ELECTRICITY AND ELECTRONICS	electricity (0.2464) electric (0.2291) electrical (0.1082) current (0.0882) flow (0.0448) magnetism (0.0329)
o	0.5091	OTHER	

Tagging Documents with Concepts

From Steyvers, Smyth, Chemudugunta, 2011

tag	P(*c*)	Concept	P(*w*|*c*)
a	0.1702	PHYSICS	electrons (0.2767) electron (0.1367) radiation (0.0899) protons (0.0723) ions (0.0532) radioactive (0.0476) proton (0.0282)
b	0.1325	CHEMICAL ELEMENTS	oxygen (0.3023) hydrogen (0.1871) carbon (0.0710) nitrogen (0.0670) sodium (0.0562) sulfur (0.0414) chlorine (0.0398)
c	0.0959	ATOMS, MOLECULES, AND SUB-ATOMIC PARTICLES	atoms (0.3009) molecules (0.2965) atom (0.2291) molecule (0.1085) ions (0.0262) isotopes (0.0135) ion (0.0105) isotope (0.0069)
d	0.0924	ELECTRICITY AND ELECTRONICS	electricity (0.2464) electric (0.2291) electrical (0.1082) current (0.0882) flow (0.0448) magnetism (0.0329)
o	0.5091	OTHER	

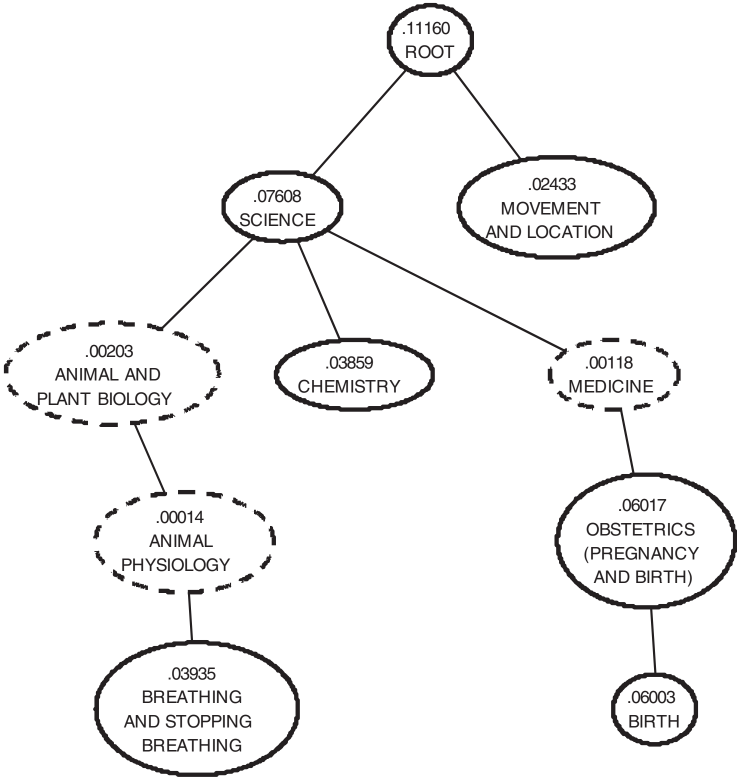
The hydrogen[b] ions[a] immediately[o] attach[o] themselves to water[o] molecules[c] to form[o] combinations[o] called[o] hydronium ions[a]. The chlorine[b] ions[a] also associate[o] with water[o] molecules[c] and become hydrated. Ordinarily[o], the positive[o] hydronium ions[a] and the negative[o] chlorine[b] ions[a] wander[o] about freely[o] in the solution[o] in all directions[o]. However, when the electrolytic cell[o] is connected[o] to a battery[o], the anode[d] becomes positively[o] charged[a] and the cathode[d] becomes negatively[o] charged[a]. The positively[o] charged[a] hydronium ions[a] are then attracted[o] toward the cathode[d] and the negatively[o] charged[a] chlorine[b] ions[a] are attracted[o] toward the anode[d]. The flow[d] of current[d] inside[o] the cell[o] therefore consists of positive[o] hydronium ions[a] flowing[d] in one direction[o] and negative[o] chlorine[b] ions[a] flowing[d] in the opposite[o] direction[o]. When the hydronium ions[a] reach[o] the cathode[d], which has an excess[o] of electrons[a], each takes[o] one electron[a] from it and thus neutralizes[o] the positively[o] charged[a] hydrogen[b] ion[a] attached[o] to it. The hydrogen[b] ions[a] thus become hydrogen[b] atoms[c] and are released[o] into the solution[o]. Here they pair[o] up to form[o] hydrogen[b] molecules[c] which gradually[o] come out of the solution[o] as bubbles[o] of hydrogen[b] gas[o]. When the chlorine[b] ions[a] reach[o] the anode[d], which has a shortage[o] of electrons[a], they give[o] up their extra[o] electrons[a] and become neutral[a] chlorine[b] atoms[c]. These pair[o] up to form[o] chlorine[b] molecules[c] which gradually[o] come out of the solution[o] as bubbles[o] of chlorine[b] gas[o]. The behavior[o] of hydrochloric acid[o] solution[o] is typical[o] of all electrolytes[o]. In general[o], when acids[o], bases[o], and salts[o] are dissolved[o] in water[o], many of their molecules[c] break[o] up into positively[o] and negatively[o] charged[a] ions[a] which are free[o] to move[o] in the solution[o].

Mapping a Document to a Thesaurus

Document in TASA Corpus

The postnatal period of development lasts from birth until death and can be divided into a neonatal period, infancy, childhood, adolescence, adulthood, and senescence. The neonatal period, which extends from birth to the end of the first four weeks, begins very abruptly at birth. Physiological adjustments must be made quickly, because the newborn must suddenly do for itself those things that the mother body has been doing for it. Thus, the newborn must carry on respiration, obtain nutrients, digest nutrients, excrete wastes, regulate body temperature, and so forth. However, its most immediate need is to obtain oxygen and excrete carbon dioxide, so its first breath is critical. The first breath must be particularly forceful, because the newborn lungs are collapsed, and the airways are small and offer considerable resistance to air movement. Also, surface tension tends to hold the moist membranes of the lungs together. Fortunately, the lungs of a full term fetus secrete surfactant, which reduces surface tension, and after the first powerful breath begins to expand the lungs, breathing becomes easier. It is not clear whether the first breath is stimulated by one or several factors. Those that may be involved include an increasing level of carbon dioxide, a decreasing ph, low oxygen concentration, a drop in body temperature and mechanical stimulation that occurs during and after the birth process. Prior to birth, the fetus depends primarily on glucose and fatty acids obtained from the mother blood as energy sources

Mapping of Document to CIDE Concept Hierarchy

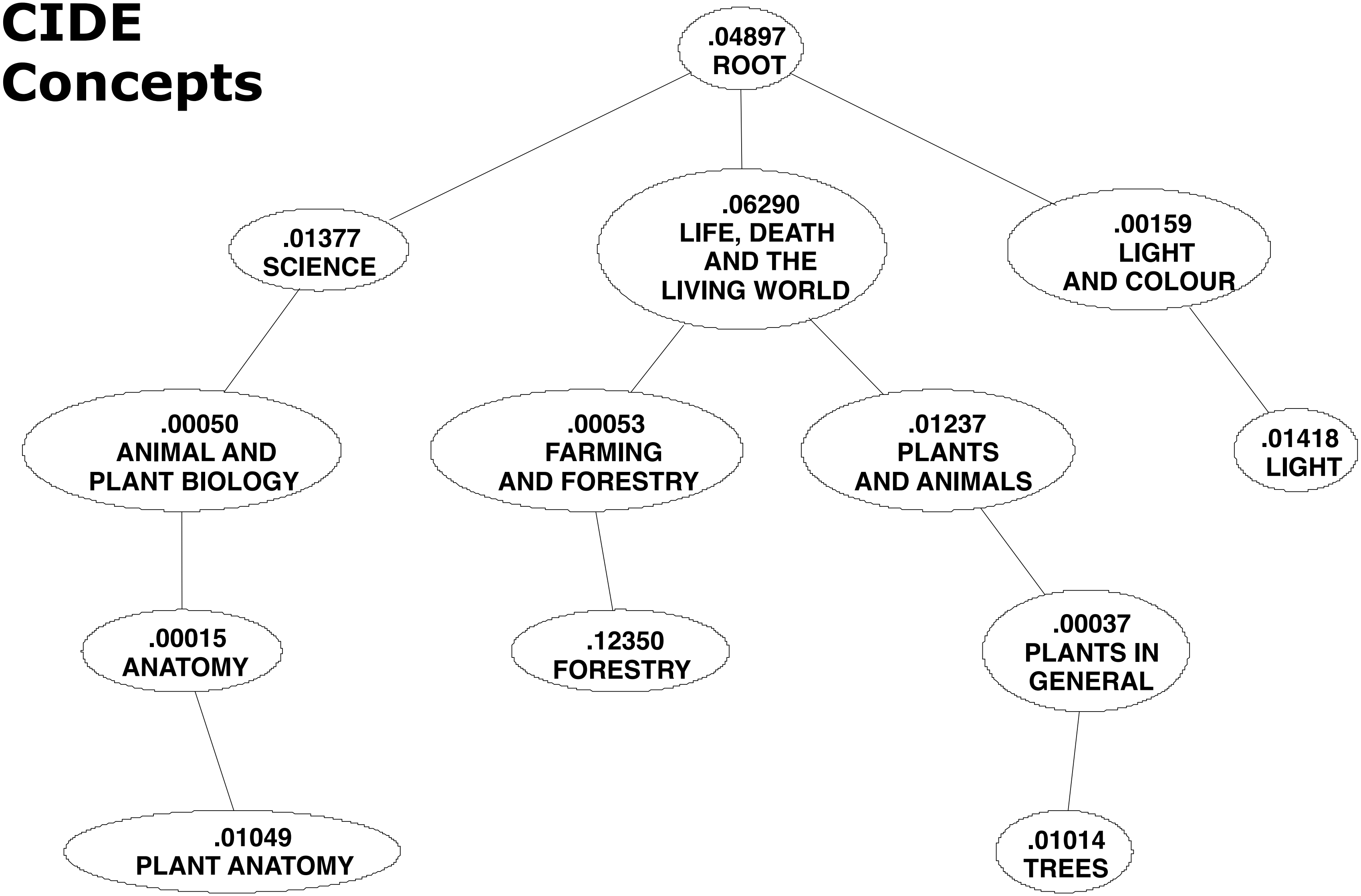


From Steyvers, Smyth, Chemudugunta, 2011

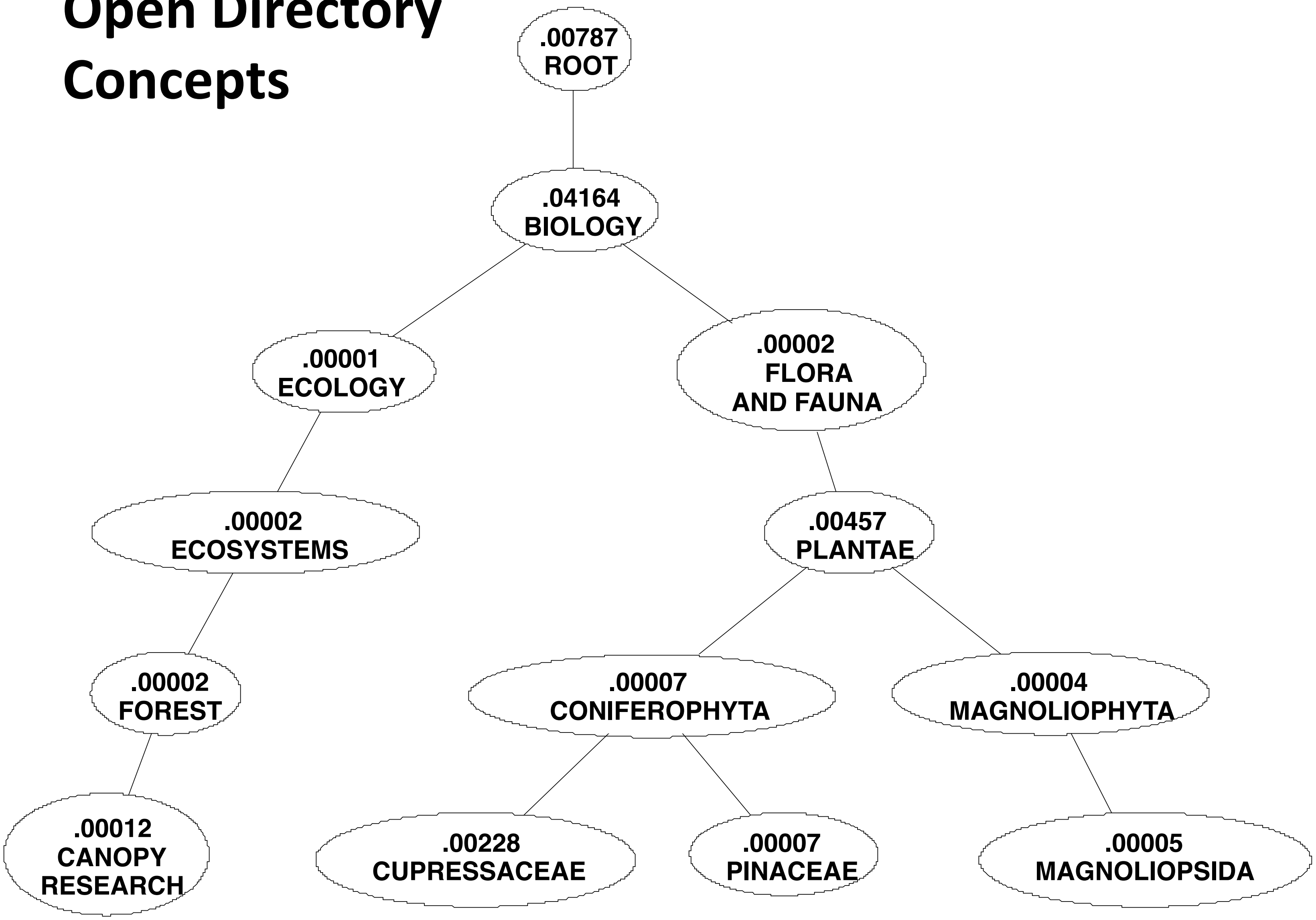
Example of Concept Tagging

Forest biomes in the temperate zone are characterized by ample rainfall, seasonal temperature changes, and day length that varies with the season. There are two types of forest biomes in North America: deciduous forest and evergreen forest biomes. Trees that lose their leaves in response to shortening periods of daylight are called deciduous trees. The deciduous forest biome contains many trees, such as maple, oak, and hickory, that lose their leaves each autumn. The fallen leaves form a thick layer of forest litter on the ground, which is slowly broken down by decomposers. Trees use large amounts of water during photosynthesis. Some water escapes through openings in the leaves. During the winter, when the ground is frozen and cannot absorb water, the leafless trees use and lose very little moisture. Losing leaves is an adaptation that helps deciduous trees stay alive through the winter. A variety of wildflowers and shrubs grow in the deciduous forest. These plants grow and bloom early each spring, before the tree leaves have grown back. The canopy of trees shades much of the sunlight from the forest floor in late spring and summer. In the deciduous forest, each layer of plant life has different adaptations. The adaptations enable plants to survive the given amounts of sunlight and moisture in each layer of the forest. For example, mosses and ferns have structures that allow them to grow successfully on the damp, shady, forest floor. The large number of producers in the deciduous forest provide food for a large number of consumers. Deer, mice, pheasants, and quail feed on the leaves, berries, and seeds of plants on the forest floor.

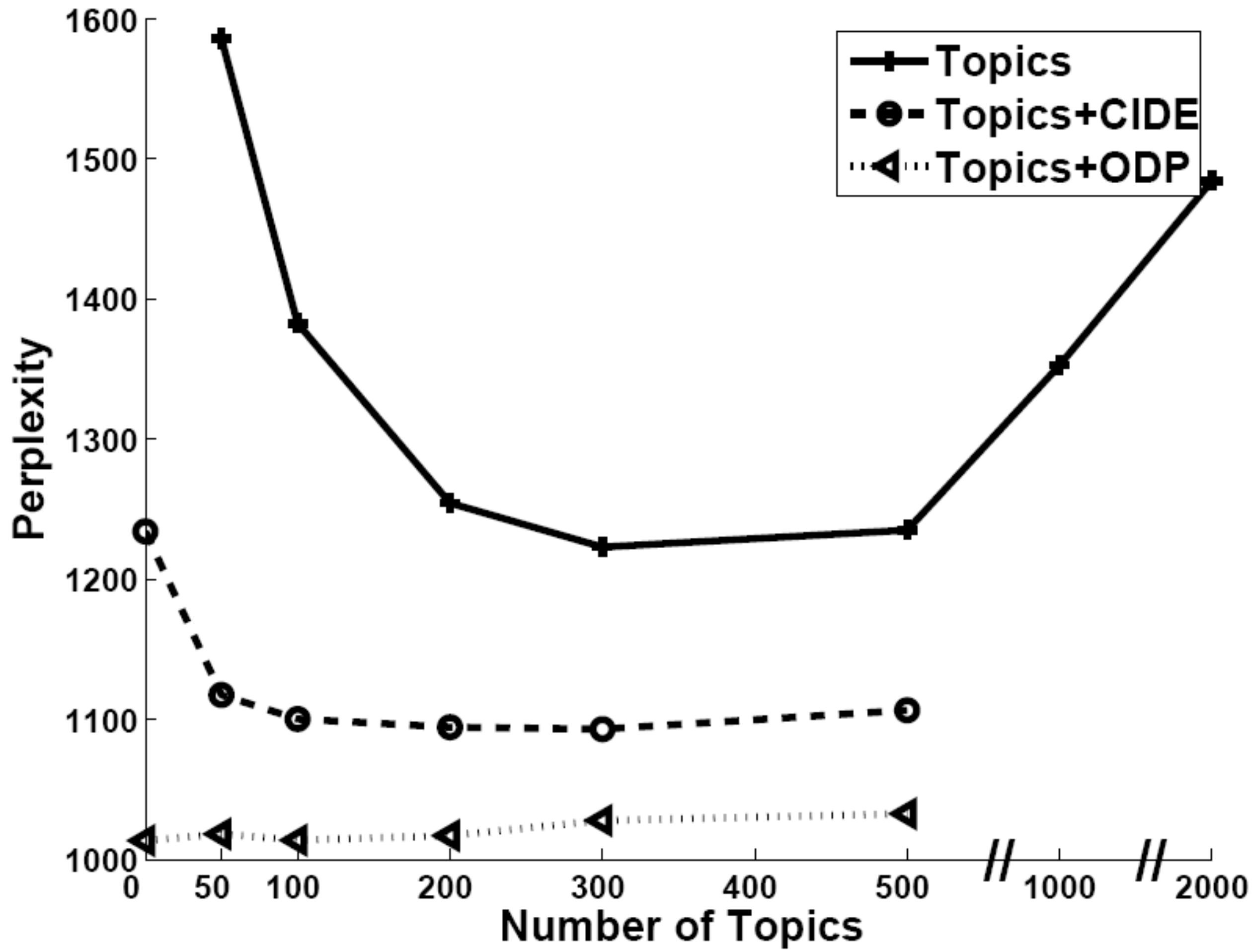
CIDE Concepts



Open Directory Concepts



Thesauri produce better Topic Models



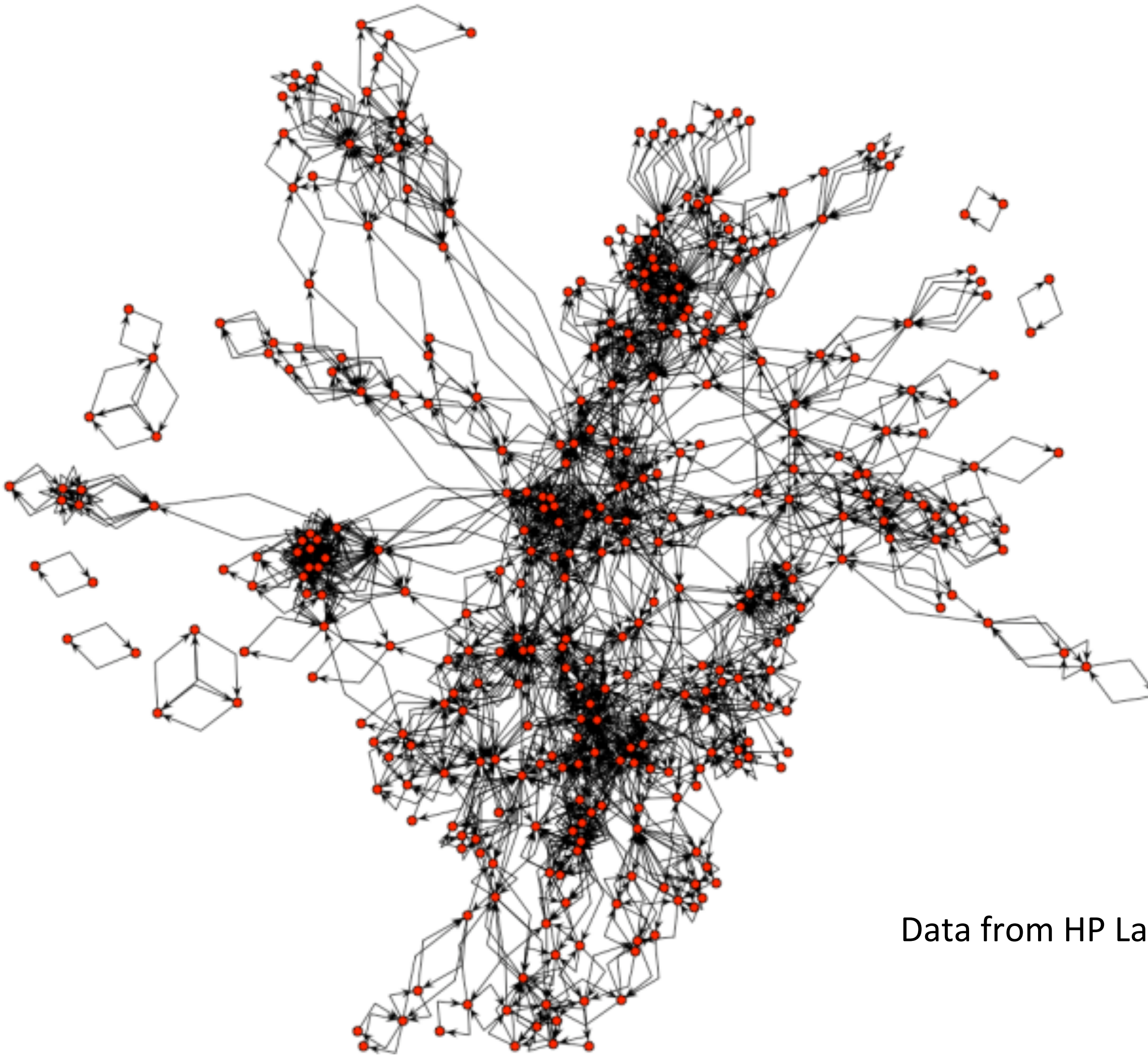

Politics
Socioeconomics
Economic Problems
Inflation
Corporate Scandals
Social Programs
Healthcare
Welfare
Terrorism
Dirty Wars
War on Terror
Green Politics
Green Anarchism
Discrimination
Predjudices
Hate Crimes
Anti-Catholicism
Development
Economic Development
Poverty
Homelessness
Malnutrition
Energy Development
National Security
Counter-Terrorism
2003 Iraq Conflict
Computer Security
Cryptography
CISSP
Human Rights
Human Rights Abuses
War Crimes
Slavery
The Holocaust
Censorship
Obscenity Law
Peace
Quakerism
United Nations
Islamism
Taliban
Al-Qaeda

Modeling Social Network Data

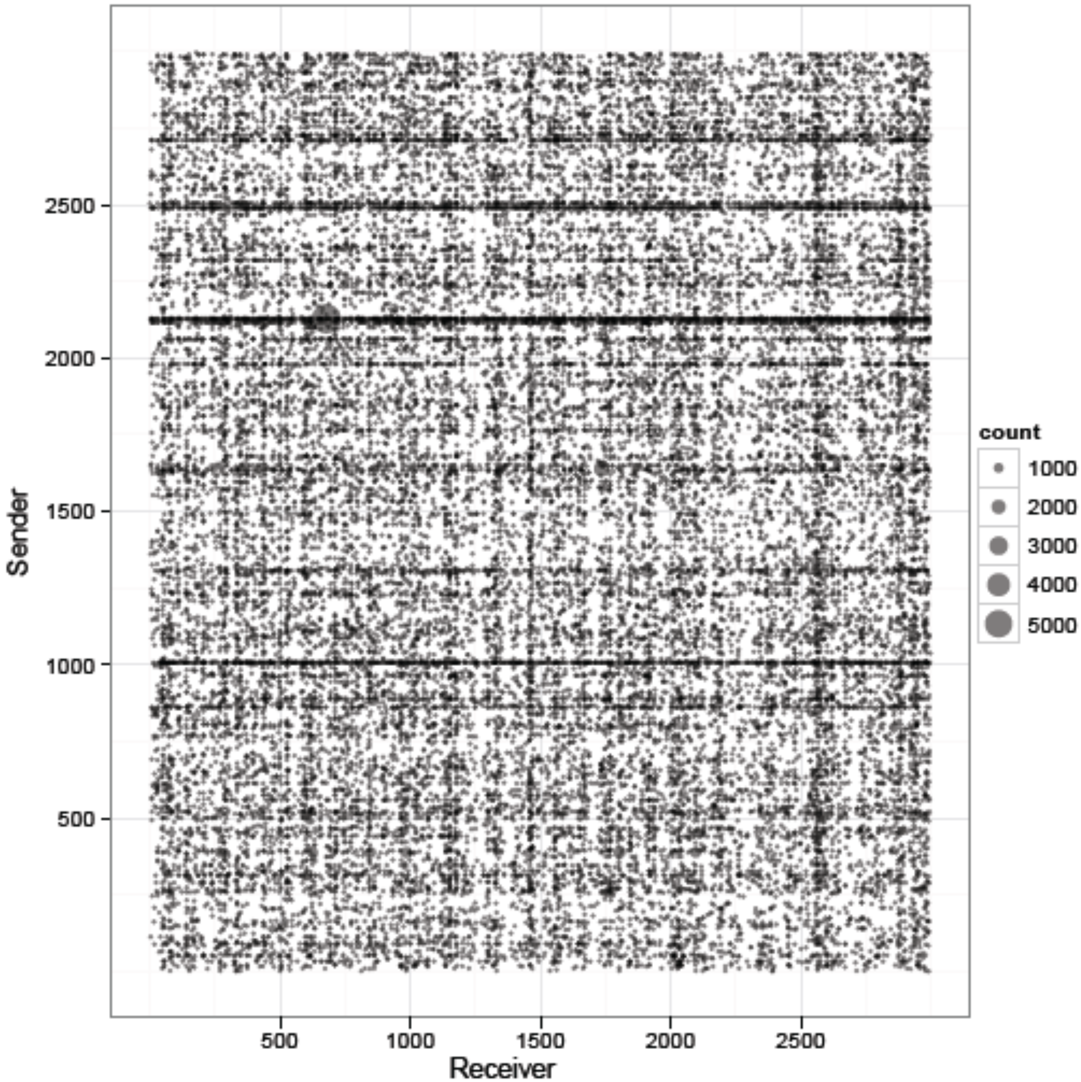
Static Network Data

- **General Notation:**
  - N actors (node set)
    - Generally assume that set of actors is known and fixed
  - Edges between actors (Y)
    - Adjacency matrix Y
    - $y_{i,j}$ indicates an edge between actor i and actor j
    - Simplest case: binary undirected/directed edges
  - Covariates/Attributes (X)
    - e.g., for each actor (e.g., age, text documents,..)
    - e.g., for each edge (e..g., numeric weights, vector of attributes, text, etc)

Email Contact Network



Data from HP Labs



Data:
Count matrix of 200,000 email messages among 3000 individuals over 3 months

Latent Variable Models for Static Networks

- **Latent-variable models for networks**
  - Assume edges can be explained by latent node characteristics
  - Latent variables chosen so that edges are conditionally independent given the latent variables

- **Comments**
  - Can be computationally much easier to work with compared to alternatives such as exponential random graph models

Example: The Latent Space Model

Hoff, Raftery, Handcock, JASA, 2002

- **Idea:**
  - Embed nodes in a latent K-dimensional Euclidean space
  - Probability of edge (i,j) = f (distance (i, j))
  - Edges are conditionally independent given K-dim locations

Example: The Latent Space Model

Hoff, Raftery, Handcock, JASA, 2002

- **Idea:**
  - Embed nodes in a latent K-dimensional Euclidean space
  - Probability of edge (i,j) = f (distance (i, j))
  - Edges are conditionally independent given K-dim locations

- **Probability model**
  - z_i = K-dim latent position vector for node i

  Log-odds (y_{ij} = 1) = log P(y_{ij} = 1)/(1 –P(y_{ij} = 1))

$$= - | z_i - z_j | + \mu + \beta x_{ij}$$

distance of nodes i and j

network density parameter

covariate effects (optional)

Example: The Latent Space Model

Hoff, Raftery, Handcock, JASA, 2002

Likelihood:

$$P(Y \mid Z, \beta, \mu) = \Pi \;\; P(\; y_{ij} \mid z_i \, , \; z_j \;, \mu \, , \; \beta \;)$$

logistic function

Estimation:

- Can maximize likelihood directly (as a function of Z,….) using gradient
- Can also be Bayesian, with priors – sample using MCMC

Computational issues

- Note that the product above is over all pairs, $O(N^2)$: poor scalability

Figure from Hoff, Raftery, Handcock, 2002

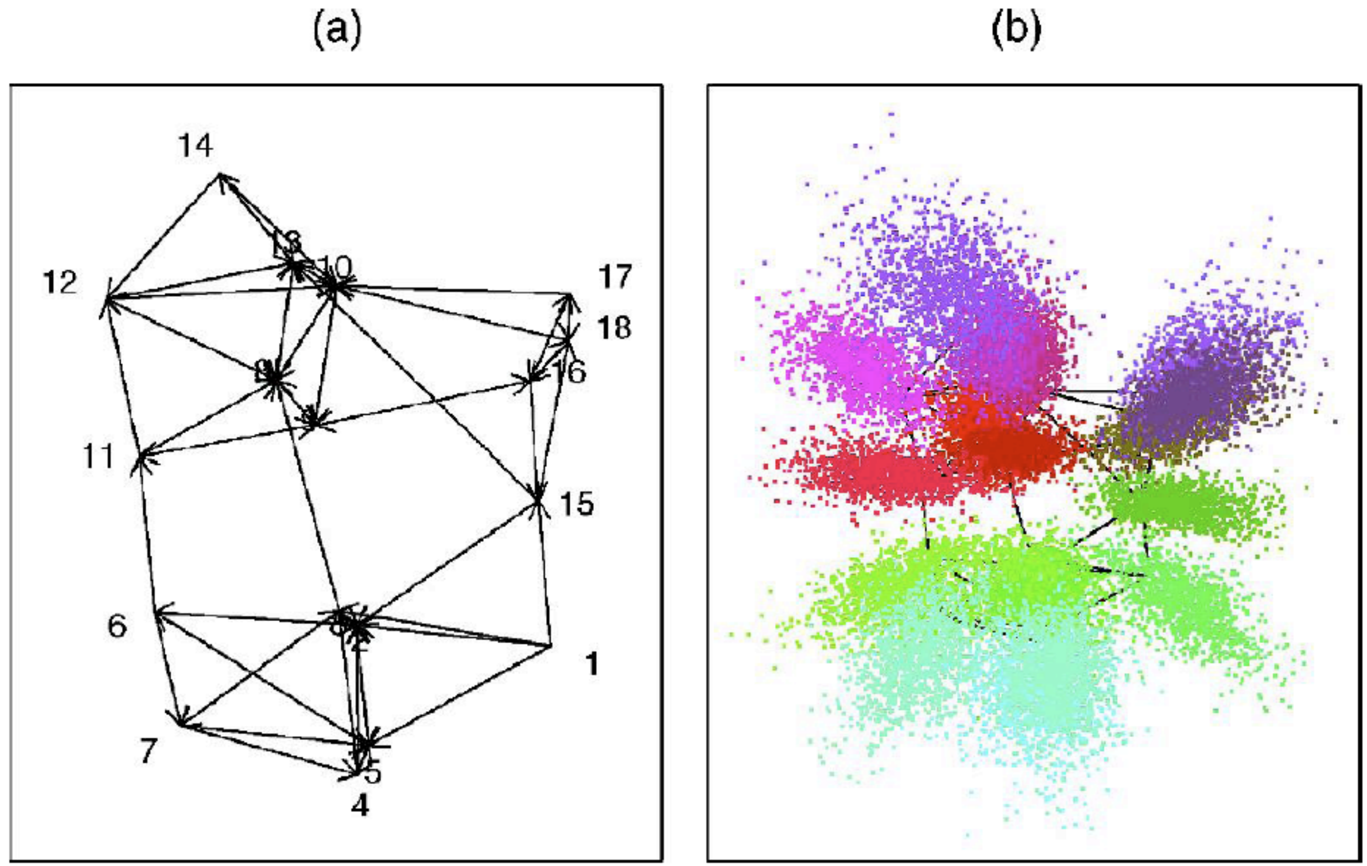


Figure 1. Maximum Likelihood Estimates (a) and Bayesian Marginal Posterior Distributions (b) for Monk Positions. The direction of a relation is indicated by an arrow.

