Hybrid Models for Text and Graphs

10/23/2012
Analysis of Social Media

Newswire Text

- Formal
- Primary purpose:
  - Inform "typical reader" about recent events
- Broad audience:
  - Explicitly establish shared context with reader
  - Ambiguity often avoided

Social Media Text

- Informal
- Many purposes:
  - Entertain, connect, persuade…
- Narrow audience:
  - Friends and colleagues
  - Shared context already established
  - Many statements are ambiguous out of social context

Newswire Text

- Goals of analysis:
  - Extract information about events from text
  - "Understanding" text requires understanding "typical reader"
    - conventions for communicating with him/her
    - Prior knowledge, background, …

Social Media Text

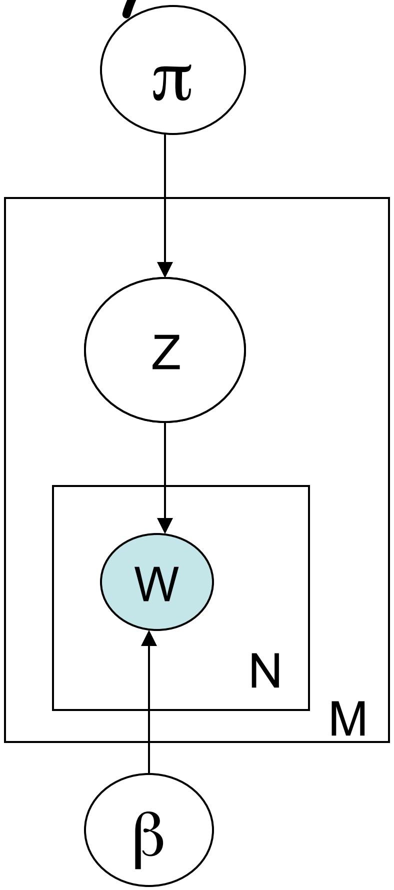
- Goals of analysis:
  - Very diverse
  - Evaluation is difficult
    - And requires revisiting often as goals evolve
  - Often "understanding" social text requires understanding a *community*

Outline

- Tools for analysis of text
  - Probabilistic models for text, communities, and time
    - Mixture models and LDA models for text
    - LDA extensions to model hyperlink structure
    - LDA extensions to model time

Introduction to Topic Models

- Mixture model: unsupervised naïve Bayes model



- Joint probability of words and classes:

$$\prod_{d=1}^{M} P(w_1,\cdots,w_{N_d},z_d|\beta,\pi) = \prod_{d=1}^{M}\left\{\pi_{z_d}\prod_{n=1}^{N_d}\beta_{z_d,w_n}\right\}$$

- But classes are not visible:

$$\prod_{d=1}^{M} P(w_1,\cdots,w_{N_d}|\pi,\beta) \quad = \quad \prod_{d=1}^{N_d}\left\{\sum_{k=1}^{K}\left(\pi_k\prod_{n=1}^{N_d}\beta_{k,w_n}\right)\right\}$$

Introduction to Topic Models

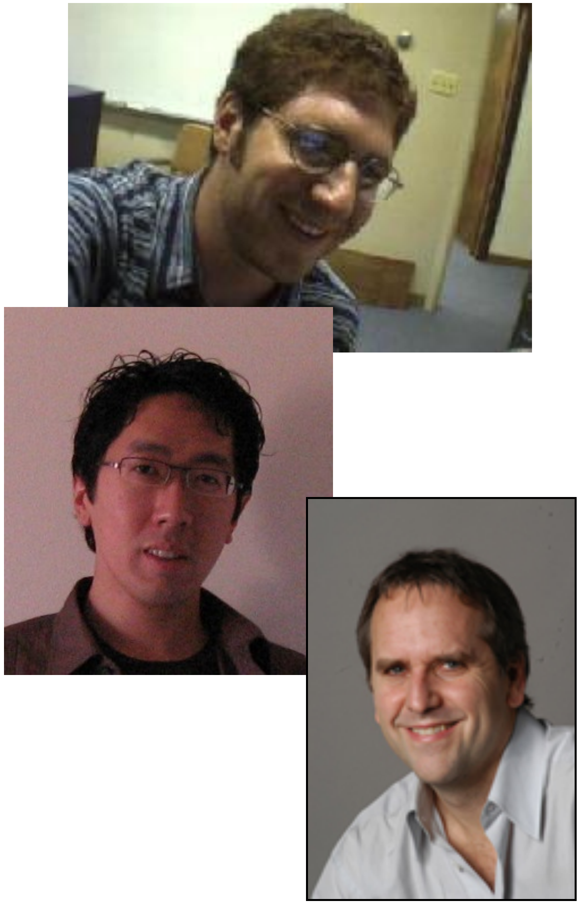
Latent Dirichlet Allocation

JMLR, 2003

David M. Blei
Computer Science Division
University of California
Berkeley, CA 94720, USA

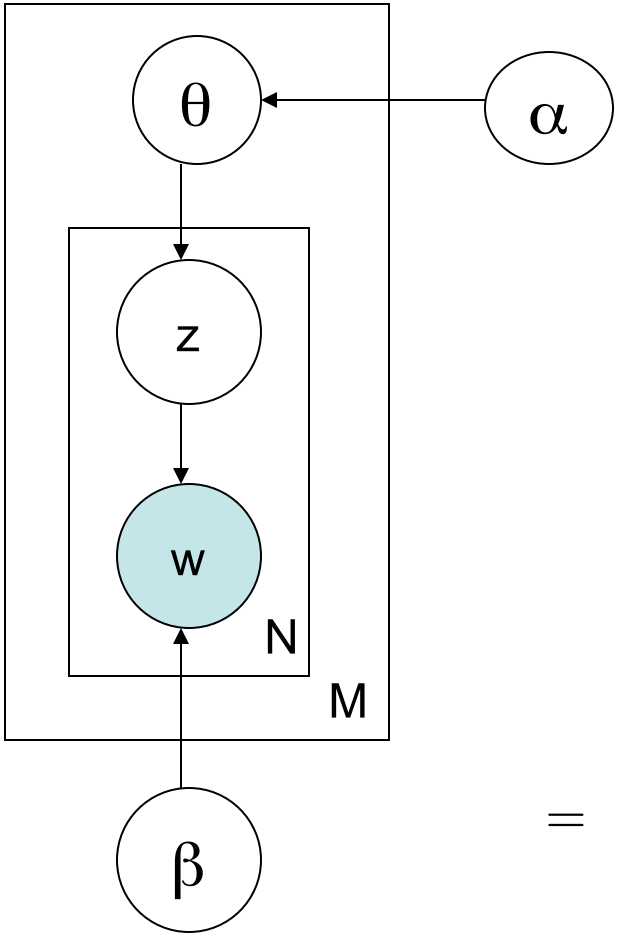
Andrew Y. Ng
Computer Science Department
Stanford University
Stanford, CA 94305, USA

Michael I. Jordan
Computer Science Division and Department of Statistics
University of California
Berkeley, CA 94720, USA



Introduction to Topic Models

- Latent Dirichlet Allocation



- For each document d = 1,···,M
  - Generate θ_d ~ Dir(.| α)
  - For each position n = 1,···, N_d
    - generate z_n ~ Mult(. | θ_d)
    - generate w_n ~ Mult(.| β_{z_n})

$$\prod_{d=1}^{N_d} P(w_1, \cdots, w_{N_d} | \beta, \alpha)$$

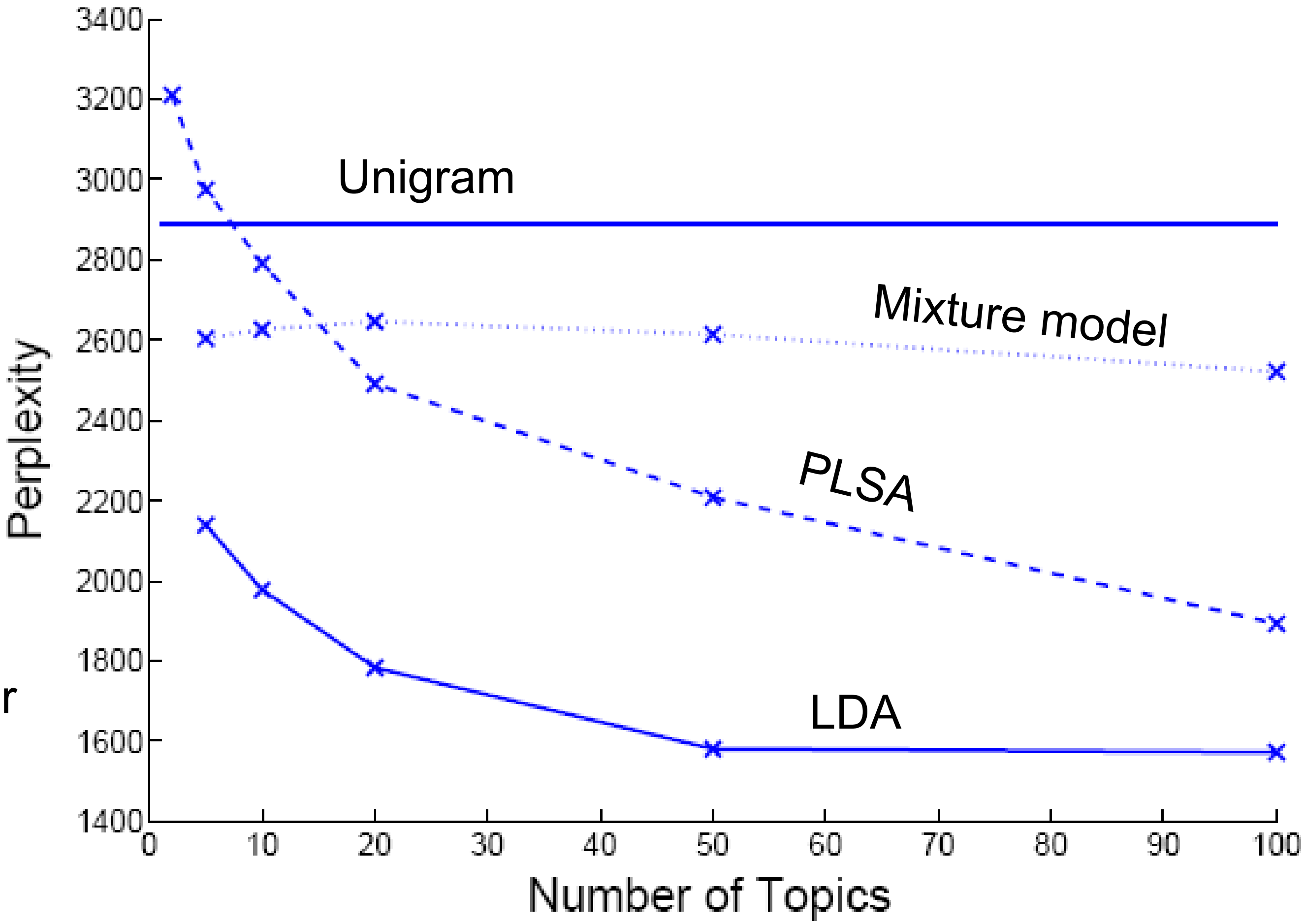
$$= \prod_{d=1}^{N_d} \int_{\theta_d} P(\theta_d|\alpha) \left\{ \prod_{n=1}^{N_d} \left(\sum_k \theta_{dk} \beta_{k w_n} \right) \right\} d\theta_d$$

Introduction to Topic Models

- Latent Dirichlet Allocation
  - Overcomes some technical issues with PLSA
    - PLSA only estimates mixing parameters for training docs
  - Parameter learning is more complicated:
    - Gibbs Sampling: easy to program, often slow
    - Variational EM

Introduction to Topic Models

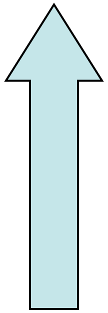
- Perplexity comparison of various models



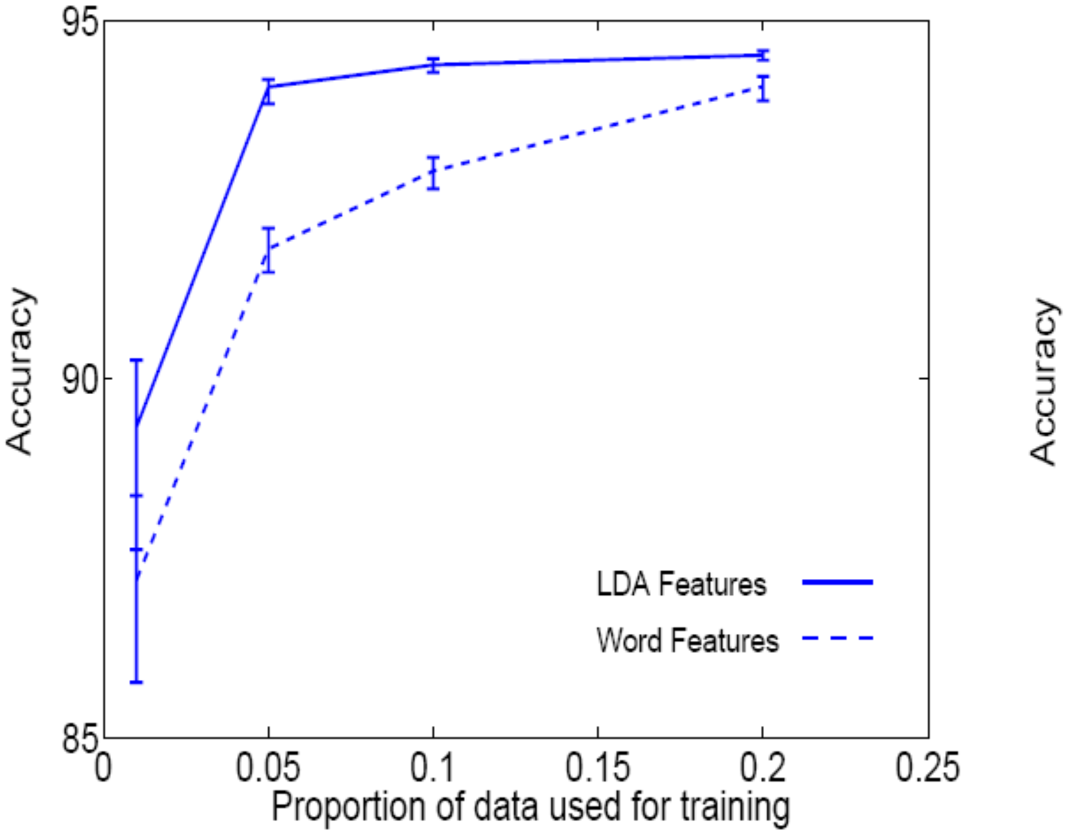
Lower is better

Introduction to Topic Models

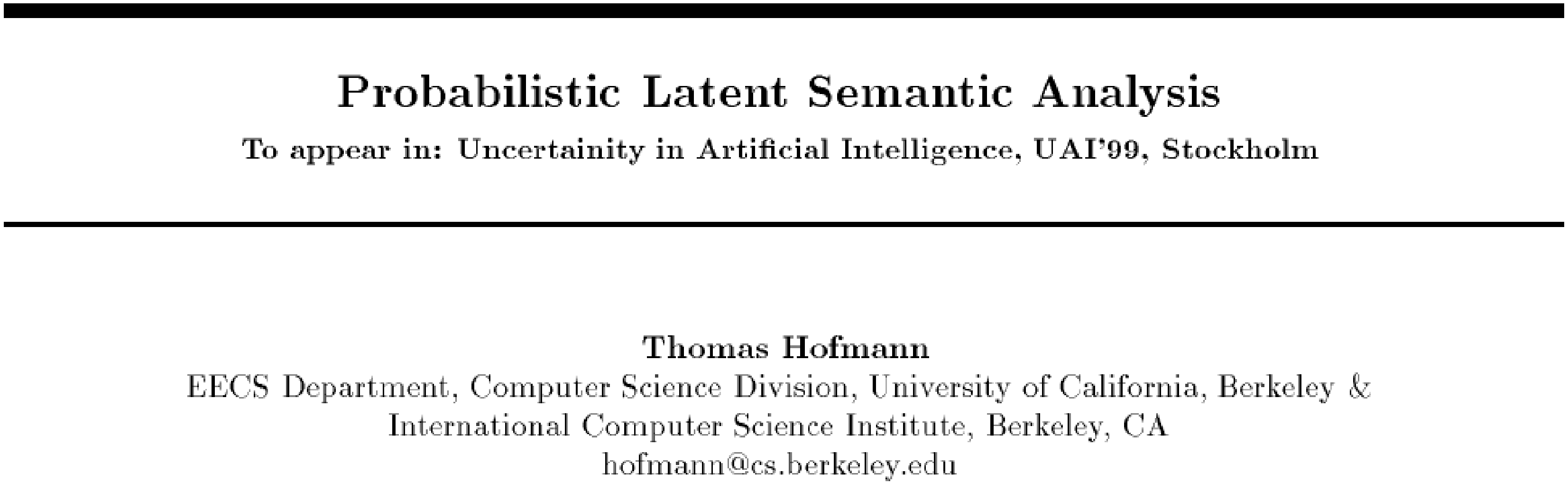
- Prediction accuracy for classification using learning with topic-models as features



Higher is better



Before LDA….LSA and pLSA



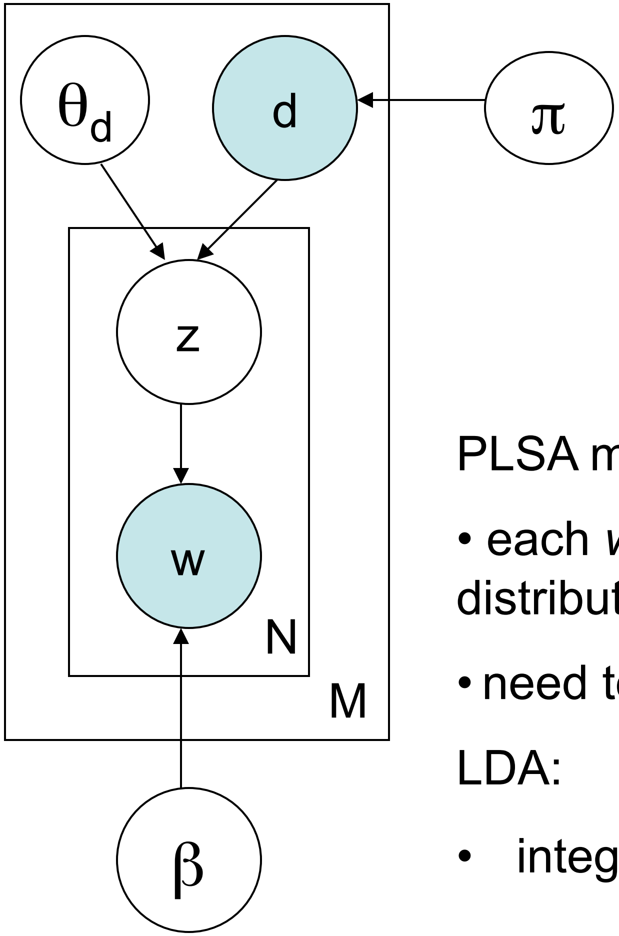
Probabilistic Latent Semantic Analysis

To appear in: Uncertainity in Artificial Intelligence, UAI'99, Stockholm

Thomas Hofmann
EECS Department, Computer Science Division, University of California, Berkeley &
International Computer Science Institute, Berkeley, CA
hofmann@cs.berkeley.edu

Introduction to Topic Models

- Probabilistic Latent Semantic Analysis Model



- Select document d ~ Mult(π)
  - For each *position* n = 1,$\cdots$, N_d
    - generate z_n ~ Mult(_ | θ_d)
    - generate w_n ~ Mult(_ | β_{z_n})

Topic distribution

PLSA model:

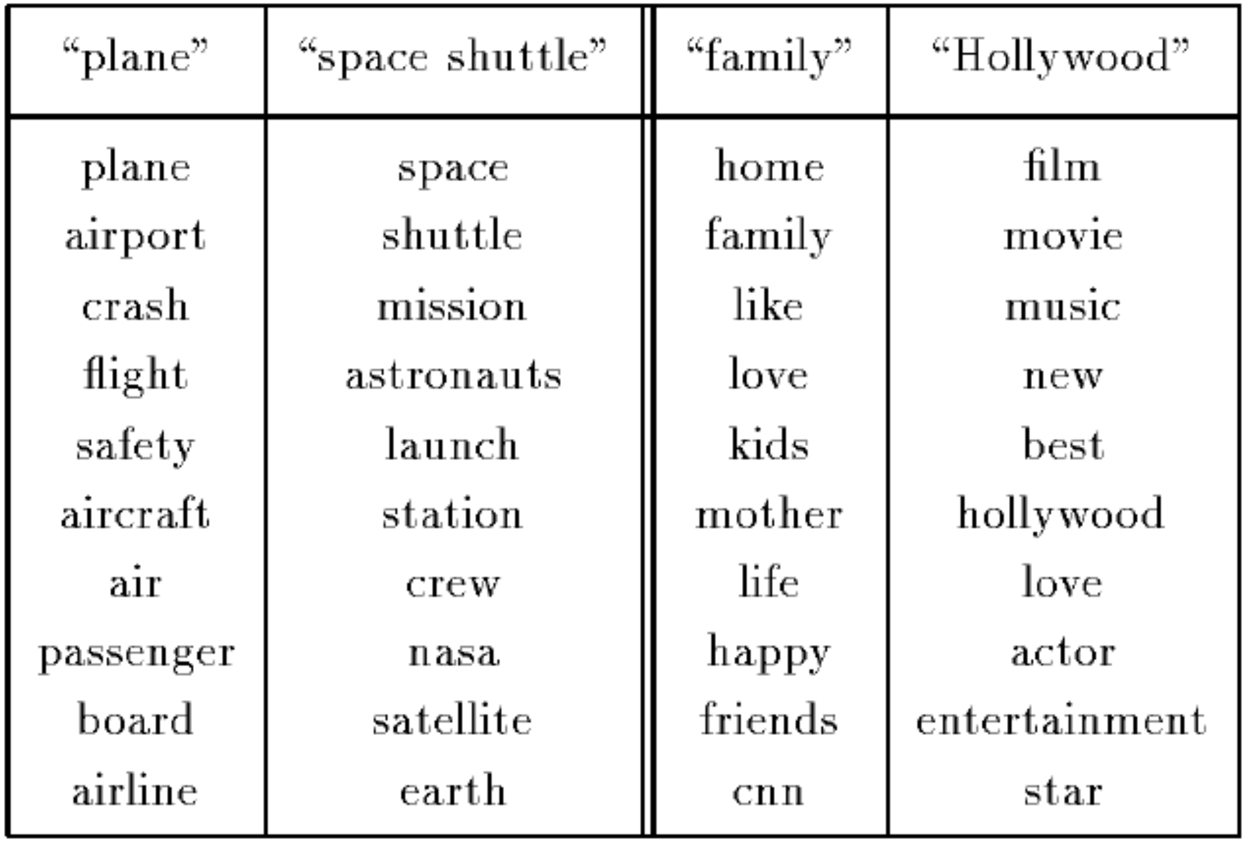
- each *word* is generated by a single unknown multinomial distribution of words, each document is mixed by θ_d
- need to estimate θ_d for each d ➔ overfitting is easy

LDA:

- integrate out θ_d and only estimate β

Introduction to Topic Models

- PLSA topics (TDT-1 corpus)

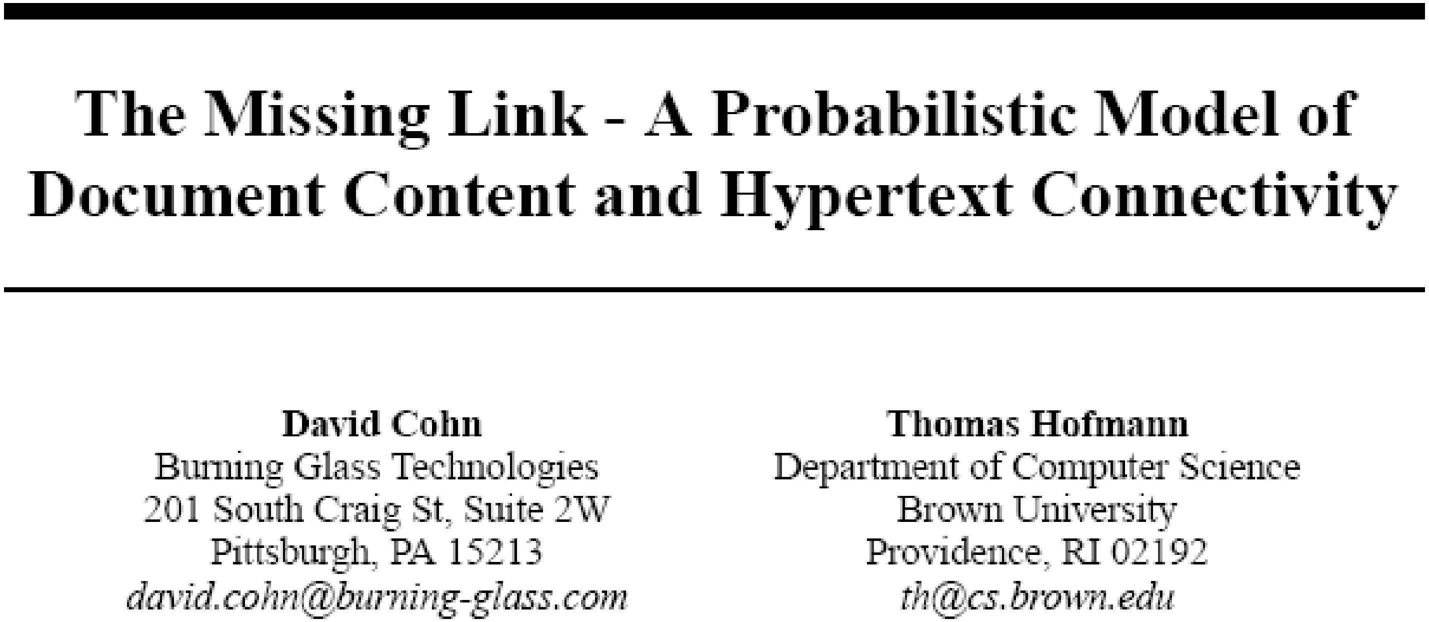


"plane"	"space shuttle"	"family"	"Hollywood"
plane	space	home	film
airport	shuttle	family	movie
crash	mission	like	music
flight	astronauts	love	new
safety	launch	kids	best
aircraft	station	mother	hollywood
air	crew	life	love
passenger	nasa	happy	actor
board	satellite	friends	entertainment
airline	earth	cnn	star