Example: Relational Topic Model

Chang and Blei, ACM SIGKDD, 2009

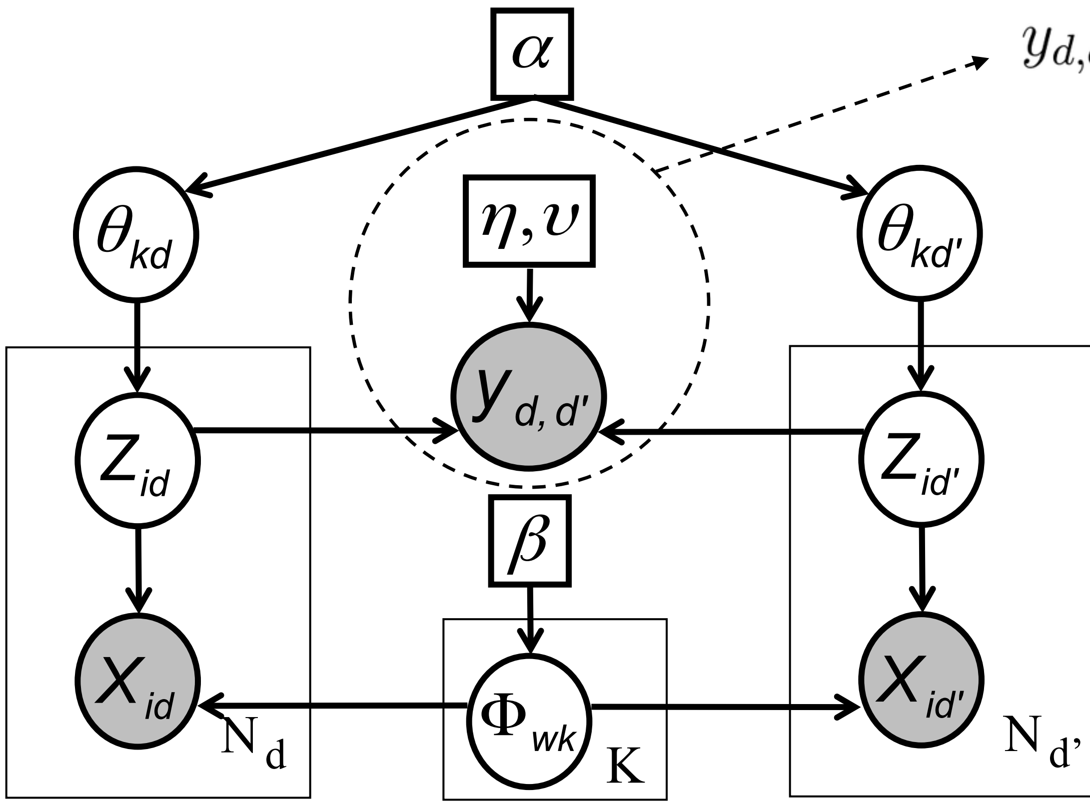
- **Nodes = documents, edges = links between documents**

- **"Standard" LDA/topic model, but ..**
  - Topics i,j influence p(edge i, j)
  - Presence of edge(i, j) encourages common topics in docs i and j

- **Model is similar to latent-space model**
  - Latent space: actors represented by k-dim location
  - Relational topics: docs represented by k-dim topic distribution
  - Both use logistic-like links for edge probabilities

Example: Relational Topic Model

Chang and Blei, ACM SIGKDD, 2009



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

"Link probability function"

Documents with similar topics are more likely to be linked

Topics influence links, and links influence topics

Example: Stochastic Block Model

Nowicki and Snijders, JASA, 2001

- **Idea:**
  - Partition the set of nodes into K "blocks" that are "structurally equivalent"
  - Model interactions at the K x K block level instead of N x N actor level

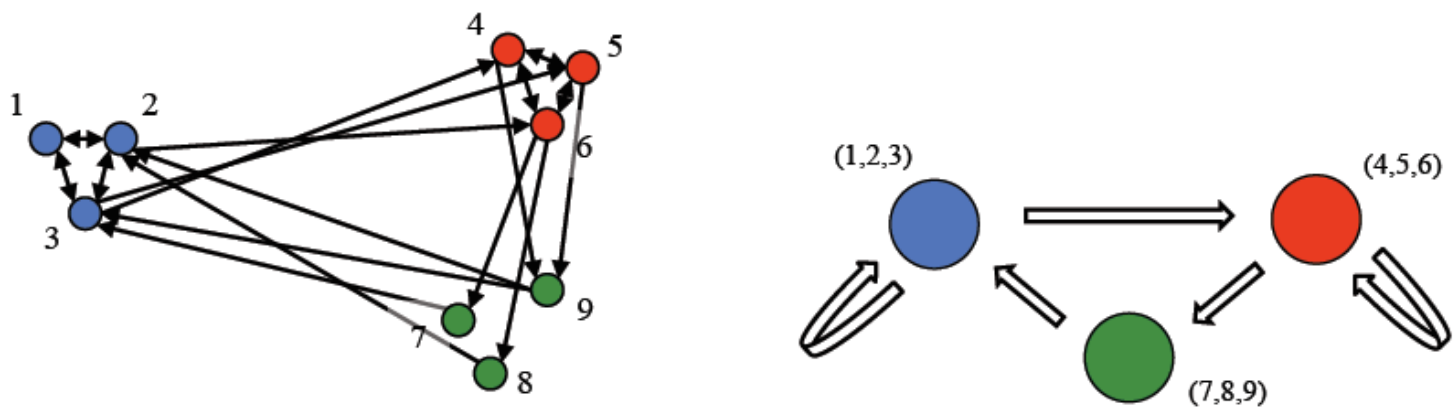
$$P(y_{ij}) = P(y_{k_i, k_j}), \quad k_i, k_j \in \{1,\ldots\ldots K\}$$

Example: Stochastic Block Model

Nowicki and Snijders, JASA, 2001

- **Idea:**
  - Partition the set of nodes into K "blocks" that are "structurally equivalent"
  - Model interactions at the K x K block level instead of N x N actor level

$$P(y_{ij}) = P(y_{k_i, k_j}), \quad k_i, k_j \in \{1, \ldots\ldots K\}$$



(Figure from Goldenberg et al, 2010)

Example: Stochastic Block Model

Nowicki and Snijders, JASA, 2001

- **Estimation**
  - 2 sets of parameters

    1. B = block-level interaction matrix, e.g., K x K matrix of Bernoullis

    2. Z = N indicator variables, mapping each node to one of K blocks

  (Can use your favorite estimation technique: EM, gradient, MCMC, etc)

- **Extensions**
  - Infinite Relational Model (IRM), Kemp et al (2006)
  - Mixed Membership Stochastic Block Model (MMSB), Airoldi et al (2008)

A Unified View…

(see also Hoff, 2009; Airoldi 2010)

e.g., logistic function

$$P(y_{ij} = 1) = \underbrace{f}_{\text{e.g., logistic function}}(\,g(z_i\,,\,z_j) + \beta x_{ij} + \mu\,)$$

A Unified View…

(see also Hoff, 2009; Airoldi 2010)

e.g., logistic function

$$P(y_{ij} = 1) = f\left(g(z_i\ ,\ z_j) + \beta x_{ij} + \mu \right)$$

function that combines
latent vectors,
with parameters θ

k x 1 latent vectors
for nodes i, j

A Unified View…

(see also Hoff, 2009; Airoldi 2010)

e.g., logistic function

covariate vector
for pairs of nodes

network
density

$$P(y_{ij} = 1) = f\left(g(z_i\ ,\ z_j) + \beta\, x_{ij} + \mu \right)$$

function that combines
latent vectors,
with parameters θ

k x 1 latent vectors
for nodes i, j

A Unified View…

(see also Hoff, 2009; Airoldi 2010)

e.g., logistic function

covariate vector
for pairs of nodes

network
density

$$P(y_{ij} = 1) = f\left(g(z_i, z_j) + \beta x_{ij} + \mu \right)$$

function that combines
latent vectors,
with parameters θ

k x 1 latent vectors
for nodes i, j

…and likelihood = product of conditionally independent edge terms

Examples

$$\text{Log-odds}\,(y_{ij} = 1) \;=\; g(z_i\,,\, z_j) + \beta\, x_{ij} + \mu$$

Latent space model:

$z_i\,,\, z_j$ = k x 1 vectors of latent positions in Euclidean space

$$g(z_i\,,\, z_j) = -\,|\, z_i - z_j \,|$$

Examples

$$\text{Log-odds}(y_{ij} = 1) = g(z_i, z_j) + \beta x_{ij} + \mu$$

Latent space model:

z_i , z_j = k x 1 vectors of latent positions in Euclidean space

$$g(z_i, z_j) = -|z_i - z_j|$$

Latent factor model: (see Hoff, 2008)

z_i = k x 1 real-valued vector

$g(z_i, z_j) = z'_i W z_j$, where W is a k x k diagonal matrix

Examples

$$\text{Log-odds}(y_{ij} = 1) = g(z_i, z_j) + \beta x_{ij} + \mu$$

Latent space model:

z_i , z_j = k x 1 vectors of latent positions in Euclidean space

$$g(z_i, z_j) = -|z_i - z_j|$$

Latent factor model: (see Hoff, 2008)

z_i = k x 1 real-valued vector

$g(z_i, z_j) = z'_i W z_j$, where W is a k x k diagonal matrix

Relational topic model:

z_i = k-dimensional topic distribution (multinomial) for document i

$g(z_i, z_j)$ = weighted element-wise product of the 2 topics

Examples

$$\text{Log-odds}(y_{ij} = 1) = g(z_i, z_j) + \beta x_{ij} + \mu$$

Latent class or stochastic blockmodel:

z_i = fixed k-dimensional binary indicator vector, e.g., (0, 0, 1, 0 , 0)

$g(z_i, z_j) = W_{zi, zj}$, where W is a k x k matrix

The indicators select which element (block) to use

Examples

$$\text{Log-odds}(y_{ij} = 1) = g(z_i, z_j) + \beta x_{ij} + \mu$$

Latent class or stochastic blockmodel:

z_i = fixed k-dimensional binary indicator vector, e.g., (0, 0, 1, 0 , 0)

$g(z_i, z_j) = W_{zi, zj}$, where W is a k x k matrix

The indicators select which element (block) to use

Mixed membership stochastic blockmodel (MMB)

Like latent class, but z_i = sampled from "actor multinomial" i

Examples

$$\text{Log-odds}(y_{ij} = 1) \;=\; g(z_i, z_j) + \beta x_{ij} + \mu$$

Latent class or stochastic blockmodel:

z_i = fixed k-dimensional binary indicator vector, e.g., (0, 0, 1, 0 , 0)

$g(z_i, z_j) = W_{zi, zj}$, where W is a k x k matrix

The indicators select which element (block) to use

Mixed membership stochastic blockmodel (MMB)

Like latent class, but z_i = sampled from "actor multinomial" i

Relational binary feature model (finite version): (Miller, Griffiths, Jordan, NIPS 2009)

z_i = k-dimensional binary vector, e.g., (1, 0, 1, 0 , 1)

$g(z_i, z_j) = z'_i \, W \, z_j$, where W is a k x k matrix

The combination of "on" features determine the pairwise effect

Dynamic Networks: Discrete Time

Case 1: discrete-time or "snapshots"

- Y_t represents the state of the network at discrete time t
- Data D = $\{Y_1 \ldots\ldots Y_t \ldots\ldots\ldots Y_T\}$

Example

- actors = students in a school
- Y_t = friendships between students in month t, t = 1, … 12

Interest is often in network dynamics and evolution
e.g., Markov models for $P(Y_{t+1} \mid Y_t)$

(See work of Tom Sneijders, Eric Xing, and others)

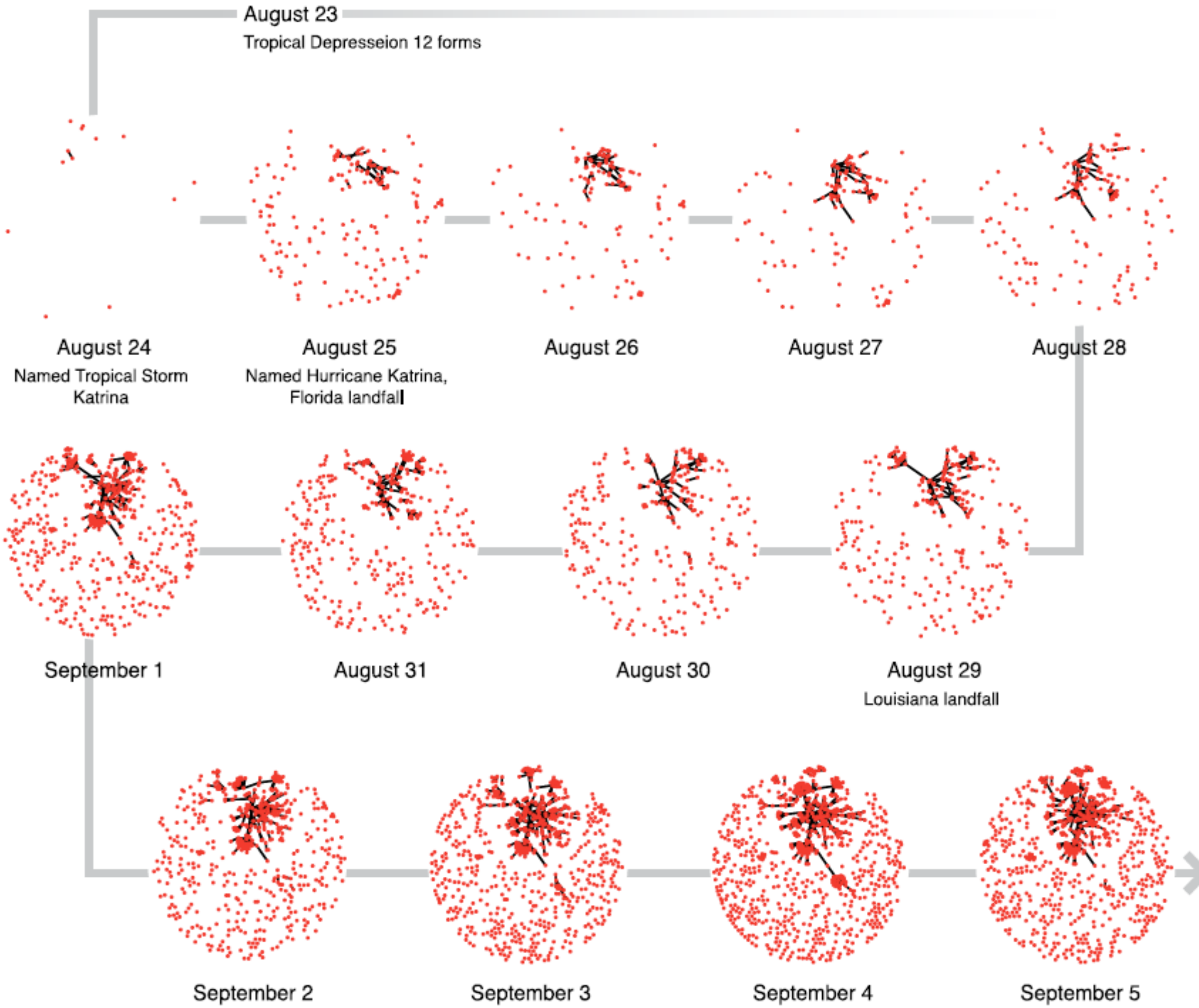


Figure from Carter Butts, UC Irvine

Latent Variable Models for Dynamic Networks

General static form for latent variable models:

$$\text{Log-odds}\,(y_{ij} = 1) = g(z_i, z_j) + \beta x_{ij} + \mu$$

One approach is to make the z's time-dependent

i.e., allow latent features of each actor change over time

An example: Gaussian linear motion models in z-space

- Sarkar and Moore (2005) for actors' latent-space positions
- Fu, Song, and Xing (2009) for actors' mixed membership vectors

Dynamic Latent Binary Feature Model

Foulds, Asuncion, DuBois, Butts, Smyth, AIStats 2011

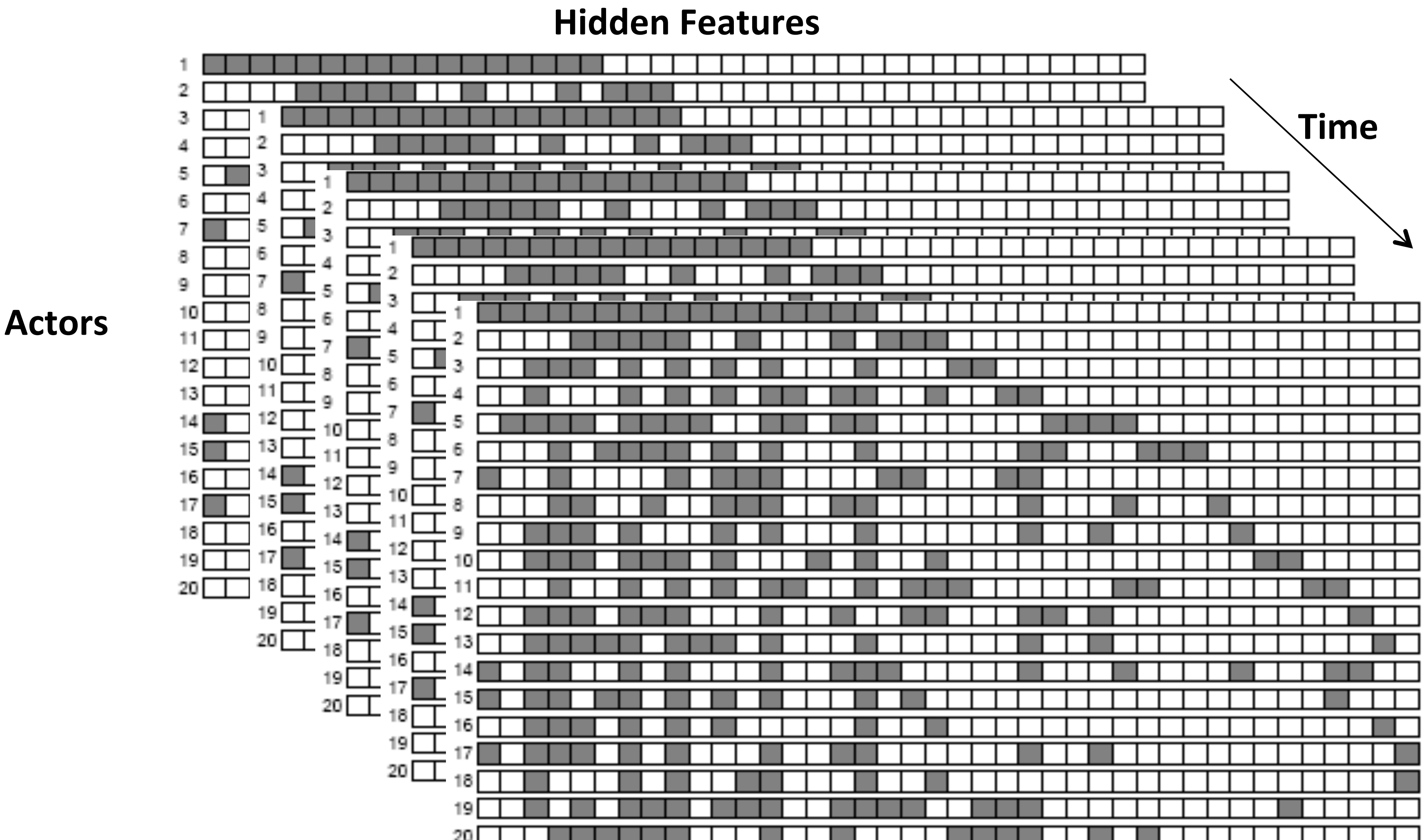
z_i = k-dimensional binary vector, e.g., (1, 0, 1, 0 , 1)

$g(z_i, z_j) = z'_i \, W \, z_j$, where W is a k x k matrix

Adding Dynamics

- The kth feature for actor i, $z_i(t)$ is a binary hidden Markov process
- Features can turn on, persist, or turn off at each time step
- For infinite version, new features can be born over time

- Inference via MCMC – tricky, but works



Presence of edge i,j at time t depends on interaction of actor i's and j's feature vectors at that time t

Examples of DRIFT Predictions

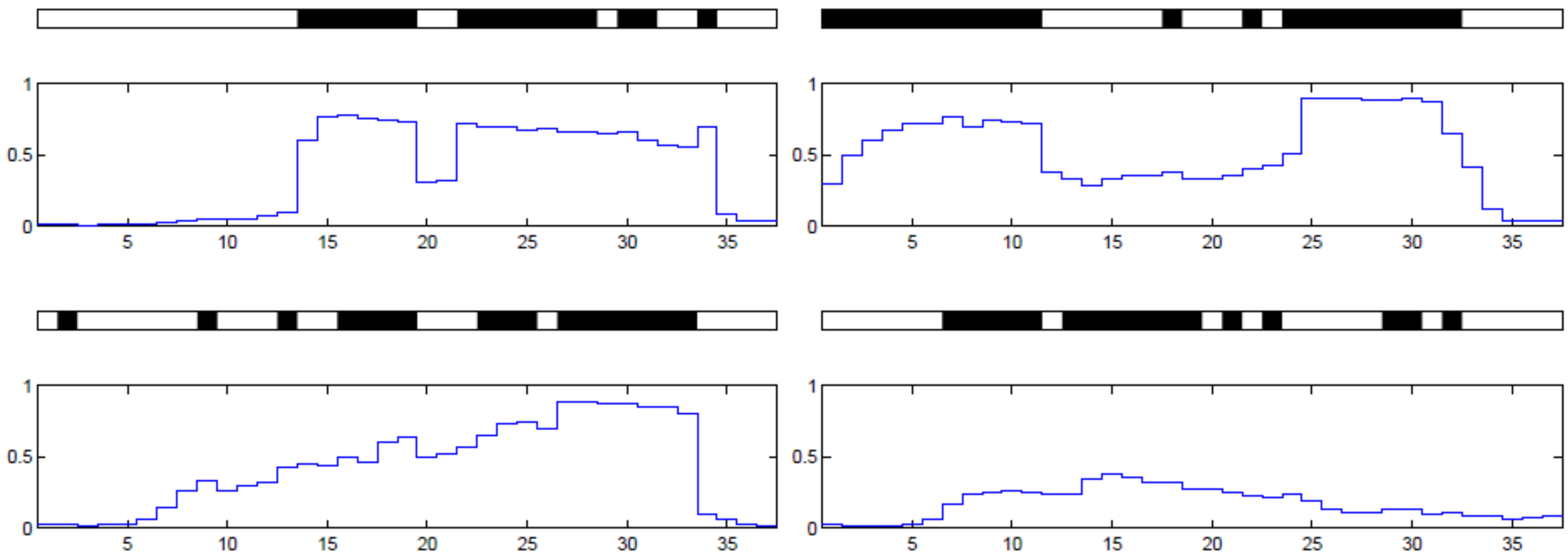
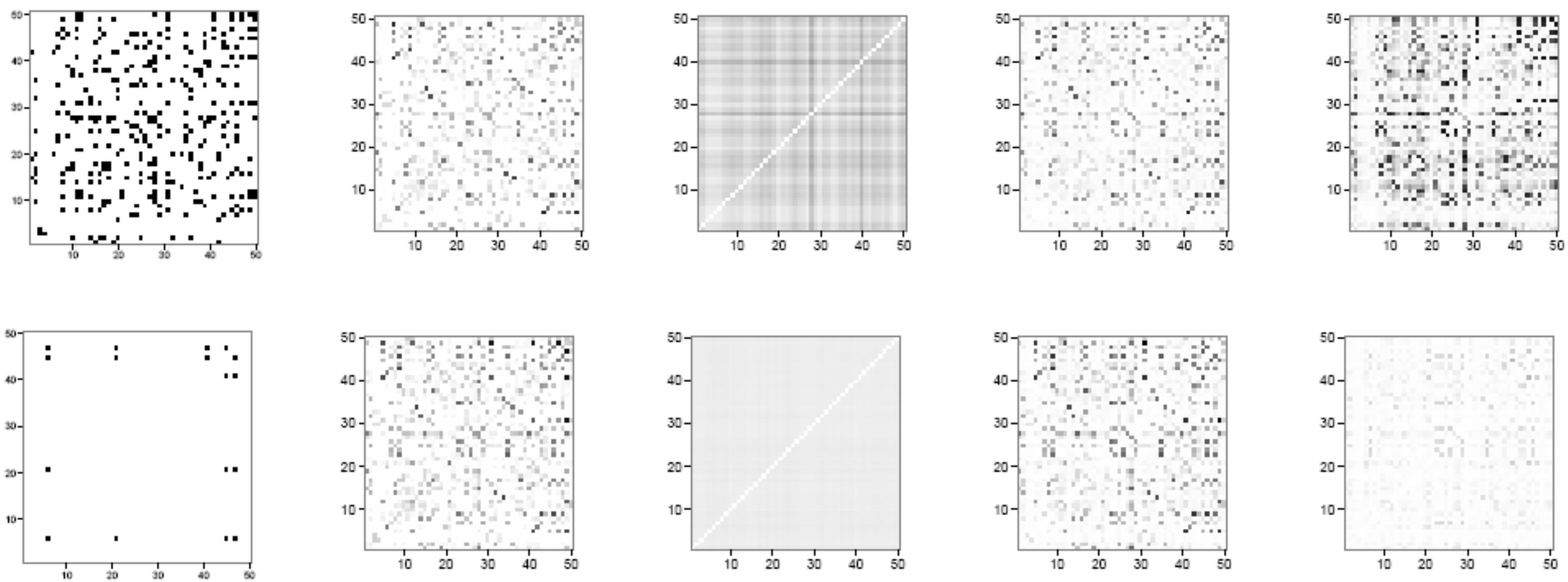


Figure 6: Estimated edge probabilities vs timestep for four pairs of actors from the Enron dataset. Above each plot the presence and absence of edges is shown, with black meaning that an edge is present.