Outline

- Tools for analysis of text
  - Probabilistic models for text, communities, and time
    - Mixture models and LDA models for text
    - **LDA extensions to model hyperlink structure**
    - LDA extensions to model time
  - Alternative framework based on graph analysis to model time & community
    - Preliminary results & tradeoffs
- Discussion of results & challenges

Hyperlink modeling using PLSA



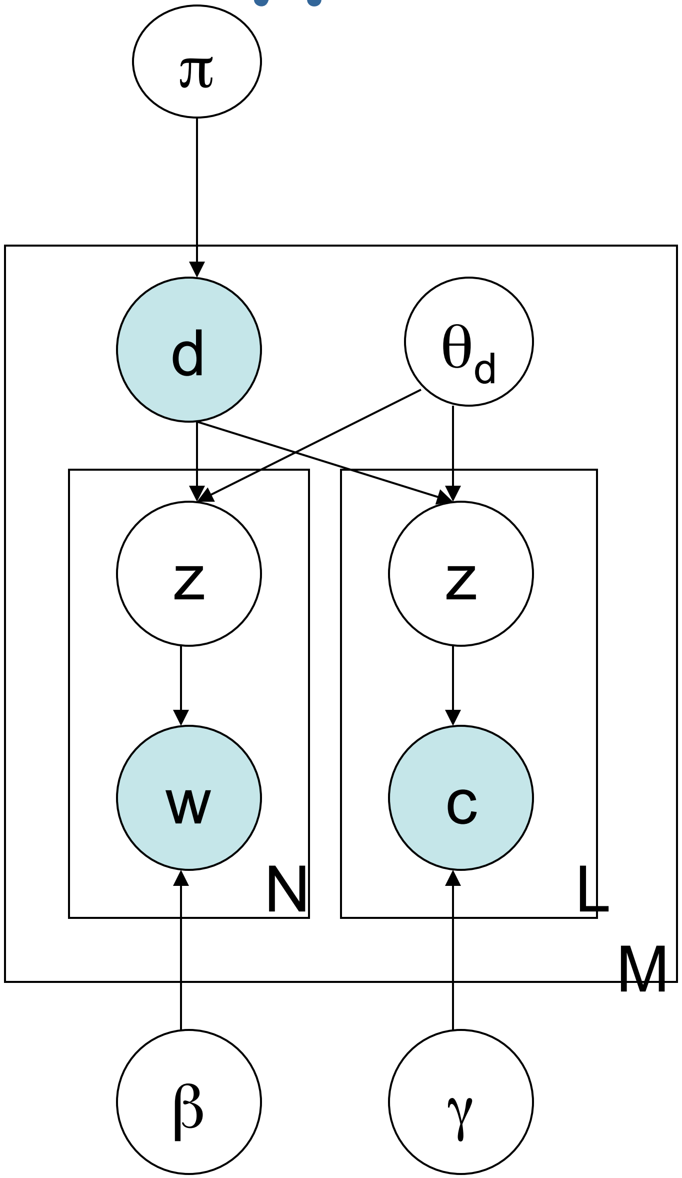
The Missing Link - A Probabilistic Model of Document Content and Hypertext Connectivity

David Cohn
Burning Glass Technologies
201 South Craig St, Suite 2W
Pittsburgh, PA 15213
david.cohn@burning-glass.com

Thomas Hofmann
Department of Computer Science
Brown University
Providence, RI 02192
th@cs.brown.edu

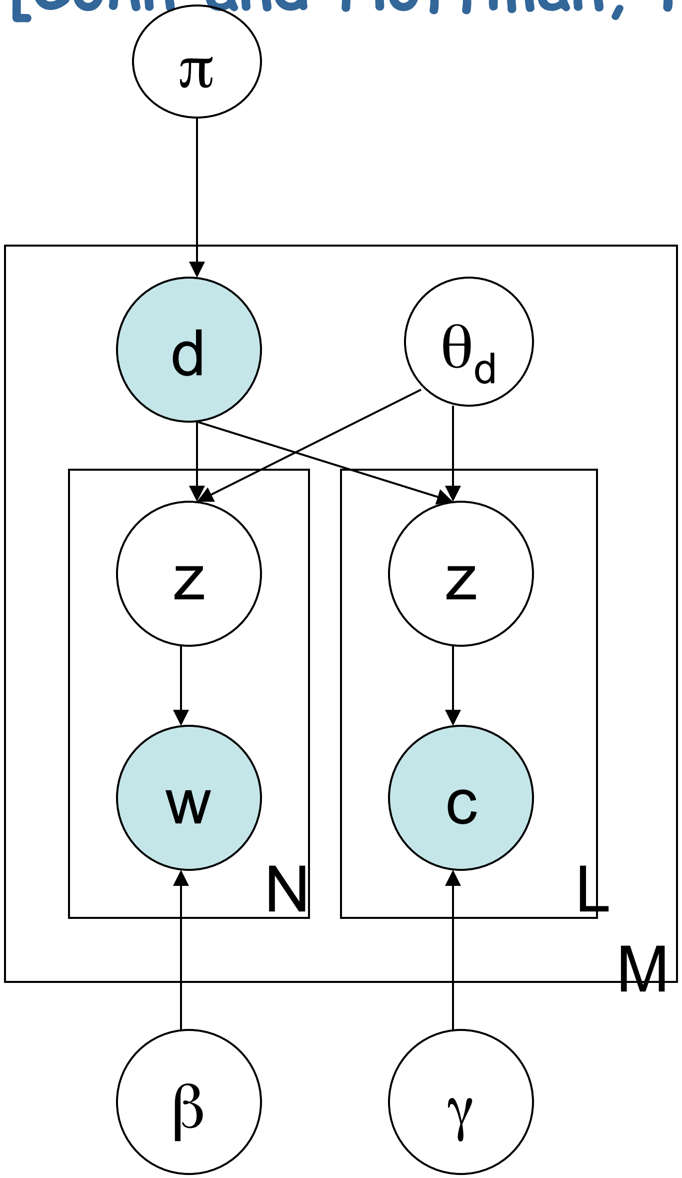
Hyperlink modeling using PLSA

[Cohn and Hoffman, NIPS, 2001]



- Select document d ~ Mult(π)
  - For each position n = 1,···, N_d
    - generate z_n ~ Mult(. | θ_d)
    - generate w_n ~ Mult(. | β_{z_n})
  - For each citation j = 1,···, L_d
    - generate z_j ~ Mult(. | θ_d)
    - generate c_j ~ Mult(. | γ_{z_j})

Hyperlink modeling using PLSA
[Cohn and Hoffman, NIPS, 2001]



PLSA likelihood:

$$\prod_{d=1}^{N_d} P(w_1,\cdots,w_{N_d},d|\theta,\beta,\pi)$$

$$= \prod_{d=1}^{N_d} \pi_d \left\{ \prod_{n=1}^{N_d} \left(\sum_k \theta_{dk}\beta_{kw_n} \right) \right\}$$

New likelihood:

$$\prod_{d=1}^{N_d} P(w_1,\cdots,w_{N_d},c_1,\cdots,c_{L_d},d|\theta,\beta,\gamma,\pi)$$

$$= \prod_{d=1}^{N_d} \pi_d \left\{ \prod_{n=1}^{N_d} \left(\sum_k \theta_{dk}\beta_{kw_n} \right) \right\} \left\{ \prod_{j=1}^{L_d} \left(\sum_k \theta_{dk}\gamma_{kc_j} \right) \right\}$$

Learning using EM

Hyperlink modeling using PLSA
[Cohn and Hoffman, NIPS, 2001]

Heuristic:

$$\prod_{d=1}^{N_d} P(w_1,\cdots,w_{N_d},c_1,\cdots,c_{L_d},d|\theta,\beta,\gamma,\pi)$$

$$= \prod_{d=1}^{N_d}\pi_d \left\{\prod_{n=1}^{N_d}\left(\sum_k \theta_{dk}\beta_{kw_n}\right)\right\}^{\alpha}\left\{\prod_{j=1}^{L_d}\left(\sum_k \theta_{dk}\gamma_{kc_j}\right)\right\}^{(1-\alpha)}$$

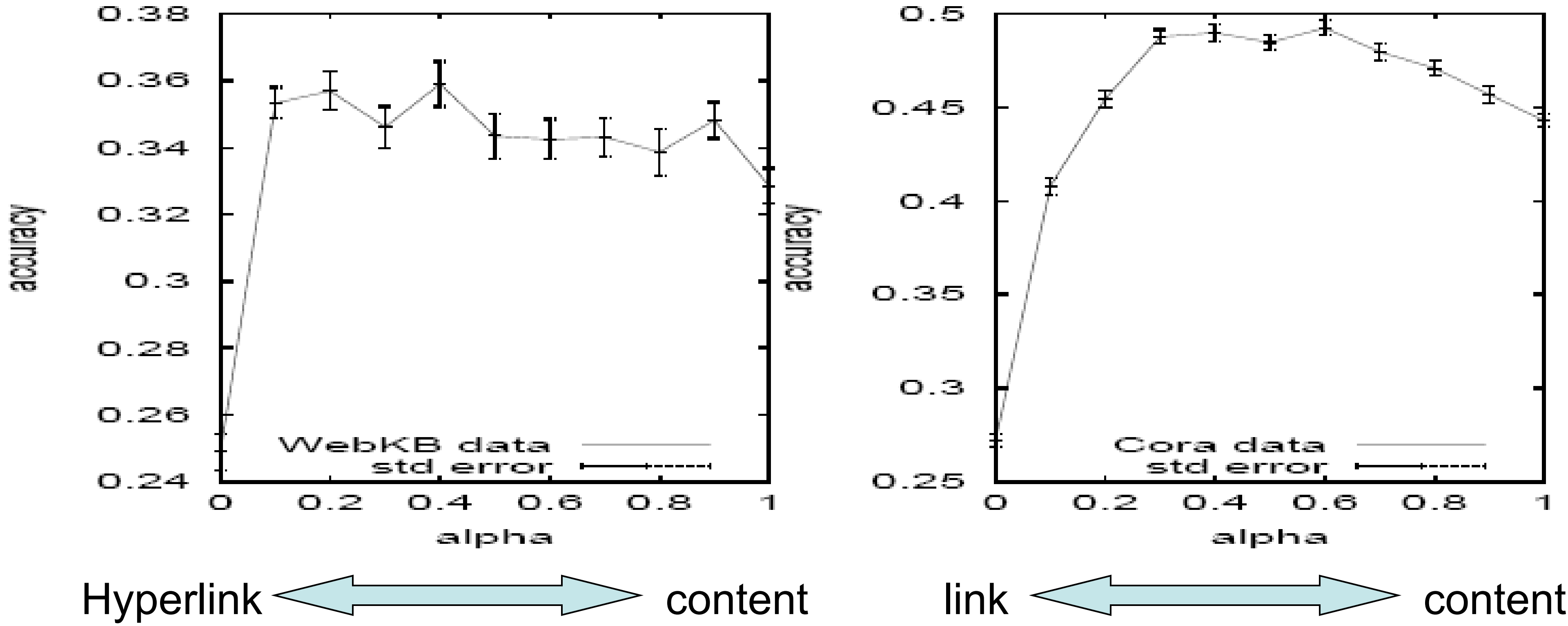
0 · α · 1 determines the relative importance of content and hyperlinks

Hyperlink modeling using PLSA [Cohn and Hoffman, NIPS, 2001]

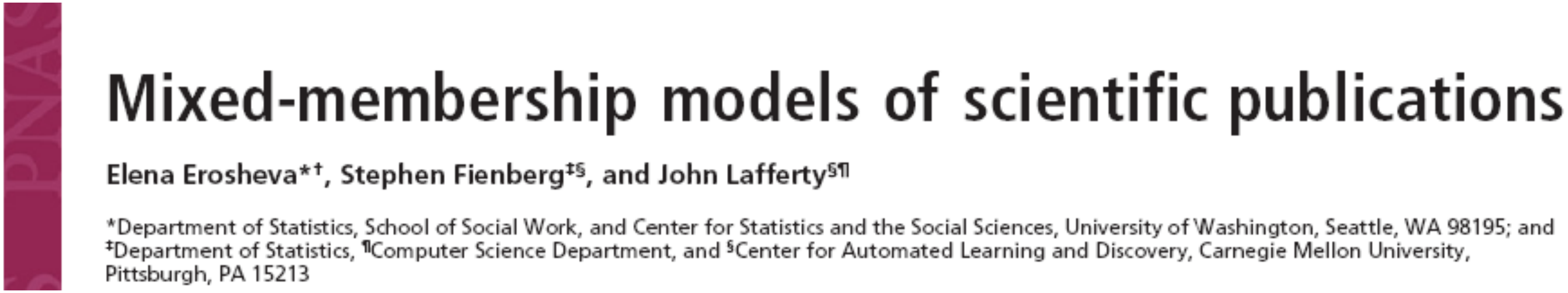
- Experiments: Text Classification
- Datasets:
  - Web KB
    - 6000 CS dept web pages with hyperlinks
    - 6 Classes: faculty, course, student, staff, etc.
  - Cora
    - 2000 Machine learning abstracts with citations
    - 7 classes: sub-areas of machine learning
- Methodology:
  - Learn the model on complete data and obtain θ_d for each document
  - Test documents classified into the label of the nearest neighbor in training set
  - Distance measured as cosine similarity in the θ space
  - Measure the performance as a function of α

Hyperlink modeling using PLSA
[Cohn and Hoffman, NIPS, 2001]

- Classification performance



Hyperlink modeling using LDA

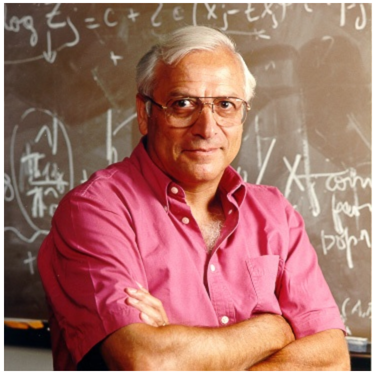


PNAS

Mixed-membership models of scientific publications

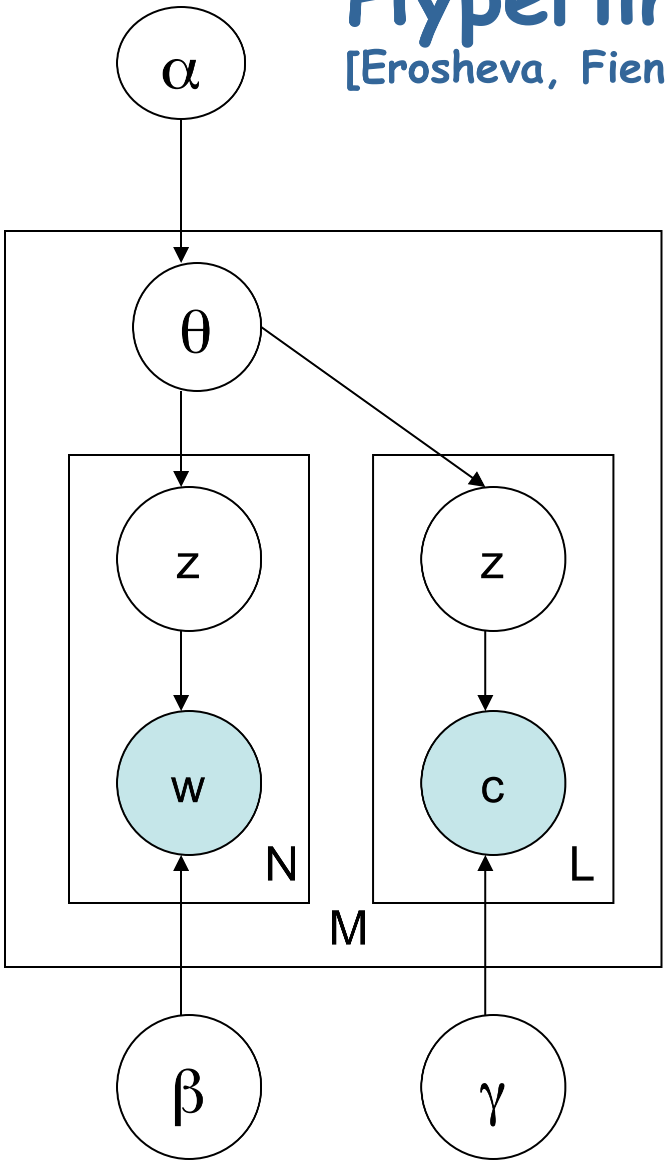
Elena Erosheva[*†], Stephen Fienberg[‡§], and John Lafferty[§¶]

[*]Department of Statistics, School of Social Work, and Center for Statistics and the Social Sciences, University of Washington, Seattle, WA 98195; and [‡]Department of Statistics, [¶]Computer Science Department, and [§]Center for Automated Learning and Discovery, Carnegie Mellon University, Pittsburgh, PA 15213



Hyperlink modeling using LinkLDA

[Erosheva, Fienberg, Lafferty, PNAS, 2004]



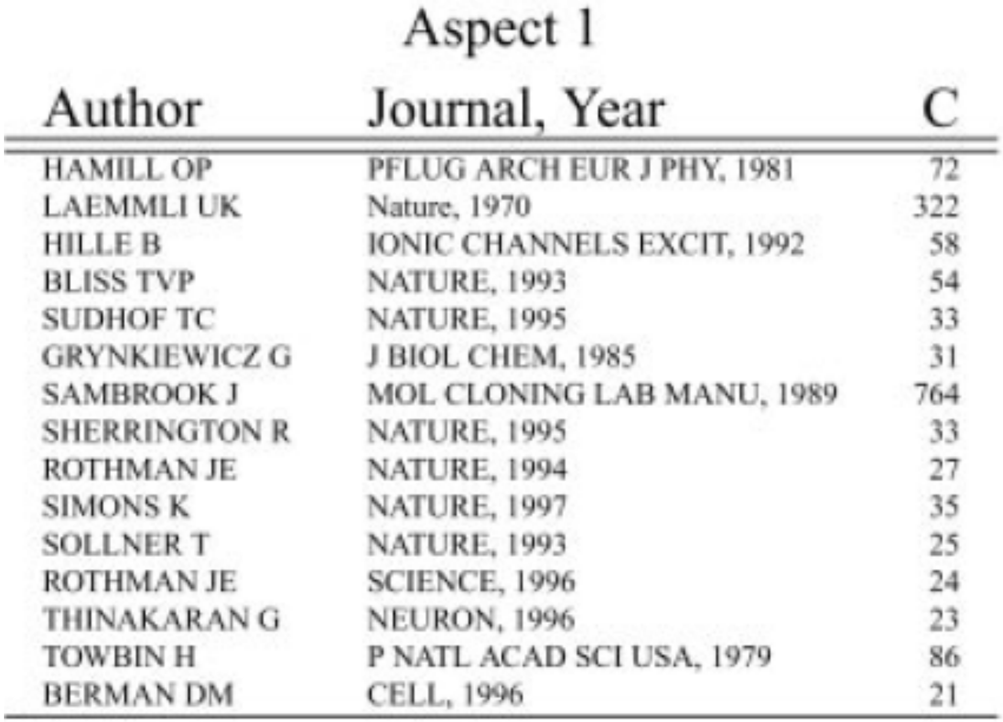
- For each document d = 1,⋯,M
  - Generate θ_d ~ Dir(¢ | α)
  - For each position n = 1,⋯, N_d
    - generate z_n ~ Mult(. | θ_d)
    - generate w_n ~ Mult(. | β_{z_n})
  - For each citation j = 1,⋯, L_d
    - generate z_j ~ Mult(. | θ_d)
    - generate c_j ~ Mult(. | γ_{z_j})

Learning using variational EM

Hyperlink modeling using LDA

[Erosheva, Fienberg, Lafferty, PNAS, 2004]

Aspect 1
Ca^{2+}
channel
membrane
channels
receptors
synaptic
neurons
G
calcium
activation
release
kinase
subunit
intracellular
acid



Aspect 1

Author	Journal, Year	C
HAMILL OP	PFLUG ARCH EUR J PHY, 1981	72
LAEMMLI UK	Nature, 1970	322
HILLE B	IONIC CHANNELS EXCIT, 1992	58
BLISS TVP	NATURE, 1993	54
SUDHOF TC	NATURE, 1995	33
GRYNKIEWICZ G	J BIOL CHEM, 1985	31
SAMBROOK J	MOL CLONING LAB MANU, 1989	764
SHERRINGTON R	NATURE, 1995	33
ROTHMAN JE	NATURE, 1994	27
SIMONS K	NATURE, 1997	35
SOLLNER T	NATURE, 1993	25
ROTHMAN JE	SCIENCE, 1996	24
THINAKARAN G	NEURON, 1996	23
TOWBIN H	P NATL ACAD SCI USA, 1979	86
BERMAN DM	CELL, 1996	21

Newswire Text

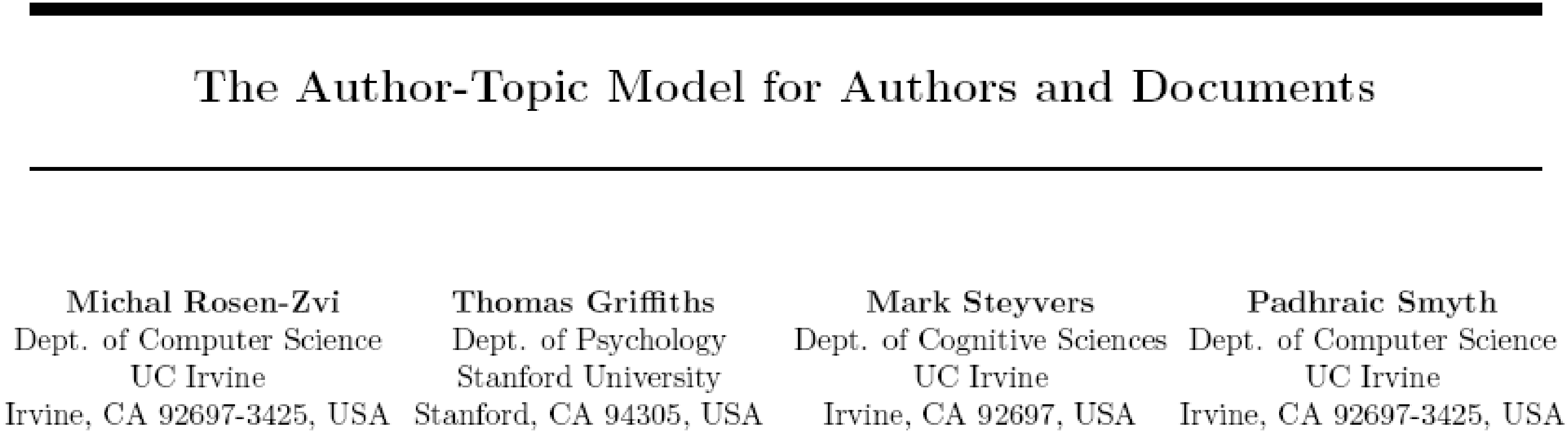
- Goals of analysis:
  - Extract information about events from text
  - "Understanding" text requires understanding "typical reader"
    - conventions for communicating with him/her
    - Prior knowledge, background, …

Social Media Text

- Goals of analysis:
  - Very diverse
  - **Evaluation is difficult**
    - And requires revisiting often as goals evolve
  - Often "understanding" social text requires understanding a *community*