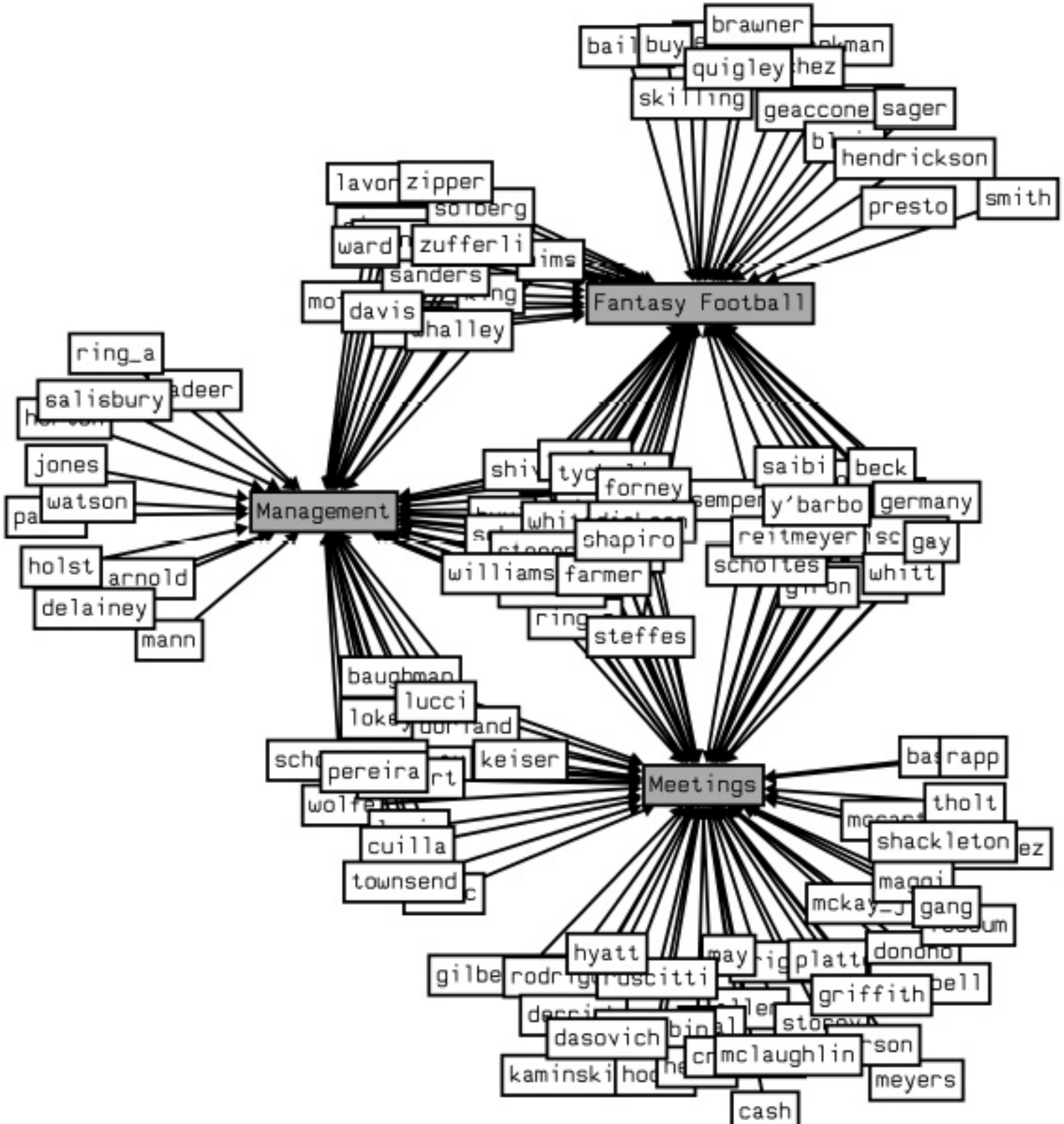
Example of DRIFT Predictions on Enron



(a) True Y (b) Baseline (c) RIFT (last) (d) RIFT (all) (e) DRIFT

Explaining Interactions with Text-based Features

(results from Enron email)



Dynamic Networks: Continuous Time

Case 2: continuous-time or "relational event" data

- y_t is an edge between some pair i and j at time t
- Birth-death edges: each y_t has start and end times
- Instantaneous edges: each y_t is (effectively) instantaneous
  - Data D = { $y_1 \ldots\ldots y_t \ldots\ldots\ldots y_T$ } - in a sense there is no graph

Example

- actors = students in a school
- y_t = text message between 2 students at time t

Interest is often in rates and patterns of communication

e.g., Poisson rates for $y_{i,j}$ given network history up to time t

Relational Event Data

- **Edge data in the form of triples: [i, j, timestamp]**

- **Can also have…**
  - Durations on edges
  - Attributes on edges (e.g., text)
  - Attributes on nodes
  - Varying node sets
  - Exogenous/external variables (e.g., other time series)

Modeling Relational Event Data

Possible Approach

$\lambda(i,j)$ = Poisson rate of edge generation between actor i and actor j

Two problems

1. rates are not static – in reality will change over time

2. modeling N^2 different rates does not scale
   (both for estimation and computation)

One approach:

Parametrize $\lambda(i,j)$ as a function of the history of the network, node attributes, etc

Relational Event Model

Butts, Social Networks, 2009

Example of model specification

$$\log \lambda^t (i,j) = \log \lambda_0 + \log \lambda_i + \log \lambda_j + \beta\, x_t(i,j)$$

p-dim vector of weights

p-dim vector of features at time t

Examples of network features include:

- persistence between pairs
- preferential attachment
- conversational behavior
- static attributes of actors i and j

Relational Event Model

Butts, Social Networks, 2009

Inhomogeneous network Poisson process

- rates λ^t (i,j) are a function of network features up to time t

- between events the rates are constant

Parameter Estimation

– Likelihood includes terms for all events that occurred and all events that did not occur, for all inter-event times
  - Various computational tricks requires to make this scalable
  - See Vu, Asuncion, Hunter, Smyth (ICML 2011, NIPS 2011)

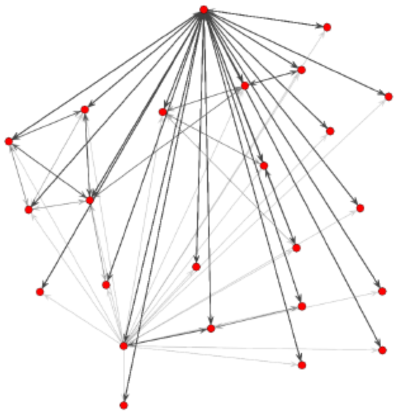
Similar ideas in event-data modeling (non-network) in statistics

see Aalen, Borgan, and Gjessing, *Survival and Event History Analysis: A Process Point of View*, 2008

Application: Classroom Dynamics

Joint work with Chris Dubois, Carter Butts, UC Irvine and Dan McFarland, Stanford

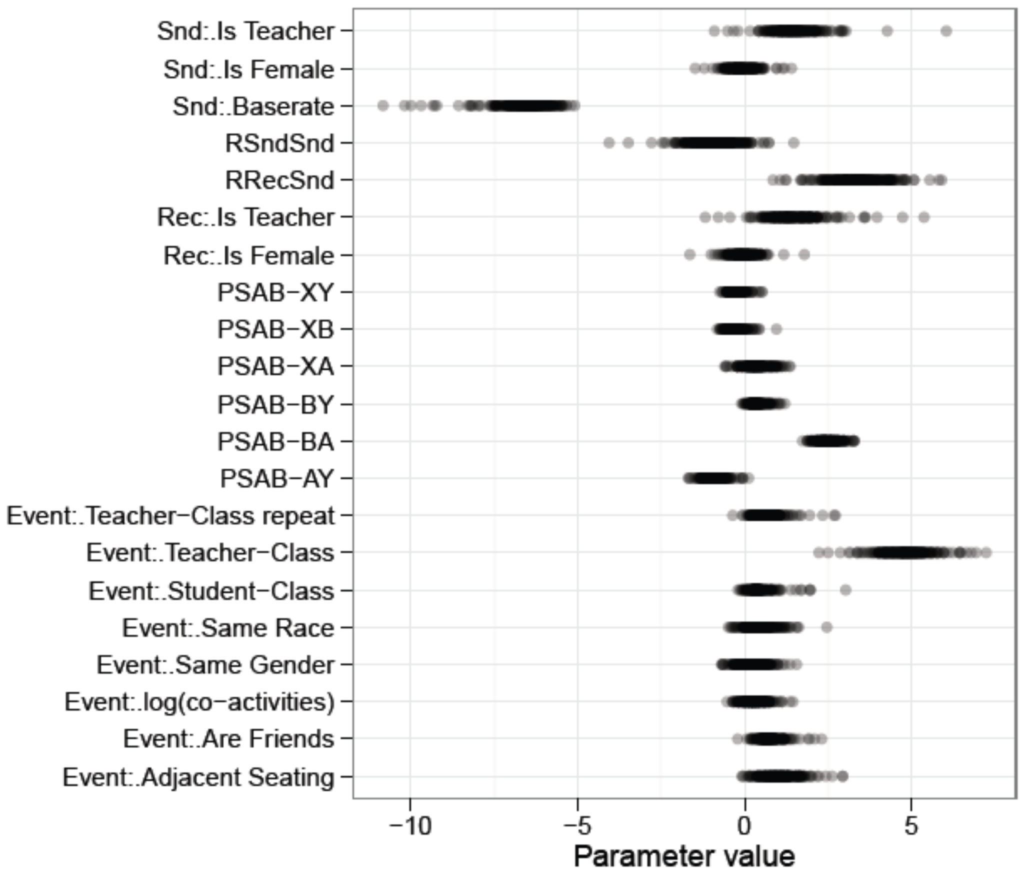
- **Education data set**
  - 600 records of classroom sessions, recording all interactions between students and teachers
  - Rich data set with student demographics, classroom variables, seating charts
  - Questions of interest: how does seating, teacher style, student age, etc, affect classroom dynamics



- **Hierarchical relational event model**
  - Model each class session with the relational event model
  - Parameter estimates shared via a Bayesian hierarchy

Application: Classroom Dynamics

Joint work with Chris Dubois, Carter Butts, UC Irvine and Dan McFarland, Stanford



Latent Variable Models for Relational Event Data

2 processes we want to model:

- Rates (e.g., Poisson)
- Choices (who connects to who)

A simple approach:

Each pair i, j (at time t) has an event rate that is Poisson λ_{ij}

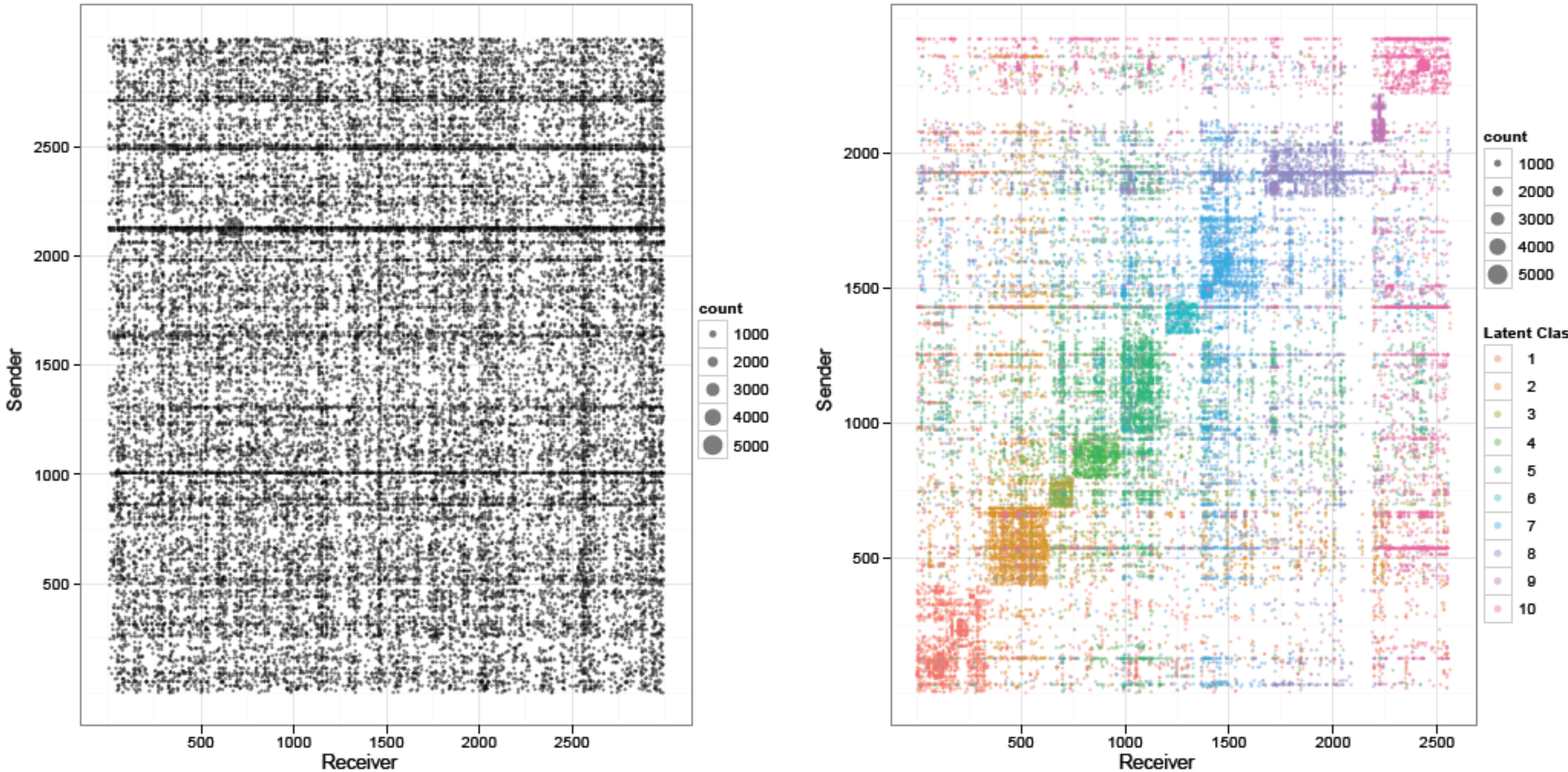
Global network rate = $\Sigma \ \lambda_{ij} = \lambda$

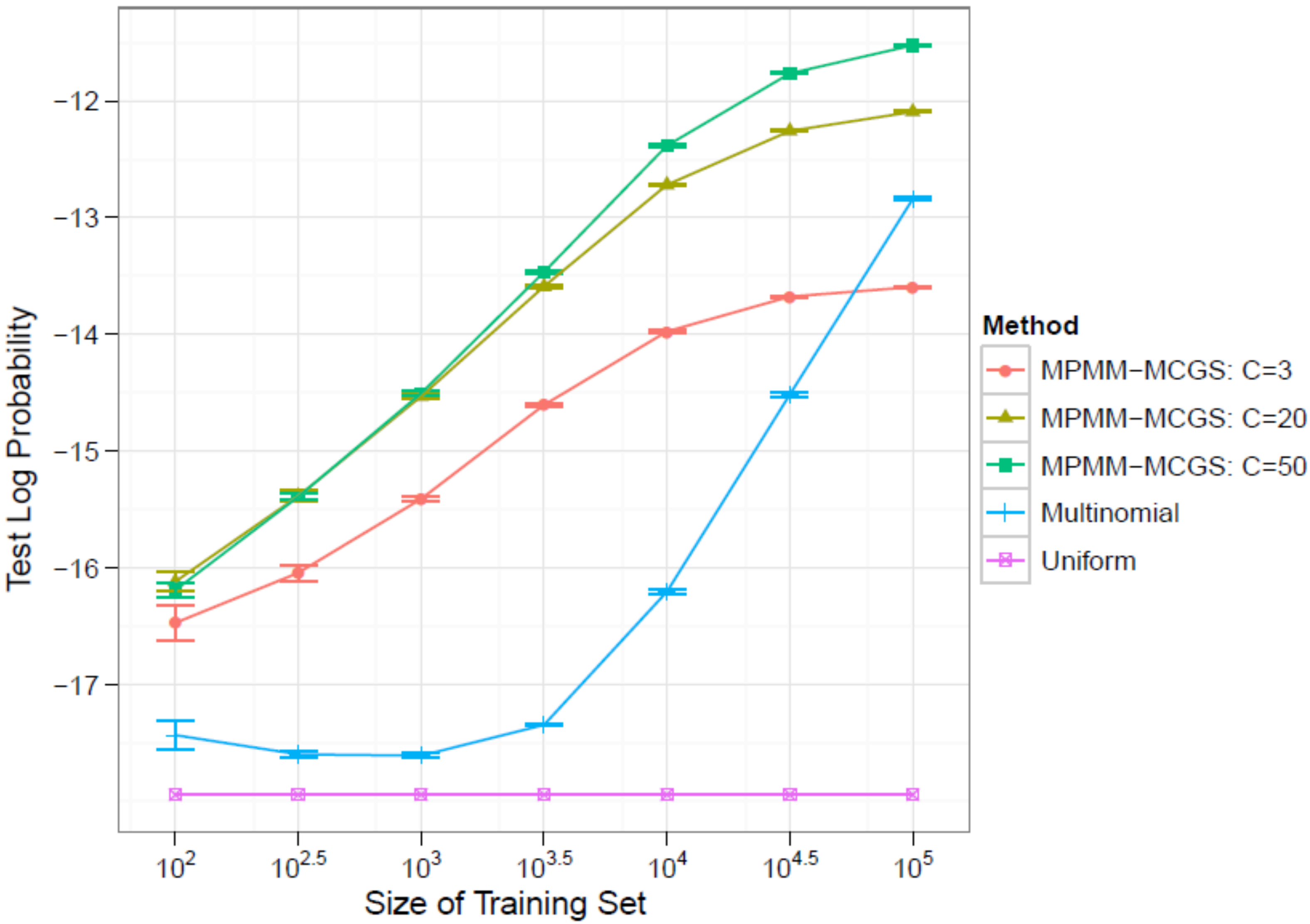
$P(y_{ij}) = \lambda_{ij} / \lambda$ or, $\lambda_{ij} = P(y_{ij}) \times \lambda$

Here $P(y_{ij})$ is a multinomial with $O(N^2)$ entries : given that an event will happen, which pair will it be?

Application to Email Data:
200,000 email messages among
3000 individuals over 3 months

Dubois and Smyth, ACM SIGKDD 2010

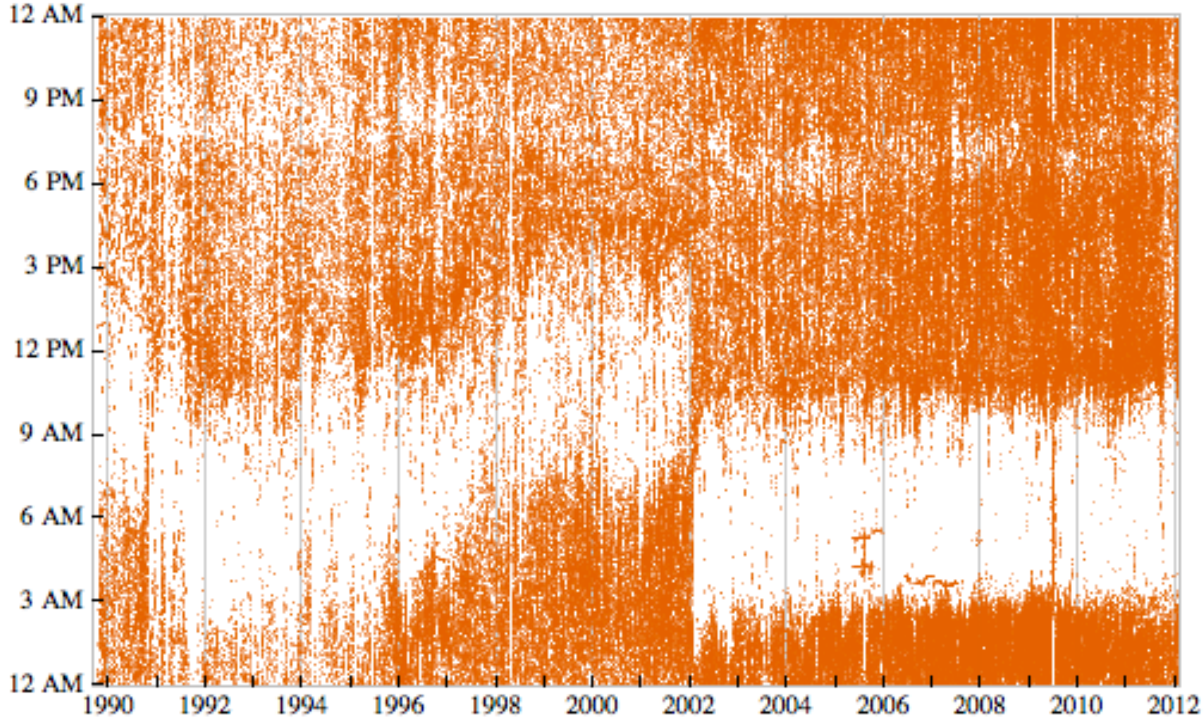


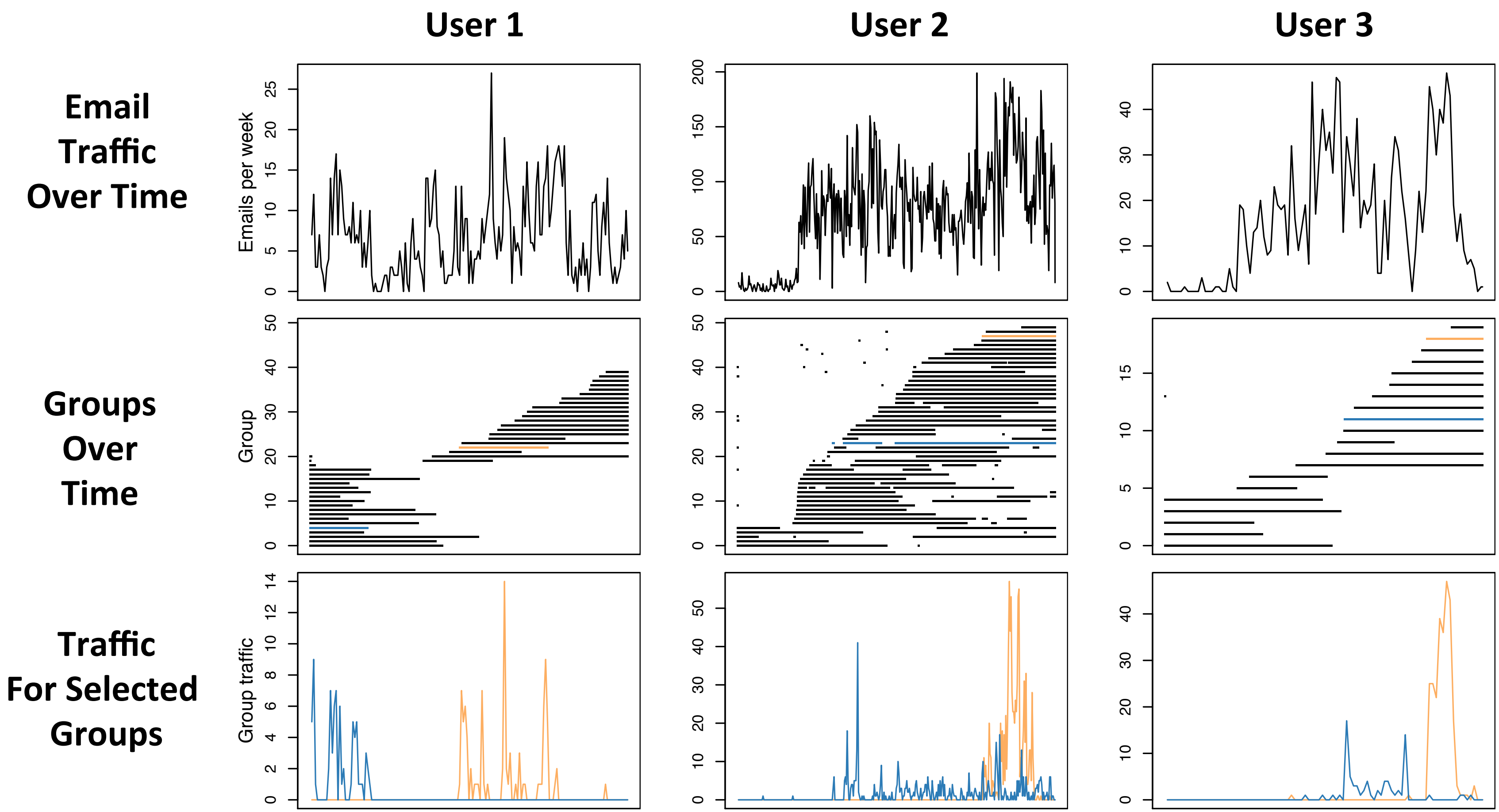


DuBois and Smyth, ACM SIGKDD 2010

Application: Personal Email Management

- What does your email history tell you about your behavior?

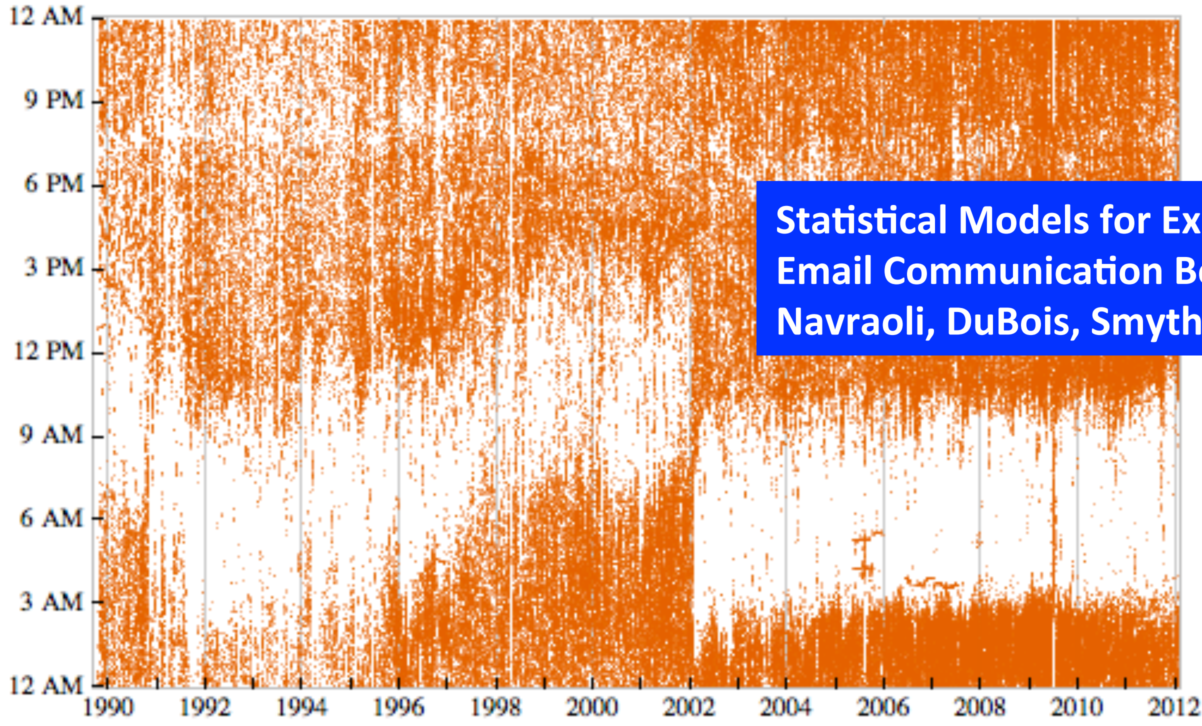




From Navaroli, DuBois, Smyth, ACML 2012

Application: Personal Email Management

- **What does your email history tell you about your behavior?**

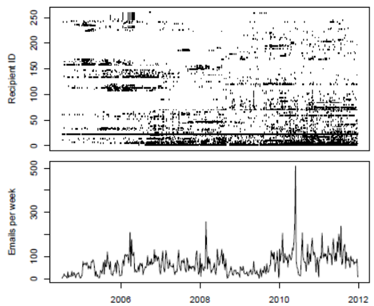


Statistical Models for Exploring Individual Email Communication Behavior
Navraoli, DuBois, Smyth, ACML 2012

Application: Personal Email Analytics

Navaroli, Dubois, Smyth, ACML 2012

- **Model both**
  - (a) Who receives your emails
  - (b) Rate at which emails are sent



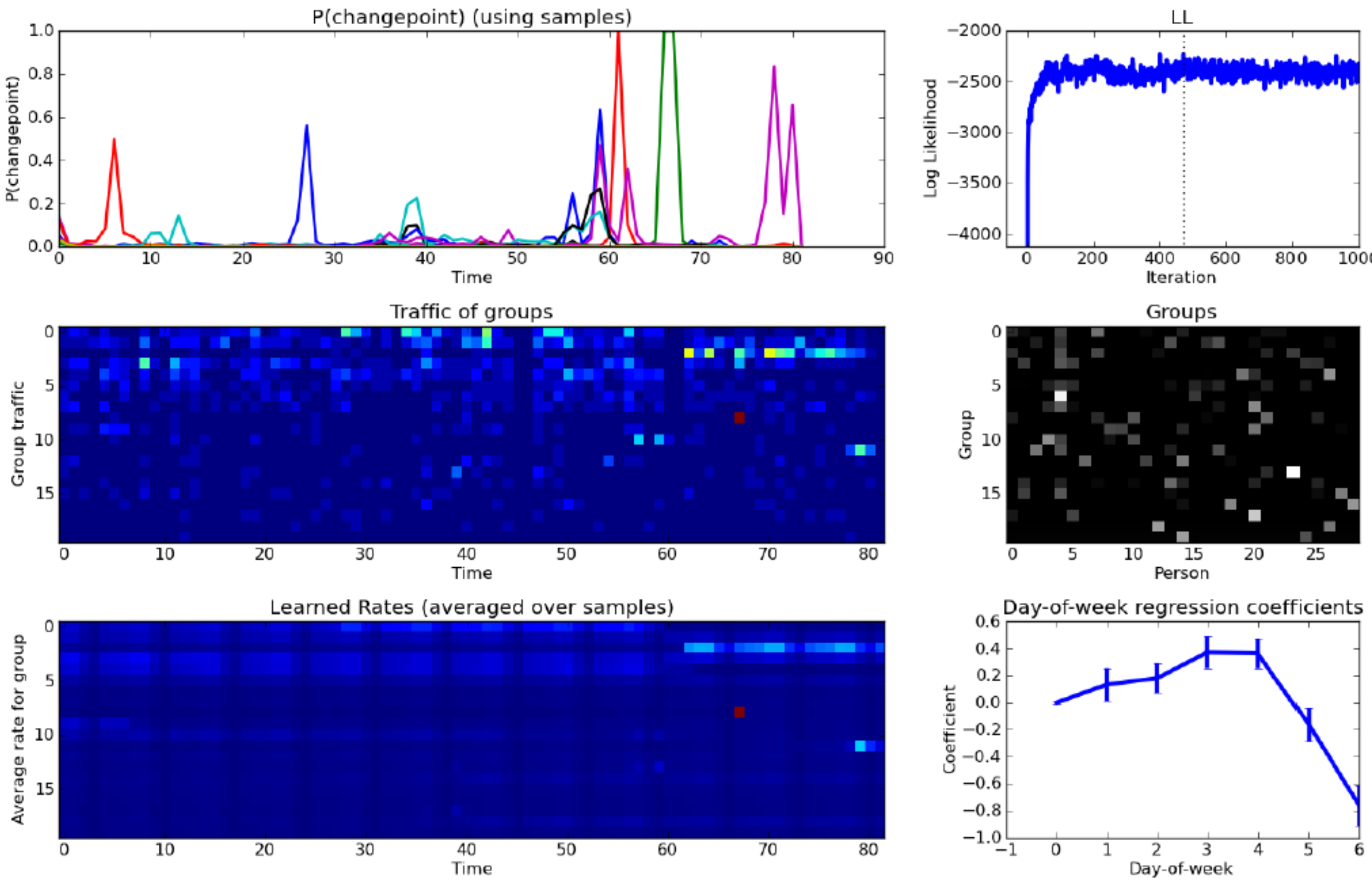
- **Group model:**
  - K latent groups, each day is "explained" by a Bernoulli-LDA model
  - Rates at which you communicate with groups can change over time

- **Rate model for each group**
  - Inhomogeneous Poisson process
    - Use Bayesian non-parametric methods to detect changepoints

Overall data is explained by sum of rates across different groups

Application: Personal Email Analytics

Joint work with Nick Navaroli, Chris DuBois



Concluding Comments

Time Complexity of Learning Algorithms

N = number of documents or actors

V = vocabulary size

K = dimensionality of latent variables

L = average doc length or average degree (in networks)

T = number of events or number of time points

Time Complexity of Learning Algorithms

N = number of documents or actors

V = vocabulary size

K = dimensionality of latent variables

L = average doc length or average degree (in networks)

T = number of events or number of time points

Time Complexity per Iteration

$O(N\,L\,K)$ (e.g., LDA, some network models)

$O(N^2 K)$ (e.g., latent space, MMSB network models)

$O(N^2\, K\, T)$ (e.g., relational event models)

Additional Aspects

- **Dimensionality of hidden variables?**
  - Non-parametric Bayesian approaches (DPs, etc)
  - Bayesian model selection methods, cross-validation
- **Hyperparameters/smoothing**
  - Can be important
  - See Wallach, Mimno, McCallum (NIPS, 2009), Asuncion et al (UAI, 2009)
- **Interpretability versus “black box prediction”?**
  - Do we want to interpret our latent variables?
    - E.g., historians, social scientists, users?
  - Or do we want to use these models as black boxes for prediction?