Science as a testbed for social text: an *open* community which we understand

Author-Topic Model for Scientific Literature



The Author-Topic Model for Authors and Documents

Michal Rosen-Zvi
Dept. of Computer Science
UC Irvine
Irvine, CA 92697-3425, USA

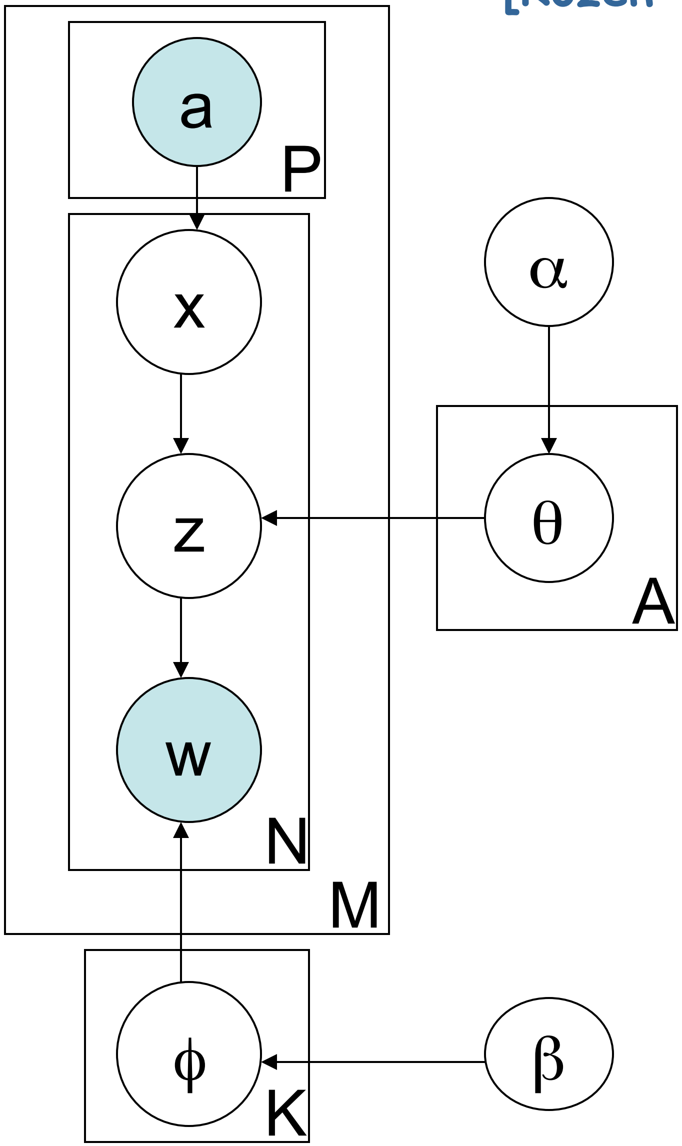
Thomas Griffiths
Dept. of Psychology
Stanford University
Stanford, CA 94305, USA

Mark Steyvers
Dept. of Cognitive Sciences
UC Irvine
Irvine, CA 92697, USA

Padhraic Smyth
Dept. of Computer Science
UC Irvine
Irvine, CA 92697-3425, USA

Author-Topic Model for Scientific Literature

[Rozen-Zvi, Griffiths, Steyvers, Smyth UAI, 2004]

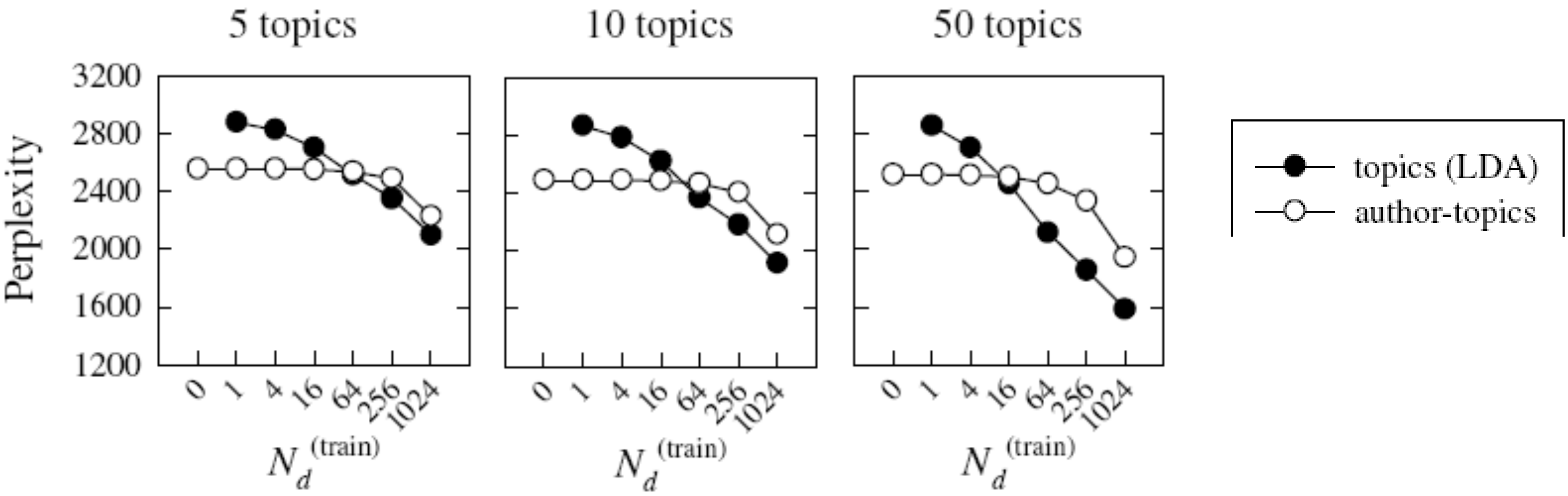


- For each author a = 1,⋯,A
  - Generate θ_a ~ Dir(. | γ)
- For each topic k = 1,⋯,K
  - Generate ϕ_k ~ Dir(. | α)
- For each document d = 1,⋯,M
  - For each position n = 1,⋯, N_d
    - Generate author x ~ Unif(. | a_d)
    - generate z_n ~ Mult(. | θ_a)
    - generate w_n ~ Mult(. | ϕ_{z_n})

Author-Topic Model for Scientific Literature

[Rozen-Zvi, Griffiths, Steyvers, Smyth UAI, 2004]

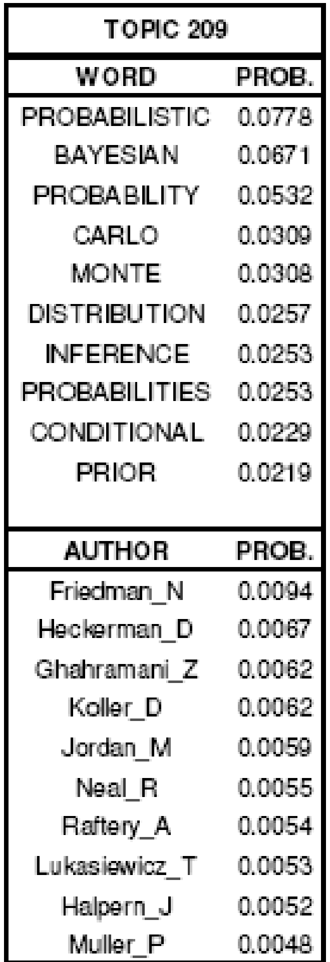
- Perplexity results



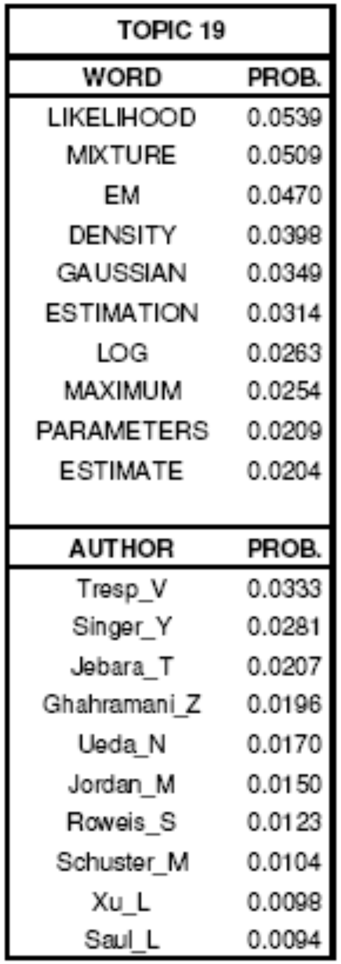
Author-Topic Model for Scientific Literature

[Rozen-Zvi, Griffiths, Steyvers, Smyth UAI, 2004]

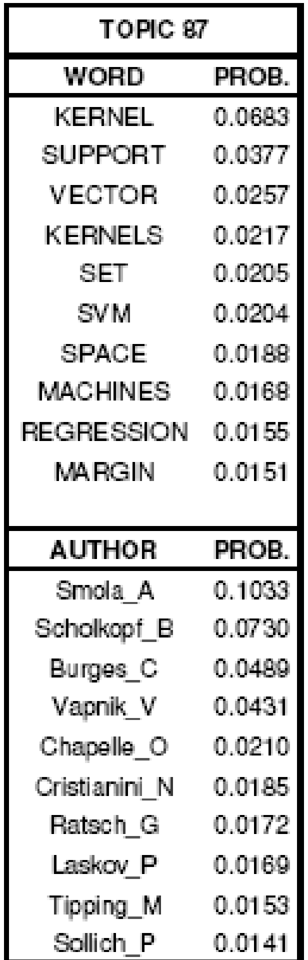
- Topic-Author visualization



TOPIC 209	
WORD	PROB.
PROBABILISTIC	0.0778
BAYESIAN	0.0671
PROBABILITY	0.0532
CARLO	0.0309
MONTE	0.0308
DISTRIBUTION	0.0257
INFERENCE	0.0253
PROBABILITIES	0.0253
CONDITIONAL	0.0229
PRIOR	0.0219
AUTHOR	PROB.
Friedman_N	0.0094
Heckerman_D	0.0067
Ghahramani_Z	0.0062
Koller_D	0.0062
Jordan_M	0.0059
Neal_R	0.0055
Raftery_A	0.0054
Lukasiewicz_T	0.0053
Halpern_J	0.0052
Muller_P	0.0048



TOPIC 19	
WORD	PROB.
LIKELIHOOD	0.0539
MIXTURE	0.0509
EM	0.0470
DENSITY	0.0398
GAUSSIAN	0.0349
ESTIMATION	0.0314
LOG	0.0263
MAXIMUM	0.0254
PARAMETERS	0.0209
ESTIMATE	0.0204
AUTHOR	PROB.
Tresp_V	0.0333
Singer_Y	0.0281
Jebara_T	0.0207
Ghahramani_Z	0.0196
Ueda_N	0.0170
Jordan_M	0.0150
Roweis_S	0.0123
Schuster_M	0.0104
Xu_L	0.0098
Saul_L	0.0094

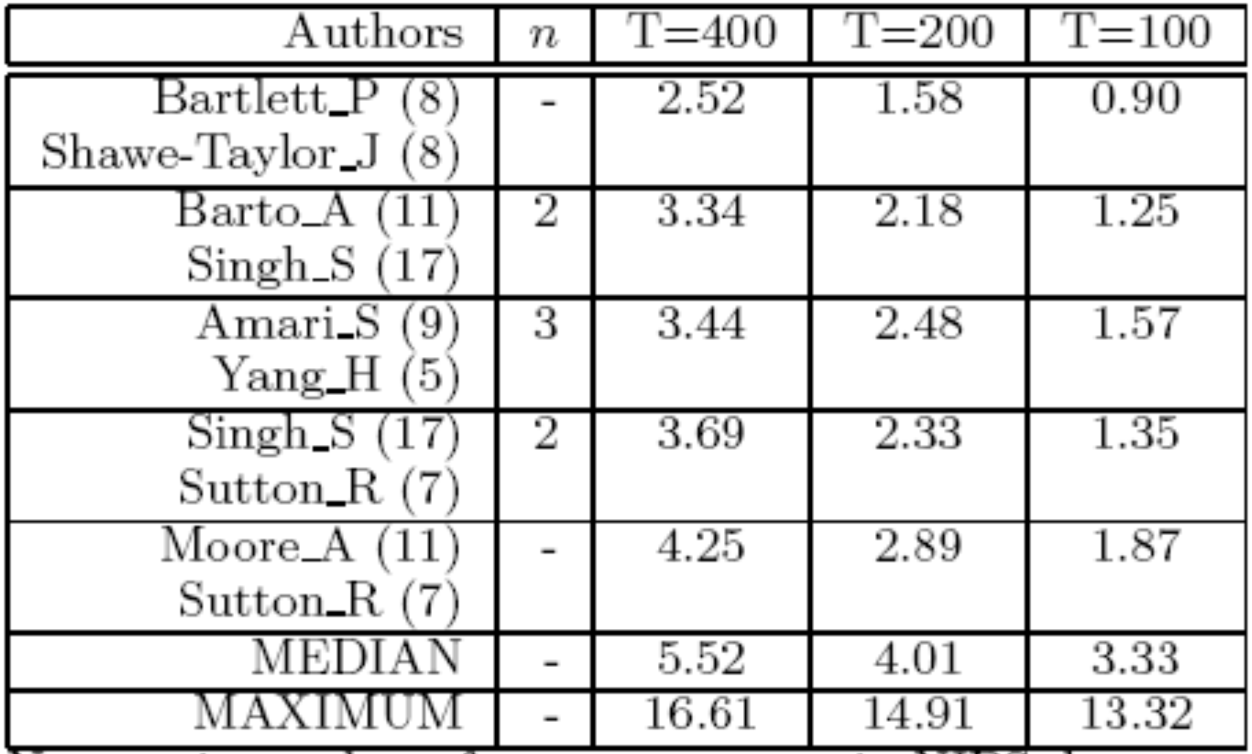


TOPIC 87	
WORD	PROB.
KERNEL	0.0683
SUPPORT	0.0377
VECTOR	0.0257
KERNELS	0.0217
SET	0.0205
SVM	0.0204
SPACE	0.0188
MACHINES	0.0168
REGRESSION	0.0155
MARGIN	0.0151
AUTHOR	PROB.
Smola_A	0.1033
Scholkopf_B	0.0730
Burges_C	0.0489
Vapnik_V	0.0431
Chapelle_O	0.0210
Cristianini_N	0.0185
Ratsch_G	0.0172
Laskov_P	0.0169
Tipping_M	0.0153
Sollich_P	0.0141

Author-Topic Model for Scientific Literature

[Rozen-Zvi, Griffiths, Steyvers, Smyth UAI, 2004]

- Application 1: Author similarity



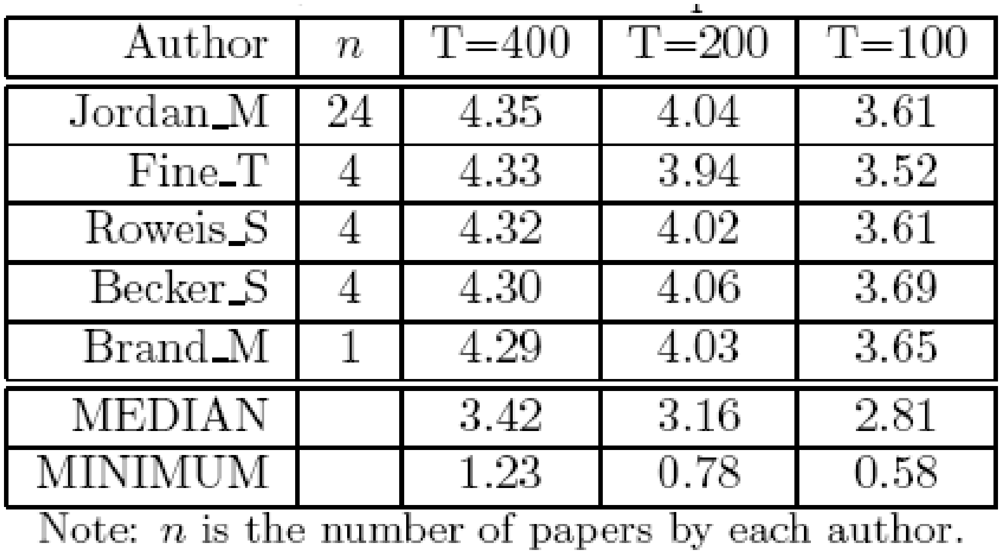
Authors	n	T=400	T=200	T=100
Bartlett_P (8) Shawe-Taylor_J (8)	-	2.52	1.58	0.90
Barto_A (11) Singh_S (17)	2	3.34	2.18	1.25
Amari_S (9) Yang_H (5)	3	3.44	2.48	1.57
Singh_S (17) Sutton_R (7)	2	3.69	2.33	1.35
Moore_A (11) Sutton_R (7)	-	4.25	2.89	1.87
MEDIAN	-	5.52	4.01	3.33
MAXIMUM	-	16.61	14.91	13.32

Note: n is number of common papers in NIPS dataset.

Author-Topic Model for Scientific Literature

[Rozen-Zvi, Griffiths, Steyvers, Smyth UAI, 2004]

- Application 2: Author entropy

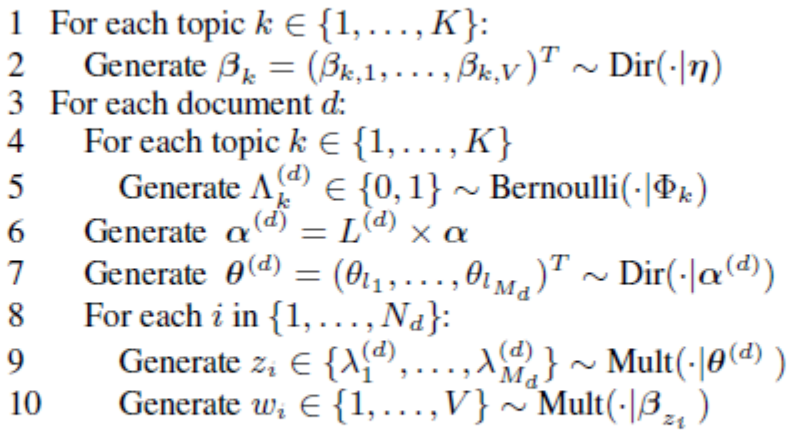


Author	n	T=400	T=200	T=100
Jordan_M	24	4.35	4.04	3.61
Fine_T	4	4.33	3.94	3.52
Roweis_S	4	4.32	4.02	3.61
Becker_S	4	4.30	4.06	3.69
Brand_M	1	4.29	4.03	3.65
MEDIAN		3.42	3.16	2.81
MINIMUM		1.23	0.78	0.58

Note: n is the number of papers by each author.

Labeled LDA:

[Ramage, Hall, Nallapati, Manning, EMNLP 2009]



```
For each topic $k \in \{1, \dots, K\}$:
    Generate $\beta_k = (\beta_{k,1}, \dots, \beta_{k,V})^T \sim \mathrm{Dir}(\cdot|\eta)$
For each document $d$:
    For each topic $k \in \{1, \dots, K\}$
        Generate $\Lambda_k^{(d)} \in \{0, 1\} \sim \mathrm{Bernoulli}(\cdot|\Phi_k)$
    Generate $\alpha^{(d)} = L^{(d)} \times \alpha$
    Generate $\theta^{(d)} = (\theta_{l_1}, \dots, \theta_{l_{M_d}})^T \sim \mathrm{Dir}(\cdot|\alpha^{(d)})$
    For each $i$ in $\{1, \dots, N_d\}$:
        Generate $z_i \in \{\lambda_1^{(d)}, \dots, \lambda_{M_d}^{(d)}\} \sim \mathrm{Mult}(\cdot|\theta^{(d)})$
        Generate $w_i \in \{1, \dots, V\} \sim \mathrm{Mult}(\cdot|\beta_{z_i})$
```

Table 1: Generative process for Labeled LDA: β_k is a vector consisting of the parameters of the multinomial distribution corresponding to the k^{th} topic, α are the parameters of the Dirichlet topic prior and η are the parameters of the word prior, while Φ_k is the label prior for topic k. For the meaning of the projection matrix $L^{(d)}$, please refer to Eq 1.

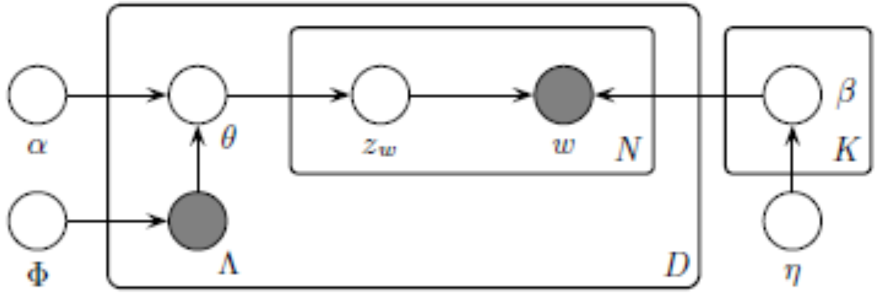
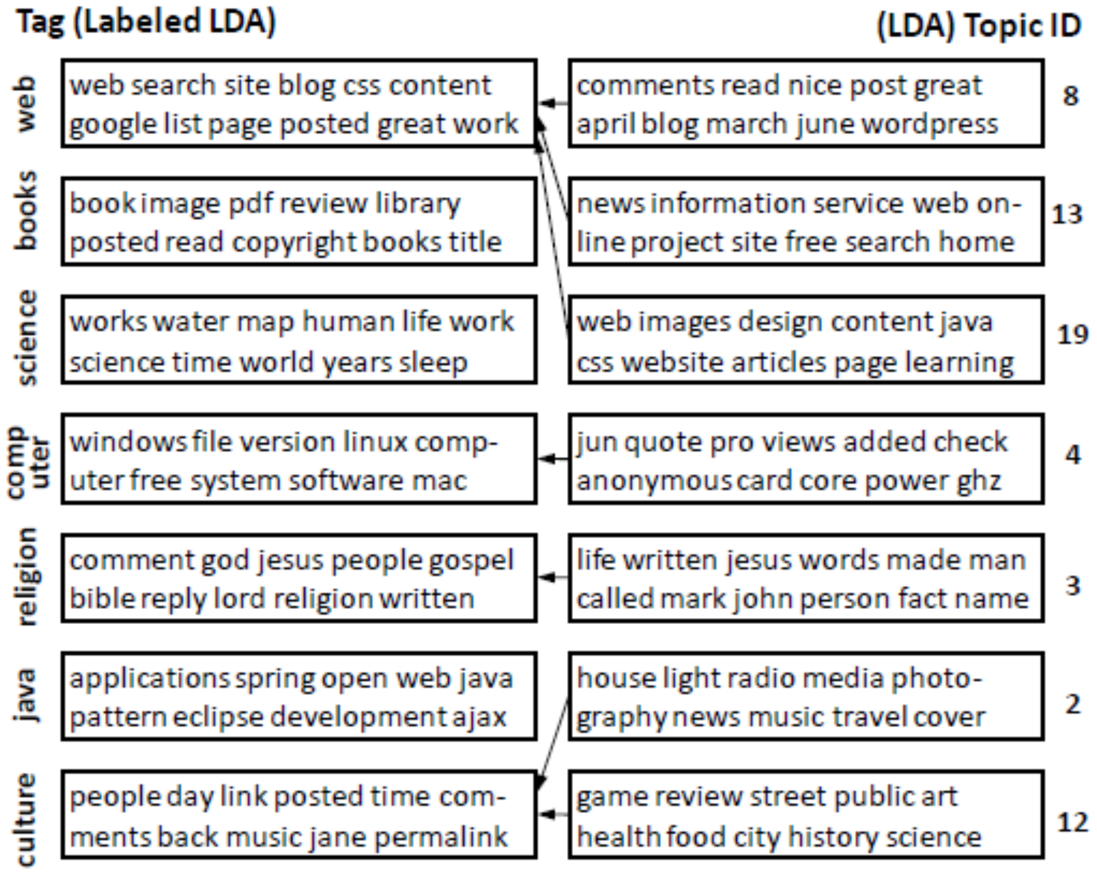


Figure 1: Graphical model of Labeled LDA: unlike standard LDA, both the label set Λ as well as the topic prior α influence the topic mixture θ.

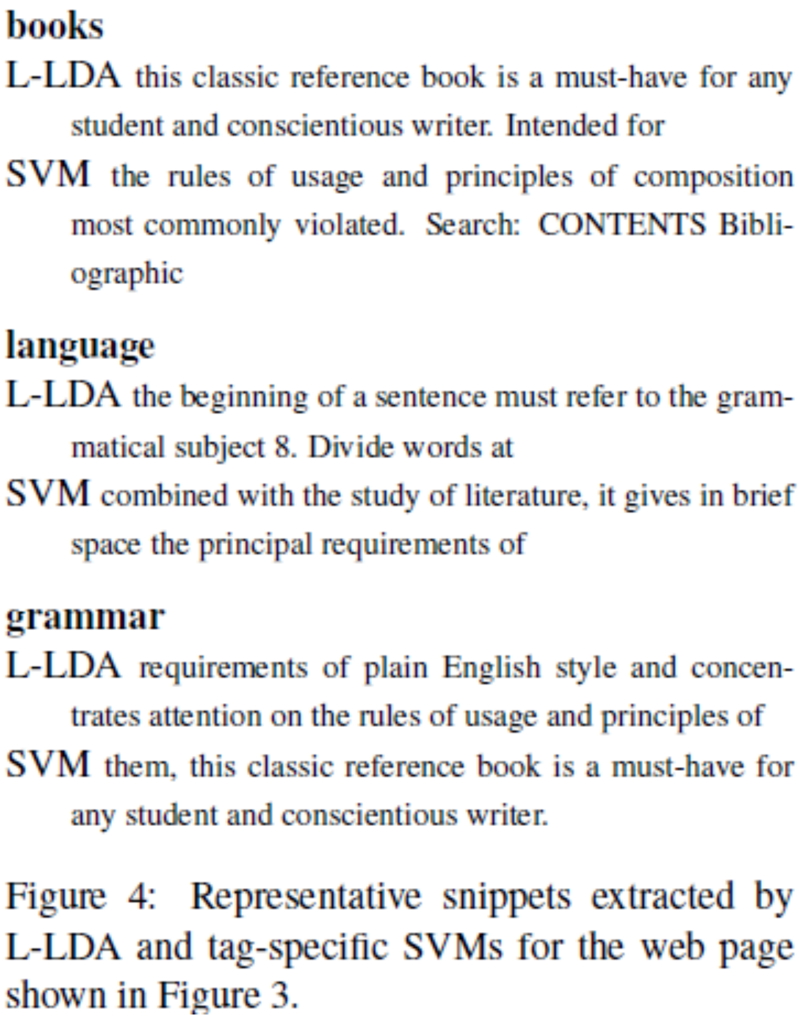
Labeled LDA

Del.icio.us tags as labels for documents



Tag (Labeled LDA)			(LDA) Topic ID
web	web search site blog css content google list page posted great work	comments read nice post great april blog march june wordpress	8
books	book image pdf review library posted read copyright books title	news information service web on-line project site free search home	13
science	works water map human life work science time world years sleep	web images design content java css website articles page learning	19
computer	windows file version linux computer free system software mac	jun quote pro views added check anonymous card core power ghz	4
religion	comment god jesus people gospel bible reply lord religion written	life written jesus words made man called mark john person fact name	3
java	applications spring open web java pattern eclipse development ajax	house light radio media photography news music travel cover	2
culture	people day link posted time comments back music jane permalink	game review street public art health food city history science	12

Labeled LDA



books

L-LDA this classic reference book is a must-have for any student and conscientious writer. Intended for

SVM the rules of usage and principles of composition most commonly violated. Search: CONTENTS Bibliographic

language

L-LDA the beginning of a sentence must refer to the grammatical subject 8. Divide words at

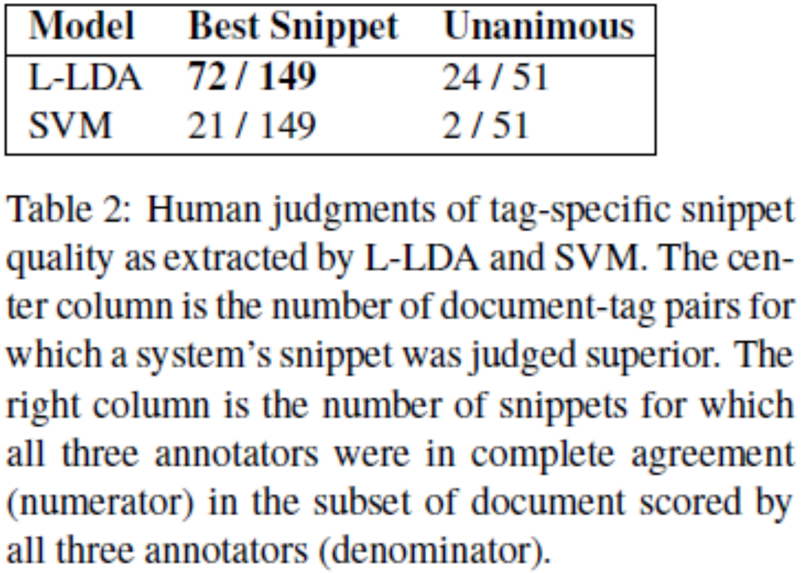
SVM combined with the study of literature, it gives in brief space the principal requirements of

grammar

L-LDA requirements of plain English style and concentrates attention on the rules of usage and principles of

SVM them, this classic reference book is a must-have for any student and conscientious writer.

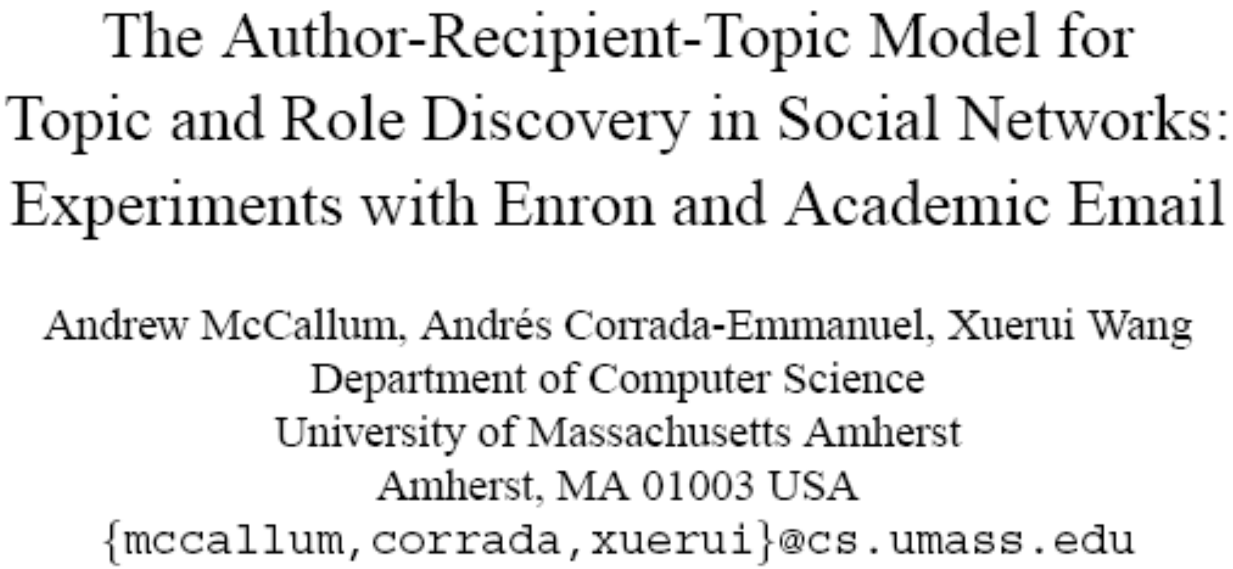
Figure 4: Representative snippets extracted by L-LDA and tag-specific SVMs for the web page shown in Figure 3.



Model	Best Snippet	Unanimous
L-LDA	**72 / 149**	24 / 51
SVM	21 / 149	2 / 51

Table 2: Human judgments of tag-specific snippet quality as extracted by L-LDA and SVM. The center column is the number of document-tag pairs for which a system's snippet was judged superior. The right column is the number of snippets for which all three annotators were in complete agreement (numerator) in the subset of document scored by all three annotators (denominator).

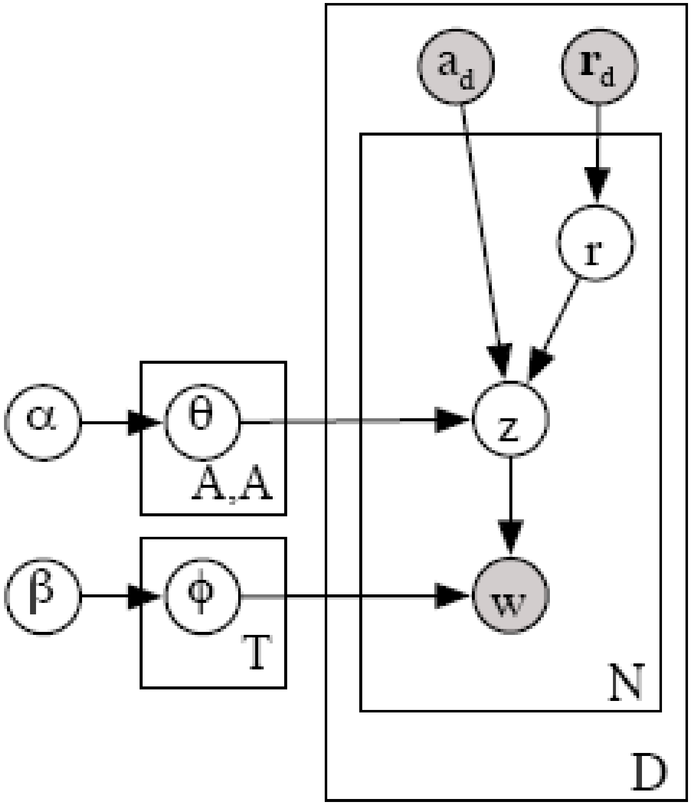
Author-Topic-Recipient model for email data [McCallum, Corrada-Emmanuel,Wang, ICJAI'05]



The Author-Recipient-Topic Model for Topic and Role Discovery in Social Networks: Experiments with Enron and Academic Email

Andrew McCallum, Andrés Corrada-Emmanuel, Xuerui Wang
Department of Computer Science
University of Massachusetts Amherst
Amherst, MA 01003 USA
{mccallum,corrada,xuerui}@cs.umass.edu

Author-Topic-Recipient model for email data [McCallum, Corrada-Emmanuel,Wang, ICJAI'05]



Gibbs sampling

$$P(z_i \,|\, \mathbf{z}_{-i}, \mathbf{x}, \mathbf{w}) \quad \propto \quad \frac{n_{z_i}^{w_v} + \beta_v}{\sum_v n_{z_i}^{w_v} + \beta_v} \frac{n_{x_i}^{z_i} + \alpha_{z_i}}{\sum_{z'} n_{x_i}^{z'} + \alpha_{z'}}$$

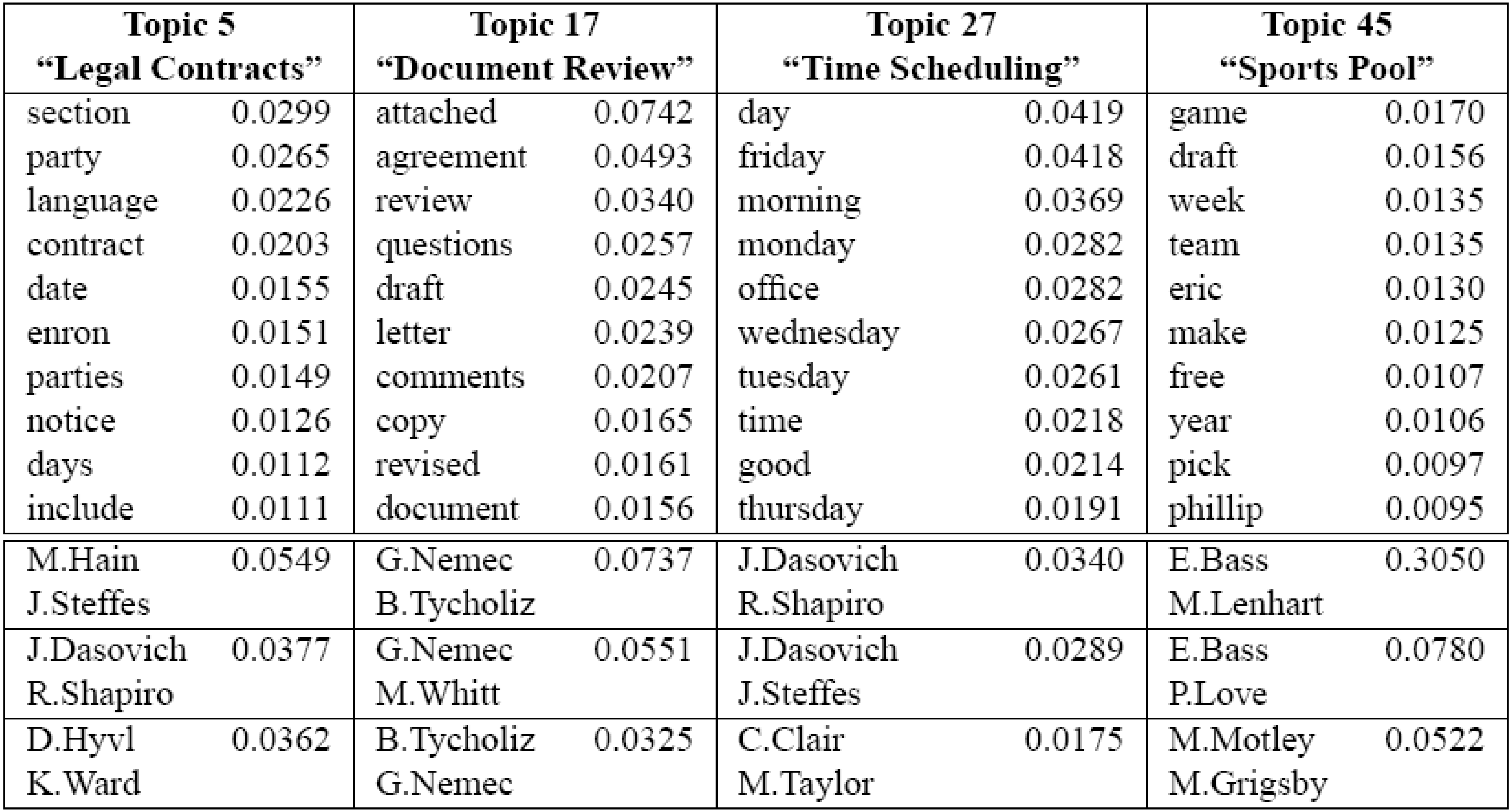
$$P(x_i \,|\, \mathbf{z}, \mathbf{x}_{-i}, \mathbf{w}) \quad \propto \quad \frac{n_{x_i}^{z_i} + \alpha_{z_i}}{\sum_{z'} n_{x_i}^{z'} + \alpha_{z'}}$$

Author-Topic-Recipient model for email data [McCallum, Corrada-Emmanuel,Wang, ICJAI'05]

- Datasets
  - Enron email data
    - 23,488 messages between 147 users
  - McCallum's personal email
    - 23,488(?) messages with 128 authors

Author-Topic-Recipient model for email data [McCallum, Corrada-Emmanuel,Wang, ICJAI'05]

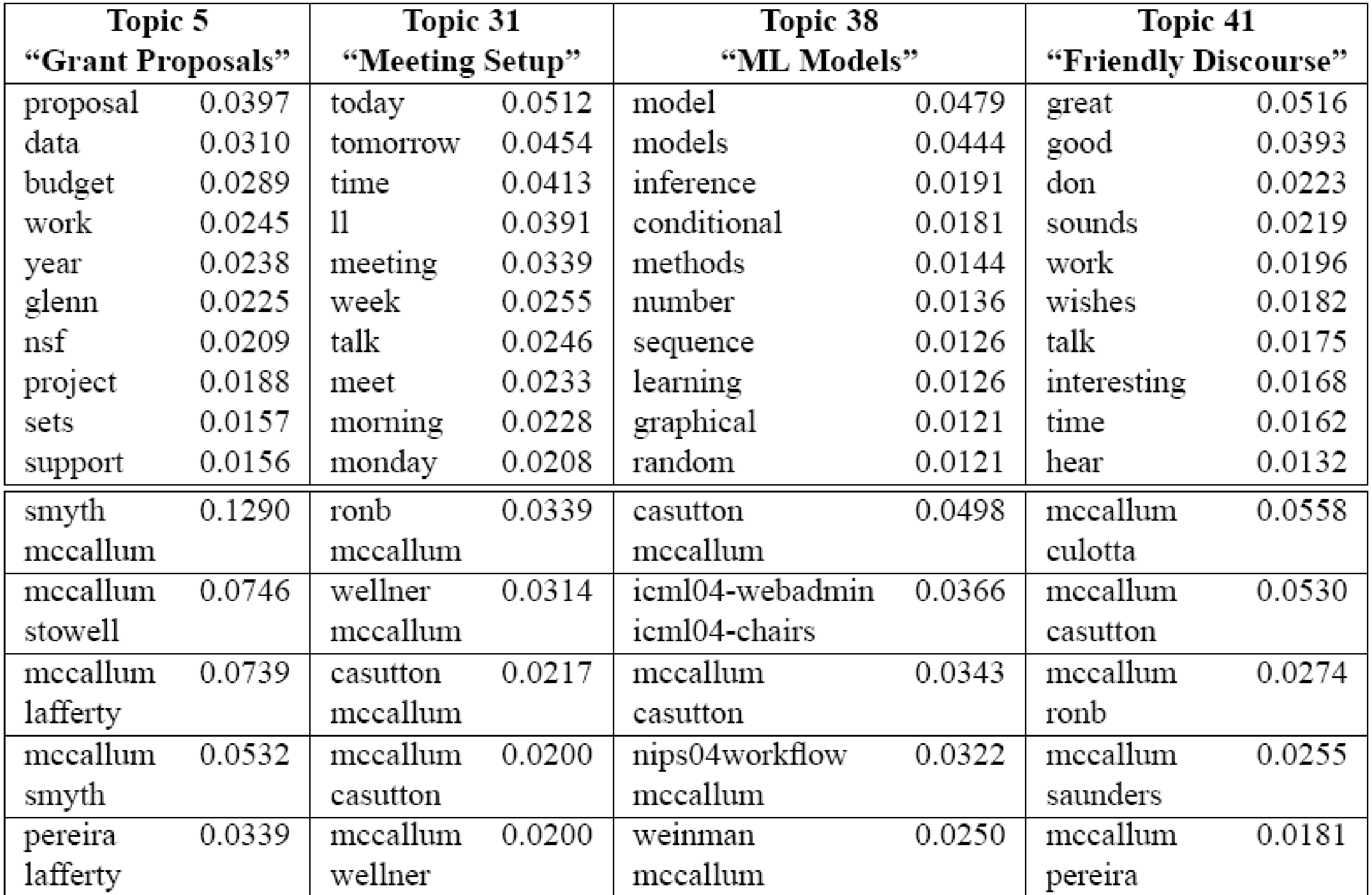
- Topic Visualization: Enron set



Topic 5 "Legal Contracts"		Topic 17 "Document Review"		Topic 27 "Time Scheduling"		Topic 45 "Sports Pool"	
section	0.0299	attached	0.0742	day	0.0419	game	0.0170
party	0.0265	agreement	0.0493	friday	0.0418	draft	0.0156
language	0.0226	review	0.0340	morning	0.0369	week	0.0135
contract	0.0203	questions	0.0257	monday	0.0282	team	0.0135
date	0.0155	draft	0.0245	office	0.0282	eric	0.0130
enron	0.0151	letter	0.0239	wednesday	0.0267	make	0.0125
parties	0.0149	comments	0.0207	tuesday	0.0261	free	0.0107
notice	0.0126	copy	0.0165	time	0.0218	year	0.0106
days	0.0112	revised	0.0161	good	0.0214	pick	0.0097
include	0.0111	document	0.0156	thursday	0.0191	phillip	0.0095
M.Hain	0.0549	G.Nemec	0.0737	J.Dasovich	0.0340	E.Bass	0.3050
J.Steffes		B.Tycholiz		R.Shapiro		M.Lenhart	
J.Dasovich	0.0377	G.Nemec	0.0551	J.Dasovich	0.0289	E.Bass	0.0780
R.Shapiro		M.Whitt		J.Steffes		P.Love	
D.Hyvl	0.0362	B.Tycholiz	0.0325	C.Clair	0.0175	M.Motley	0.0522
K.Ward		G.Nemec		M.Taylor		M.Grigsby	

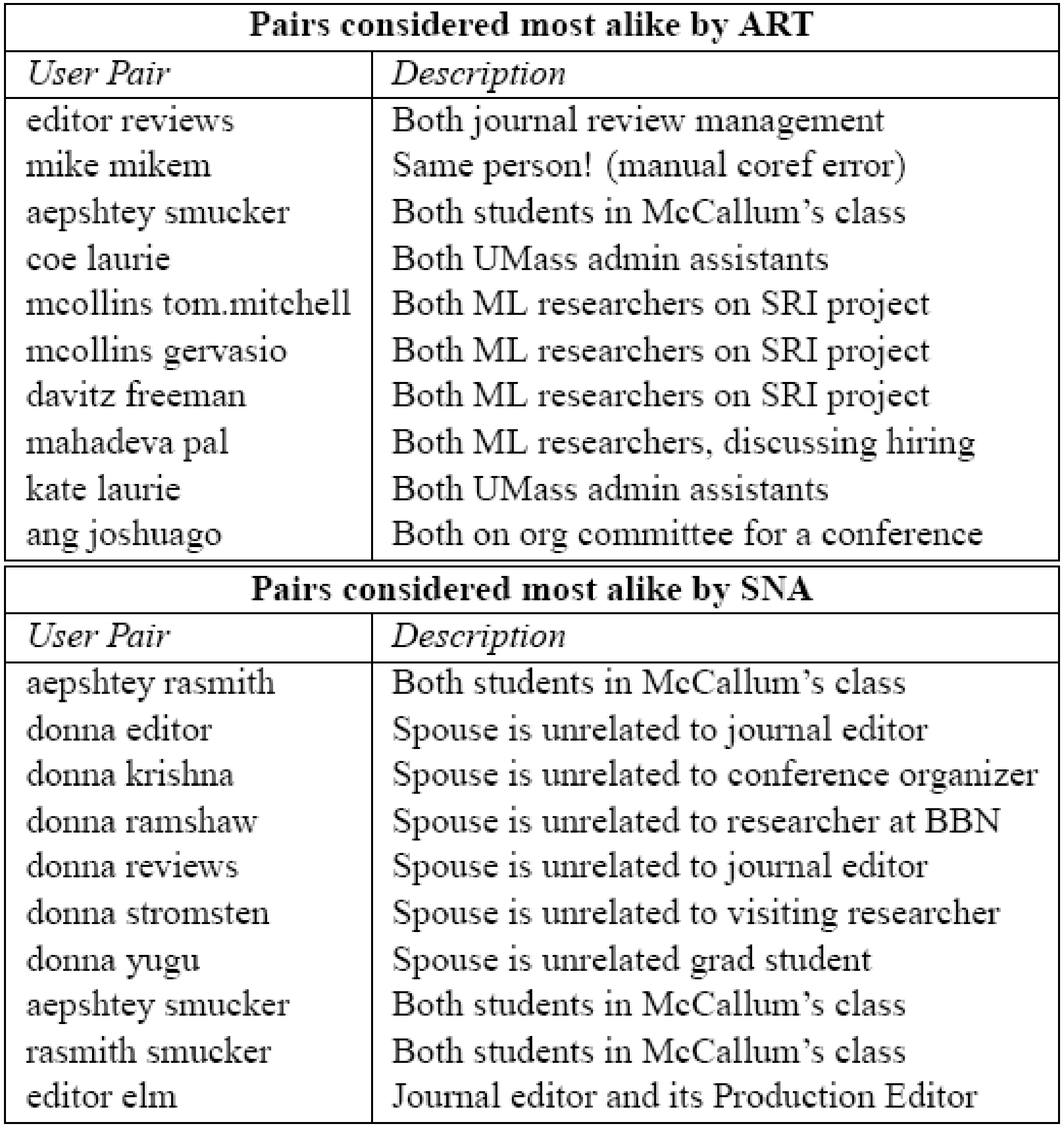
Author-Topic-Recipient model for email data [McCallum, Corrada-Emmanuel,Wang, ICJAI'05]

- Topic Visualization: McCallum's data



Topic 5 "Grant Proposals"		Topic 31 "Meeting Setup"		Topic 38 "ML Models"		Topic 41 "Friendly Discourse"	
proposal	0.0397	today	0.0512	model	0.0479	great	0.0516
data	0.0310	tomorrow	0.0454	models	0.0444	good	0.0393
budget	0.0289	time	0.0413	inference	0.0191	don	0.0223
work	0.0245	ll	0.0391	conditional	0.0181	sounds	0.0219
year	0.0238	meeting	0.0339	methods	0.0144	work	0.0196
glenn	0.0225	week	0.0255	number	0.0136	wishes	0.0182
nsf	0.0209	talk	0.0246	sequence	0.0126	talk	0.0175
project	0.0188	meet	0.0233	learning	0.0126	interesting	0.0168
sets	0.0157	morning	0.0228	graphical	0.0121	time	0.0162
support	0.0156	monday	0.0208	random	0.0121	hear	0.0132
smyth mccallum	0.1290	ronb mccallum	0.0339	casutton mccallum	0.0498	mccallum culotta	0.0558
mccallum stowell	0.0746	wellner mccallum	0.0314	icml04-webadmin icml04-chairs	0.0366	mccallum casutton	0.0530
mccallum lafferty	0.0739	casutton mccallum	0.0217	mccallum casutton	0.0343	mccallum ronb	0.0274
mccallum smyth	0.0532	mccallum casutton	0.0200	nips04workflow mccallum	0.0322	mccallum saunders	0.0255
pereira lafferty	0.0339	mccallum wellner	0.0200	weinman mccallum	0.0250	mccallum pereira	0.0181