Summary

- **Latent variables are a useful tool in analyzing high-dimensional structured data sets**

- **Significant progress in the past 10 years in terms of**
  - Representation, e.g., multi-membership
  - Learning, e.g., collapsed Gibbs sampling, DP methods, etc

- **Many different models....with a few key underlying ideas**

- **New research opportunities**
  - Semi-supervised learning
  - Dynamic networks
  - Applications, e.g., in humanities and social sciences

Additional Reading

A survey of statistical network models

A. Goldenberg, A. Zheng, S. Fienberg, E. Airoldi, *Foundations and Trends in Machine Learning*, 2009

Multiplicative latent factor models for description and prediction of social networks

P. D. Hoff, *Computational and Mathematical Organization Theory* , 2009.

Latent factor models for relational arrays and network data

P. D. Hoff, Tutorial notes, NIPS, 2010

Discrete component analysis

W. Buntine and A. Jakulin, *Proceedings of SFSS Workshop*, 2005

(+ various papers mentioned in the slides)

What Next?

Historically, social science applications of network analysis has focused on understanding rather than prediction per se

For data miners/computer scientists, predictive modeling plays a much more important role

Key question: what are the important applications/problems that network/graph models can solve, that can't be solved by other means?

Acknowledgements

- **PhD students:**
  - Arthur Asuncion, Chris DuBois, Jimmy Foulds, Nick Navaroli

- **Collaborators**
  - Carter Butts, Dave Hunter

- **Funding**
  - National Science Foundation
  - Office of Naval Research (MURI grant)
  - Yahoo!, Google, IBM, Microsoft, Experian

Ordinal Version of Relational Events

- If we don't have time-stamps, but do have the order of the events…
- Can use the fact that "choice probability" can be written as a ratio of rates (using Poisson superposition:

$$P(i, j) = \lambda^t (i,j) \ / \ \Sigma \ \lambda^t (i,j)$$

- Can still learn the model from sequence of events, with relative rates
  - Overall network rate λ_0 is unspecified

Marginal Product Mixture Model (MPMM)

- **Likelihood**

$$
\begin{aligned}
P(D|\Phi) &= \prod_{i=1}^{T} \sum_{c=1}^{C} P(s_i|\theta, c) P(r_i|\phi, c) P(a_i|\psi, c) P(c|\pi) \\
&= \prod_{i=1}^{T} \sum_{c=1}^{C} \theta_{c,s_i} \phi_{c,r_i} \psi_{c,a_i} \pi_c
\end{aligned}
$$

- **Estimation**
  - Can use EM or Collapsed Gibbs Sampling
    - Both are fast – only need to loop over observed events (can ignore pairs where no events occurred)

- **Extensions**
  - Modulate "choice process" with time-varying network rate
  - Different types of events ("actions")
  - Markov (hidden) dependence on selection of event class

Comparing MPMM and MMB

- **MMB model**
  - For every pair of actors
    - Sample latent class for i and j
    - Given latent classes, sample a binary edge, or a count (e.g., Poisson)

- **MPMM model**
  - For every event
    - Select latent class of event
    - Given latent class, sample i and j

- **Differences**
  - MMB models whole graphs, but not individual events
    - So dynamics are from graph to graph (e.g., Fu, Song, Xing 2009)
  - MPMM models individual events, not whole graphs
    - Allows dynamics at the event level (e.g., Markov dependence of events)
  - And inference in MPMM is much more tractable….

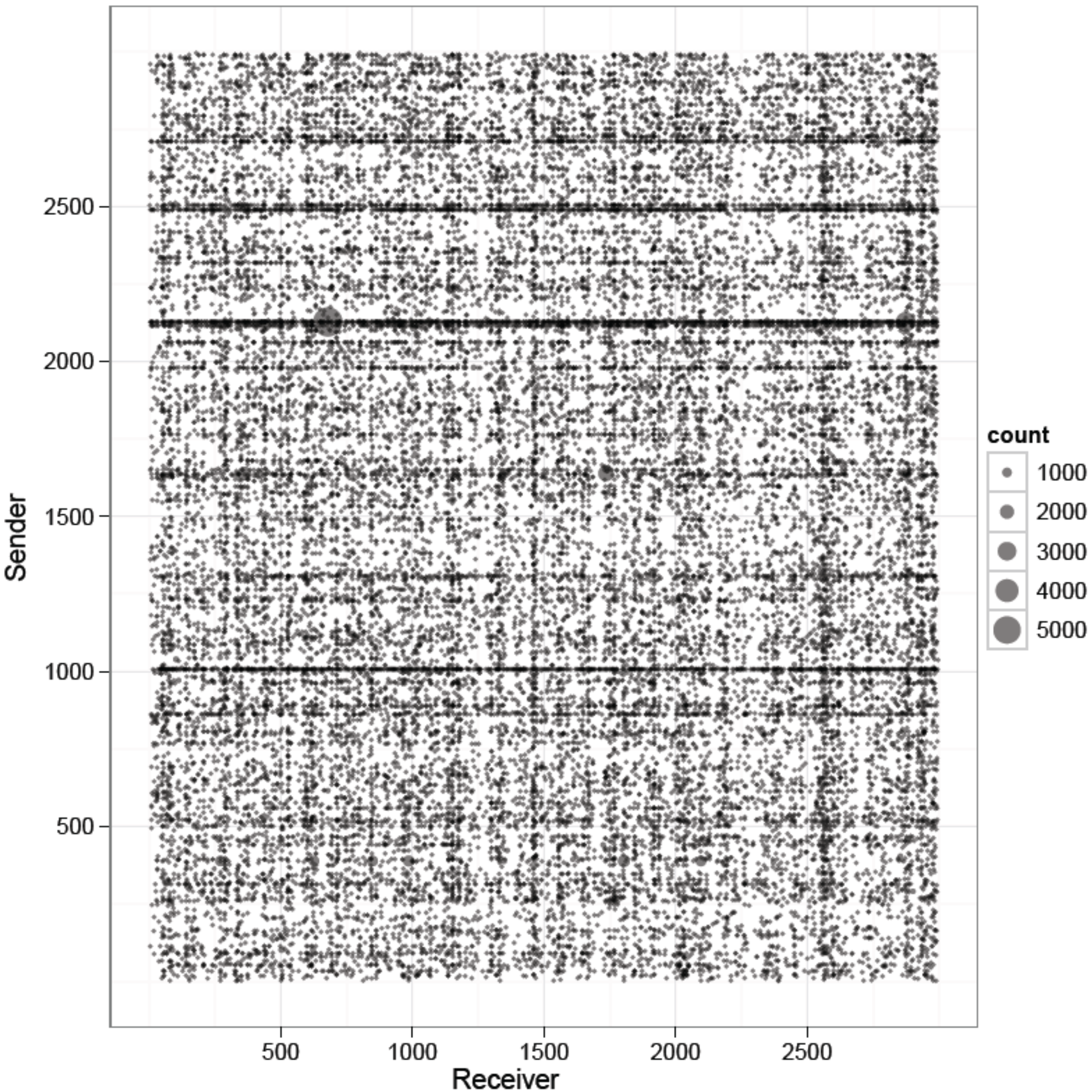
Estimation

- **EM**
  - Straightforward and fast
- **MCMC, Collapsed Gibbs sampling**
  - Also straightforward and fast
- **Both EM and Gibbs scale linearly in the number of observed events (edges)**
  - Easy to apply to large data sets

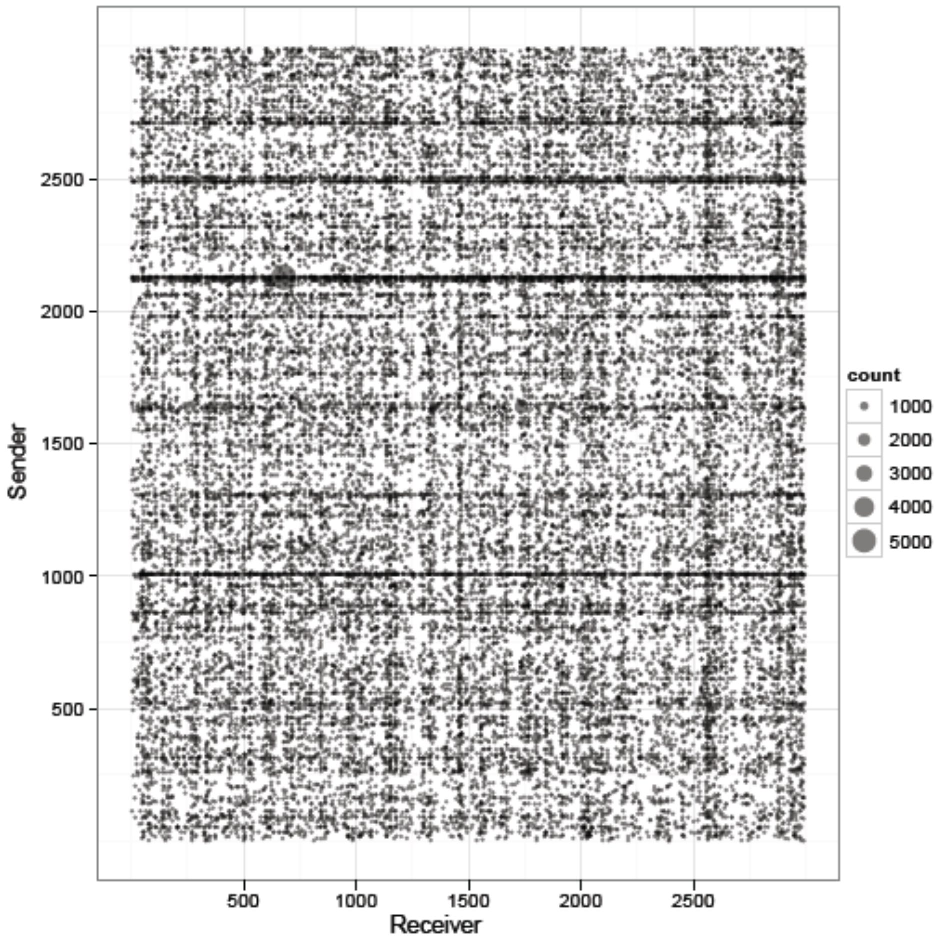
Eckmann Email Data Set

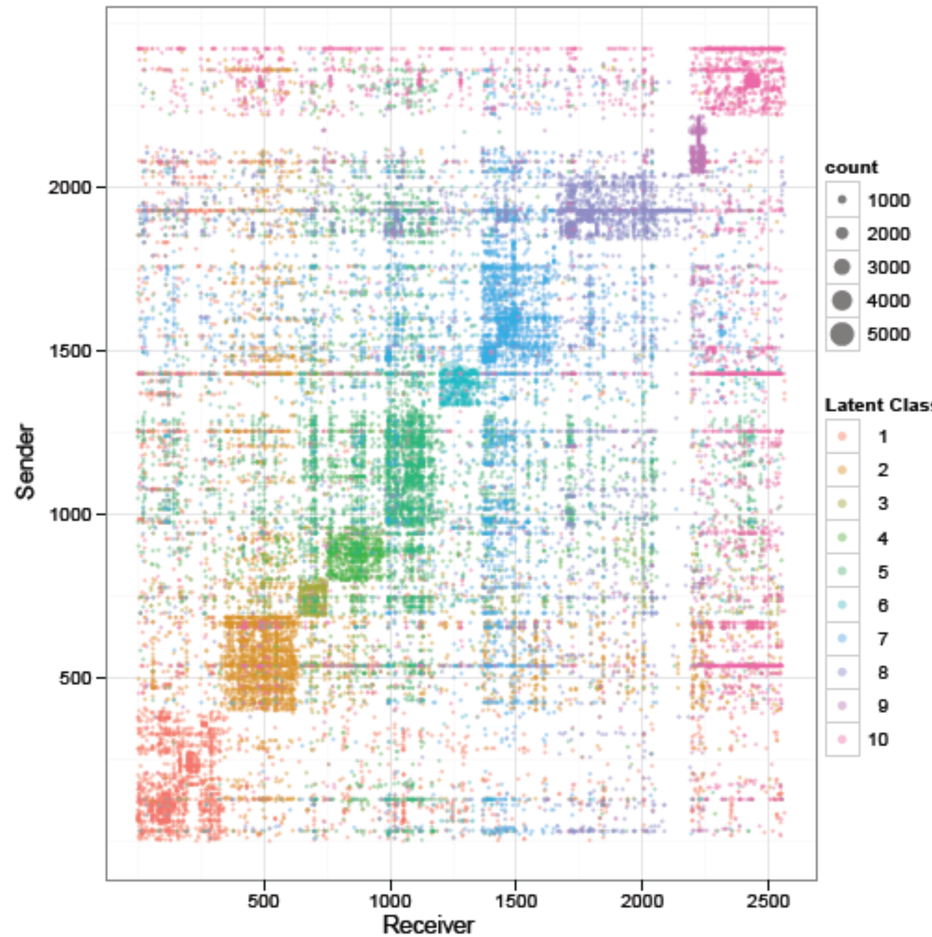
200,000 emails

2997 individuals, 82 days



Email Data (Eckmann)



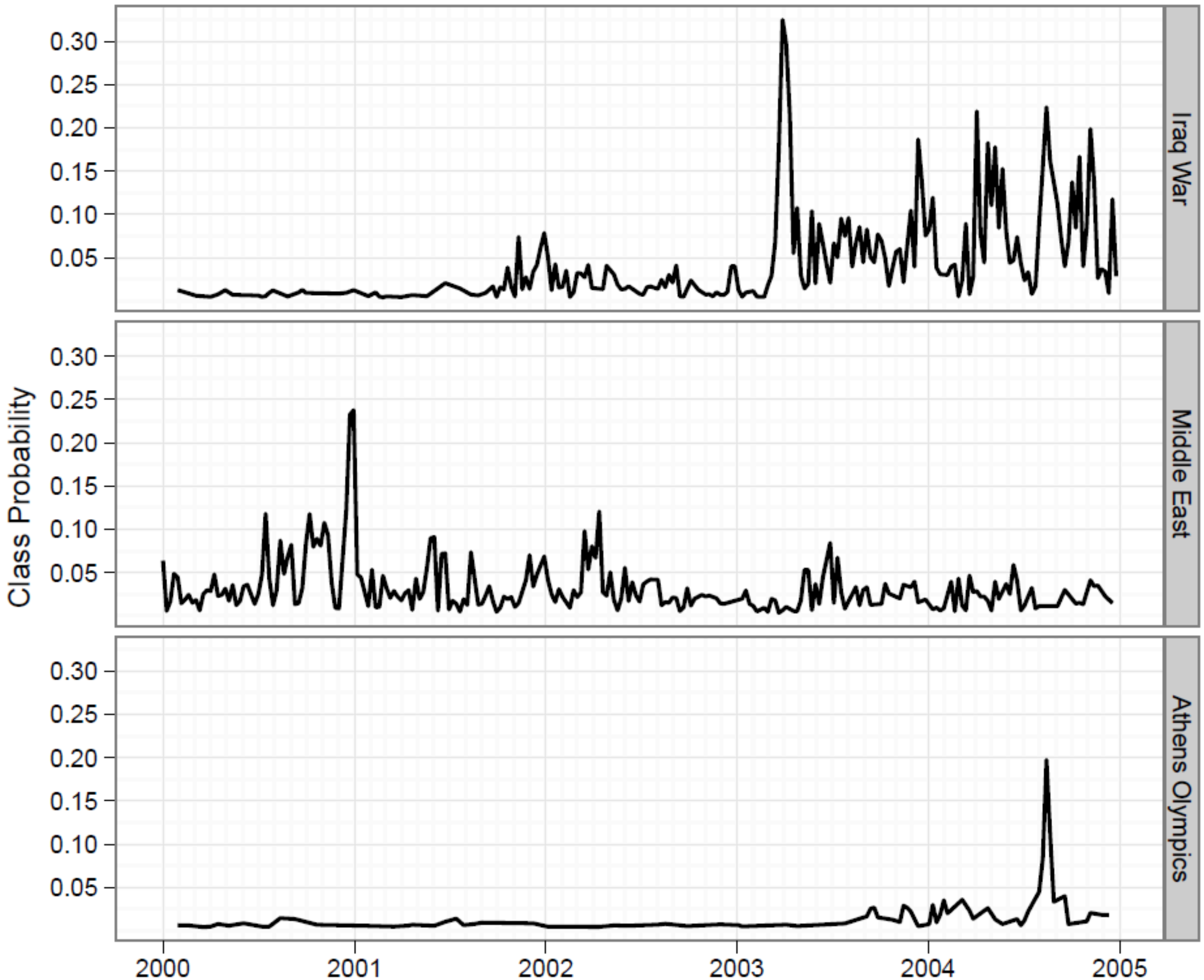


International Relations Data

(King, 2003)

40,000 events
2700 actors
171 action types

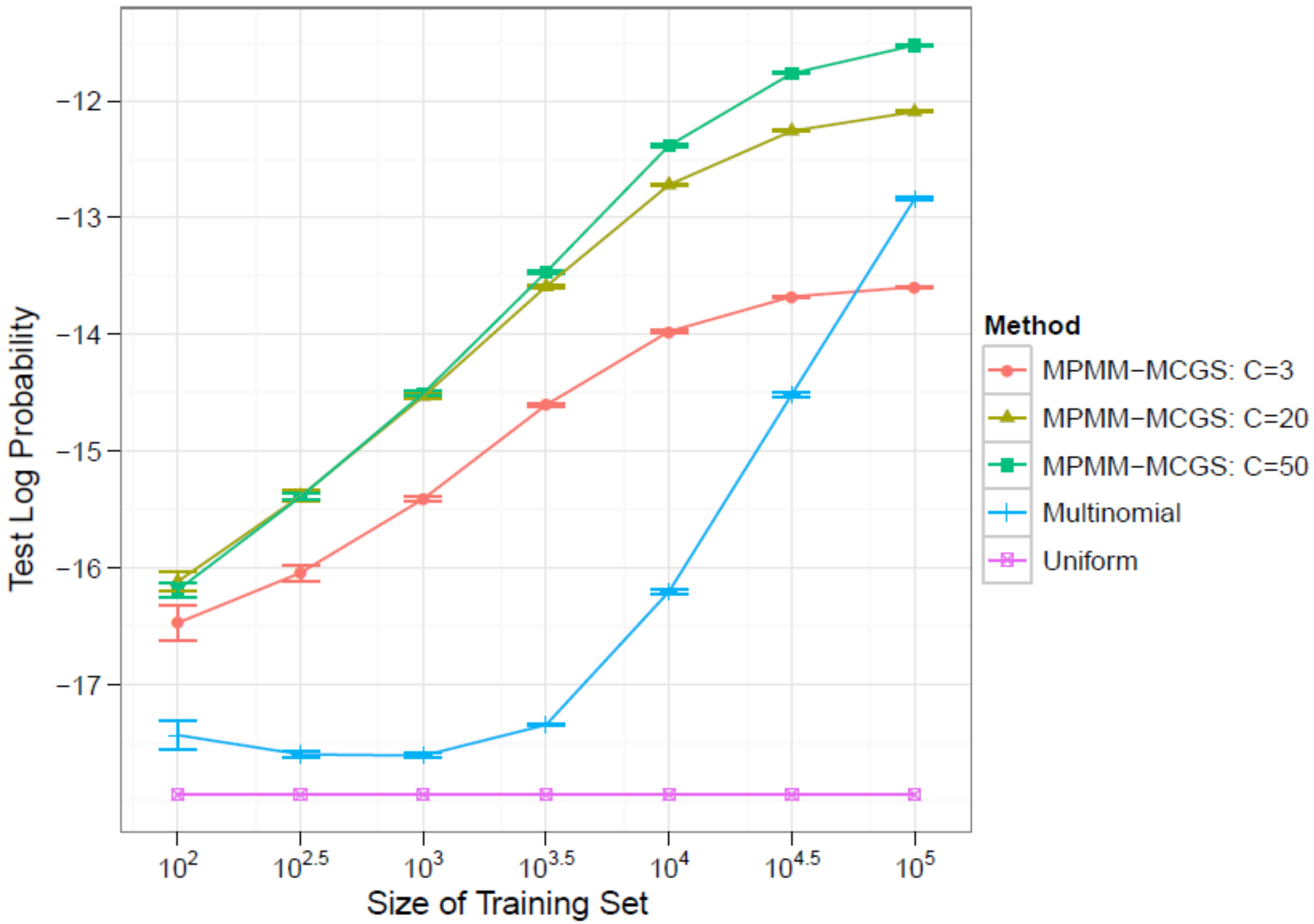
Top Senders	Pr.	Top Receivers	Pr.	Top Actions	Pr.
Class A					
United States : Government agents	0.47	Greece : NA	0.05	Sports contest	0.59
United States : Athletes	0.29	Australia : Government agents	0.02	Agree or accept	0.14
United States : Nominal agents	0.04	United Kingdom : NA	0.02	Optimistic comment	0.04
United States : Police	0.04	Canada : Government agents	0.02	Comment	0.03
United States : Occupations	0.04	France : NA	0.01	Control crowds	0.03
United States : Ethnic agents	0.03	Belgium : Government agents	0.01	Improve relations	0.01
Class B					
United States : Military	0.88	Iraq : Government agents	0.17	Comment	0.19
United States : Government agents	0.08	Iraq : National executive	0.07	Military raid	0.14
United States : Military hardware	0.01	Iraq : Military	0.05	Military clash	0.10
United States : Officials	0.00	Iraq : Ethnic agents	0.05	Military occupation	0.10
United States : Police	0.00	Iraq : Intangible things	0.04	Shooting	0.10
United States : Motor vehicles	0.00	NA : Insurgents	0.04	Political arrests and detentions	0.04
Class C					
Top Senders	Pr.	Top Receivers	Pr.	Top Actions	Pr.
United States : National executive	0.73	Palestine : National executive	0.22	Discussions	0.44
United States : Diplomats	0.15	Israel : National executive	0.12	NA	0.22
United States : Government agents	0.06	Israel : Government agents	0.09	Call for action	0.09
United States : Human actions	0.01	Egypt : National executive	0.06	Demand	0.04
United States : Artists	0.01	Palestine : Government agents	0.04	Collaborate	0.03
United States : Occupations	0.01	India : Government agents	0.03	Host a meeting	0.03


Class Probability
Iraq War
Middle East
Athens Olympics
0.30
0.25
0.20
0.15
0.10
0.05
2000
2001
2002
2003
2004
2005

Prediction and Evaluation

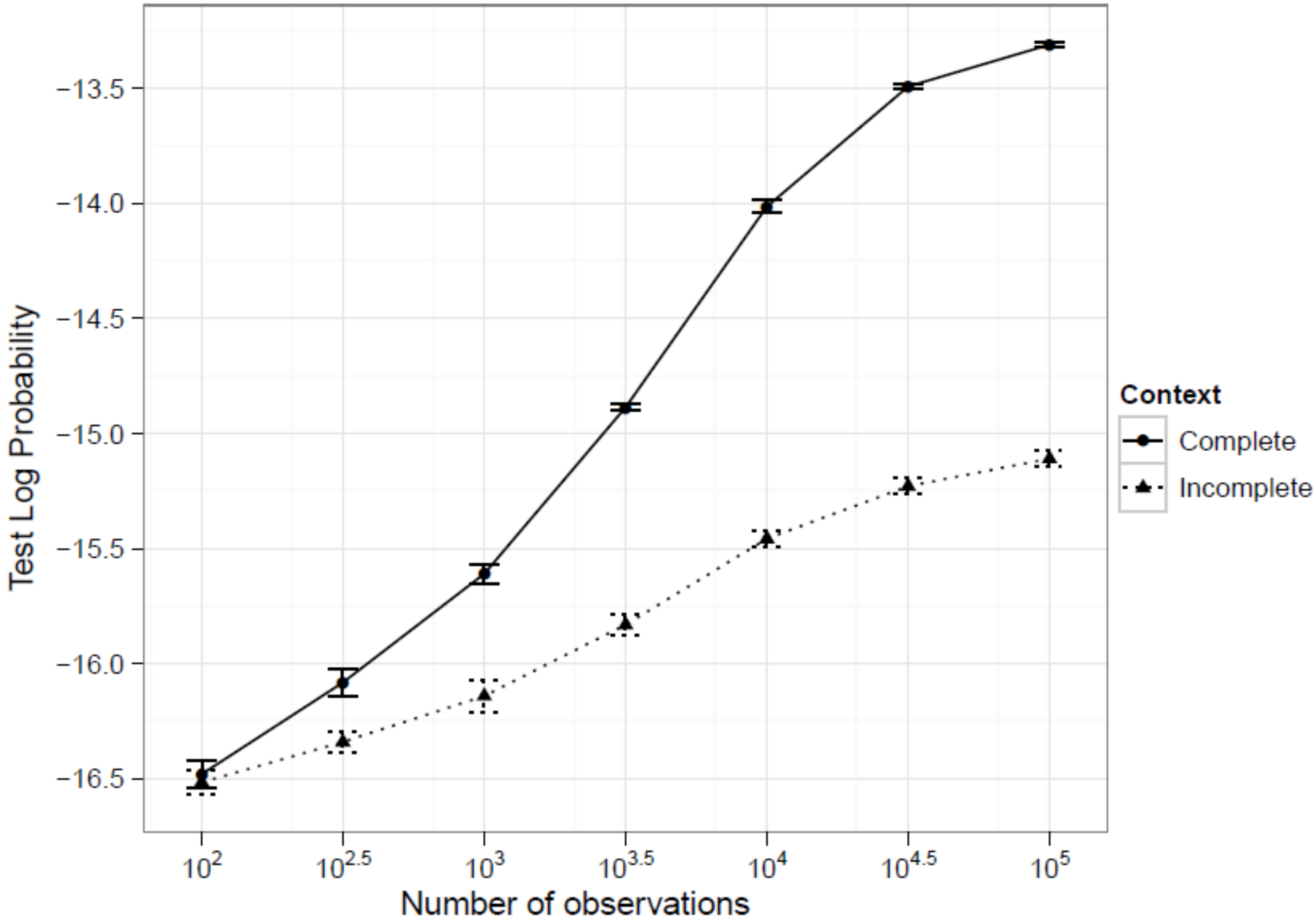
- **Use future data to evaluate predictive power and compare models**

- **Metrics**
  - Log score = log probability of events that actually occurred
  - Brier/MSE style scores
  - Ranking/ROC scores



Test Log Probability
−12
−13
−14
−15
−16
−17
10^2
$10^{2.5}$
10^3
$10^{3.5}$
10^4
$10^{4.5}$
10^5
Size of Training Set
Method
MPMM−MCGS: C=3
MPMM−MCGS: C=20
MPMM−MCGS: C=50
Multinomial
Uniform

UCIRVINE | UNIVERSITY of CALIFORNIA



Comments on Evaluation

- **Prediction on independent test data is critical**
  - Relatively easy to do with dynamic networks
  - Tricky to do with static networks (but see Hoff, 2009)

- **Caveat**
  - For link (or link probability) prediction it can be very difficult to beat relatively simple baselines, e.g.,
    - Graph(t+1) = Graph(t)
    - p(event) = smoothed estimate based on historical frequency of that pair

- **Solution?**
  - More interesting questions than just predicting what happens next, e.g
    - How likely is that group A will communicate with group B in the next k days?
    - If we have events with missing information, can we infer sender/receiver?
    - Can we detect significant shifts/non-stationarity?

What Next?