Author-Topic-Recipient model for email data [McCallum, Corrada-Emmanuel,Wang, ICJAI’05]



Pairs considered most alike by ART	
User Pair	*Description*
editor reviews	Both journal review management
mike mikem	Same person! (manual coref error)
aepshtey smucker	Both students in McCallum’s class
coe laurie	Both UMass admin assistants
mcollins tom.mitchell	Both ML researchers on SRI project
mcollins gervasio	Both ML researchers on SRI project
davitz freeman	Both ML researchers on SRI project
mahadeva pal	Both ML researchers, discussing hiring
kate laurie	Both UMass admin assistants
ang joshuago	Both on org committee for a conference

Pairs considered most alike by SNA	
User Pair	*Description*
aepshtey rasmith	Both students in McCallum’s class
donna editor	Spouse is unrelated to journal editor
donna krishna	Spouse is unrelated to conference organizer
donna ramshaw	Spouse is unrelated to researcher at BBN
donna reviews	Spouse is unrelated to journal editor
donna stromsten	Spouse is unrelated to visiting researcher
donna yugu	Spouse is unrelated grad student
aepshtey smucker	Both students in McCallum’s class
rasmith smucker	Both students in McCallum’s class
editor elm	Journal editor and its Production Editor

Models of hypertext for blogs [ICWSM 2008]

Ramesh Nallapati

Amr Ahmed

Eric Xing

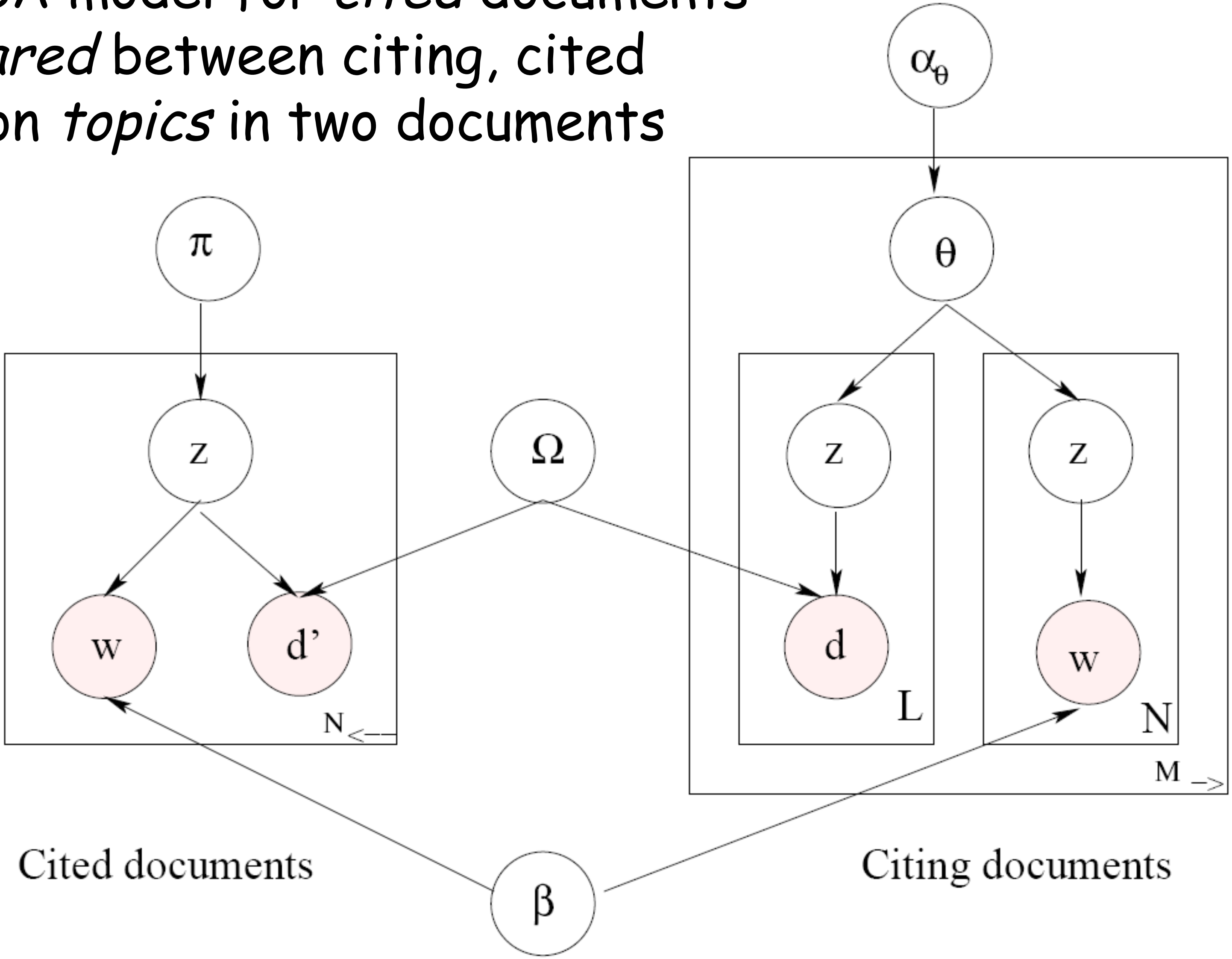
me

LinkLDA model for *citing* documents

Variant of PLSA model for *cited* documents

Topics are *shared* between citing, cited

Links depend on *topics* in two documents



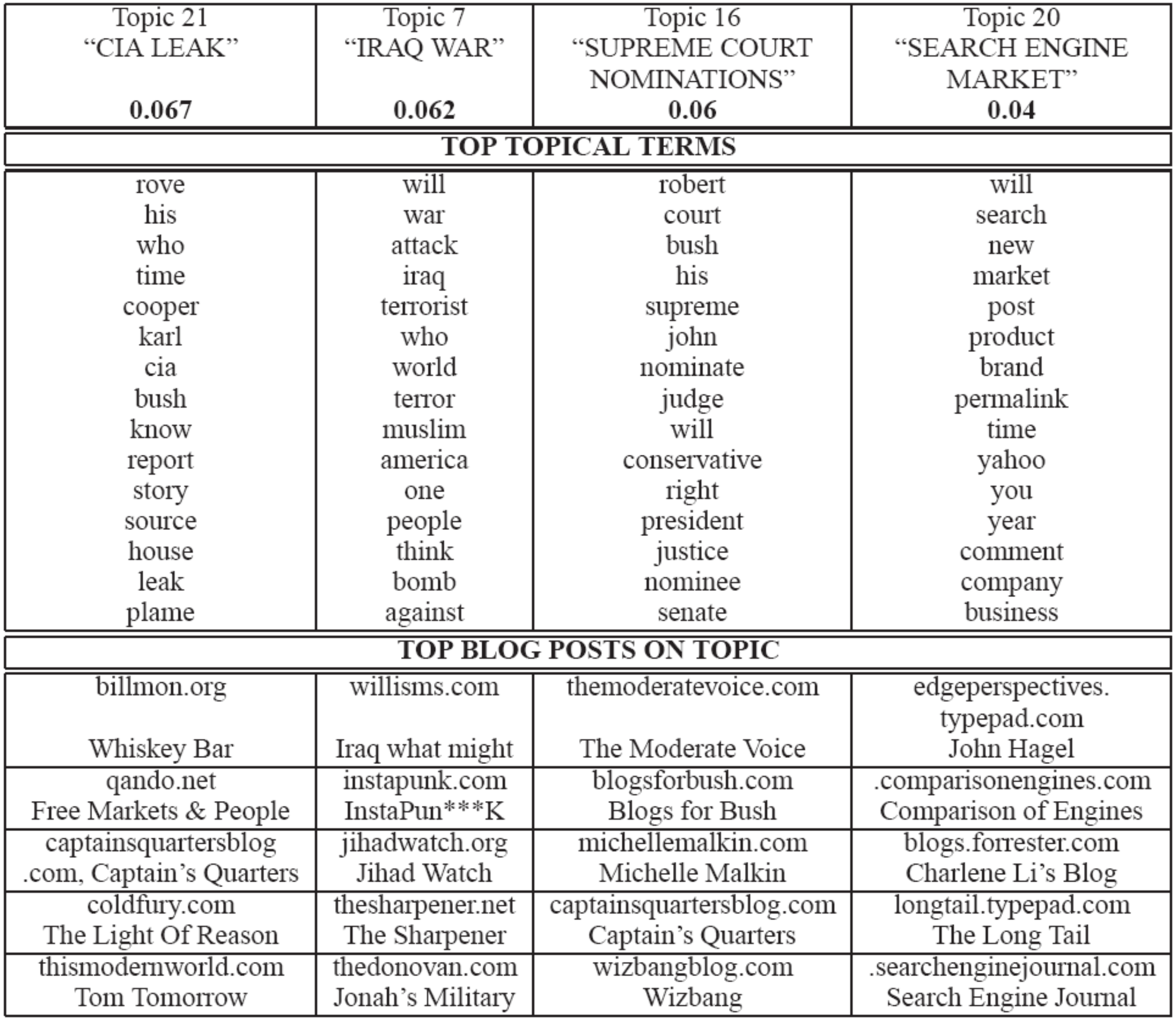
Link-PLSA-LDA

Experiments

- 8.4M blog postings in Nielsen/Buzzmetrics corpus
  - Collected over three weeks summer 2005
- Selected all postings with >=2 inlinks or >=2 outlinks
  - 2248 citing (2+ outlinks), 1777 cited documents (2+ inlinks)
  - Only 68 in both sets, which are duplicated
- Fit model using variational EM

Topics in blogs

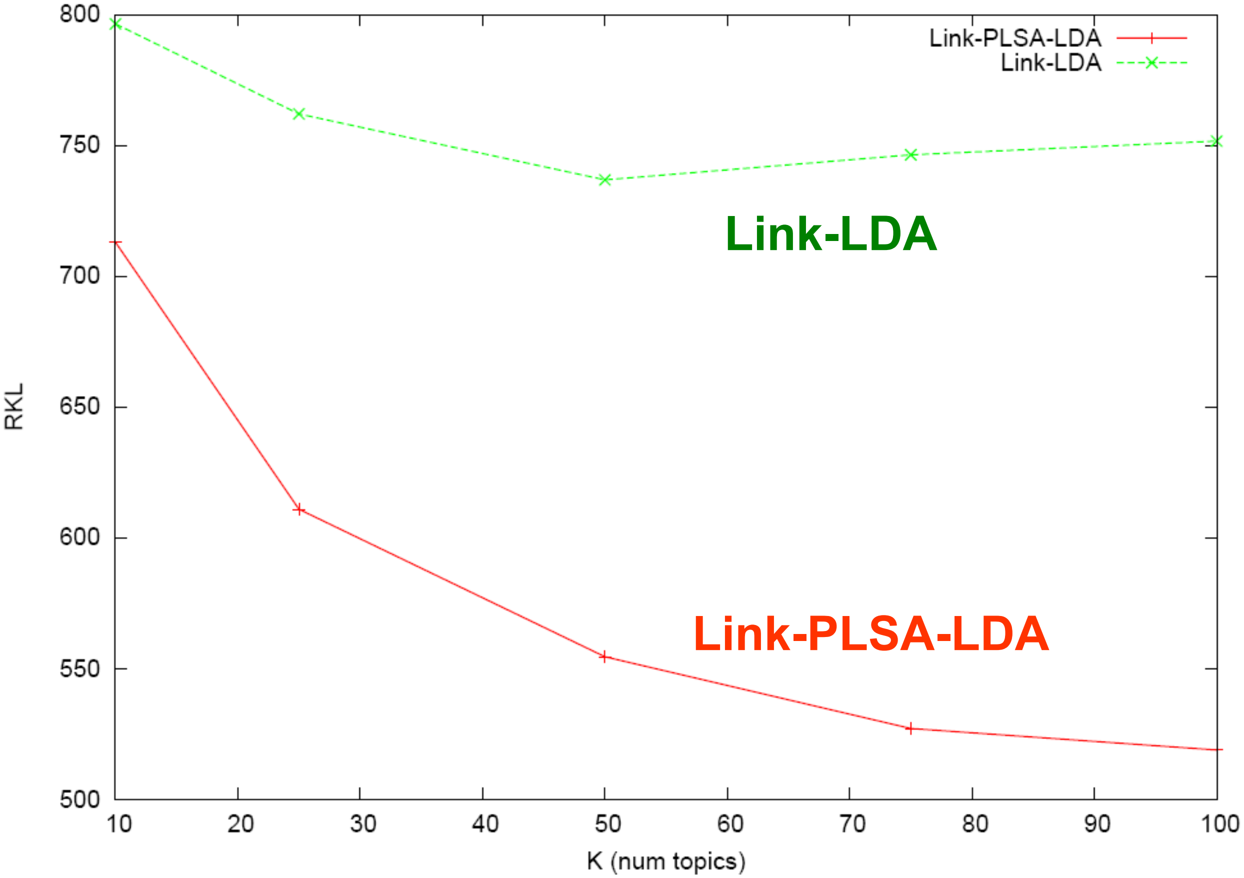
Model can answer questions like: which blogs are most likely to be cited when discussing topic *z*?

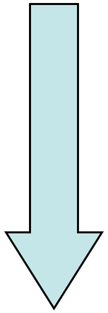


Topic 21 "CIA LEAK" **0.067**	Topic 7 "IRAQ WAR" **0.062**	Topic 16 "SUPREME COURT NOMINATIONS" **0.06**	Topic 20 "SEARCH ENGINE MARKET" **0.04**
TOP TOPICAL TERMS			
rove	will	robert	will
his	war	court	search
who	attack	bush	new
time	iraq	his	market
cooper	terrorist	supreme	post
karl	who	john	product
cia	world	nominate	brand
bush	terror	judge	permalink
know	muslim	will	time
report	america	conservative	yahoo
story	one	right	you
source	people	president	year
house	think	justice	comment
leak	bomb	nominee	company
plame	against	senate	business
TOP BLOG POSTS ON TOPIC			
billmon.org Whiskey Bar	willisms.com Iraq what might	themoderatevoice.com The Moderate Voice	edgeperspectives.typepad.com John Hagel
qando.net Free Markets & People	instapunk.com InstaPun***K	blogsforbush.com Blogs for Bush	.comparisonengines.com Comparison of Engines
captainsquartersblog .com, Captain's Quarters	jihadwatch.org Jihad Watch	michellemalkin.com Michelle Malkin	blogs.forrester.com Charlene Li's Blog
coldfury.com The Light Of Reason	thesharpener.net The Sharpener	captainsquartersblog.com Captain's Quarters	longtail.typepad.com The Long Tail
thismodernworld.com Tom Tomorrow	thedonovan.com Jonah's Military	wizbangblog.com Wizbang	.searchenginejournal.com Search Engine Journal

Topics in blogs

Model can be evaluated by predicting which links an author will include in a an article

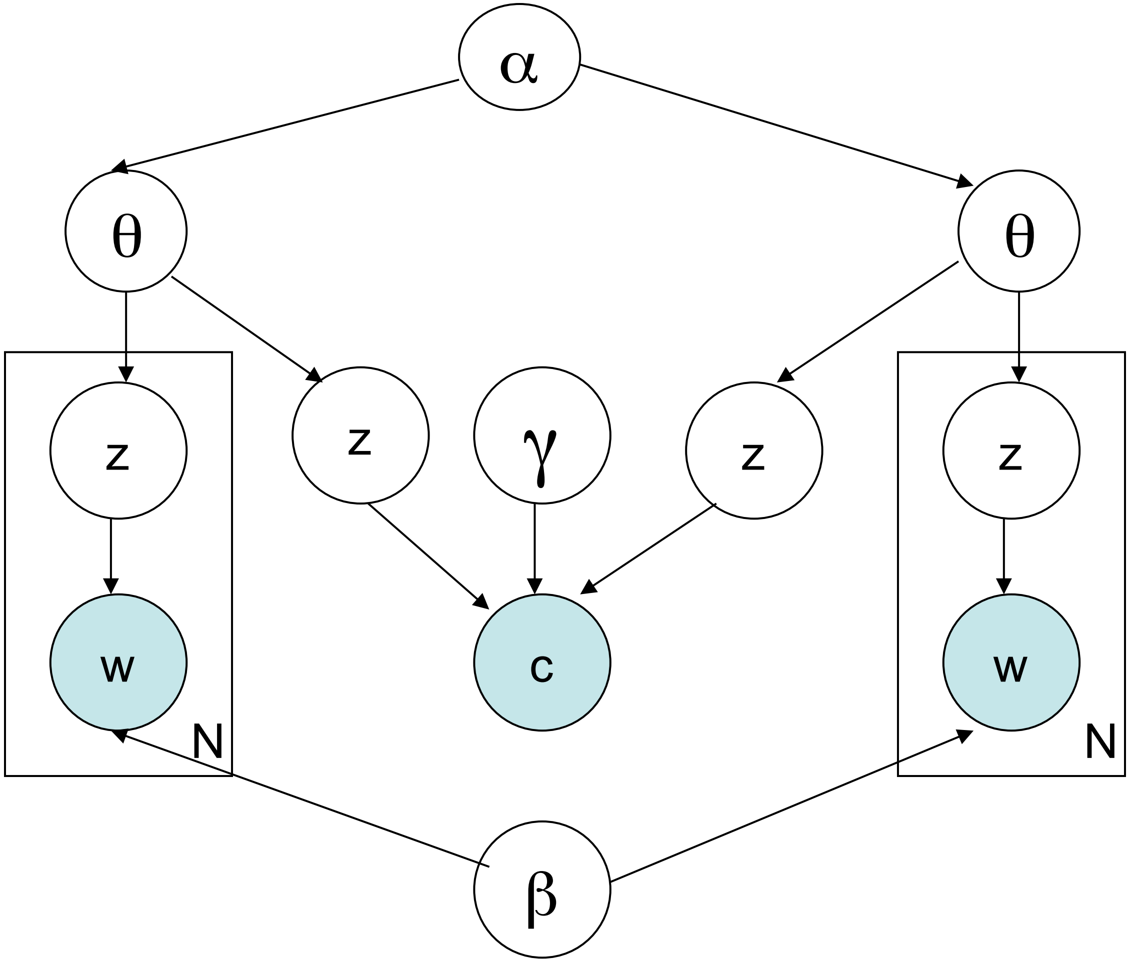




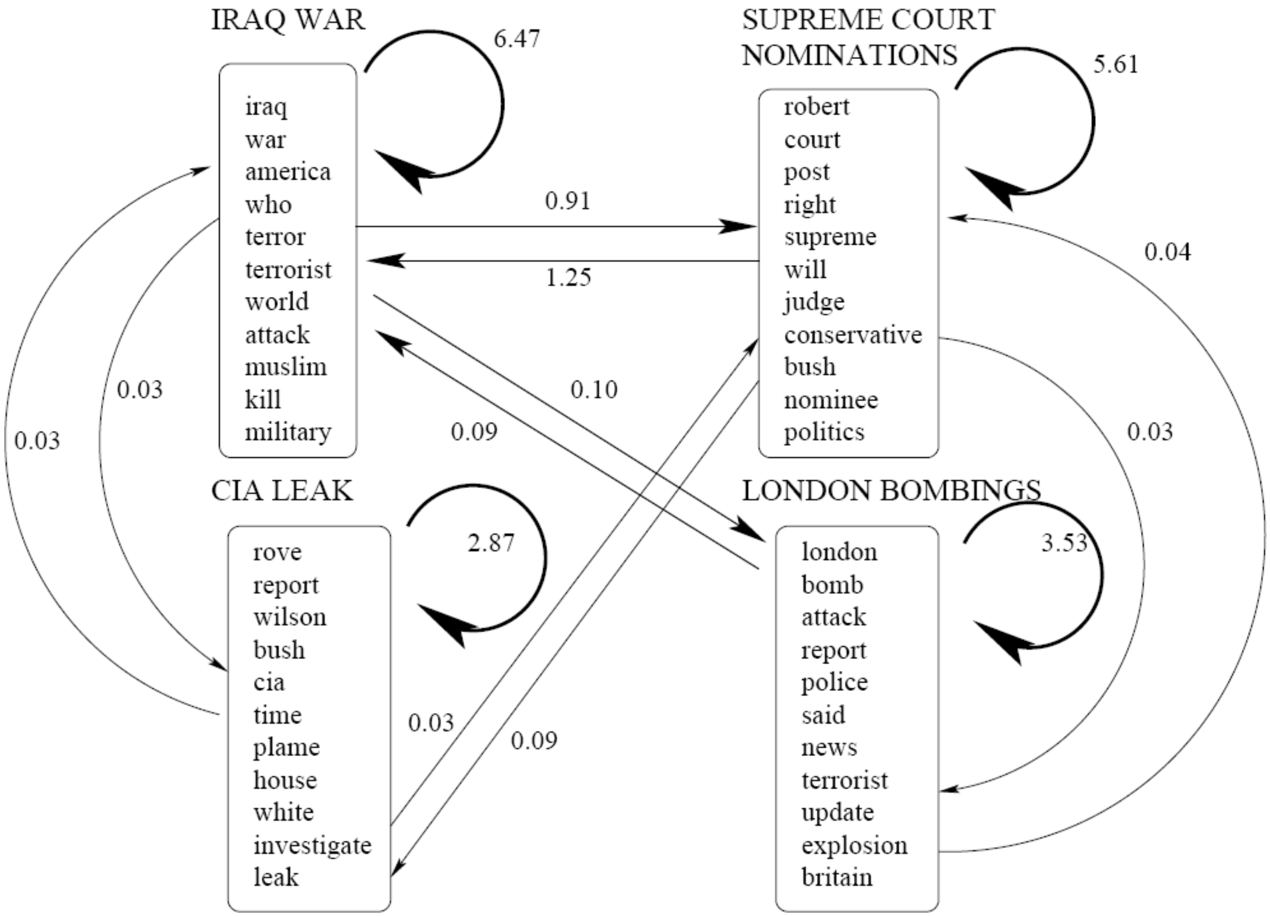
Lower is better

Another model: Pairwise Link-LDA

- LDA for both cited and citing documents
- Generate an *indicator* for *every pair* of docs
  - *Vs.* generating pairs of docs
- Link depends on the mixing components (θ's)
  - *stochastic block model*



Pairwise Link-LDA supports new inferences…



…but doesn’t perform better on link prediction

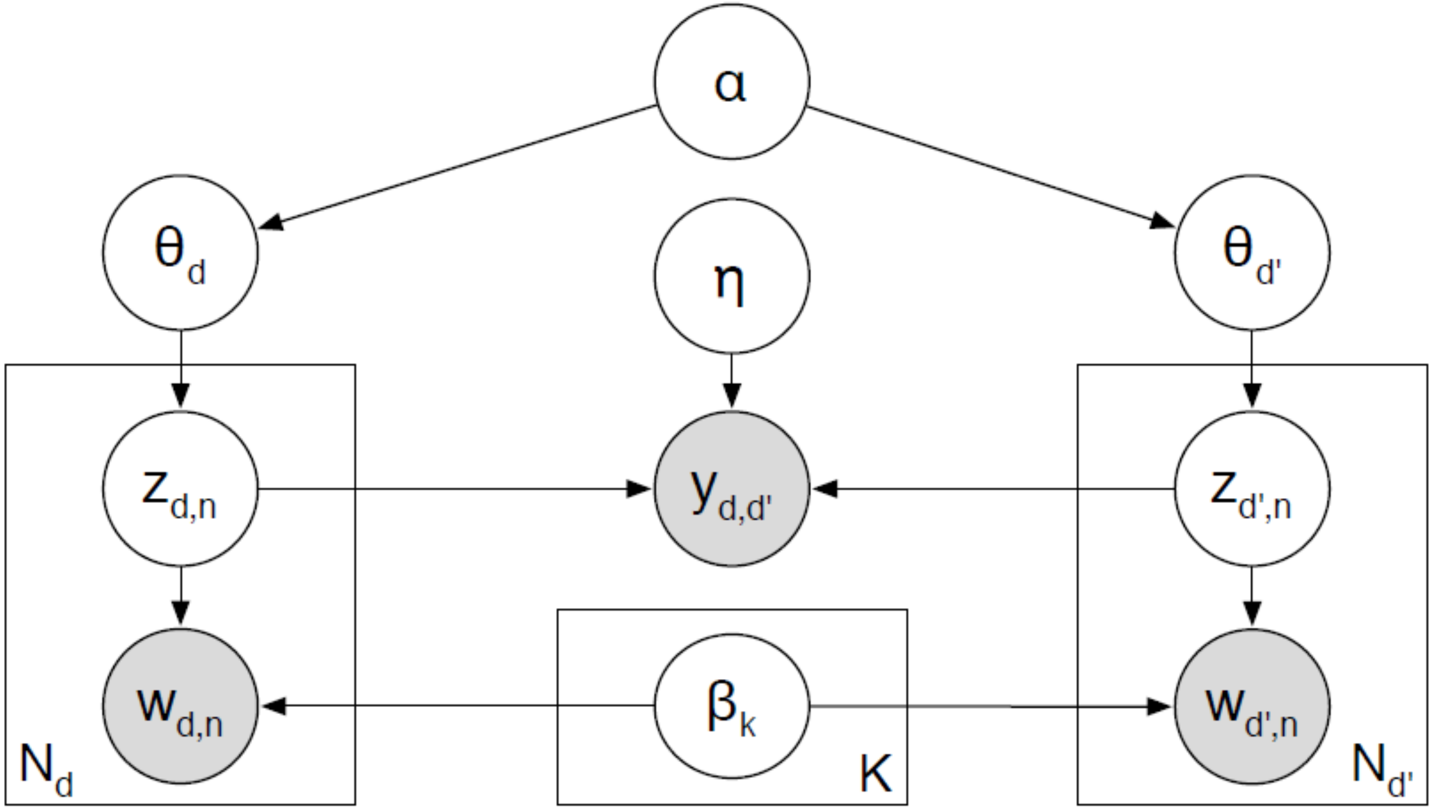
Outline

- Tools for analysis of text
  - Probabilistic models for text, communities, and time
    - Mixture models and LDA models for text
    - **LDA extensions to model hyperlink structure**
      - **Observation: these models can be used for many purposes…**
    - LDA extensions to model time
  - Alternative framework based on graph analysis to model time & community
- Discussion of results & challenges

Relational Topic Models for Document Networks

Jonathan Chang
Department of Electrical Engineering
Princeton University
Princeton, NJ 08544
jcone@princeton.edu

David M. Blei
Department of Computer Science
Princeton University
35 Olden St.
Princeton, NJ 08544
blei@cs.princeton.edu



$$\psi_\sigma(y=1) = \sigma(\boldsymbol{\eta}^{\mathrm{T}}(\overline{\mathbf{z}}_d \circ \overline{\mathbf{z}}_{d'}) + \nu),$$

Authors are using a number of clever tricks for inference….