- Historically, social science applications of network analysis focused on understanding rather than prediction per se

- For data miners/computer scientists, predictive modeling plays a much more important role

- Key question: what are the important applications/problems that network/graph models can solve, that can't be solved by other means?
  - Candidates?
  - e.g., tools for egocentric modeling/analysis/management of personal communication data (email, social media, etc)
    - Change detection
    - Ranking of incoming communication events
    - ……

Summary

- **Latent variable models are useful for network modeling/prediction**
  - Broad “toolbox” of building blocks
  - May be scalable to large data sets
- **Latent models for dynamic data show promise**
- **Dynamic network data comes in multiple forms**
  - Aggregated/longitudinal data
  - Time-stamped event data
    (quite different in nature)
- **Models need to be evaluated via prediction on test data**

Resources

A survey of statistical network models

A. Goldenberg, A. Zheng, S. Fienberg, E. Airoldi, *Foundations and Trends in Machine Learning*, 2009

Multiplicative latent factor models for description and prediction of social networks

P. D. Hoff, *Computational and Mathematical Organization Theory , 2009.*

Random effects models for network data

P. D. Hoff, in *Dynamic Social Network Modeling and Analysis, 2003*

A relational event model for social action

C. E. Butts, *Sociological Methodology*, 2008

Slides from 2010 Whistler Summer School on Social Networks

http://people.cs.ubc.ca/~murphyk/pims2010Whistler/

Comments on "Relational Event Data"

No aggregation of edges in "time bins"

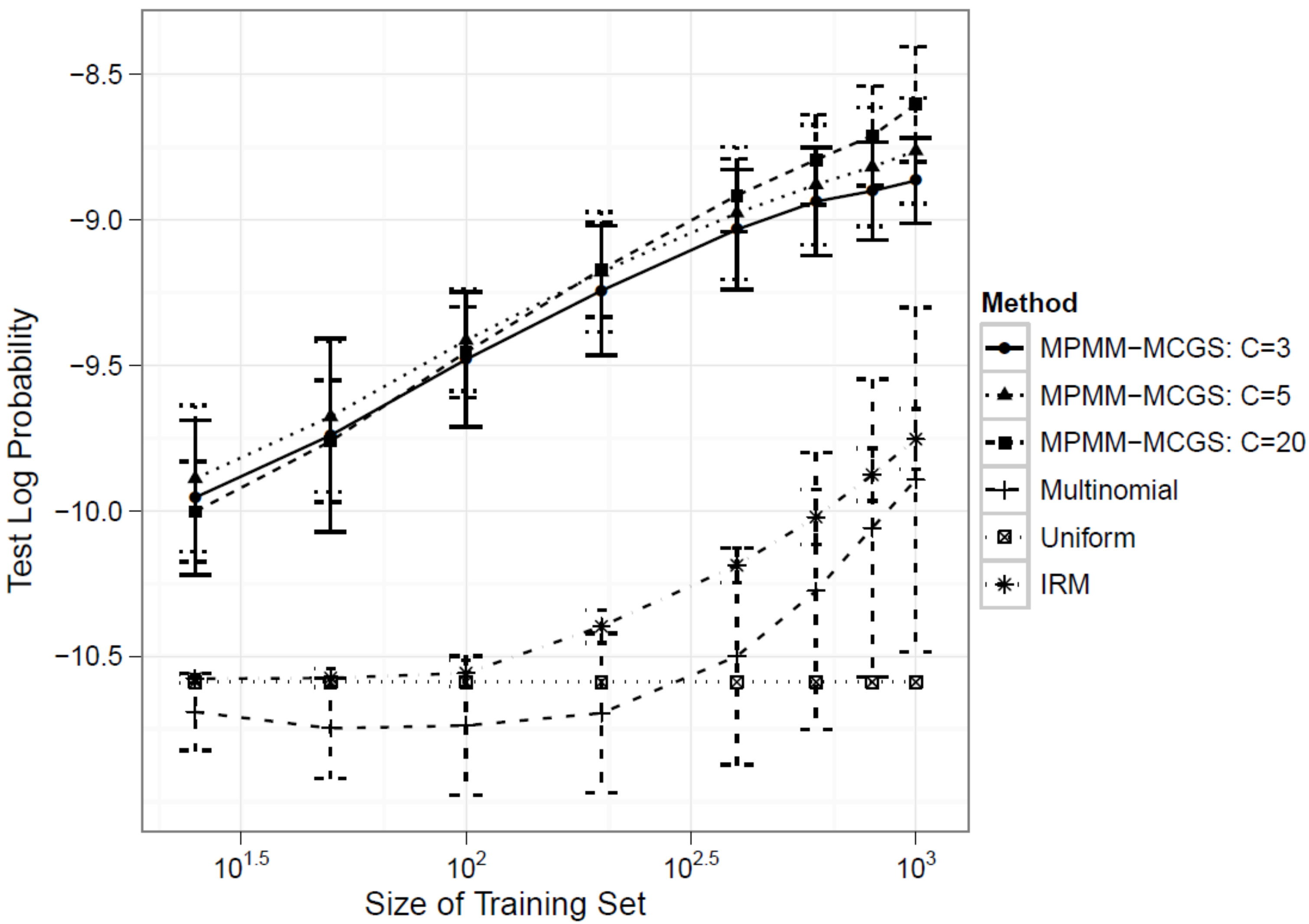
- In effect, there is no graph or network unless we aggregate

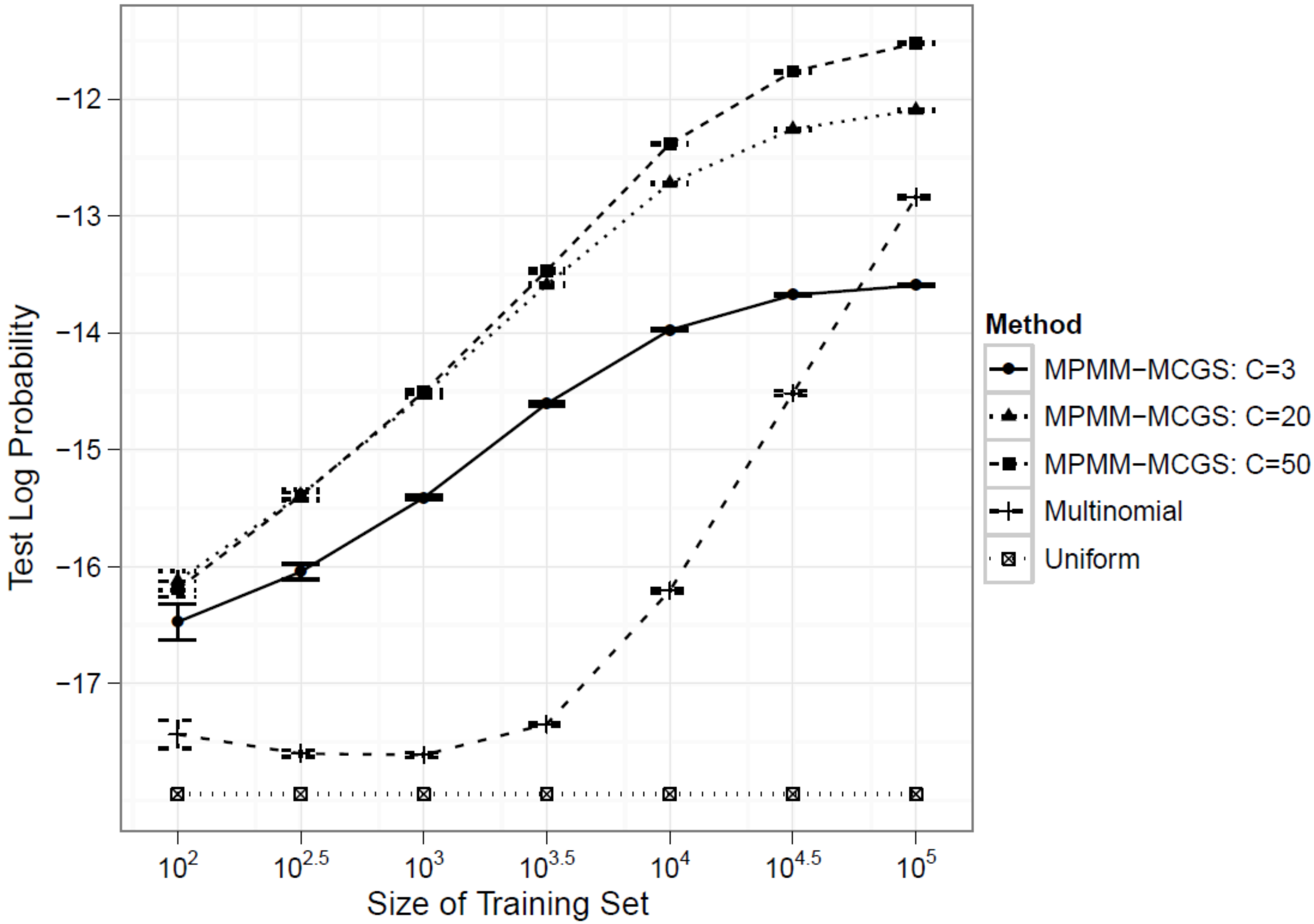
Example: Enron email data

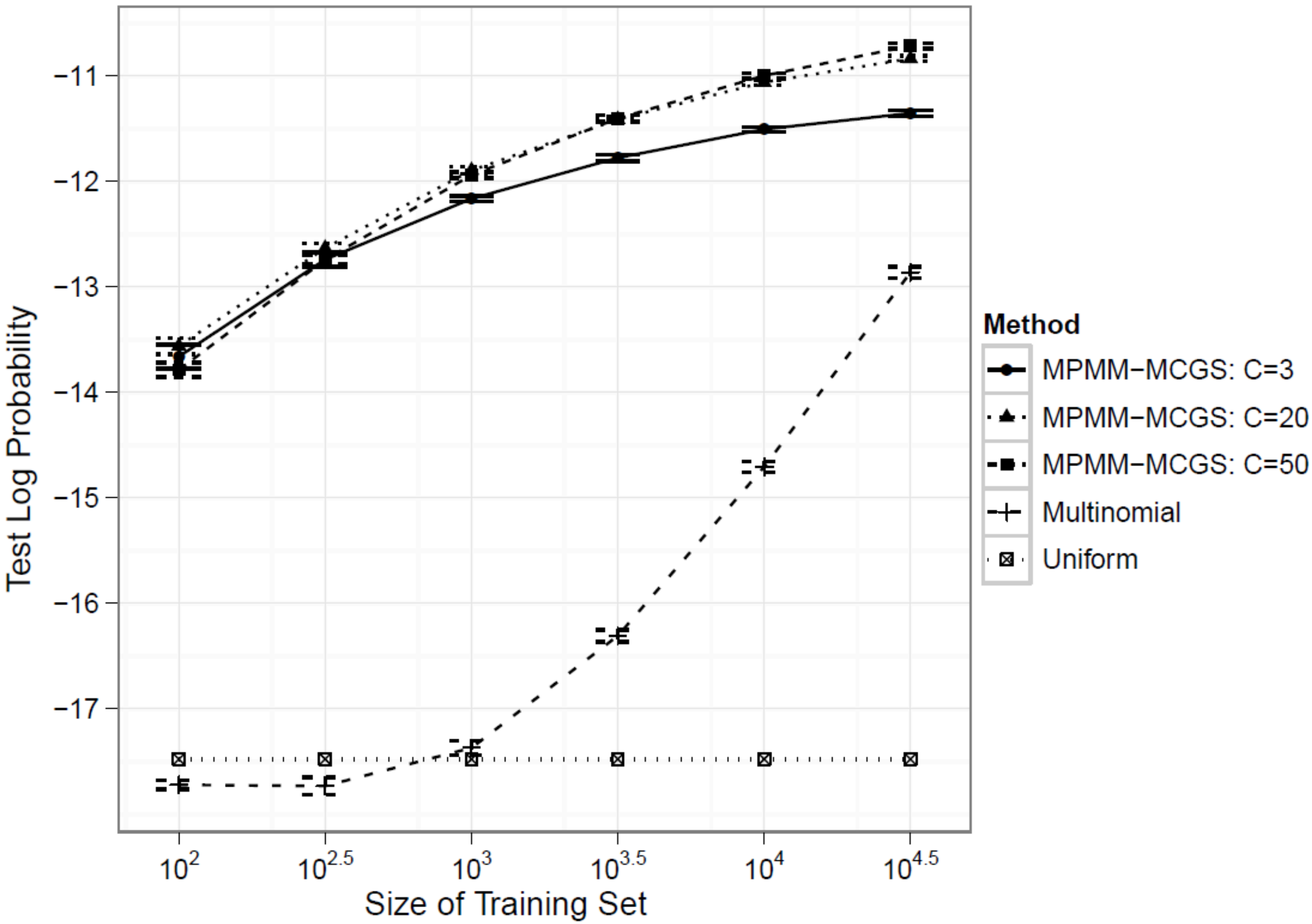
- "raw data" = set of instantaneous network events
- we can work with this event-level data directly,….

….. or we can aggregate to time-bins, e.g.,

- G_t = graph representing aggregate of emails in month t, where $e_{i,j}$ = 1 if i and j exchange at least k emails, and 0 otherwise


Test Log Probability
−8.5
−9.0
−9.5
−10.0
−10.5
$10^{1.5}$
10^2
$10^{2.5}$
10^3
Size of Training Set
Method
MPMM−MCGS: C=3
MPMM−MCGS: C=5
MPMM−MCGS: C=20
Multinomial
Uniform
IRM


Test Log Probability
−12
−13
−14
−15
−16
−17
10^2
$10^{2.5}$
10^3
$10^{3.5}$
10^4
$10^{4.5}$
10^5
Size of Training Set
Method
MPMM−MCGS: C=3
MPMM−MCGS: C=20
MPMM−MCGS: C=50
Multinomial
Uniform



Test Log Probability
−11
−12
−13
−14
−15
−16
−17
10^2
$10^{2.5}$
10^3
$10^{3.5}$
10^4
$10^{4.5}$
Size of Training Set
Method
MPMM−MCGS: C=3
MPMM−MCGS: C=20
MPMM−MCGS: C=50
Multinomial
Uniform

DRIFT Model

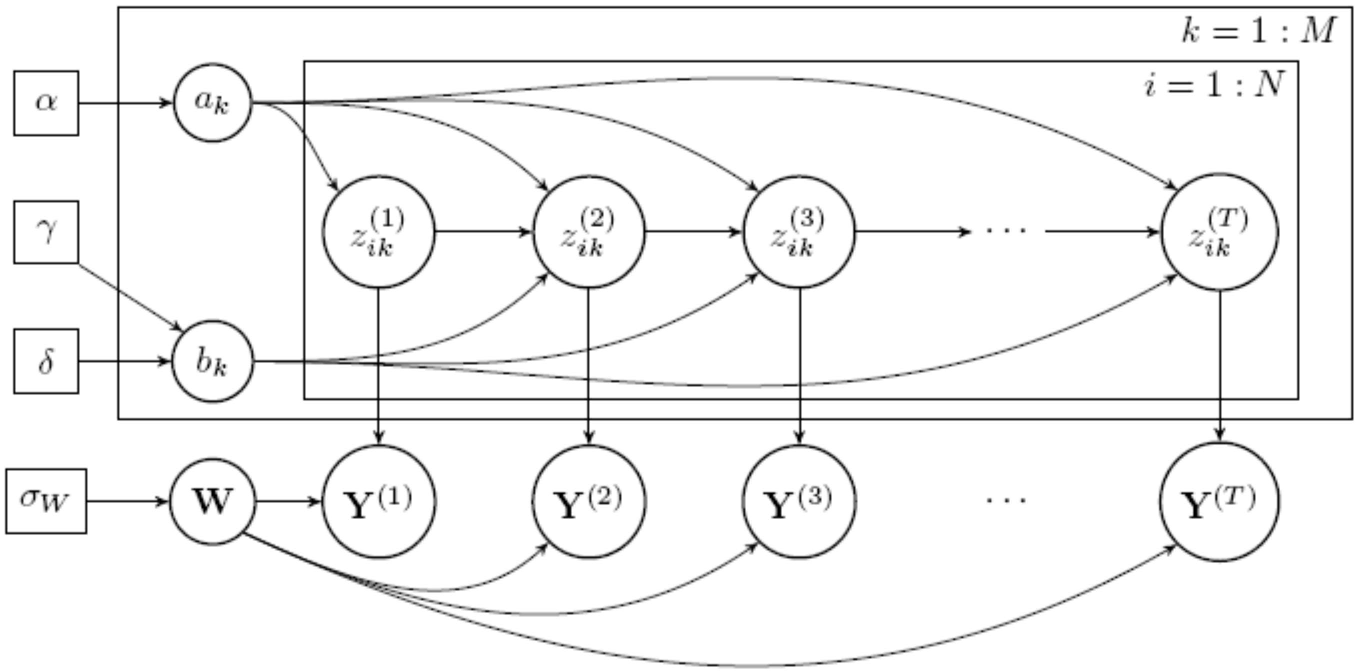


Figure 1: Graphical model for DRIFT.

Marginal Mixture Model

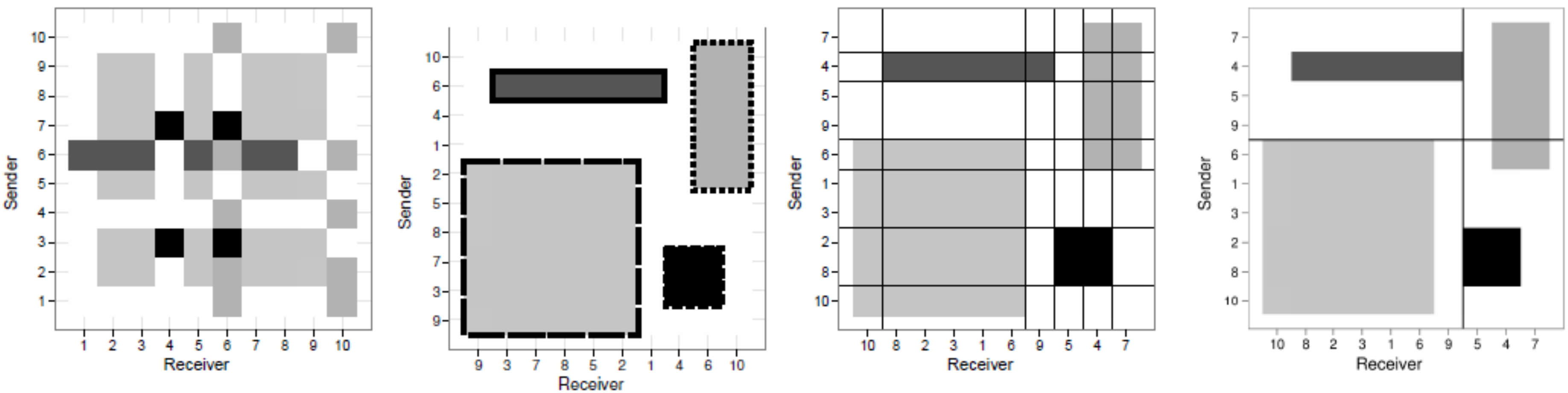
Talk by Chris DuBois, Tuesday

Sender, receiver, action type **cond. ind.** given a latent class

- For each event
  1. Draw $c \sim \text{Categorical}(\vec{\pi})$, the event's class
  2. Draw $s|c \sim \text{Categorical}(\vec{\theta}_c)$, the event's sender
  3. Draw $r|c \sim \text{Categorical}(\vec{\phi}_c)$, the event's receiver
  4. Draw $a|c \sim \text{Categorical}(\vec{\psi}_c)$, the event's type

- Likelihood:

$$
\begin{aligned}
P(D|\Phi) &= \prod_{i=1}^{T} \sum_{c=1}^{C} P(s_i|\theta, c)P(r_i|\phi, c)P(a_i|\psi, c)P(c|\pi) \\
&= \prod_{i=1}^{T} \sum_{c=1}^{C} \theta_{c,s_i}\phi_{c,r_i}\psi_{c,a_i}\pi_c
\end{aligned}
$$


Sender
10
9
8
7
6
5
4
3
2
1
Receiver
1 2 3 4 5 6 7 8 9 10
Sender
10
6
4
1
2
5
8
7
3
9
Receiver
9 3 7 8 5 2 1 4 6 10
Sender
7
4
5
9
6
1
3
2
8
10
Receiver
10 8 2 3 1 6 9 5 4 7
Sender
7
4
5
9
6
1
3
2
8
10
Receiver
10 8 2 3 1 6 9 5 4 7

(a) (b) (c) (d)

Estimation/Inference

Use your favorite techniques

- EM, gradient descent
- Variational Bayes
- (Collapsed) Gibbs sampling
- More complex MCMC schemes

Estimation is non-trivial

e.g., latent space and RTM models have $O(N^2)$ terms in the likelihood, product over all edges and non-edges => scale poorly

-> could ignore non-edges (assume missing), e.g., Chang and Blei

-> could approximate non-edge terms – see Raftery et al, 2010

Issues

- Representation
- Interpretation
- Estimation
- Evaluation

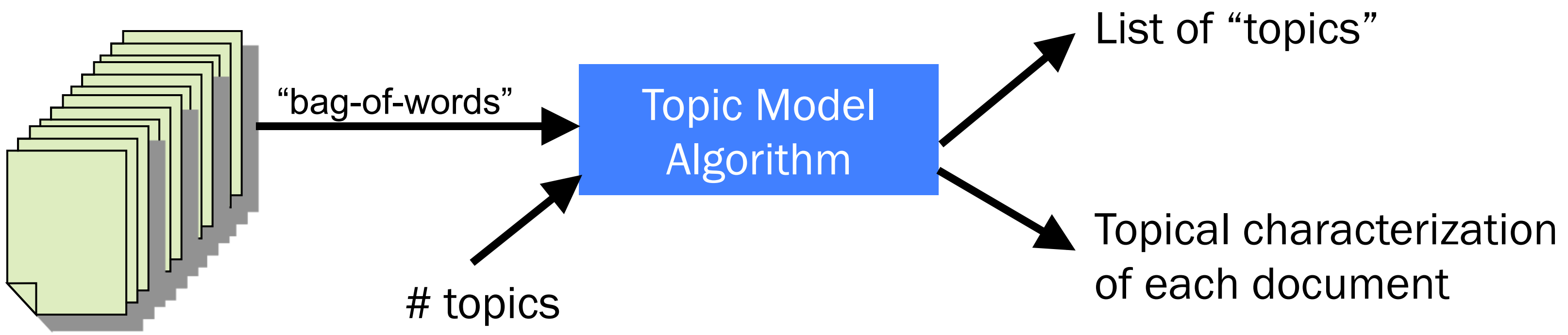
Motivation

- **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 <u>w</u>**

- **Topic = multinomial distribution over <u>w</u>**

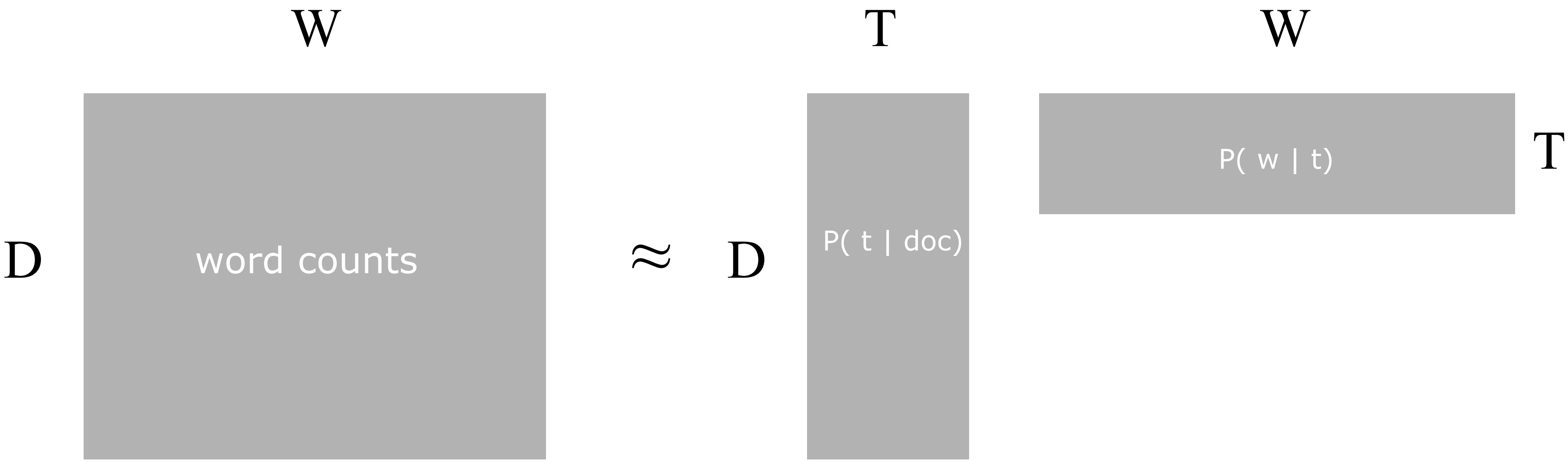
  $= P(w_1, w_2, \ldots\ldots, w_W \mid t)$

- **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

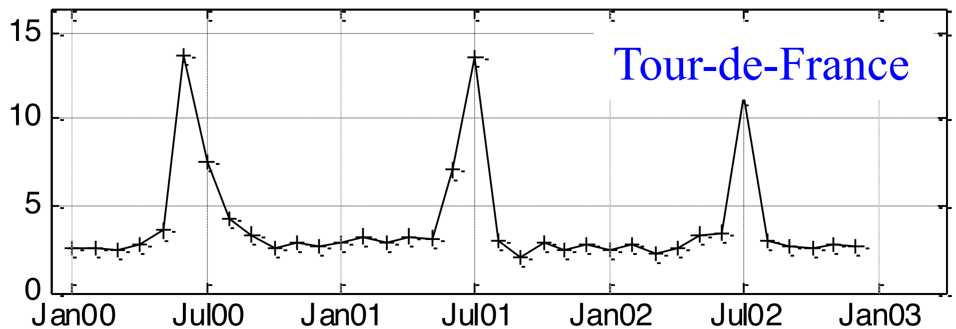
TOPIC 114	
WORD	**PROB.**
COMMITTEE	0.0197
BILL	0.0189
HOUSE	0.0169
WASHINGTON	0.0140
SENATE	0.0135
POLITICIAN X	0.0114
CONGRESS	0.0112
PRESIDENT	0.0105
LEGISLATION	0.0099
DC	0.0093
SENDER	**PROB.**
***	0.0696
***	0.0453
***	0.0255
***	0.0173
***	0.0317