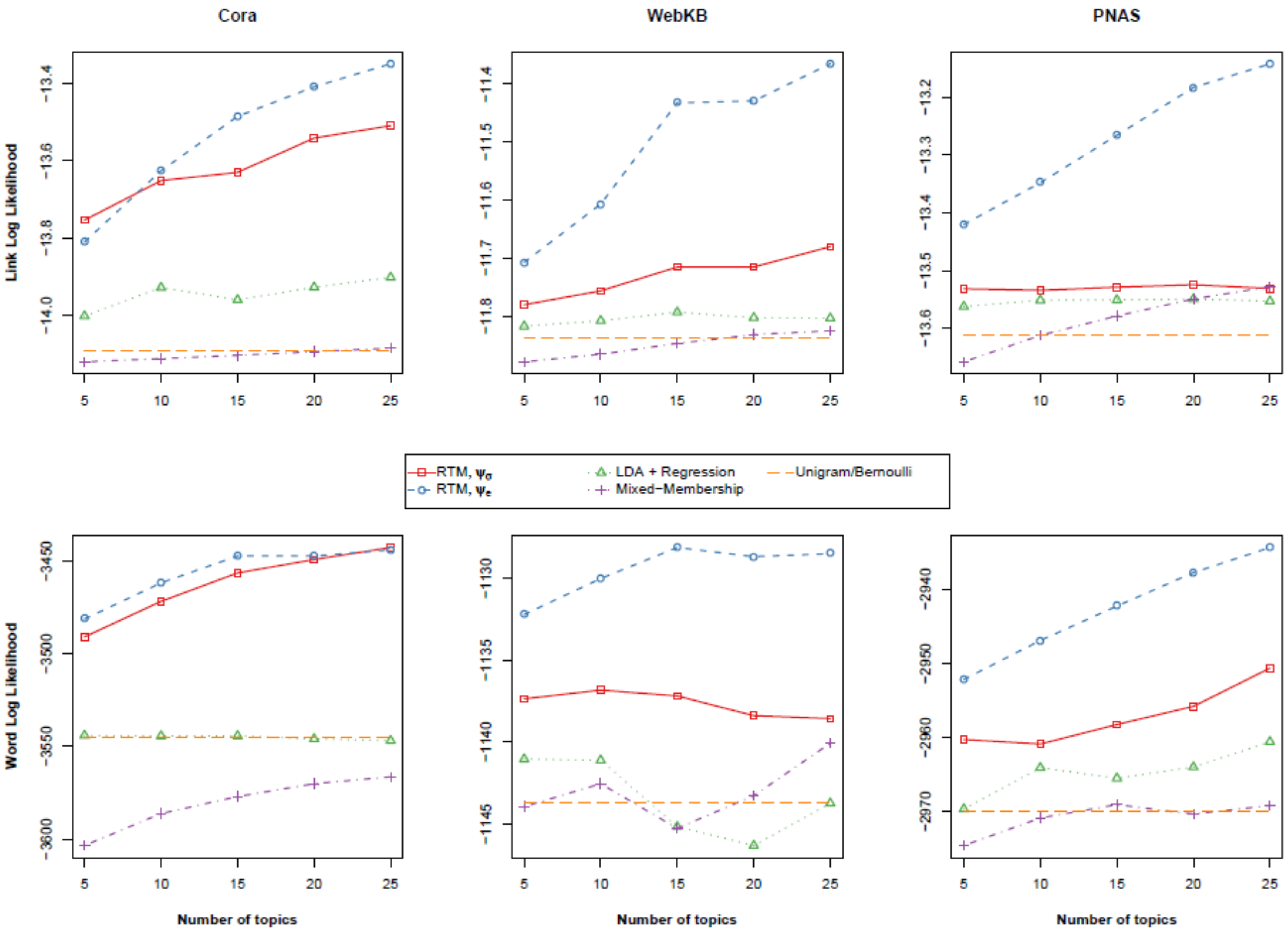
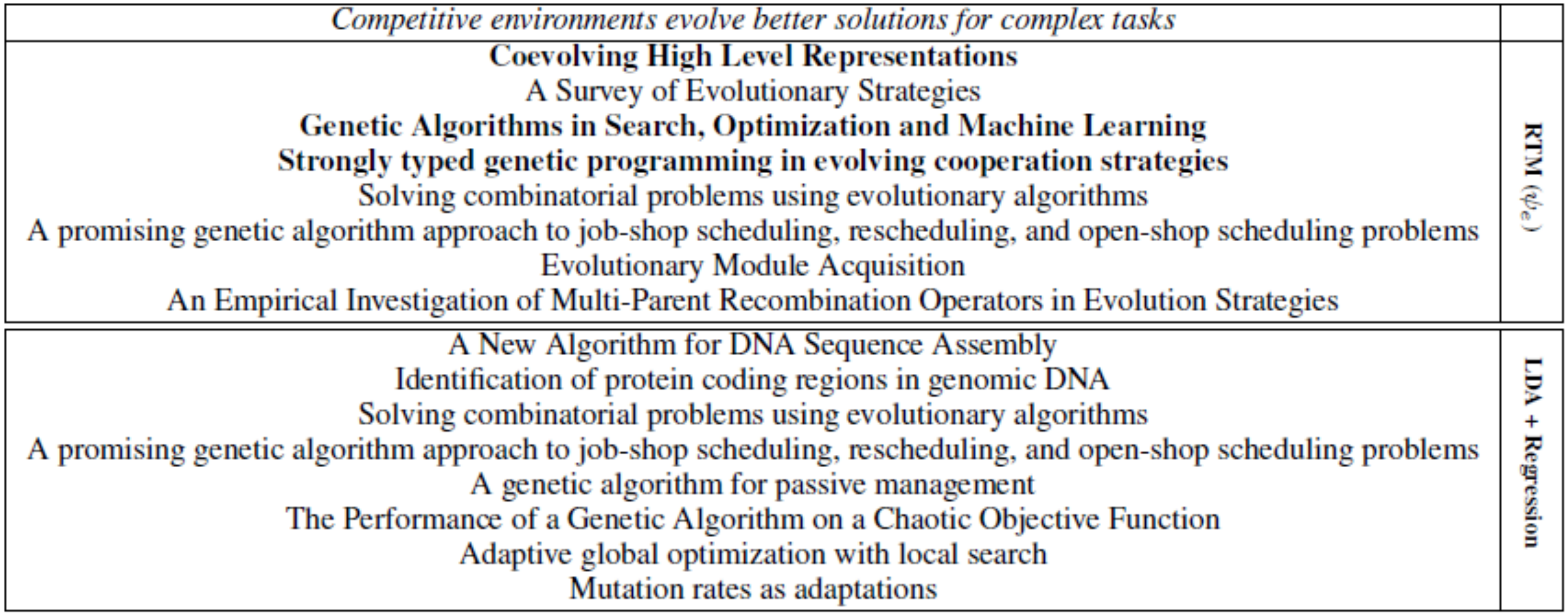
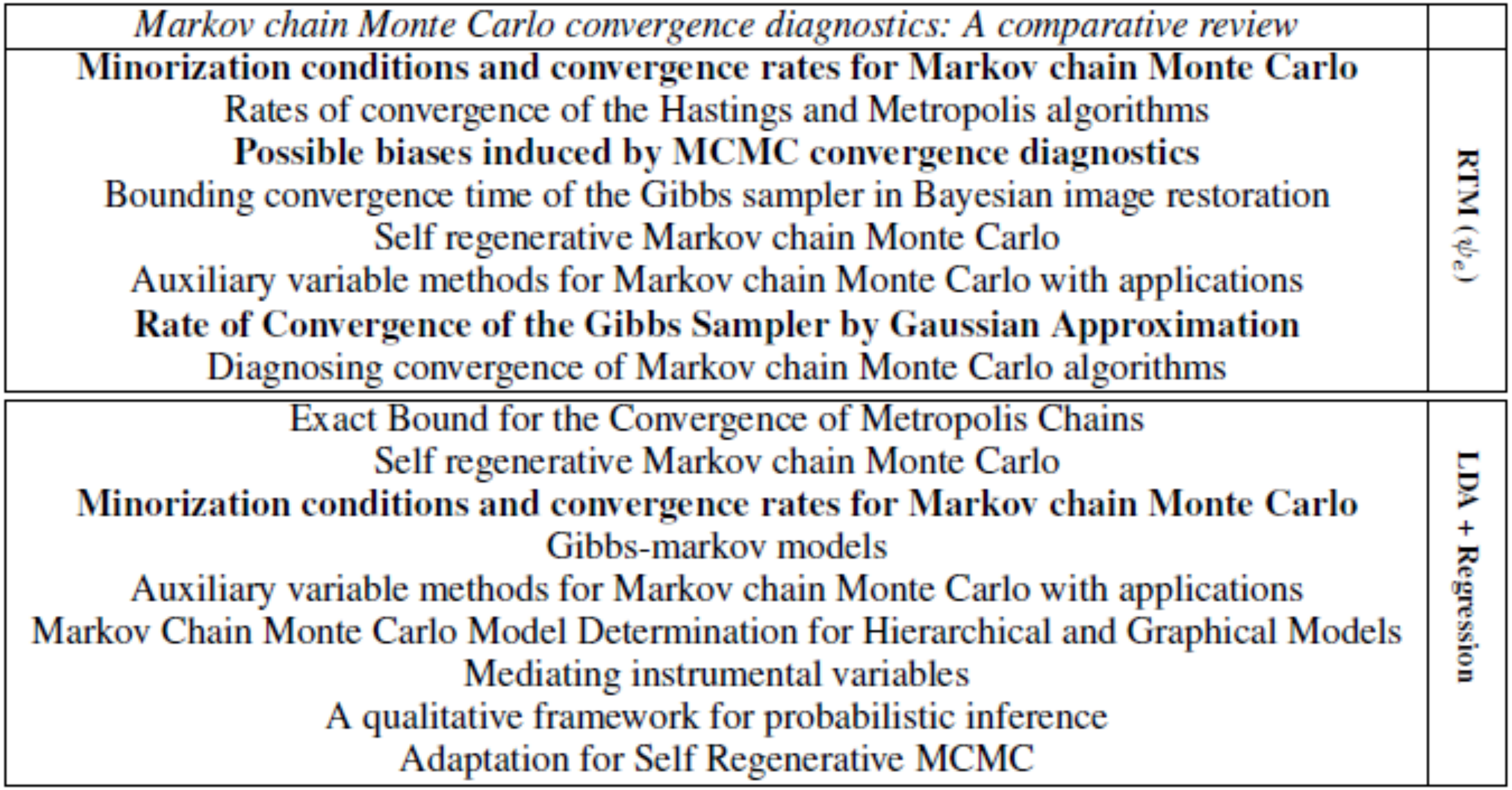


Figure 3: Average held-out predictive link log likelihood (top) and word log likelihood (bottom) as a function of the number of topics. For all three corpora, RTMs outperform baseline unigram, LDA, and "Mixed-Membership," which is the model of Nallapati et al. (2008).



Competitive environments evolve better solutions for complex tasks	
Coevolving High Level Representations	RTM (ψ_e)
A Survey of Evolutionary Strategies	
Genetic Algorithms in Search, Optimization and Machine Learning	
Strongly typed genetic programming in evolving cooperation strategies	
Solving combinatorial problems using evolutionary algorithms	
A promising genetic algorithm approach to job-shop scheduling, rescheduling, and open-shop scheduling problems	
Evolutionary Module Acquisition	
An Empirical Investigation of Multi-Parent Recombination Operators in Evolution Strategies	
A New Algorithm for DNA Sequence Assembly	LDA + Regression
Identification of protein coding regions in genomic DNA	
Solving combinatorial problems using evolutionary algorithms	
A promising genetic algorithm approach to job-shop scheduling, rescheduling, and open-shop scheduling problems	
A genetic algorithm for passive management	
The Performance of a Genetic Algorithm on a Chaotic Objective Function	
Adaptive global optimization with local search	
Mutation rates as adaptations	

Table 1: Top eight link predictions made by RTM (ψ_e) and LDA + Regression for two documents (italicized) from *Cora*. The models were trained with 10 topics. Boldfaced titles indicate actual documents cited by or citing each document. Over the whole corpus, RTM improves precision over LDA + Regression by 80% when evaluated on the first 20 documents retrieved.



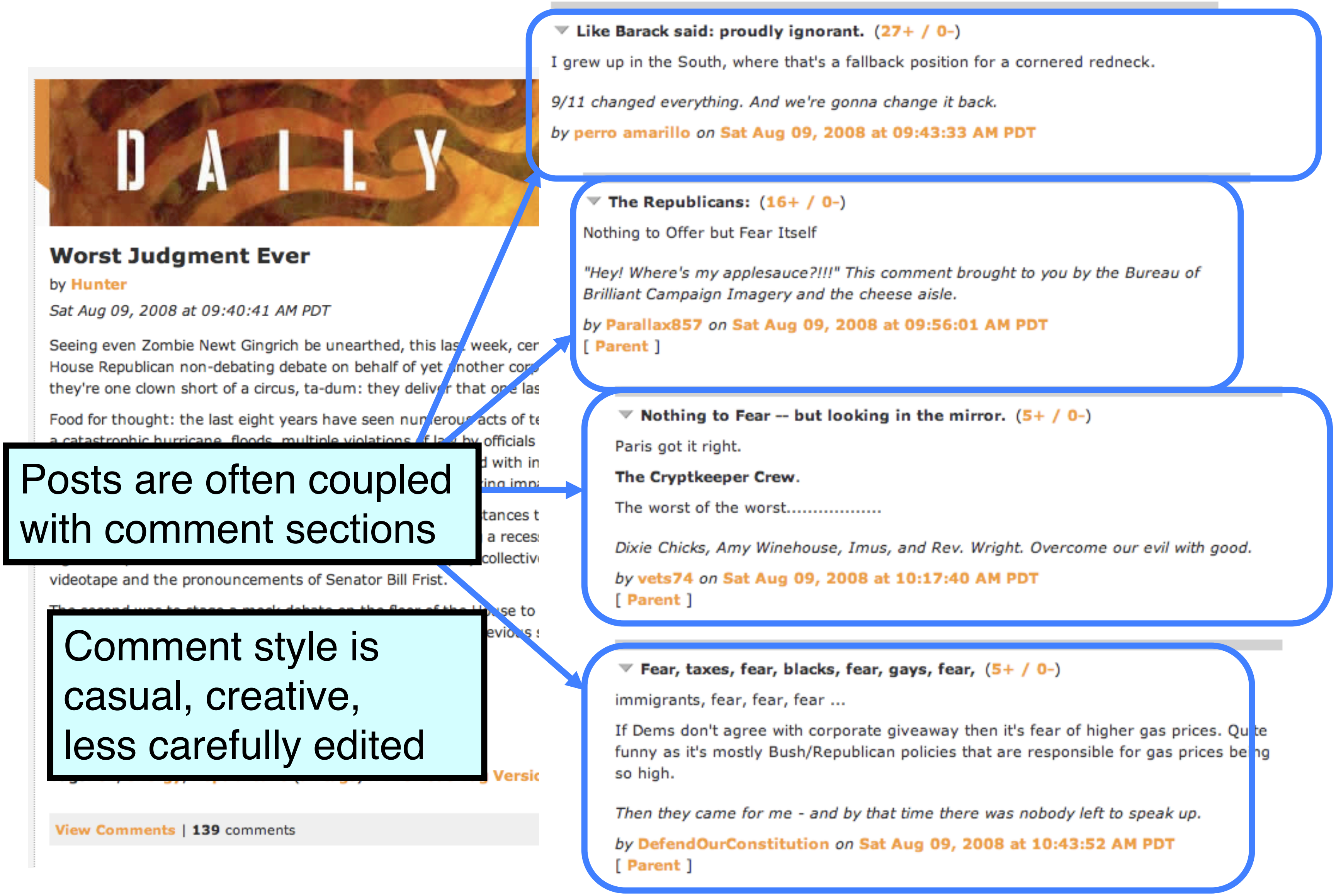
Markov chain Monte Carlo convergence diagnostics: A comparative review	
Minorization conditions and convergence rates for Markov chain Monte Carlo Rates of convergence of the Hastings and Metropolis algorithms **Possible biases induced by MCMC convergence diagnostics** Bounding convergence time of the Gibbs sampler in Bayesian image restoration Self regenerative Markov chain Monte Carlo Auxiliary variable methods for Markov chain Monte Carlo with applications **Rate of Convergence of the Gibbs Sampler by Gaussian Approximation** Diagnosing convergence of Markov chain Monte Carlo algorithms	**RTM (ψ_e)**
Exact Bound for the Convergence of Metropolis Chains Self regenerative Markov chain Monte Carlo **Minorization conditions and convergence rates for Markov chain Monte Carlo** Gibbs-markov models Auxiliary variable methods for Markov chain Monte Carlo with applications Markov Chain Monte Carlo Model Determination for Hierarchical and Graphical Models Mediating instrumental variables A qualitative framework for probabilistic inference Adaptation for Self Regenerative MCMC	**LDA + Regression**

Predicting Response to Political Blog Posts with Topic Models [NAACL ’09]

Tae Yano

Noah Smith

Political blogs and and comments



DAILY

Worst Judgment Ever

by Hunter

Sat Aug 09, 2008 at 09:40:41 AM PDT

Seeing even Zombie Newt Gingrich be unearthed, this las week, cer House Republican non-debating debate on behalf of yet another corp they're one clown short of a circus, ta-dum: they deliver that one las

Food for thought: the last eight years have seen numerous acts of te a catastrophic hurricane, floods, multiple violations of law by officials d with in ing imp tances t a reces collectiv videotape and the pronouncements of Senator Bill Frist.

The second was to stage a mock debate on the floor of the House to evious s

Versio

View Comments | 139 comments

Like Barack said: proudly ignorant. (27+ / 0-)

I grew up in the South, where that's a fallback position for a cornered redneck.

9/11 changed everything. And we're gonna change it back.

by perro amarillo *on* Sat Aug 09, 2008 at 09:43:33 AM PDT

The Republicans: (16+ / 0-)

Nothing to Offer but Fear Itself

"Hey! Where's my applesauce?!!!" This comment brought to you by the Bureau of Brilliant Campaign Imagery and the cheese aisle.

by Parallax857 *on* Sat Aug 09, 2008 at 09:56:01 AM PDT
[Parent]

Nothing to Fear -- but looking in the mirror. (5+ / 0-)

Paris got it right.

The Cryptkeeper Crew.

The worst of the worst..................

Dixie Chicks, Amy Winehouse, Imus, and Rev. Wright. Overcome our evil with good.

by vets74 *on* Sat Aug 09, 2008 at 10:17:40 AM PDT
[Parent]

Fear, taxes, fear, blacks, fear, gays, fear, (5+ / 0-)

immigrants, fear, fear, fear ...

If Dems don't agree with corporate giveaway then it's fear of higher gas prices. Quite funny as it's mostly Bush/Republican policies that are responsible for gas prices being so high.

Then they came for me - and by that time there was nobody left to speak up.

by DefendOurConstitution *on* Sat Aug 09, 2008 at 10:43:52 AM PDT
[Parent]

Posts are often coupled with comment sections

Comment style is casual, creative, less carefully edited

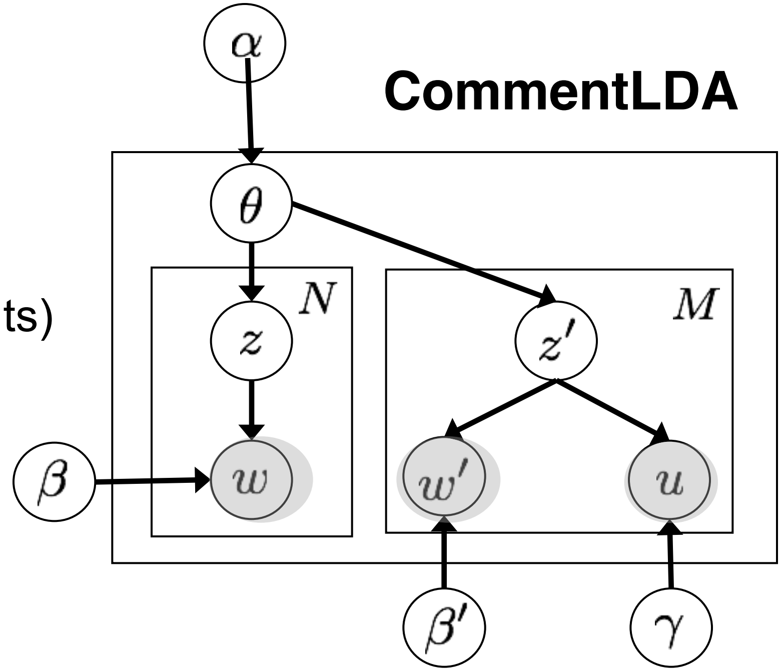
Political blogs and comments

- Most of the text associated with large "A-list" community blogs is comments
  - 5-20x as many words in comments as in text for the 5 sites considered in Yano et al.
- A large part of socially-created commentary in the blogosphere is comments.
  - Not blog → blog hyperlinks
- Comments *do not* just echo the post

Modeling political blogs

Our political blog model:

z, z` = topic
w = word (in post)
w`= word (in comments)
u = user

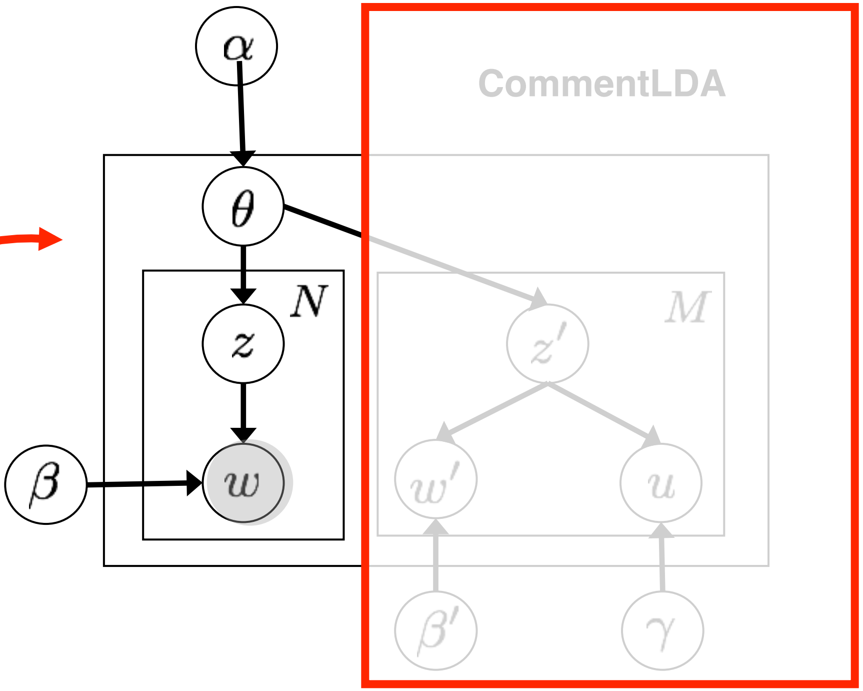


D = # of documents; **N** = # of words in post; **M** = # of words in comments

Modeling political blogs

Our proposed political blog model:

LHS is vanilla **LDA**

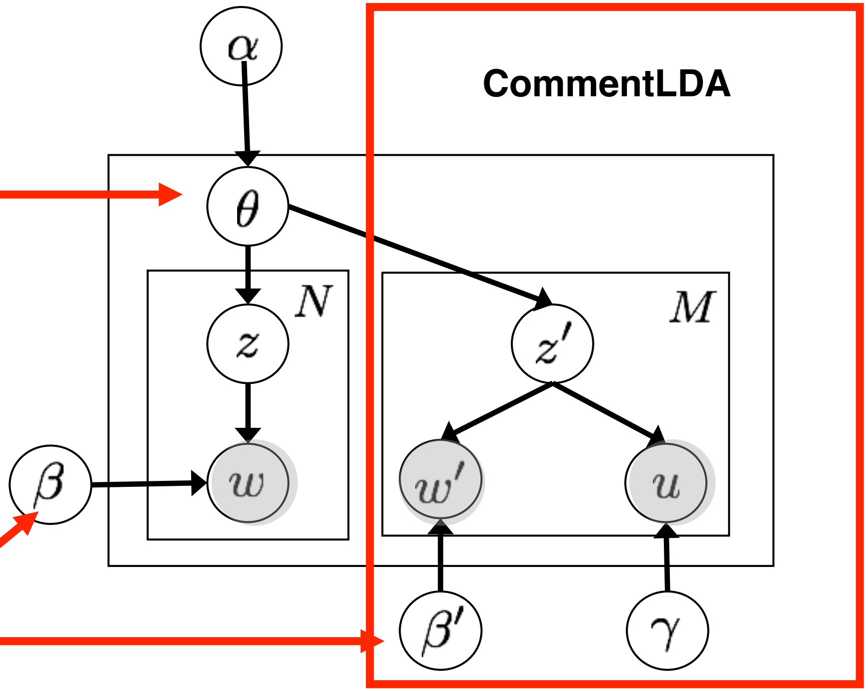


D = # of documents; **N** = # of words in post; **M** = # of words in comments

Modeling political blogs

Our proposed political blog model:

RHS to capture the generation of reaction separately from the post body



Two chambers share the same topic-mixture

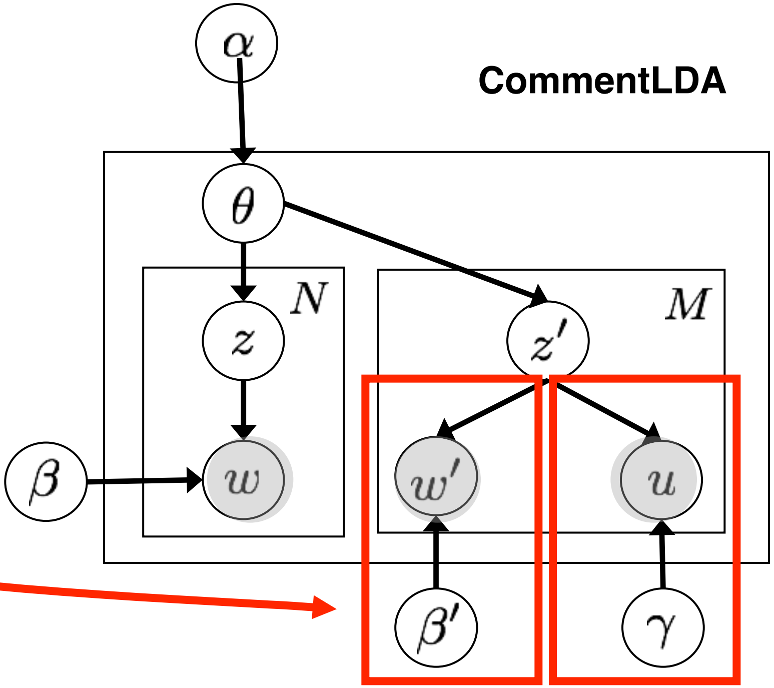
Two separate sets of word distributions

D = # of documents; **N** = # of words in post; **M** = # of words in comments

Modeling political blogs

Our proposed political blog model:

User IDs of the commenters as a part of comment text



generate the words in the comment section

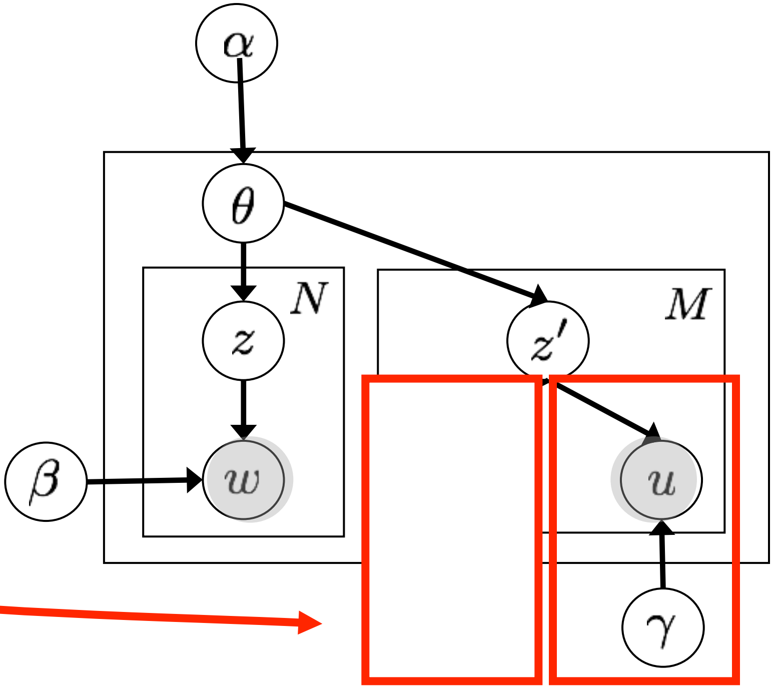
D = # of documents; **N** = # of words in post; **M** = # of words in comments

Modeling political blogs

Another model we tried:

Took out the words from the comment section!

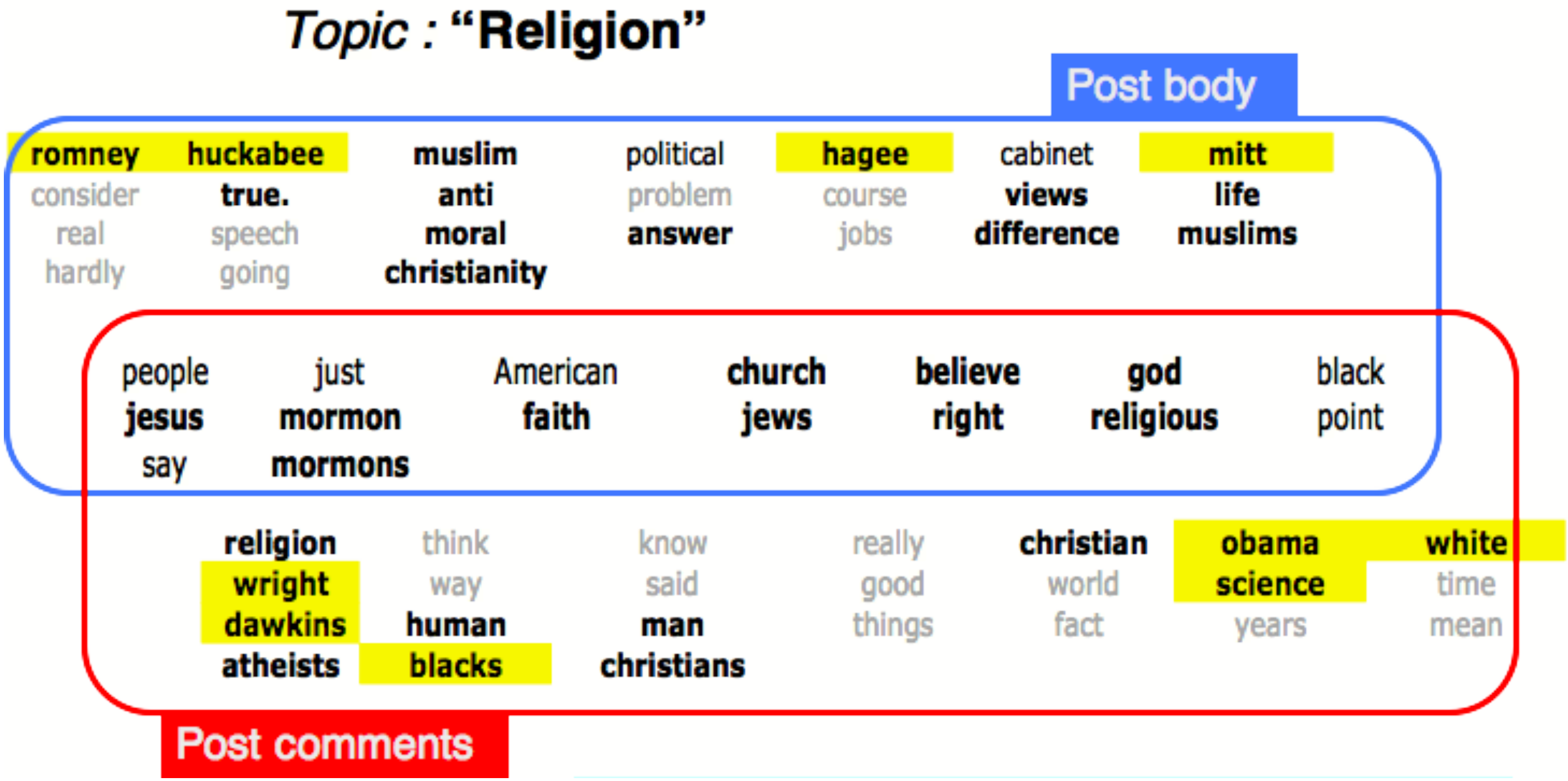
This is a model agnostic to the words in the comment section!



The model is *structurally* equivalent to the LinkLDA from (Erosheva et al., 2004)

D = # of documents; **N** = # of words in post; **M** = # of words in comments

Topic discovery - Matthew Yglesias (MY) site



Topic : **"Religion"**

Post body

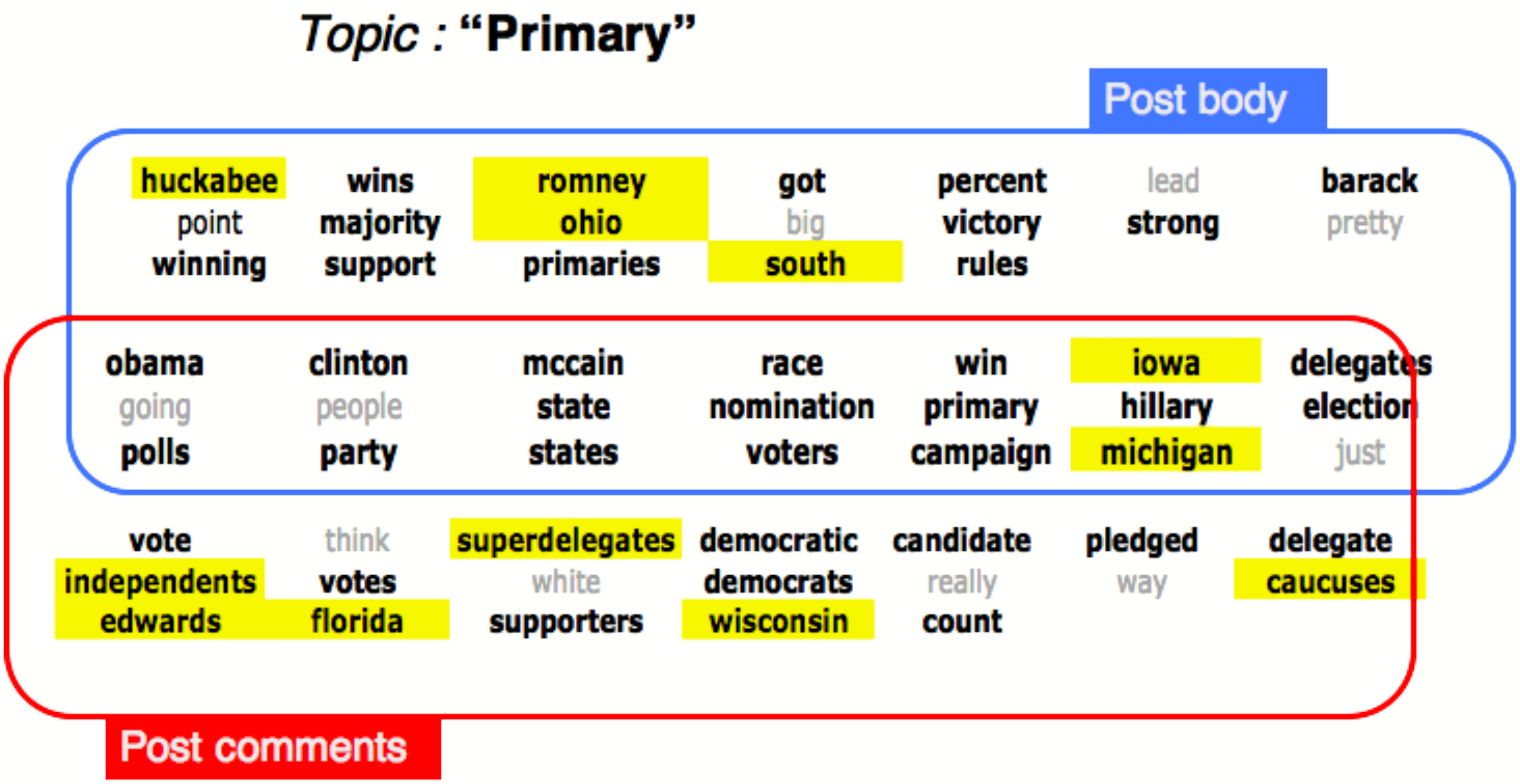
romney huckabee muslim political hagee cabinet mitt
consider true. anti problem course views life
real speech moral answer jobs difference muslims
hardly going christianity

people just American church believe god black
jesus mormon faith jews right religious point
say mormons

religion think know really christian obama white
wright way said good world science time
dawkins human man things fact years mean
atheists blacks christians

Post comments

Topic discovery - Matthew Yglesias (MY) site



Topic : "Primary"

Post body

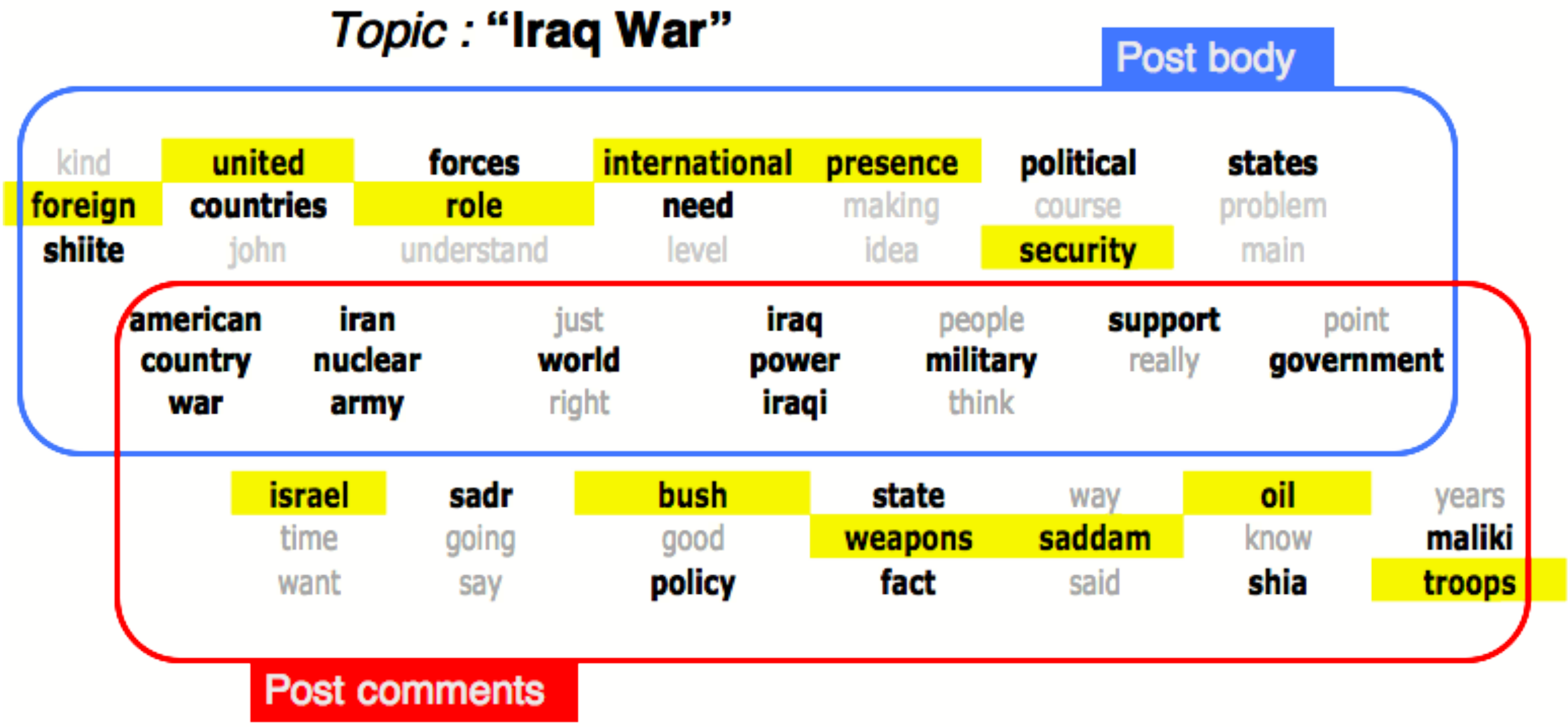
huckabee point winning
wins majority support
romney ohio primaries
got big south
percent victory rules
lead strong
barack pretty

obama going polls
clinton people party
mccain state states
race nomination voters
win primary campaign
iowa hillary michigan
delegates election just

vote independents edwards
think votes florida
superdelegates white supporters
democratic democrats wisconsin
candidate really count
pledged way
delegate caucuses

Post comments

Topic discovery - Matthew Yglesias (MY) site



Topic : **"Iraq War"**

Post body

kind united forces international presence political states
foreign countries role need making course problem
shiite john understand level idea security main

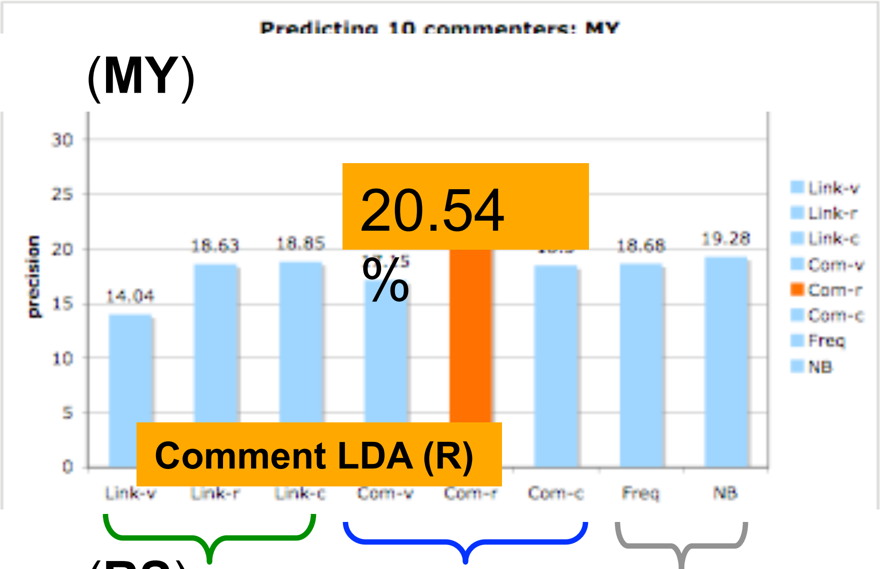
american iran just iraq people support point
country nuclear world power military really government
war army right iraqi think

israel sadr bush state way oil years
time going good weapons saddam know maliki
want say policy fact said shia troops

Post comments

Comment prediction

(MY)

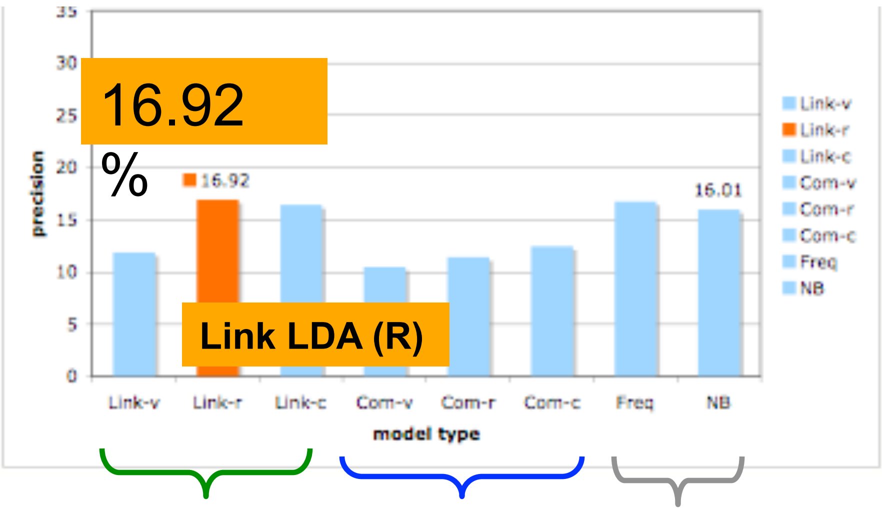


20.54 %

Comment LDA (R)

- **LinkLDA** and **CommentLDA** consistently outperform **baseline** models
- Neither consistently outperforms the other.