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

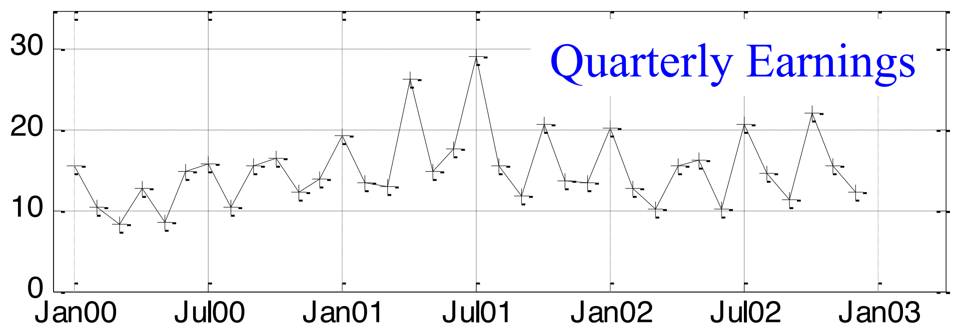
Topic trends from New York Times


The New York Times
Japanese Commuter Train Kills at Least 71

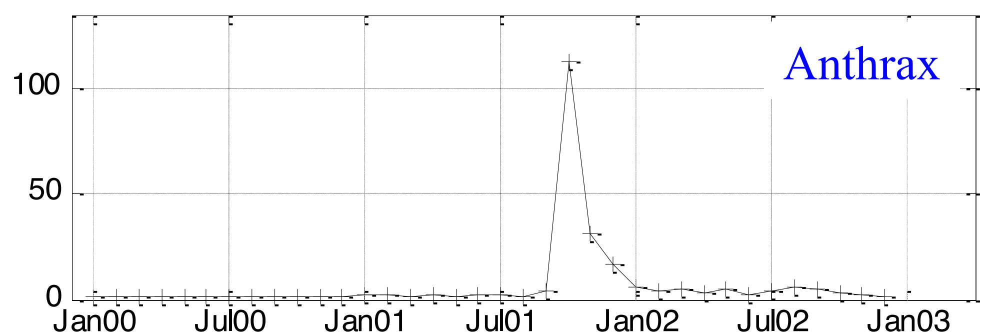
330,000 articles
2000-2002



TOUR
RIDER
LANCE_ARMSTRONG
TEAM
BIKE
RACE
FRANCE



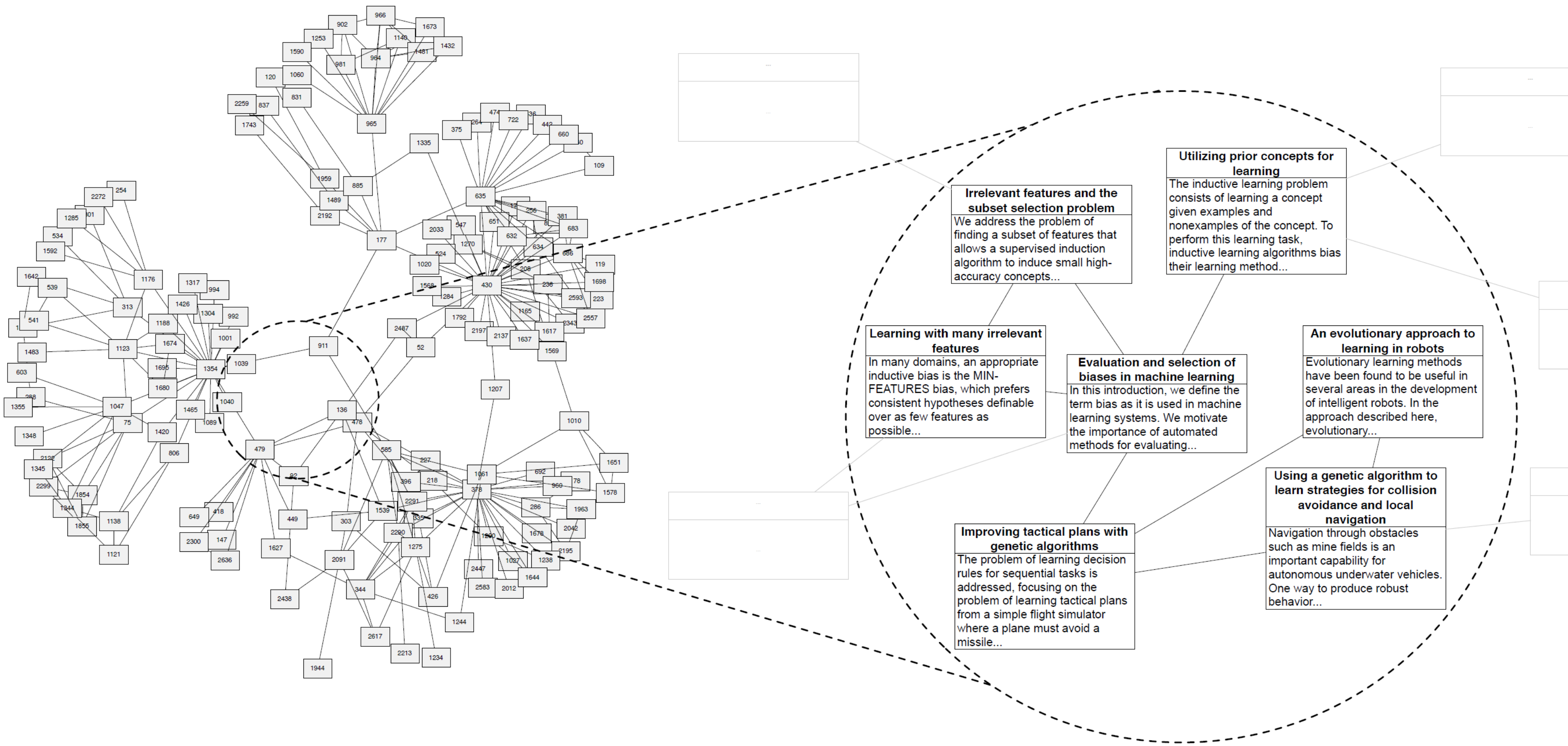
COMPANY
QUARTER
PERCENT
ANALYST
SHARE
SALES
EARNING



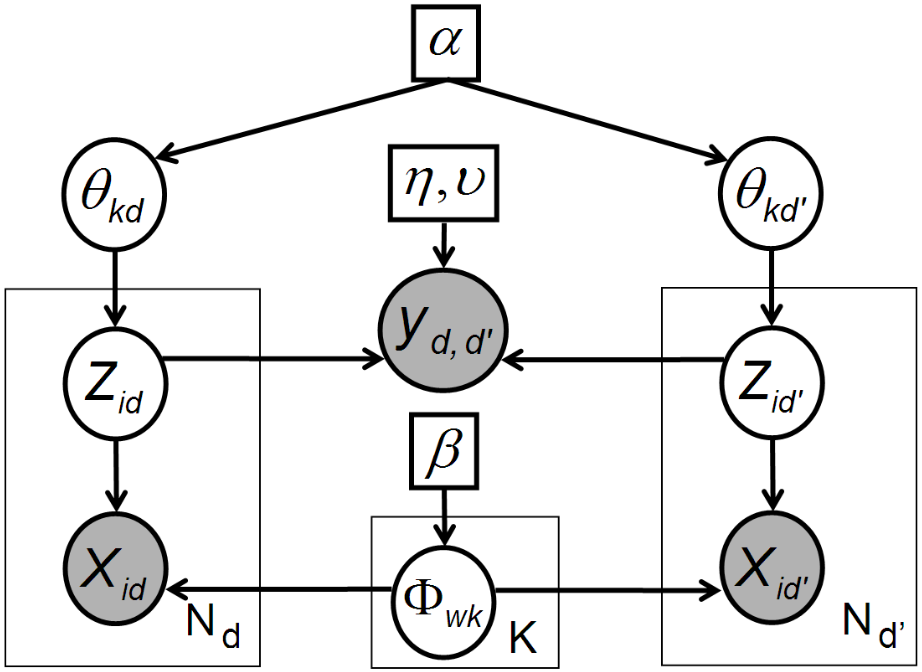
ANTHRAX
LETTER
MAIL
WORKER
OFFICE
SPORES
POSTAL
BUILDING

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 <u>very costly</u> 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

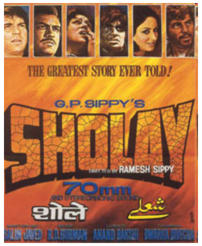
ALGORITHM	MEAN LINK RANK OF PREDICTIONS
Random Guessing	5000
LDA + Regression	2321
Ignoring Non-Edges	1955
Fast Approximation	2089
Subsampling 5% + Caching	1739

Examples of Movie Data Topics

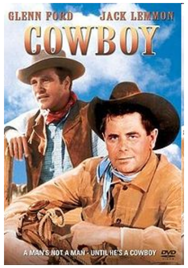
POLICE:	[t2] police agent kill gun action escape car film
DISNEY:	[t4] disney film animated movie christmas cat animation story
AMERICAN:	[t5] president war american political united states government against
CHINESE:	[t6] film kong hong chinese chan wong china link
WESTERN:	[t7] western town texas sheriff 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

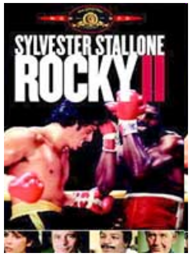
- **'Sholay'**
  - Indian film, 45% of words belong to topic 24 (Hindi topic)
  - Top 5 most probable movie links in training set:
    - 'Laawaris'
    - 'Hote Hote Pyaar Ho Gaya'
    - 'Trishul'
    - 'Mr. Natwarlal'
    - 'Rangeela'



- **'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'



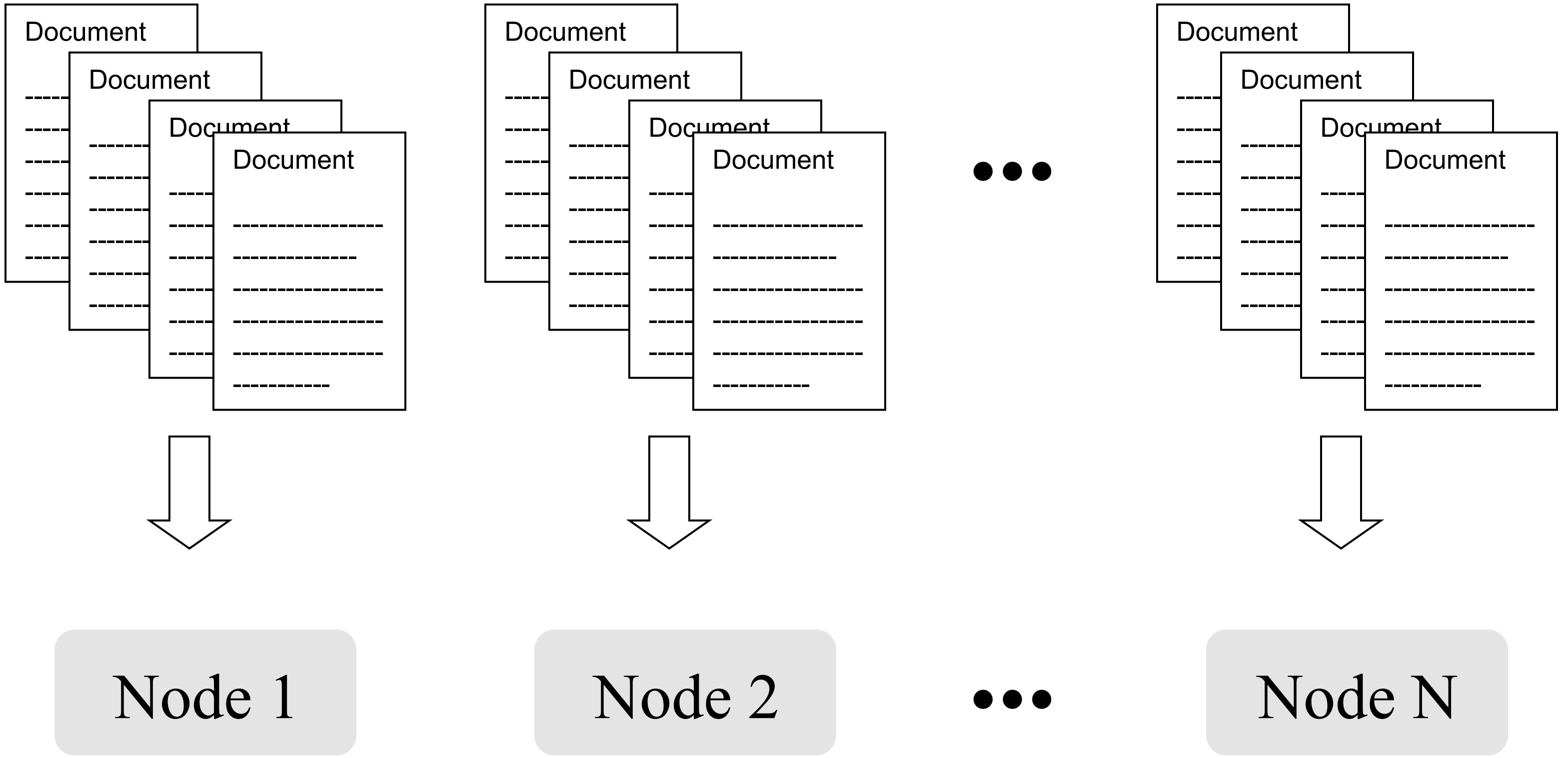
Scalability

- **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

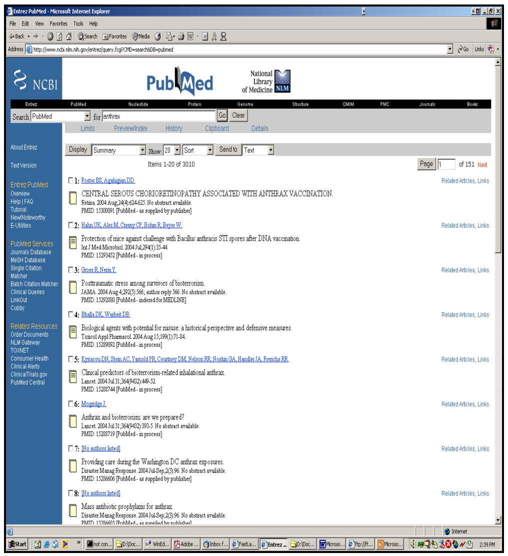
Distributed Topic Modeling

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

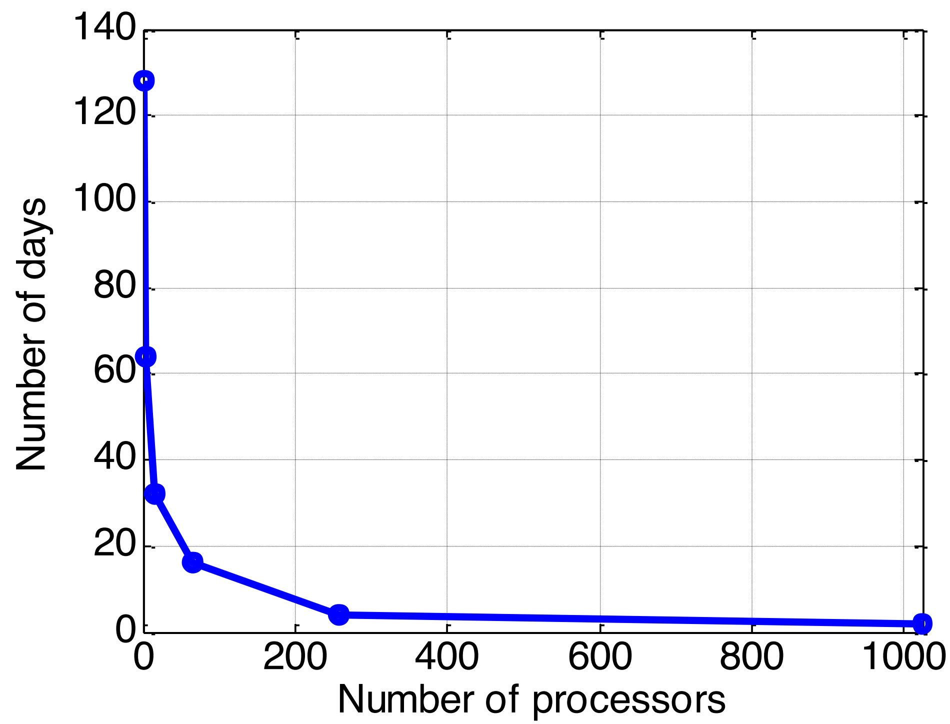


Global synchronization of statistics after each local sampling pass

Large Scale Experiments

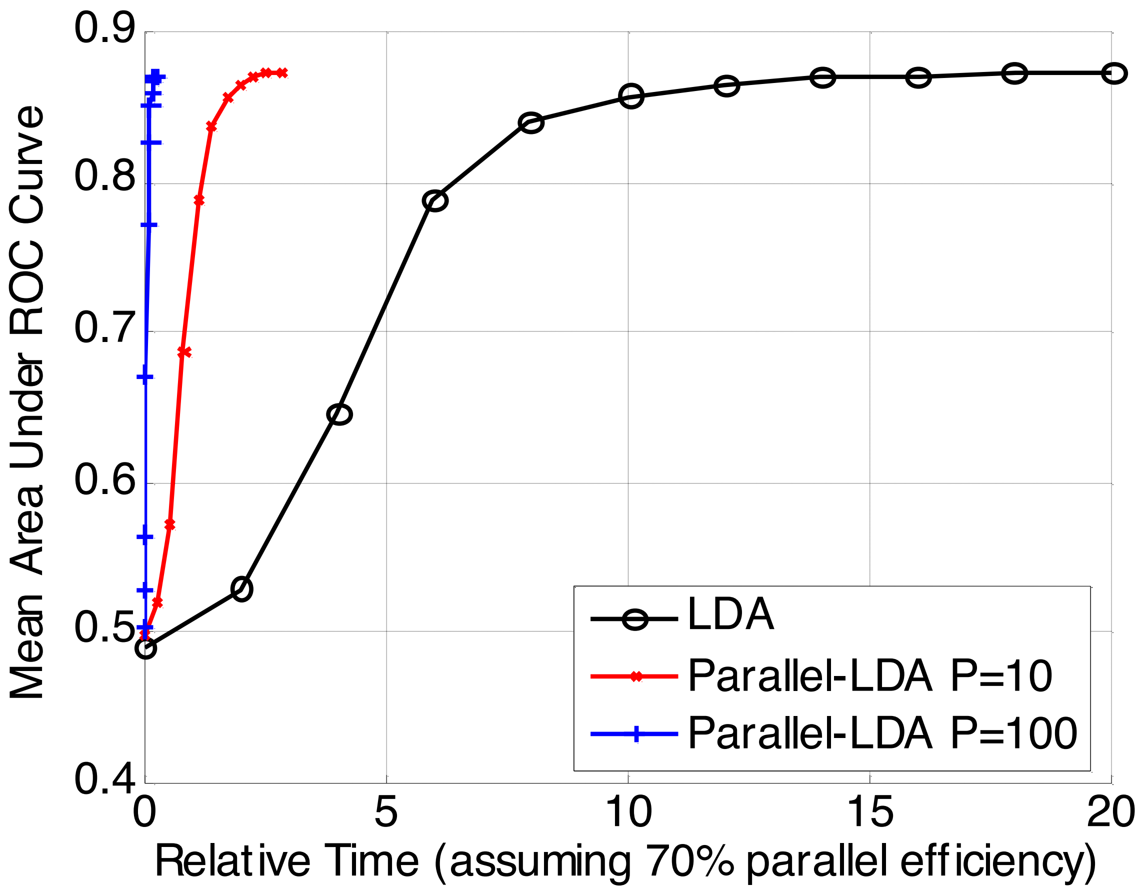


MEDLINE
8 million abstracts
1 billion words
2000 topics



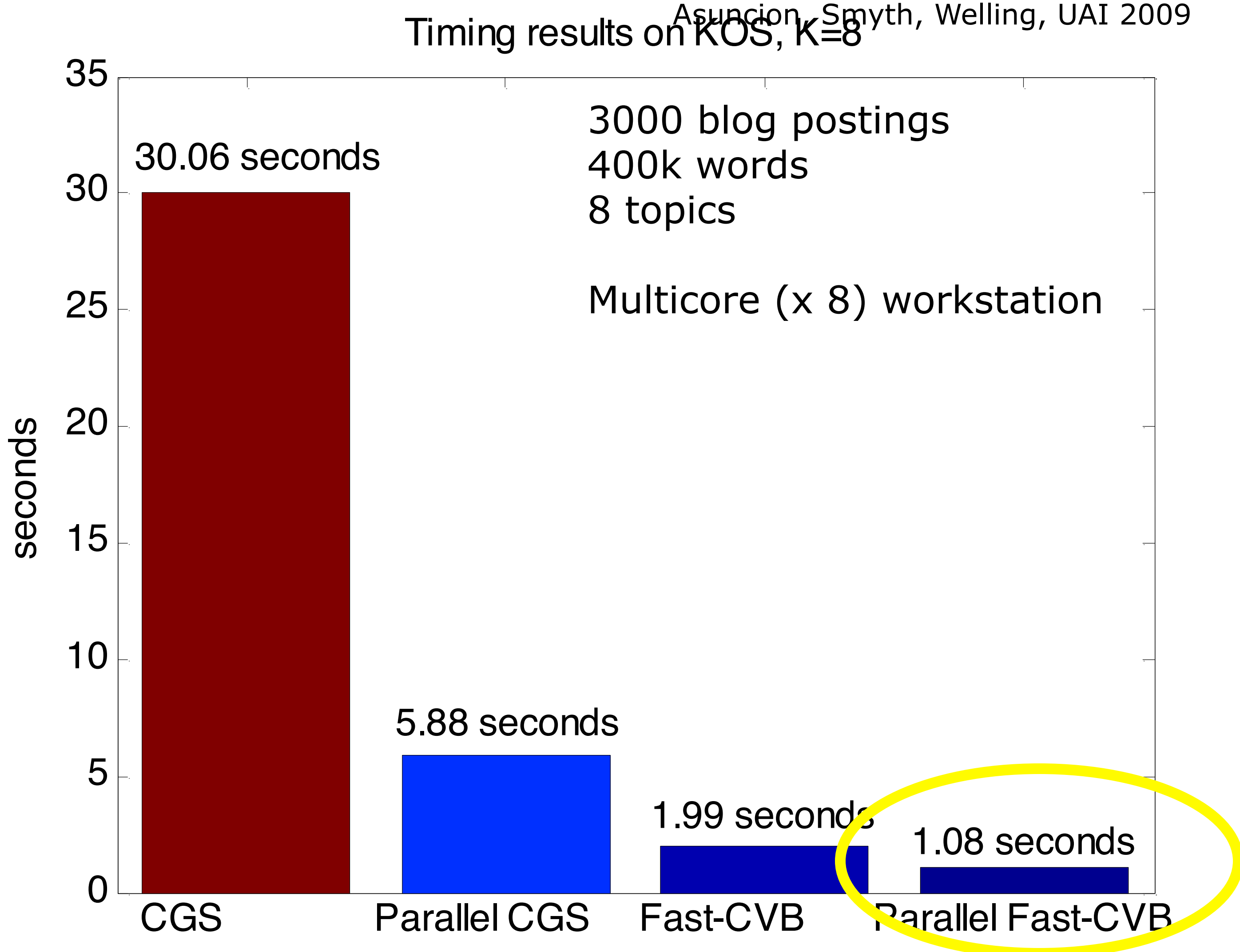
Experiments with 1000 processors at the San Diego Supercomputing Center (SDSC)

Speed up during Learning

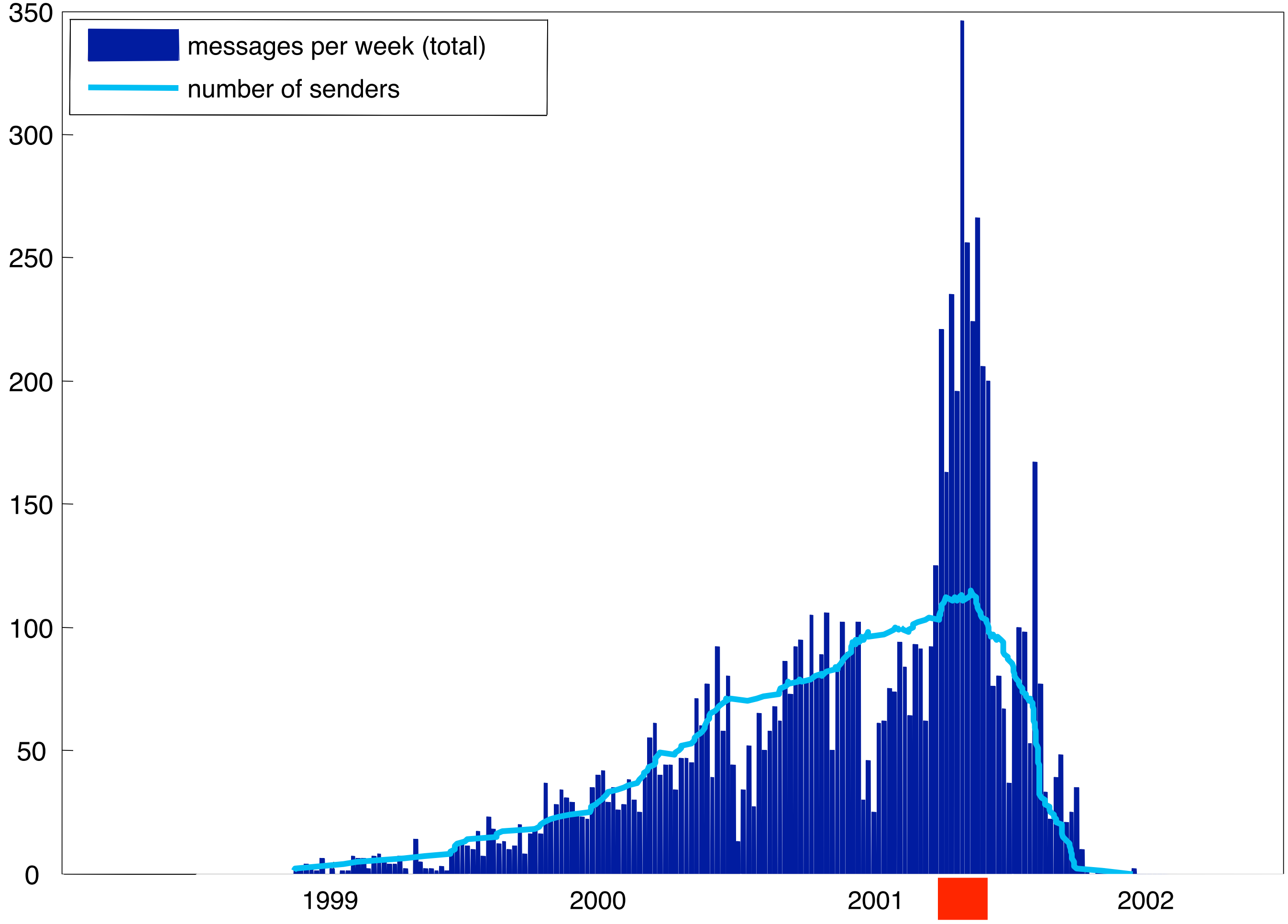


Real-Time Topic Modeling

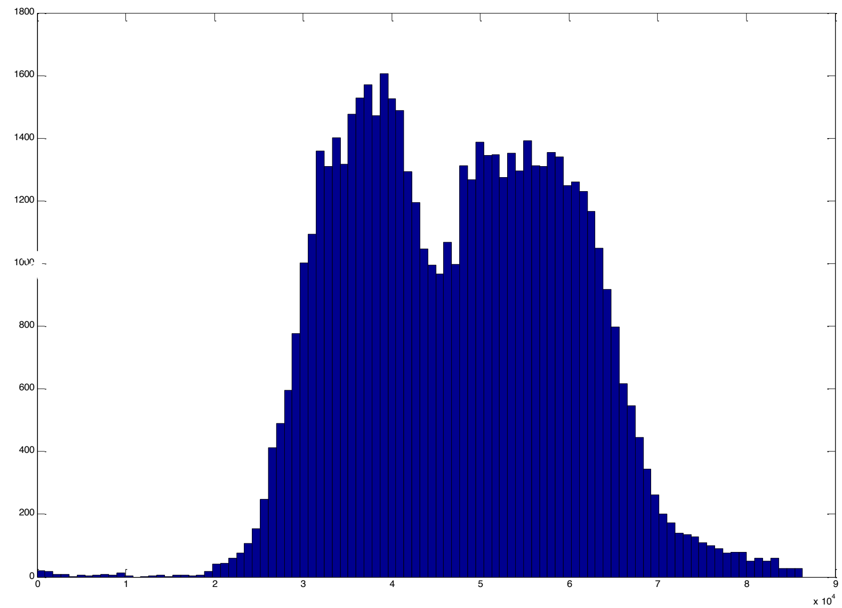
Asuncion, Smyth, Welling, UAI 2009

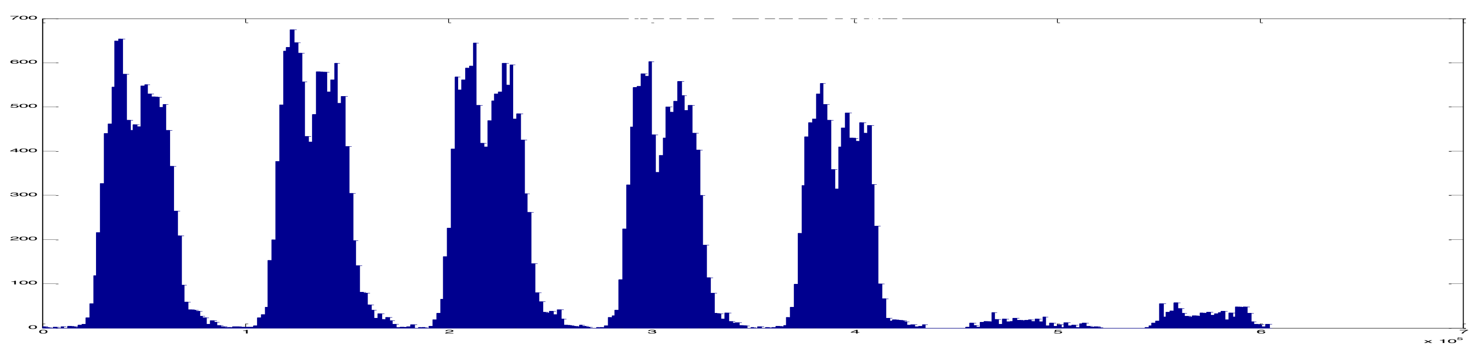


Enron email dataset



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: $P(w \mid z_k)$

$P(w \mid doc)$

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

Modes for Events

Mode: $P(m_k \mid time)$

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

$P(s, r \mid time)$

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

Similarities to Topic Model

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

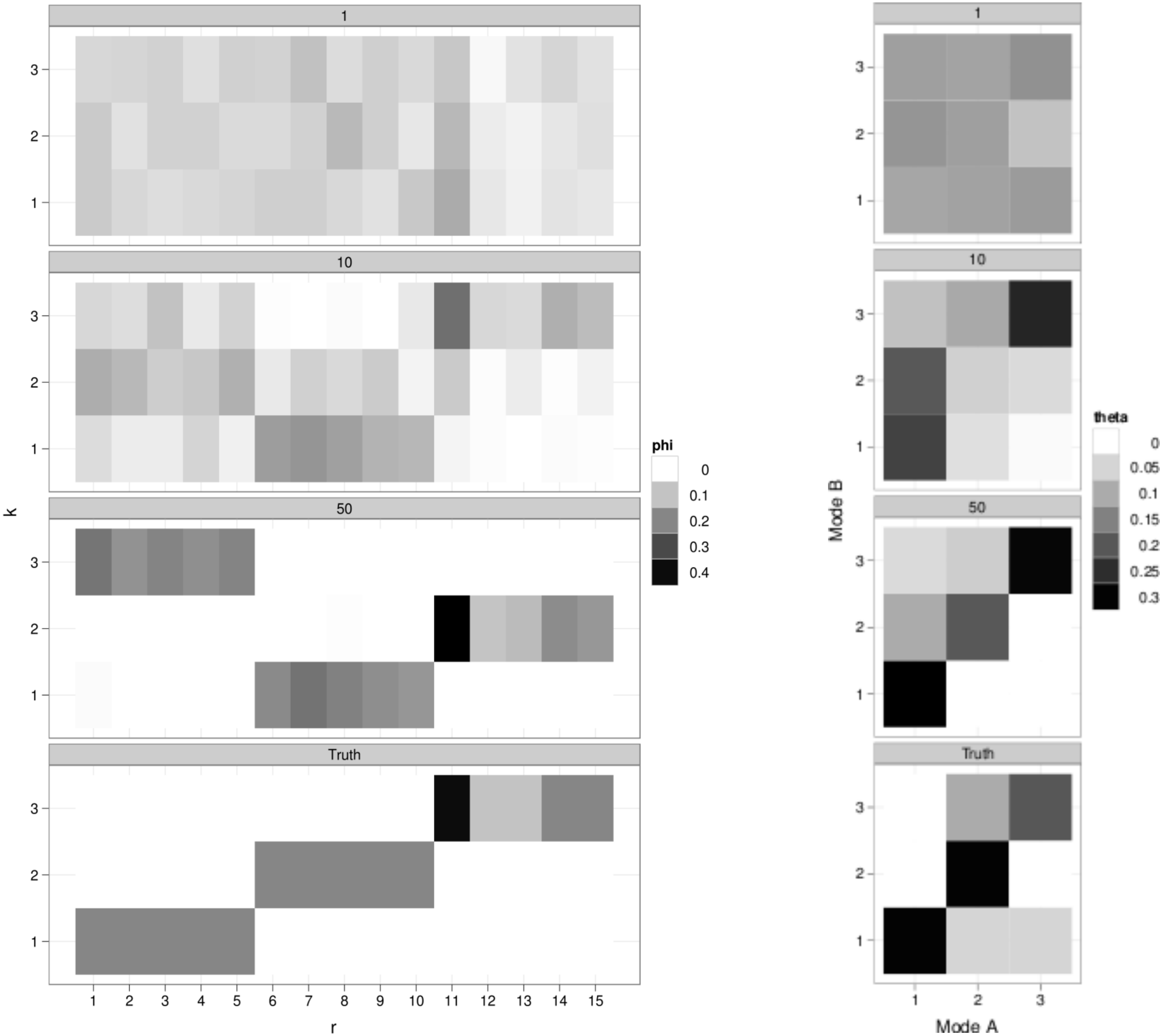
Similarities to Topic Model

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

Similarities to Topic Model

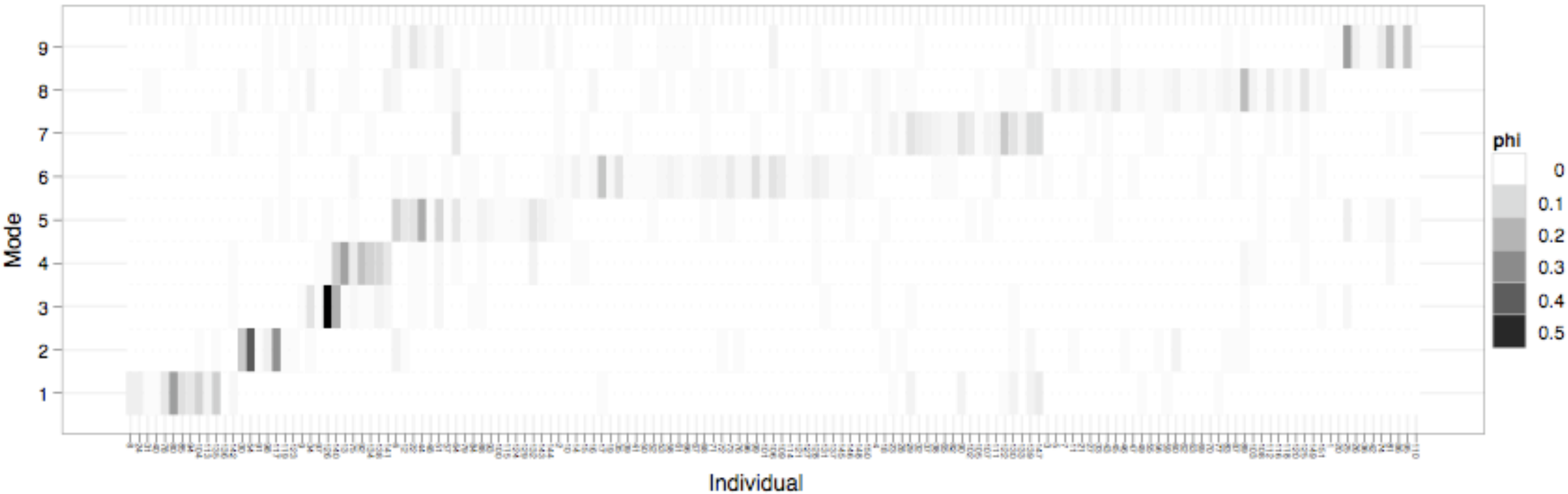
Topics for Text	**Modes for Events**
Topic: $P(z_k \mid doc)$	Mode: $P(m_k \mid time)$
Word: $P(w \mid z_k)$	Event: $P(s, r \mid m_k)$
$P(w \mid doc) = \sum P(w \mid z_k) P(z_k \mid doc)$	$P(s, r \mid time) = \sum P(s, r \mid m_k) P(m_k \mid time)$

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

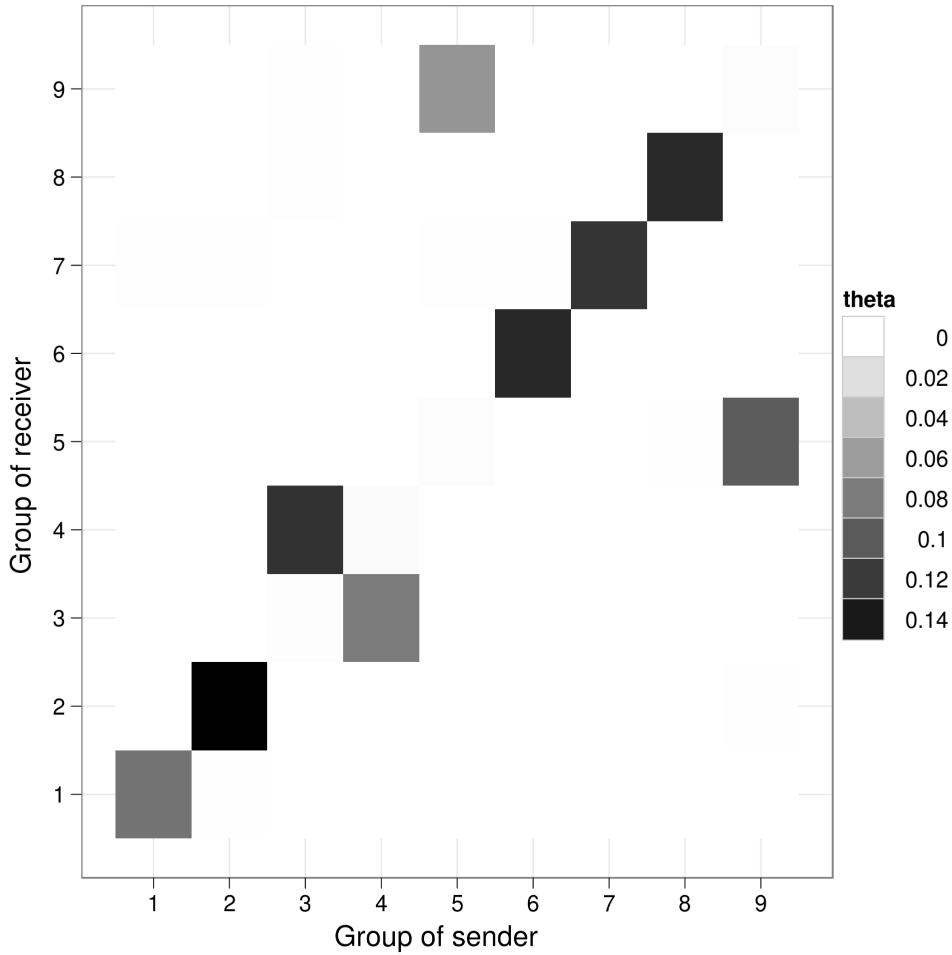


1
10
50
Truth
k
r
phi
0
0.1
0.2
0.3
0.4
Mode B
Mode A
theta
0
0.05
0.1
0.15
0.2
0.25
0.3

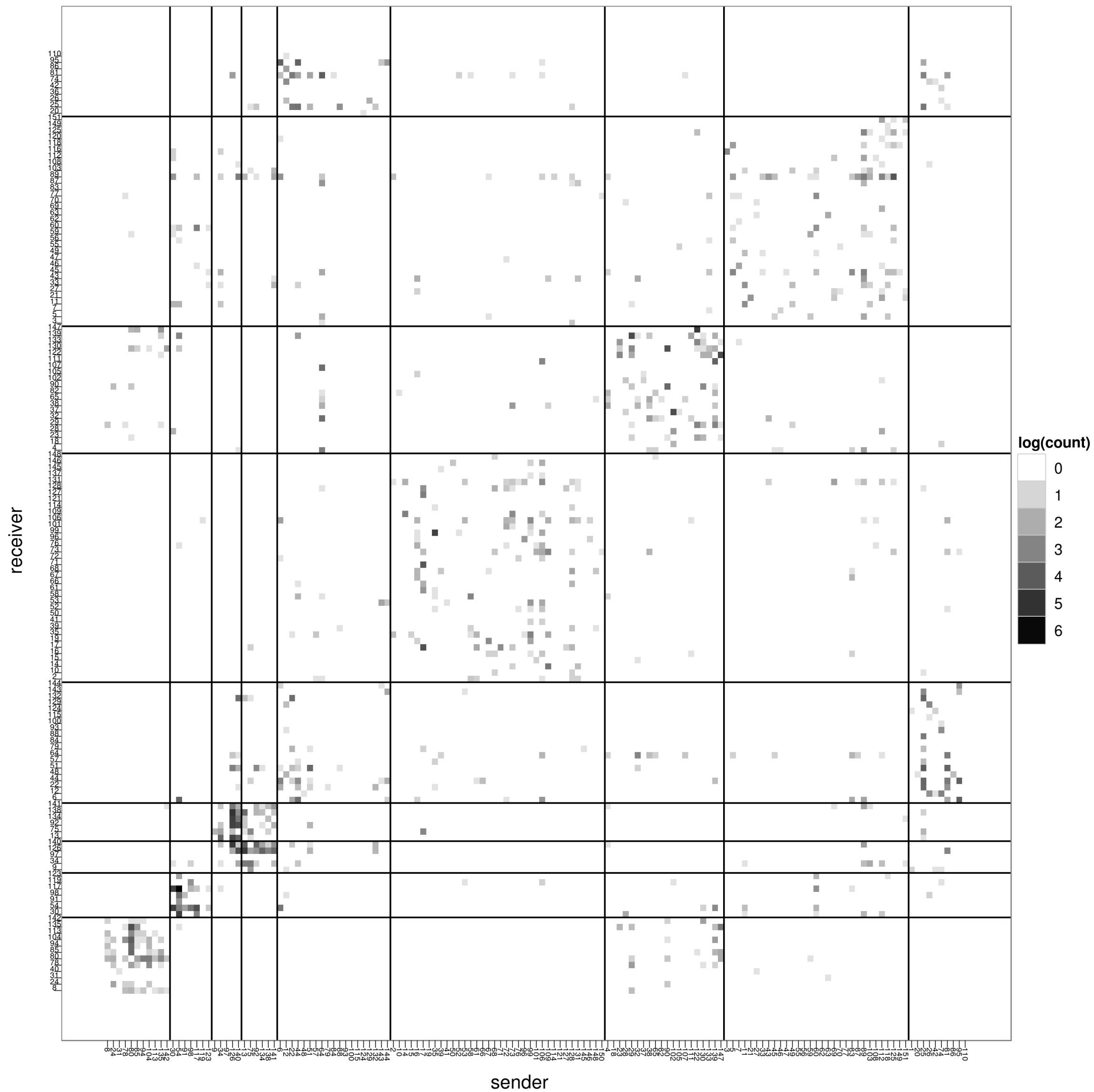
Enron: Mode Probabilities for Senders and Receivers



Enron: Joint Sender-Receiver Mode Probabilities



Number of emails sent between individuals, grouped by modes.



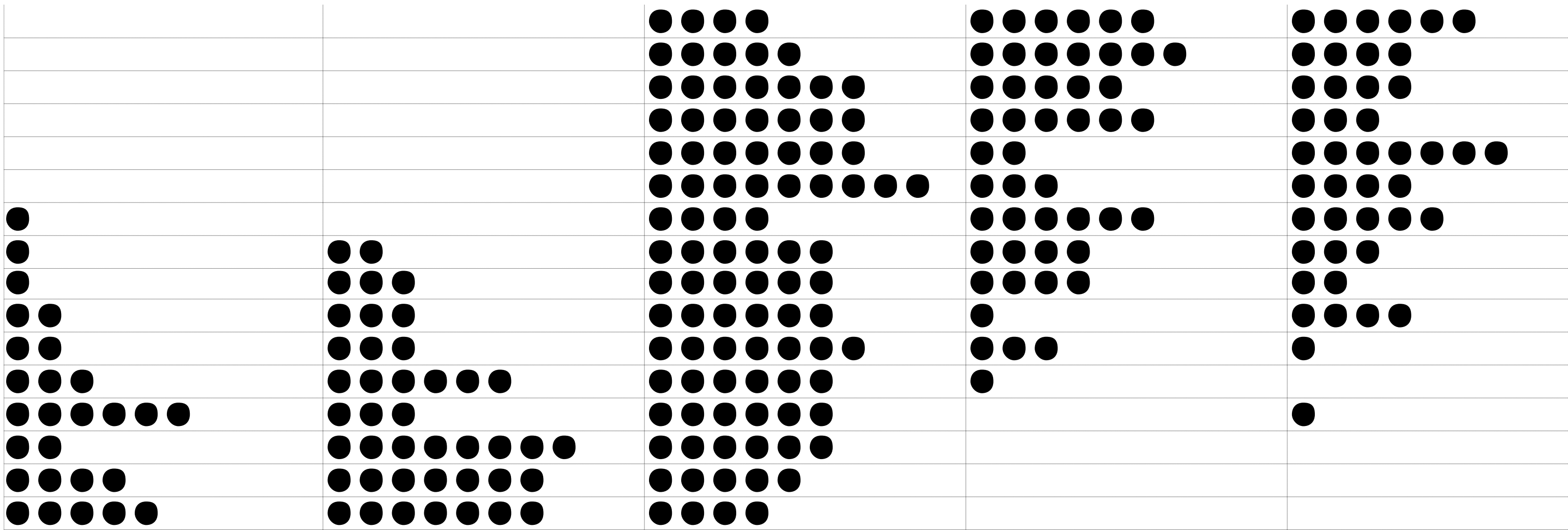
Ongoing and Future Work

- **Add Markov dependence to the modes**
  - P(m_k | 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

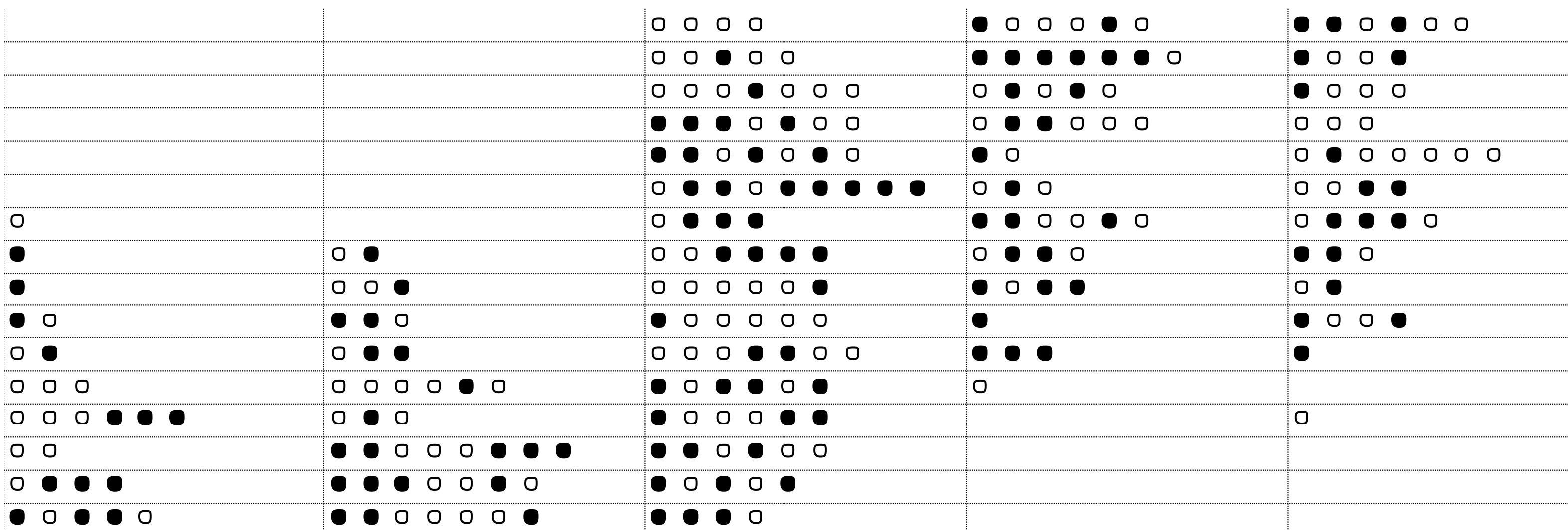
- **Integrate events with text**
  - Joint models over events and text associated with events

Word/Document counts for 16 Artificial Documents

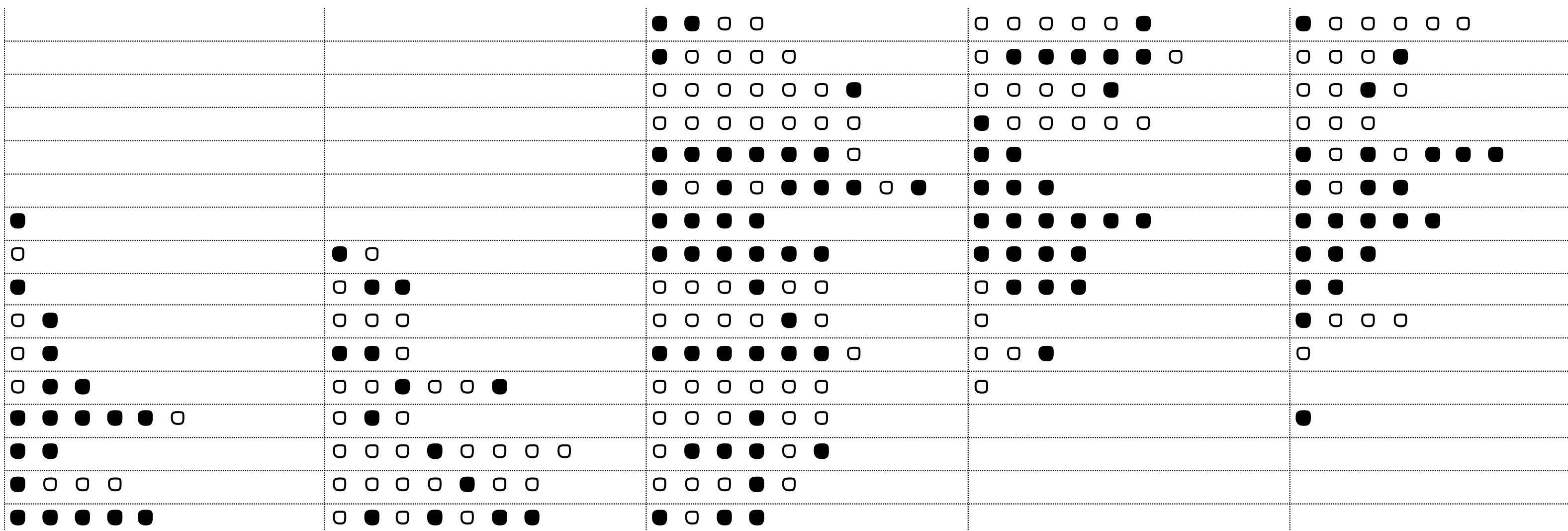


Example of Gibbs Sampling

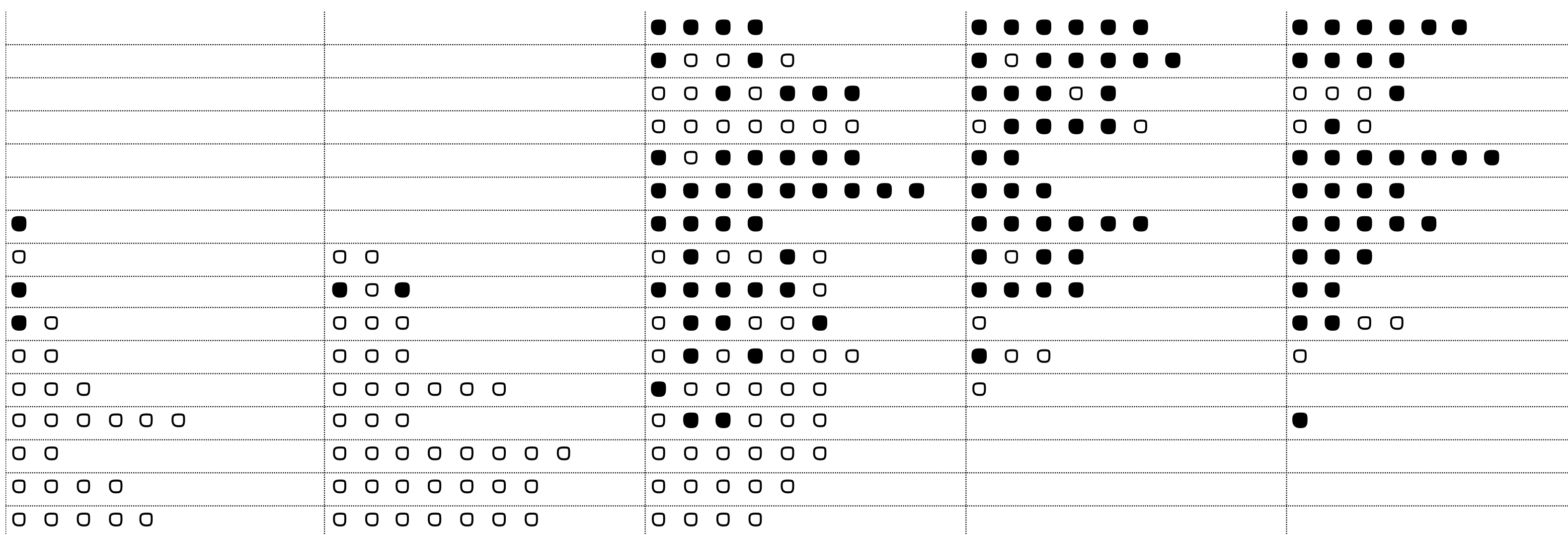
- **Assign word tokens randomly to 2 topics:**



After 1 iteration



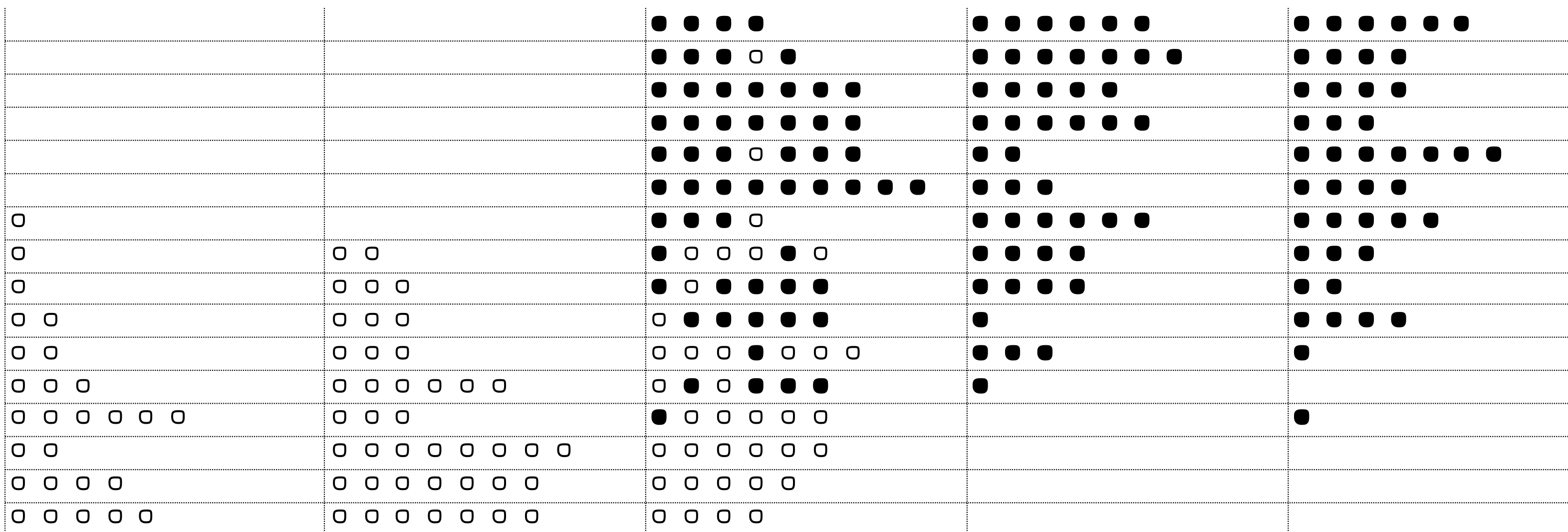
After 4 iterations



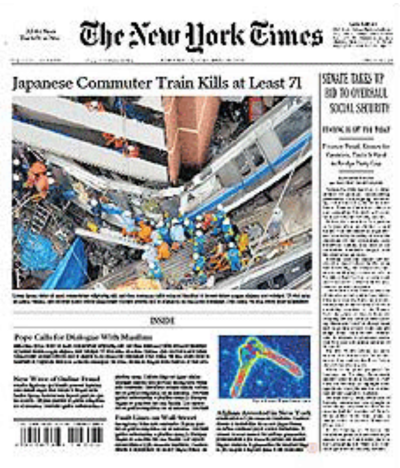
After 32 iterations

topic 1	
stream	.40
bank	.35
river	.25

topic 2	
bank	.39
money	.32
loan	.29

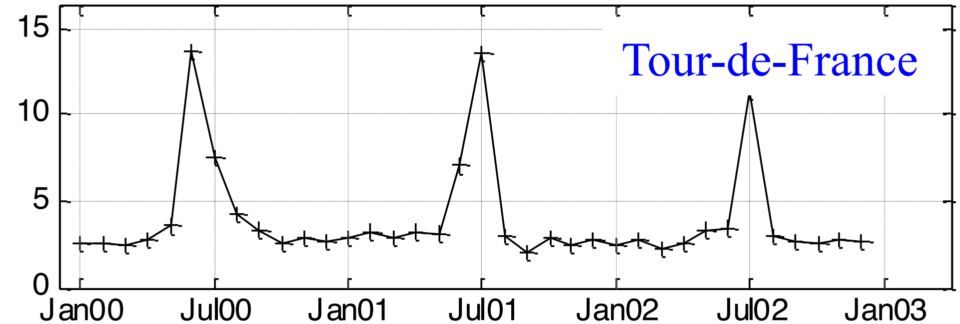


Topic trends from New York Times

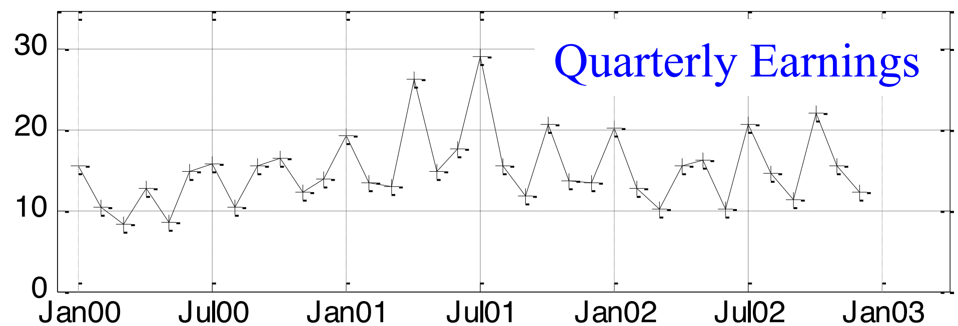

The New York Times

Japanese Commuter Train Kills at Least 71

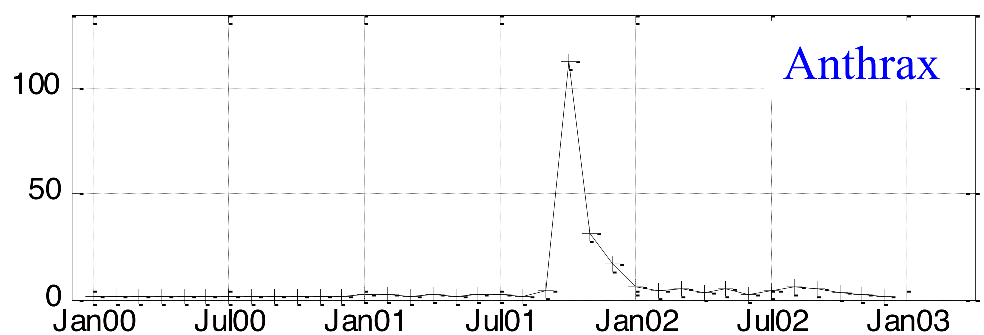
330,000 articles
2000-2002



TOUR
RIDER
LANCE_ARMSTRONG
TEAM
BIKE
RACE
FRANCE



COMPANY
QUARTER
PERCENT
ANALYST
SHARE
SALES
EARNING



ANTHRAX
LETTER
MAIL
WORKER
OFFICE
SPORES
POSTAL
BUILDING

Binary Feature Relational Model

Miller, Griffiths, Jordan, NIPS 2009

- **Based on idea of Indian Buffet Process**
  - Represent each actor by a set of latent binary features
  - Non-parametric: infinite number of features

    ….but in practice, given data, only a finite number are inferred

- **Motivation:**
  - Classes defined over combinatorial number of binary features

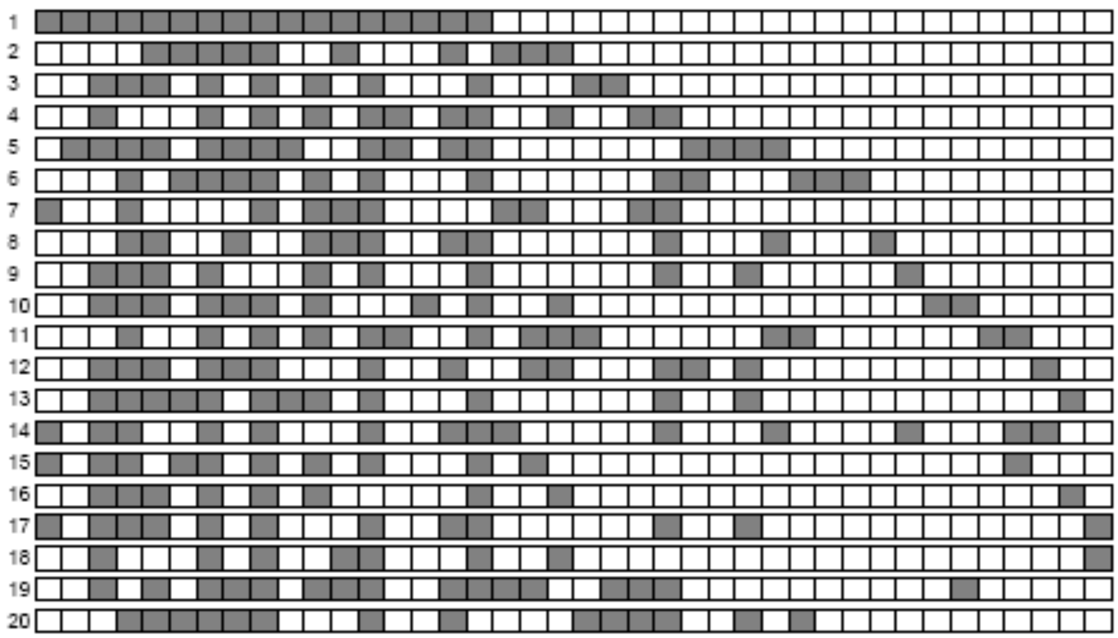
- **Edge process**
  - p(edge i, j) is a function of i and j's latent binary features
  - Learn binary features that explain well the observed edges

Relational Binary Feature Model for Networks

Miller, Griffiths, Jordan, NIPS 2009

Hidden Features

Actors



Presence of edge between actor i and actor j is (e.g.)
a logistic function of a weighted sum of features they have in common

Estimation: based on MCMC

Additional Modeling Aspects….