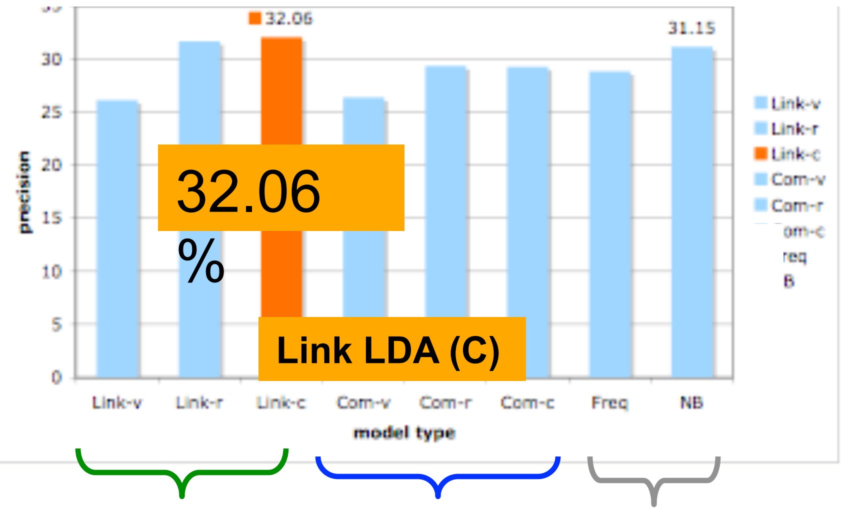
(RS)



16.92 %

Link LDA (R)

(CB)



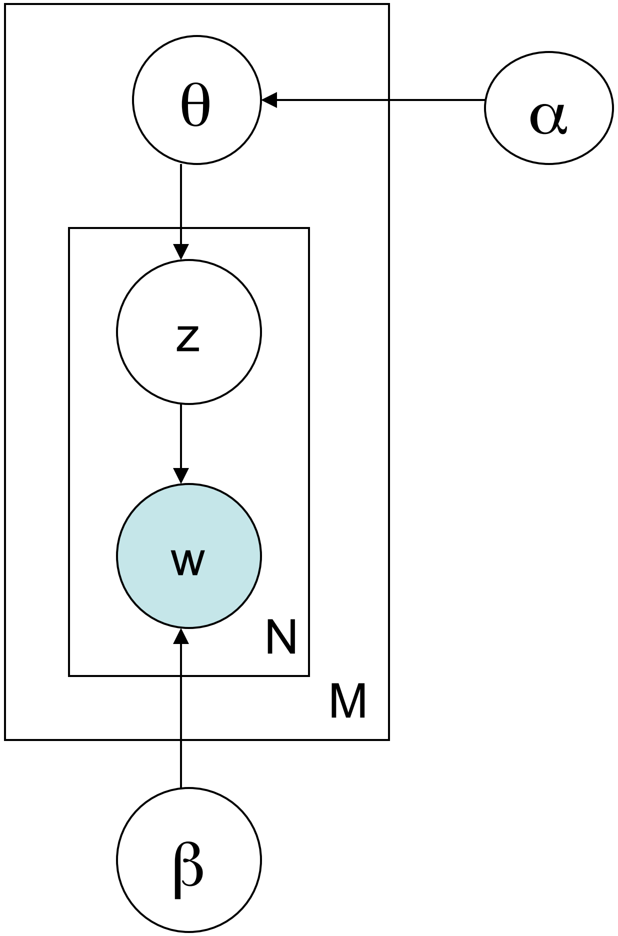
32.06 %

Link LDA (C)

user prediction: Precision at top 10

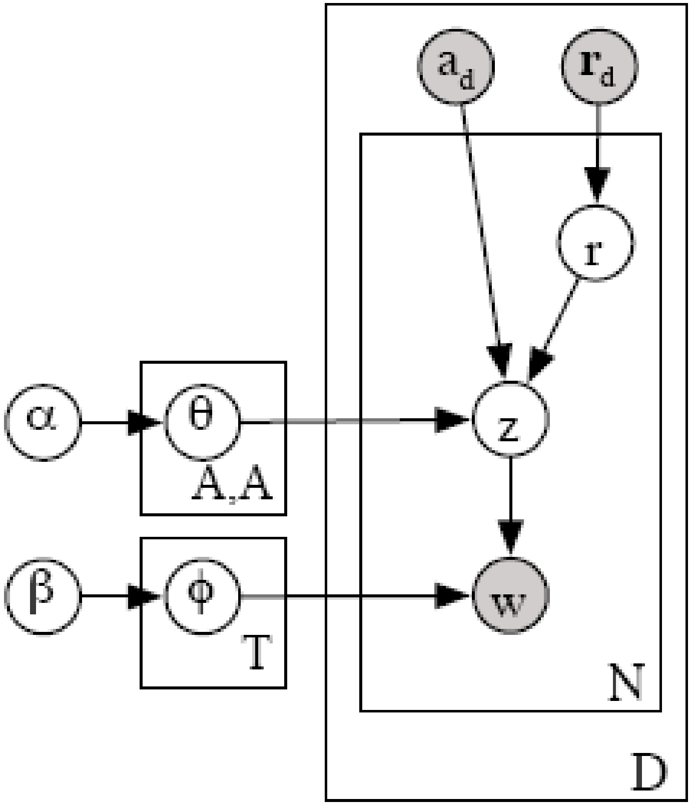
From left to right: Link LDA(-v, -r,-c) Cmnt LDA (-v, -r, -c), Baseline (Freq, NB)

Document modeling with Latent Dirichlet Allocation (LDA)



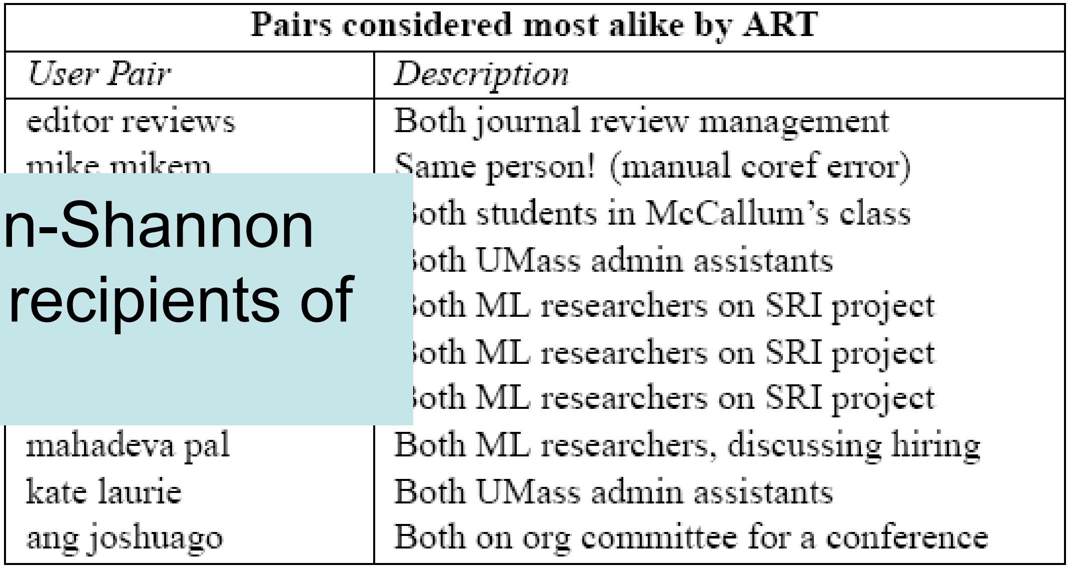
- For each document d = 1,···,M
  - Generate θ_d ~ Dir(. | α)
  - For each position n = 1,···, N_d
    - generate z_n ~ Mult(. | θ_d)
    - generate w_n ~ Mult(. | β_{z_n})

Author-Topic-Recipient model for email data [McCallum, Corrada-Emmanuel,Wang, ICJAI'05]

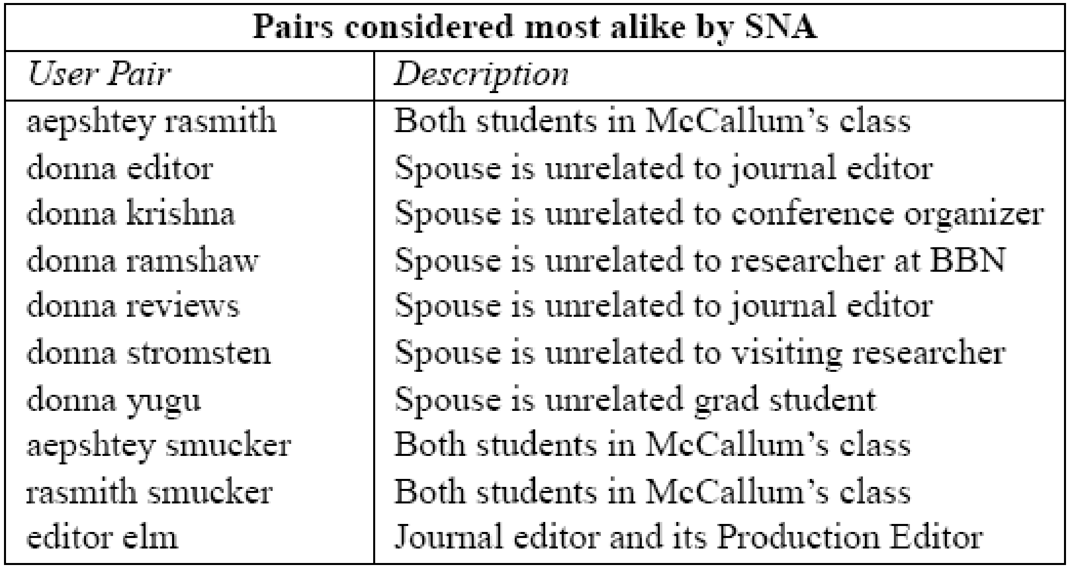


Author-Topic-Recipient model for email data [McCallum, Corrada-Emmanuel,Wang, ICJAI'05]

“SNA” = Jensen-Shannon divergence for recipients of messages



Pairs considered most alike by ART	
User Pair	*Description*
editor reviews	Both journal review management
mike mikem	Same person! (manual coref error)
	oth students in McCallum's class
	oth UMass admin assistants
	oth ML researchers on SRI project
	oth ML researchers on SRI project
	oth ML researchers on SRI project
mahadeva pal	Both ML researchers, discussing hiring
kate laurie	Both UMass admin assistants
ang joshuago	Both on org committee for a conference



Pairs considered most alike by SNA	
User Pair	*Description*
aepshtey rasmith	Both students in McCallum's class
donna editor	Spouse is unrelated to journal editor
donna krishna	Spouse is unrelated to conference organizer
donna ramshaw	Spouse is unrelated to researcher at BBN
donna reviews	Spouse is unrelated to journal editor
donna stromsten	Spouse is unrelated to visiting researcher
donna yugu	Spouse is unrelated grad student
aepshtey smucker	Both students in McCallum's class
rasmith smucker	Both students in McCallum's class
editor elm	Journal editor and its Production Editor

Modeling Citation Influences

Unsupervised Prediction of Citation Influences

Laura Dietz

Steffen Bickel

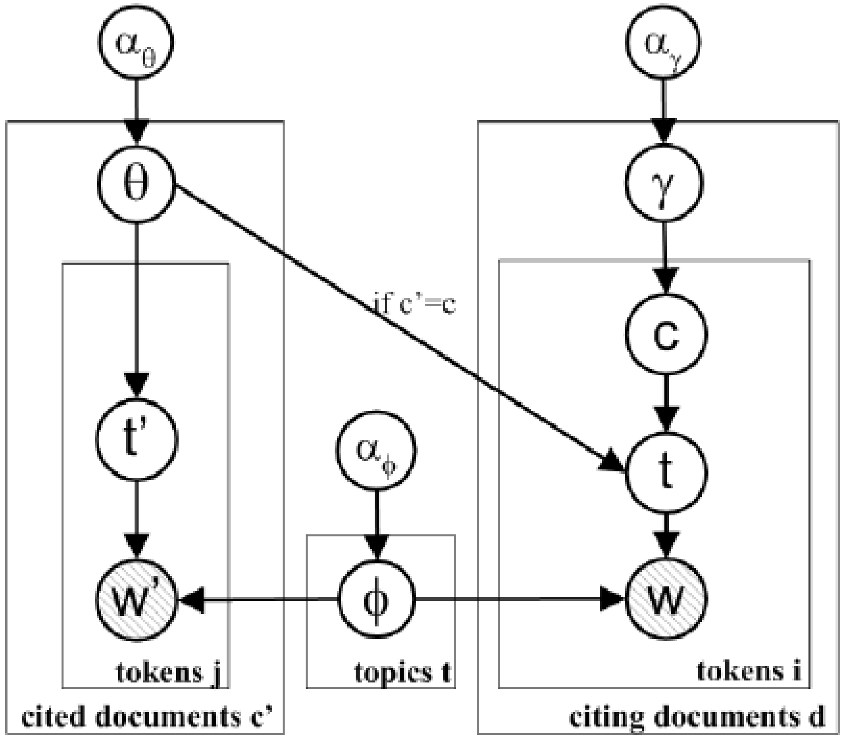
Tobias Scheffer

Max Planck Institute for Computer Science, Saarbrücken, Germany

Modeling Citation Influences

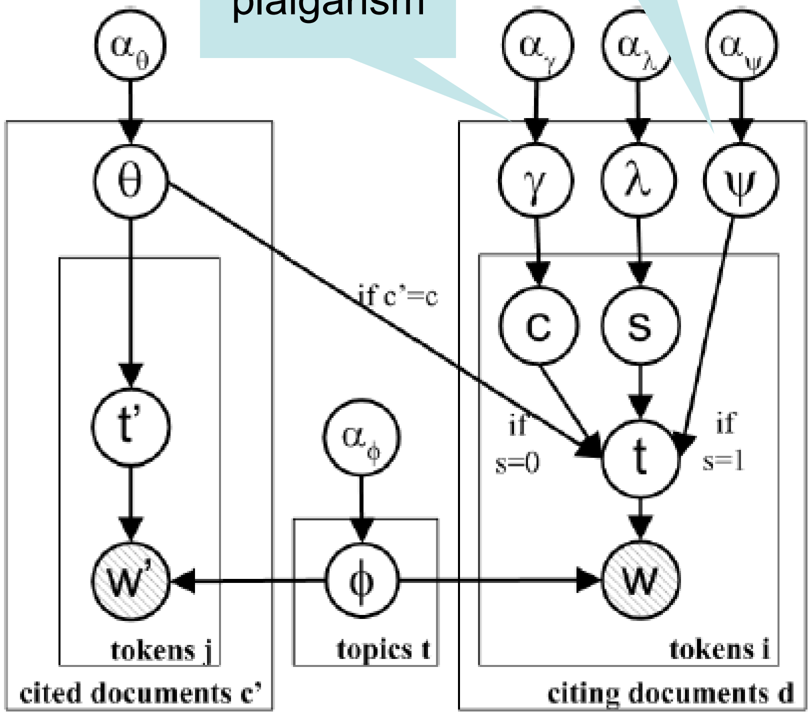
[Dietz, Bickel, Scheffer, ICML 2007]

- Copycat model of citation influence



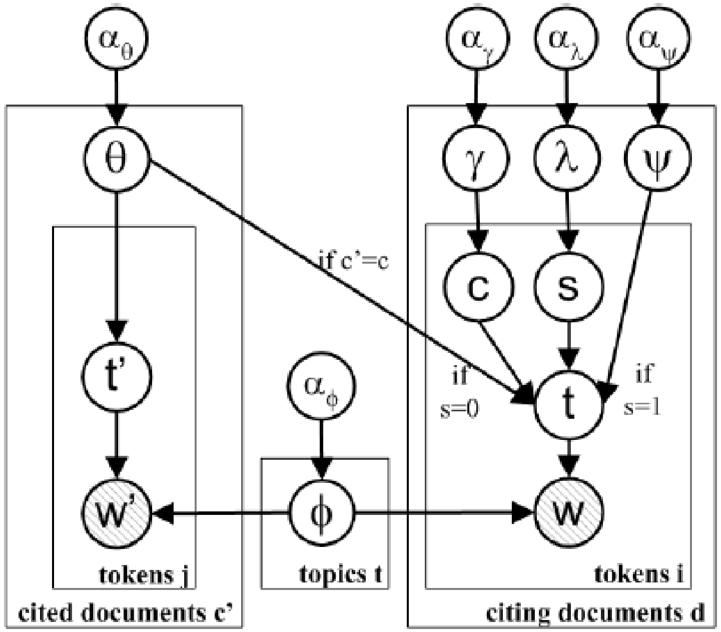
innovation

plaigarism



c is a cited document

s is a coin toss to mix γ and ψ



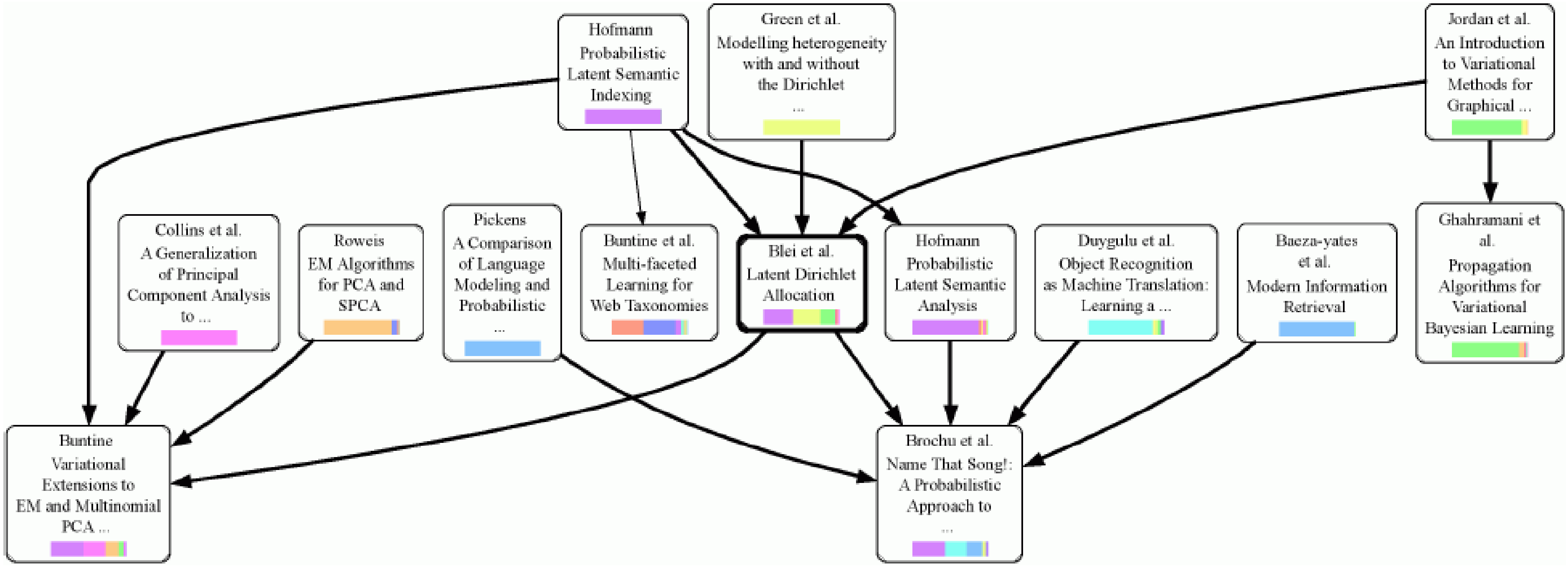
s is a coin toss to mix γ and ψ

- for all citing documents $d \in D$ do
  - draw a citation mixture $\gamma_d = p(c|d)|_{L(d)} \sim dirichlet(\vec{\alpha}_\gamma)$[1] restricted to the publications c cited by this publication d
  - draw an innovation topic mixture $\psi_d = p(t|d) \sim dirichlet(\vec{\alpha}_\psi)$
  - draw the proportion between tokens associated with citations and those associated with the innovation topic mixture $\lambda_d = p(s = 0|d) \sim beta(\alpha_{\lambda_\theta}, \alpha_{\lambda_\psi})$
  - for all tokens i do
    - toss a coin $s_{d,i} \sim bernoulli(\lambda_d)$
    - if $s_{d,i} = 0$
      - draw a cited document $c_{d,i} \sim multi(\gamma_d)$
      - draw a topic $t_{d,i} \sim multi(\theta_{c_{d,i}})$ from the cited document's topic mixture
    - else $(s_{d,i} = 1)$
      - draw the topic $t_{d,i} \sim multi(\psi_d)$ from the innovation topic mixture
    - draw a word $w_{d,i} \sim multi(\phi_{t_{d,i}})$ from the topic specific word distribution

Modeling Citation Influences

[Dietz, Bickel, Scheffer, ICML 2007]

- Citation influence graph for LDA paper



Modeling Citation Influences

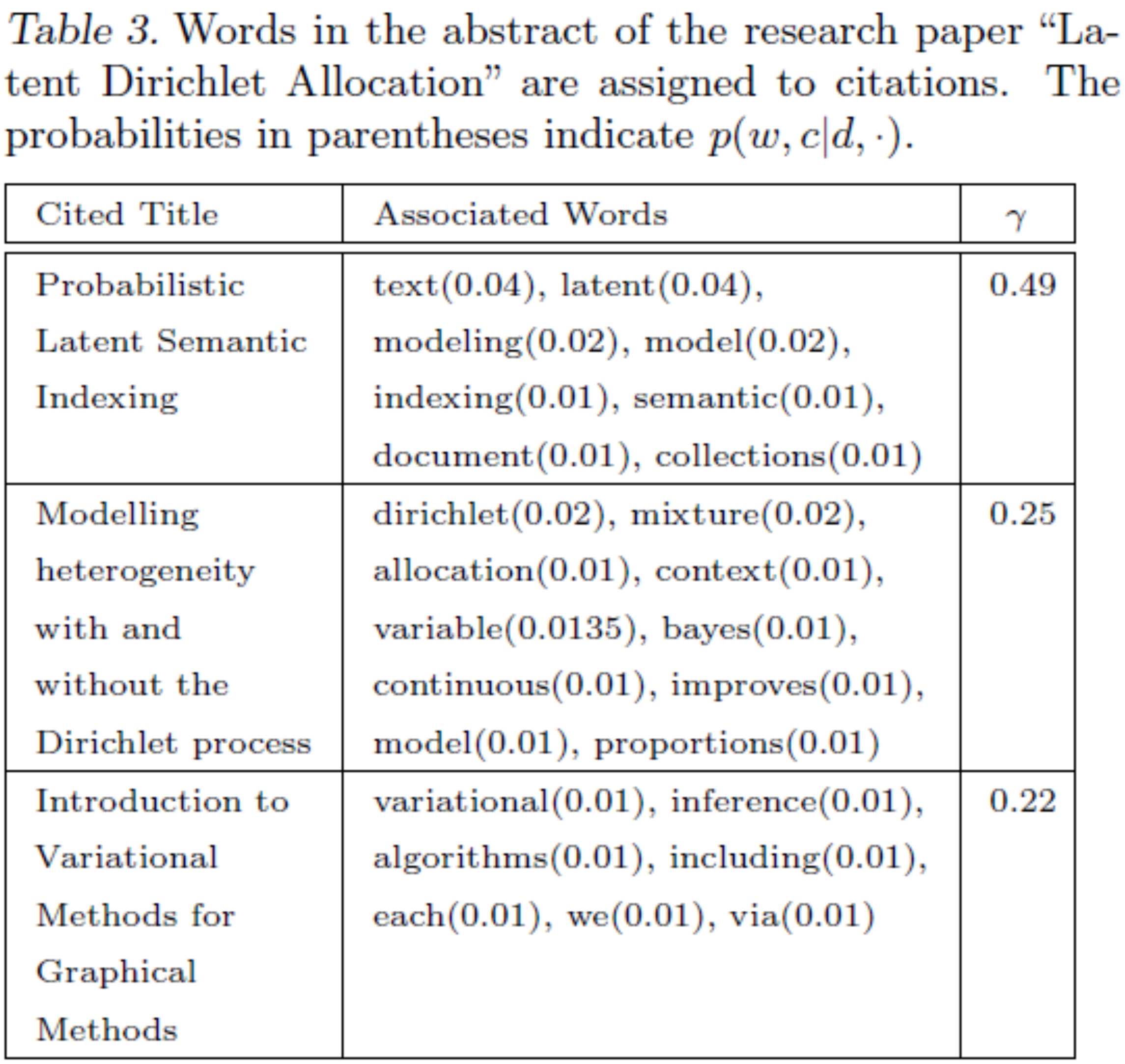


Table 3. Words in the abstract of the research paper "Latent Dirichlet Allocation" are assigned to citations. The probabilities in parentheses indicate $p(w, c|d, \cdot)$.

Cited Title	Associated Words	γ
Probabilistic Latent Semantic Indexing	text(0.04), latent(0.04), modeling(0.02), model(0.02), indexing(0.01), semantic(0.01), document(0.01), collections(0.01)	0.49
Modelling heterogeneity with and without the Dirichlet process	dirichlet(0.02), mixture(0.02), allocation(0.01), context(0.01), variable(0.0135), bayes(0.01), continuous(0.01), improves(0.01), model(0.01), proportions(0.01)	0.25
Introduction to Variational Methods for Graphical Methods	variational(0.01), inference(0.01), algorithms(0.01), including(0.01), each(0.01), we(0.01), via(0.01)	0.22

Modeling Citation Influences

User study: self-reported citation influence on Likert scale

LDA-post is Prob(cited doc| paper)

LDA-js is Jensen-Shannon dist in topic space

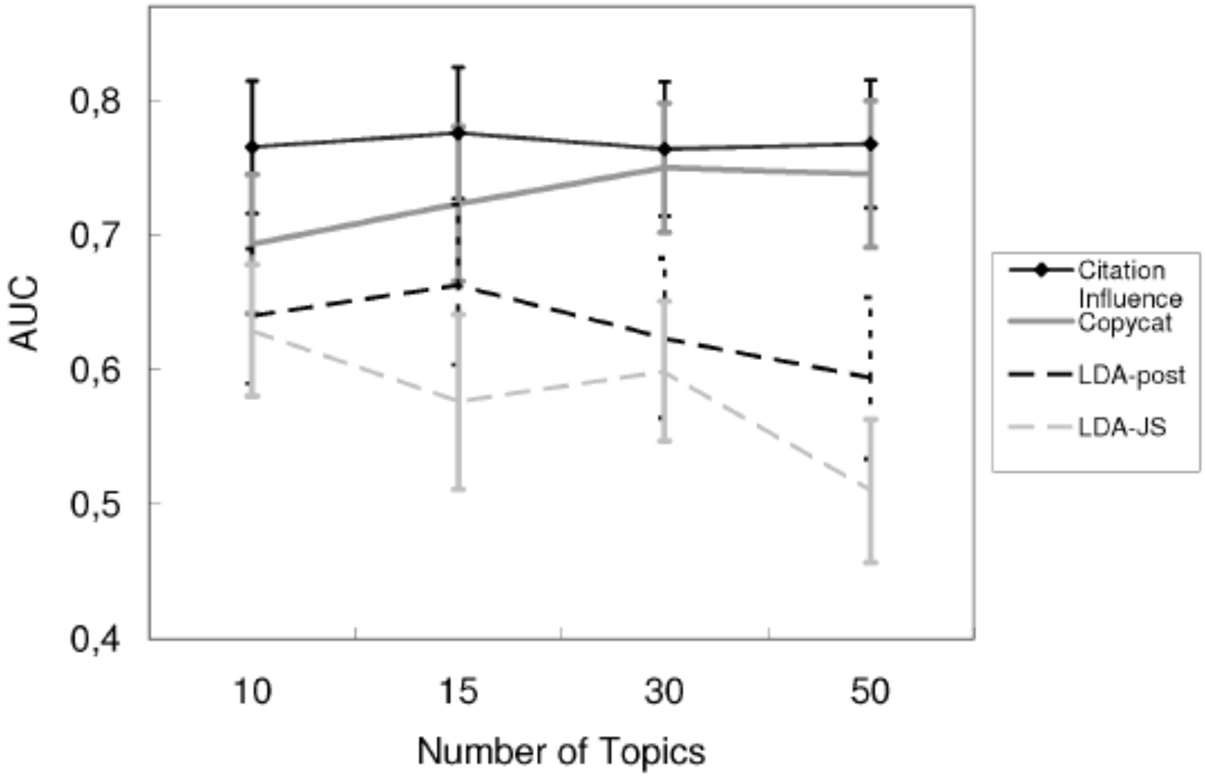


Figure 4. Predictive performance of the models. The error bars indicate the standard error of the AUC values averaged over the citing publications.