- **As with static networks….**
  - Actor attributes, e.g., actor age
  - Edge (event) attributes, e.g., text of an email
- **Can also have time-dependent covariates/attributes**
  - E.g., actor attributes changing over time
  - Network level "external" covariates
    - Calendar effects: time of day, day of week, time of year
    - External events – exogenous time-series

Mixtures for Relational Events

DuBois and Smyth, ACM SIGKDD 2010

- **Mixture model over events**
  - First choose event class k, k = 1, …. K

    $k \sim \pi$

    y_{ij} : $i \sim \phi$ (sender nodes | k), $j \sim \phi$ (receiver nodes | k)

  - Parameters

    π : K x 1 multinomial = relative likelihood of different event classes

    ϕ (sender nodes | k), ϕ (receiver nodes | k)
    - 2K multinomials, each of size N

- **Simple (but effective) model**
  - Quite similar in spirit to LDA/topic model for documents

Simple Baseline

- **We could predict the likelihood of i and j communicating based directly on i and j's history**
  - Multinomial with $O(N^2)$ entries
  - Can use smoothing to combat sparsity

- **Problems**
  - Data can be extremely sparse for large N – smoothing is non-informative, and does not "borrow strength" from the graph

- **Nonetheless this is a useful baseline when evaluating predictions**
  - Historically, few papers evaluate models predictively
  - Even fewer compare their models to simple baselines

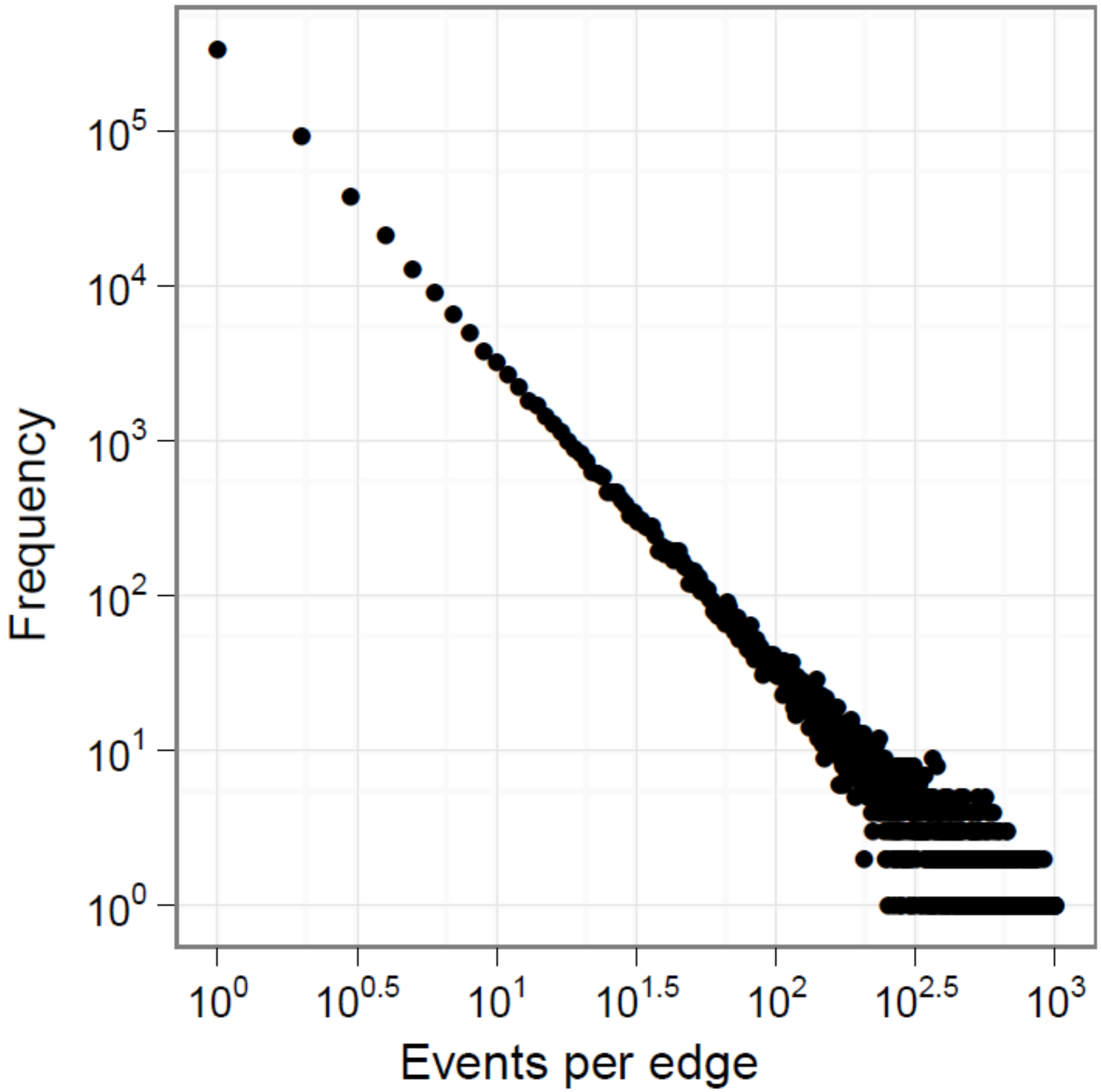
Some Comments on Evaluation

- **Prediction on independent test data is critical**
  - Relatively easy to do with dynamic networks
  - Tricky to do with static networks (but see Hoff, 2009)

- **Caveat**
  - For link (or link probability) prediction it can be very difficult to beat relatively simple baselines, e.g.,
    - Graph(t+1) = Graph(t)
    - p(event) = smoothed estimate based on historical frequency of that pair

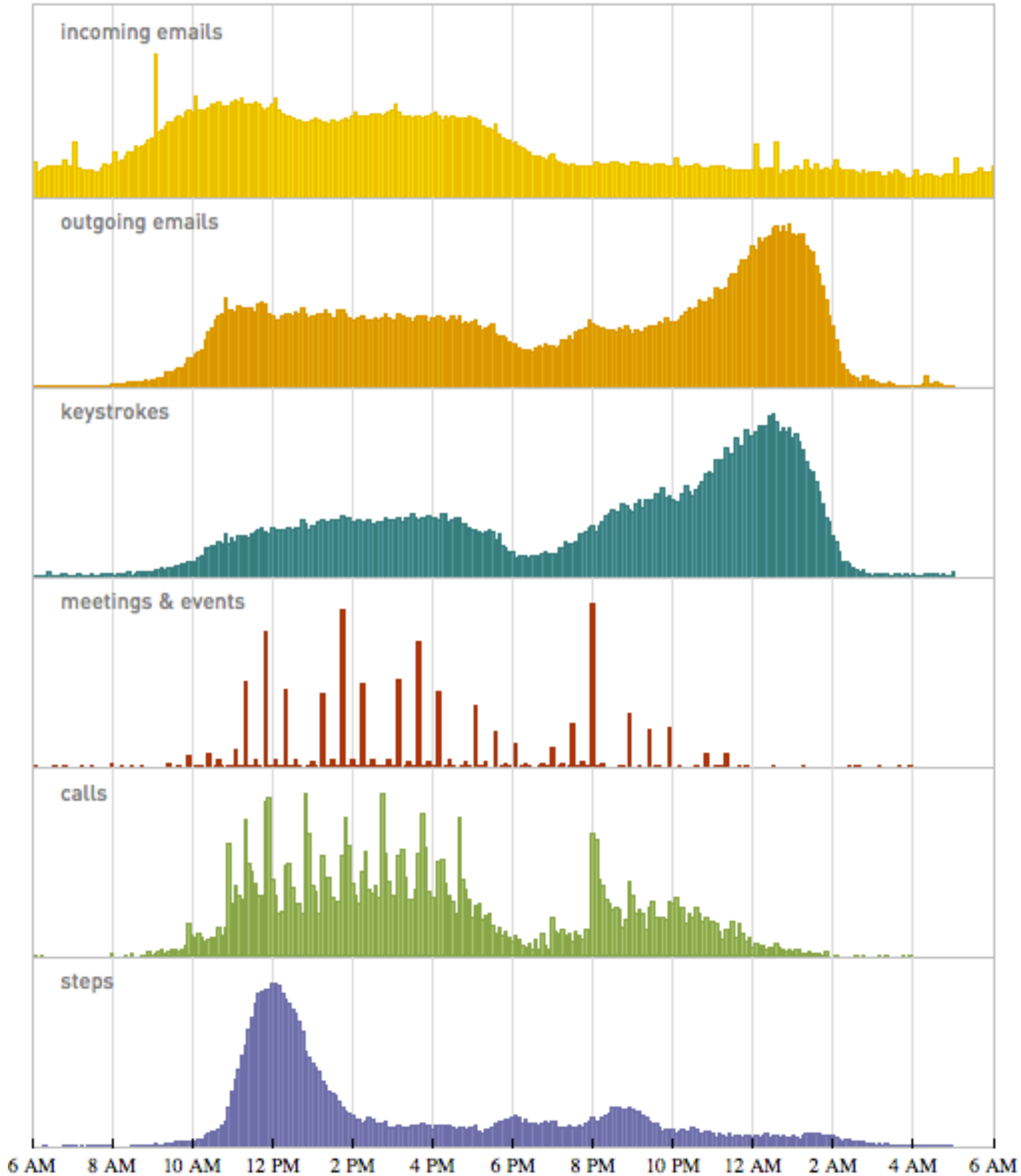
- **Solution?**
  - More interesting questions than just predicting what happens next, e.g
    - How likely is that group A will communicate with group B in the next k days?
    - If we have events with missing information, can we infer sender/receiver?
    - Can we detect significant shifts/non-stationarity?

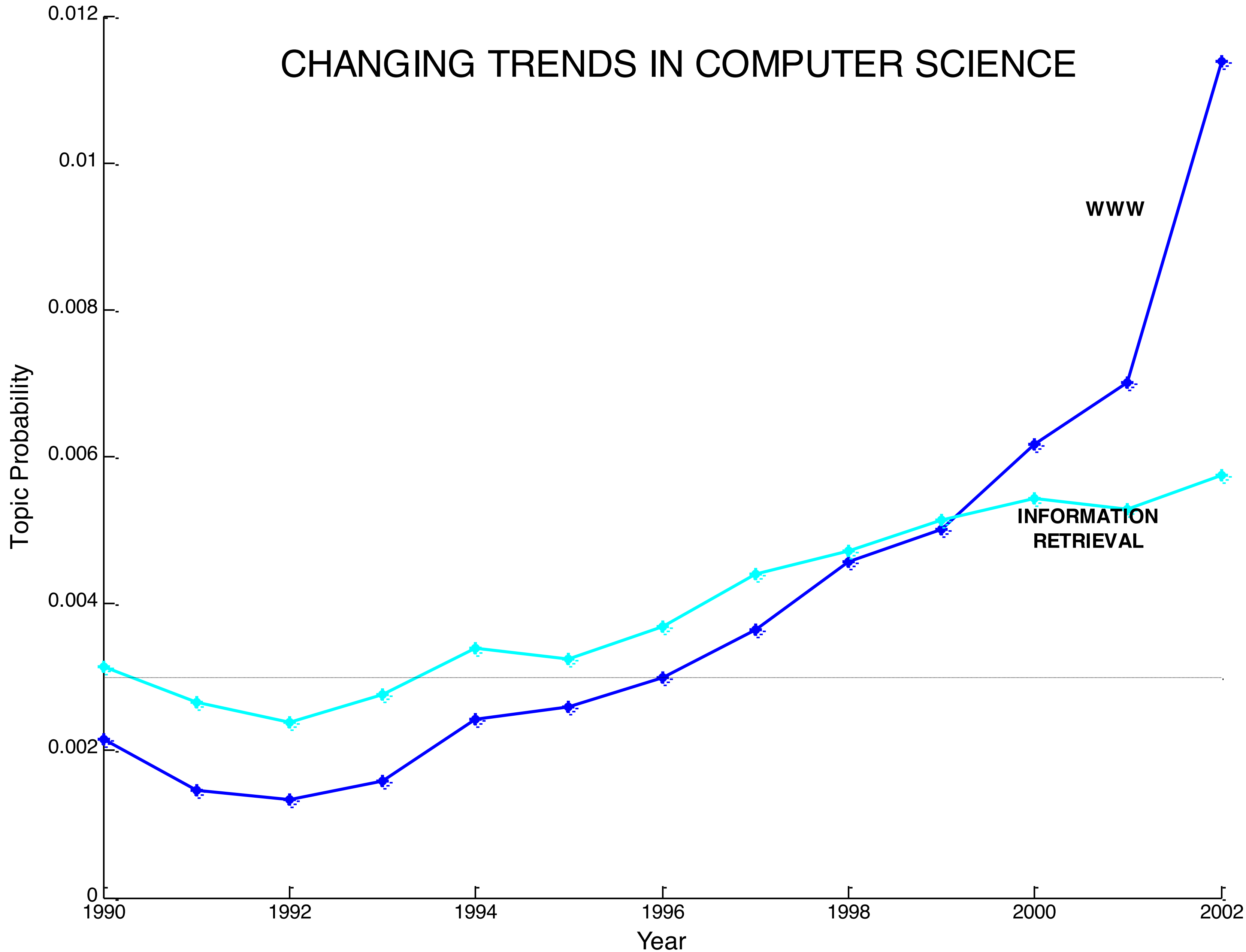
Illustration of Sparsity: Frequency of Events per pair of Actors



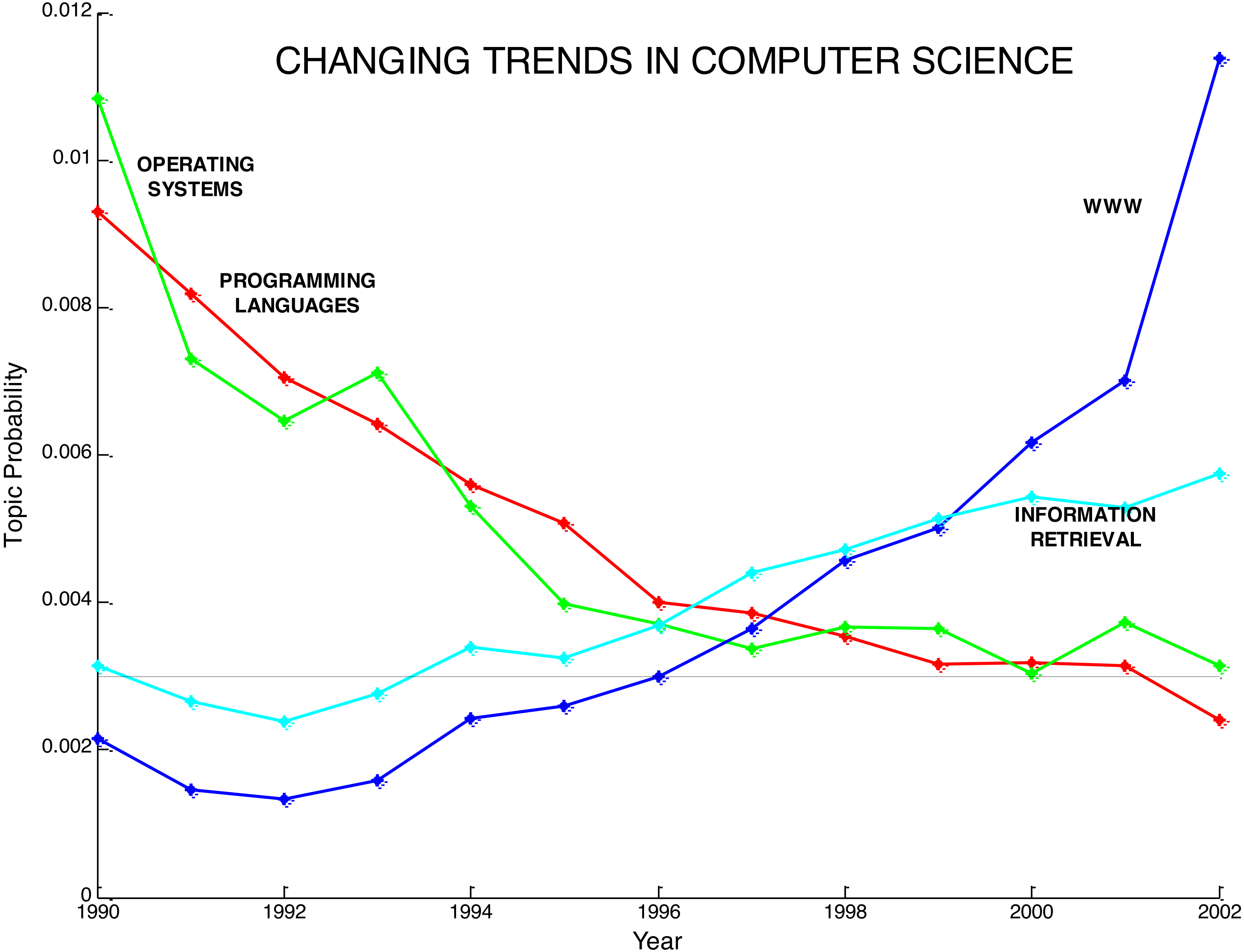
International Political Events data
King, 2003

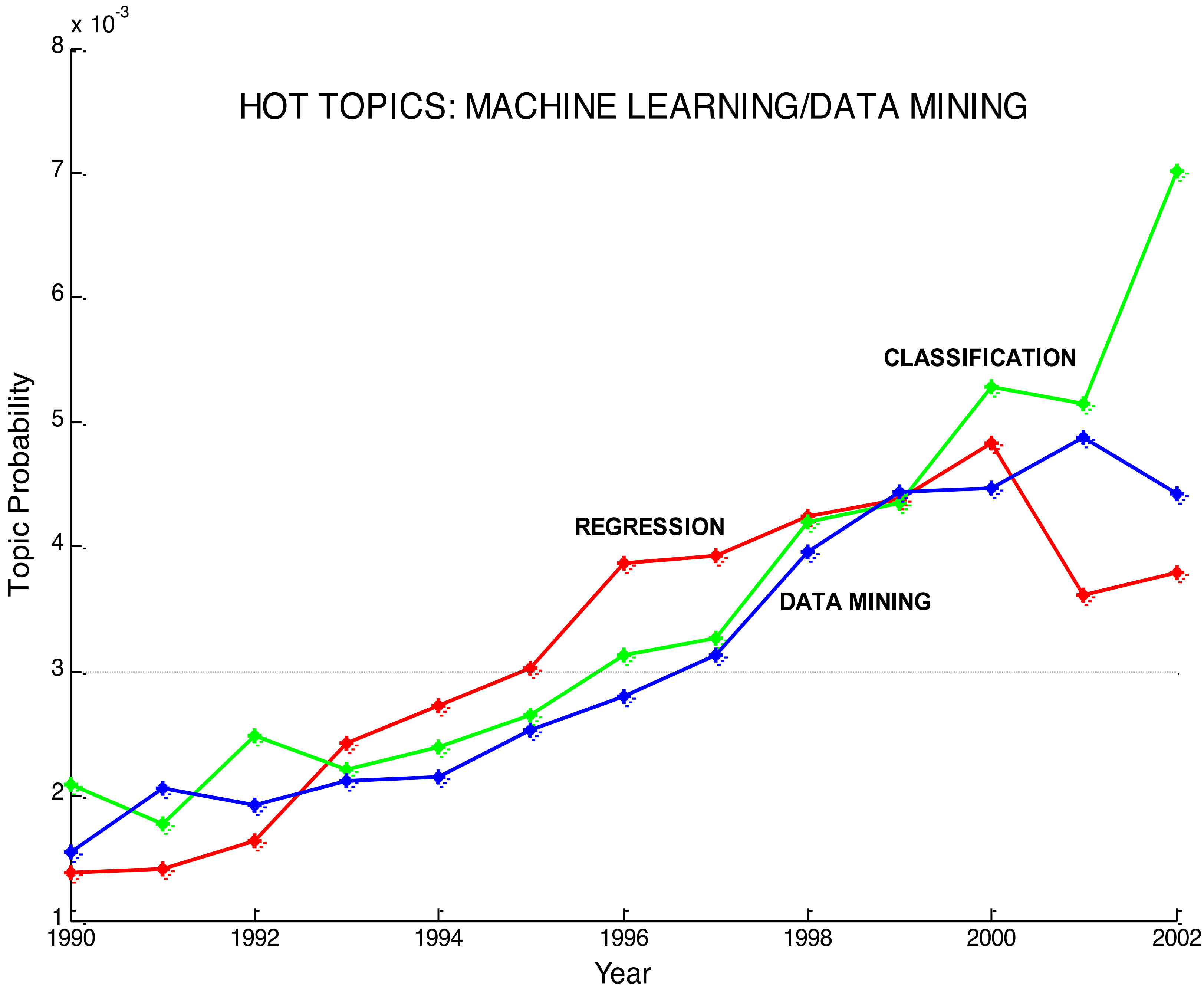
blog.stephenwolfram.com
The Personal Analytics of My Life
March 8th 2012



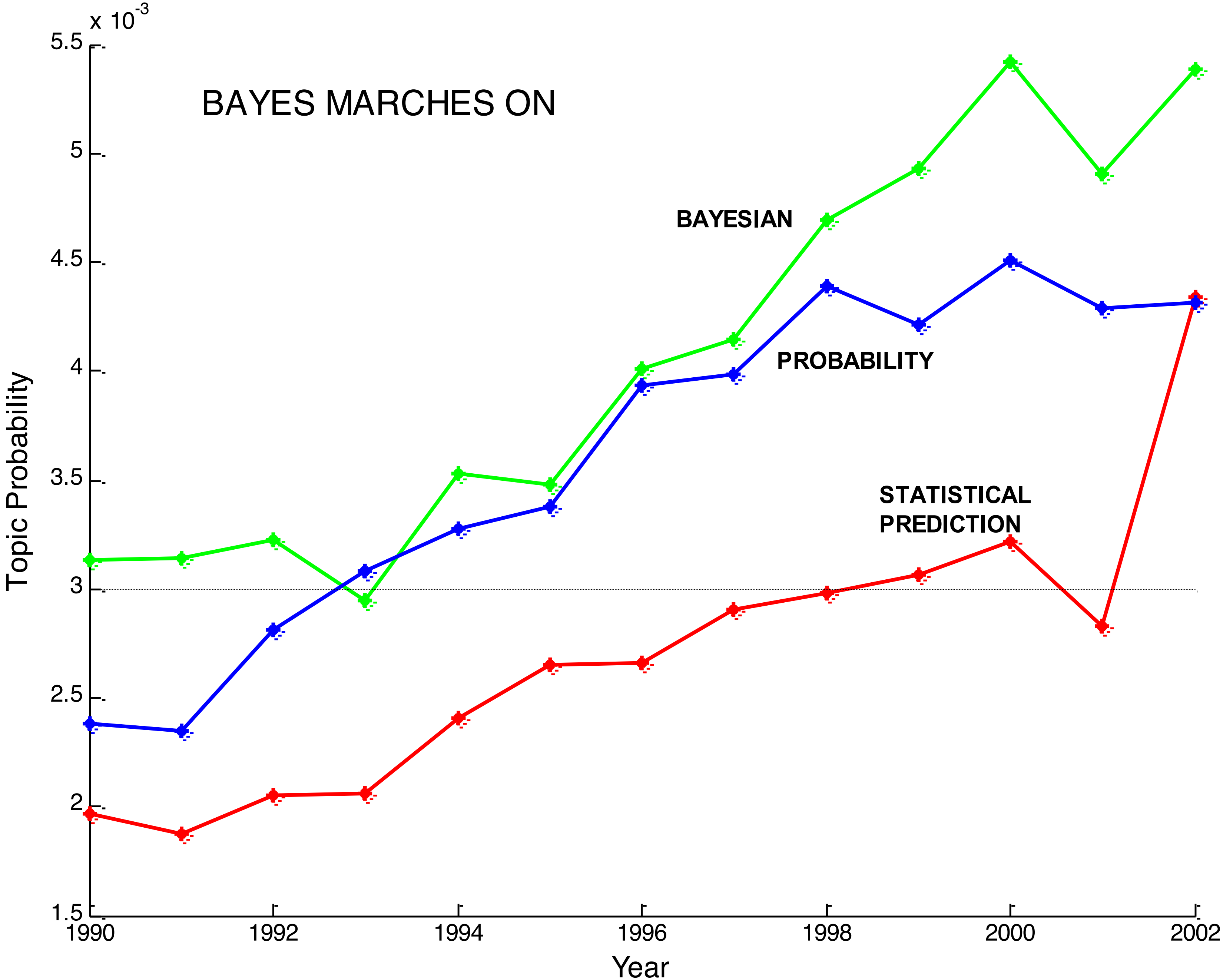


CHANGING TRENDS IN COMPUTER SCIENCE
WWW
INFORMATION RETRIEVAL
Topic Probability
0.012
0.01
0.008
0.006
0.004
0.002
0
1990
1992
1994
1996
1998
2000
2002
Year


CHANGING TRENDS IN COMPUTER SCIENCE
OPERATING SYSTEMS
PROGRAMMING LANGUAGES
WWW
INFORMATION RETRIEVAL
Topic Probability
Year
0
0.002
0.004
0.006
0.008
0.01
0.012
1990
1992
1994
1996
1998
2000
2002

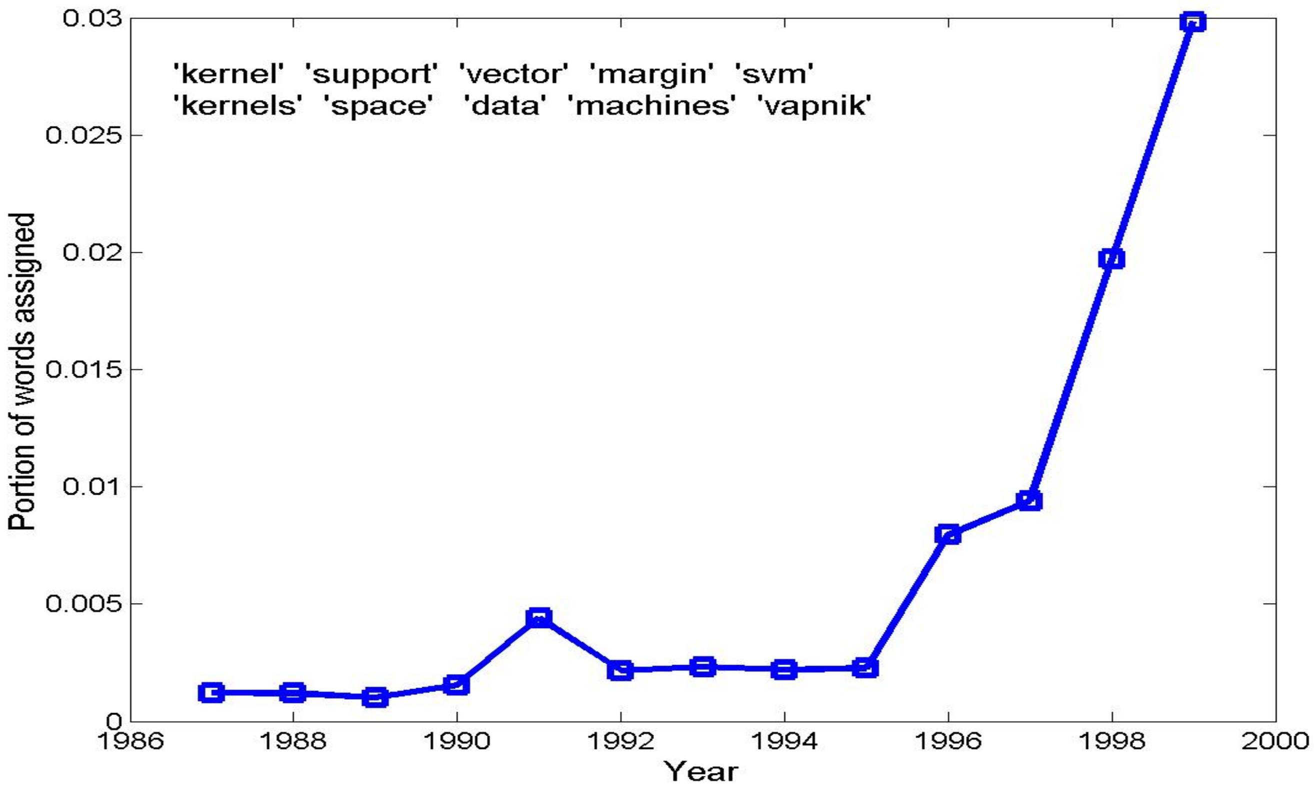


x 10^-3
HOT TOPICS: MACHINE LEARNING/DATA MINING
CLASSIFICATION
REGRESSION
DATA MINING
Topic Probability
Year
1
2
3
4
5
6
7
8
1990
1992
1994
1996
1998
2000
2002



x 10^-3
5.5
5
4.5
4
3.5
3
2.5
2
1.5
Topic Probability
BAYES MARCHES ON
BAYESIAN
PROBABILITY
STATISTICAL
PREDICTION
1990
1992
1994
1996
1998
2000
2002
Year

NIPS: support vector topic



NIPS: neural network topic