Latent-variable models of influence and decision-making

Sean Gerrish
Princeton University
Computer Science Department

29 February 2012

"Most of the Big Data surge is data in the wild – unruly stuff like words, images and video on the Web... It is called unstructured data and is not typically grist for traditional databases."

The Age of Big Data. The New York Times. Feb 11, 2012

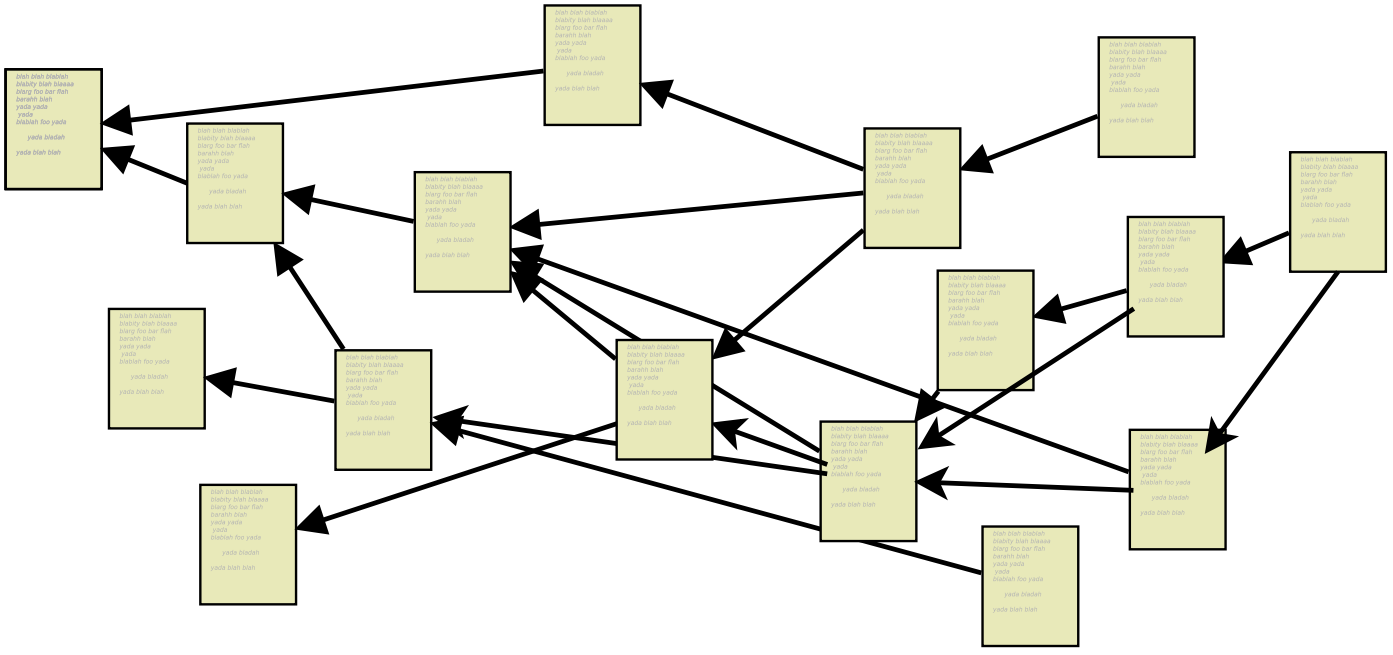
Overview

Machine learning is making it possible to understand what is going on in large collections of text documents.

- Understanding the main patterns of influence and events in collections of text documents
  - How to find influential text documents
  - How to discover international relations using news articles

- Better understanding government
  - Understanding lawmakers' issue preferences
  - How to predict votes on new bills

Common approach: citations

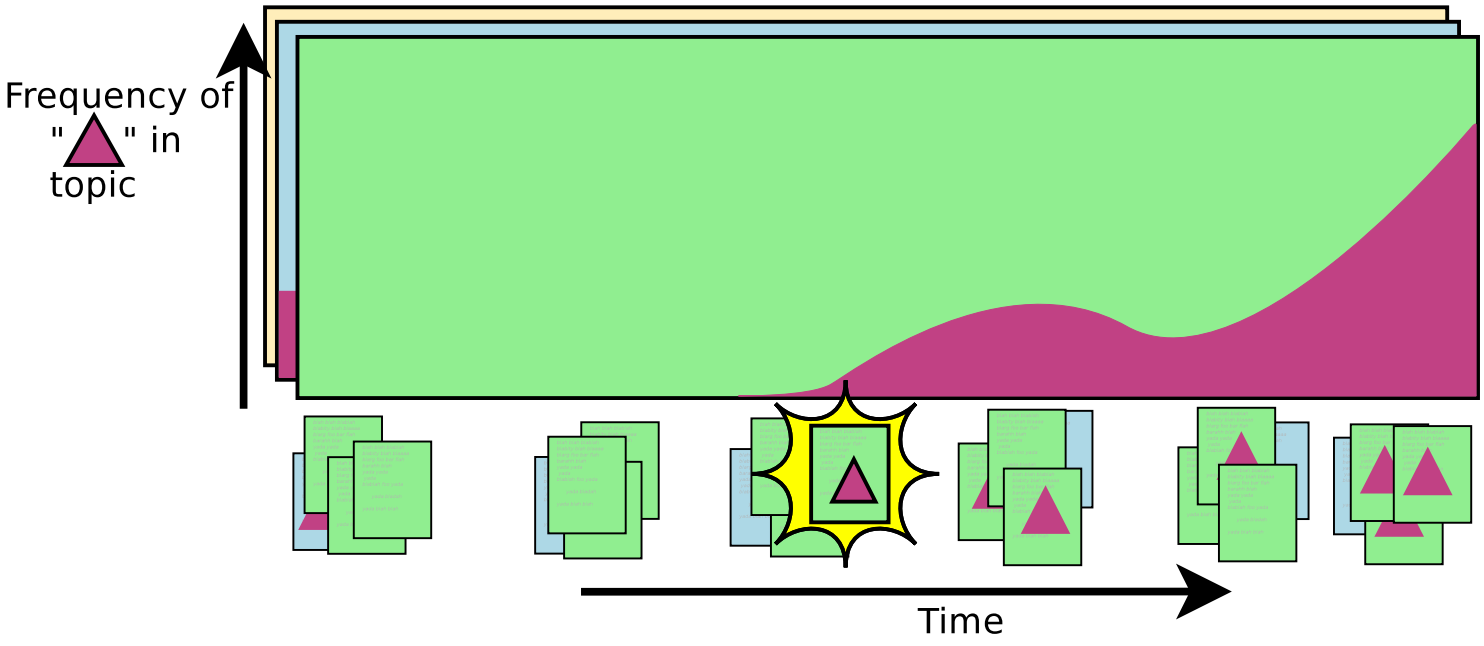


- Traditionally, citations are used to identify influential documents.
- E.g.: *A novel progressive spongiform encephalopathy in cattle*, [Wells et al., 1987] (Cited by 709)

Common approach: citations

- Citations are limited
  - Hard to get
  - May not exist
  - Describe only one kind of influence
- A language-based approach enables us to measure the influence of new kinds of documents
  - News articles
  - Historic manuscripts

Methodology

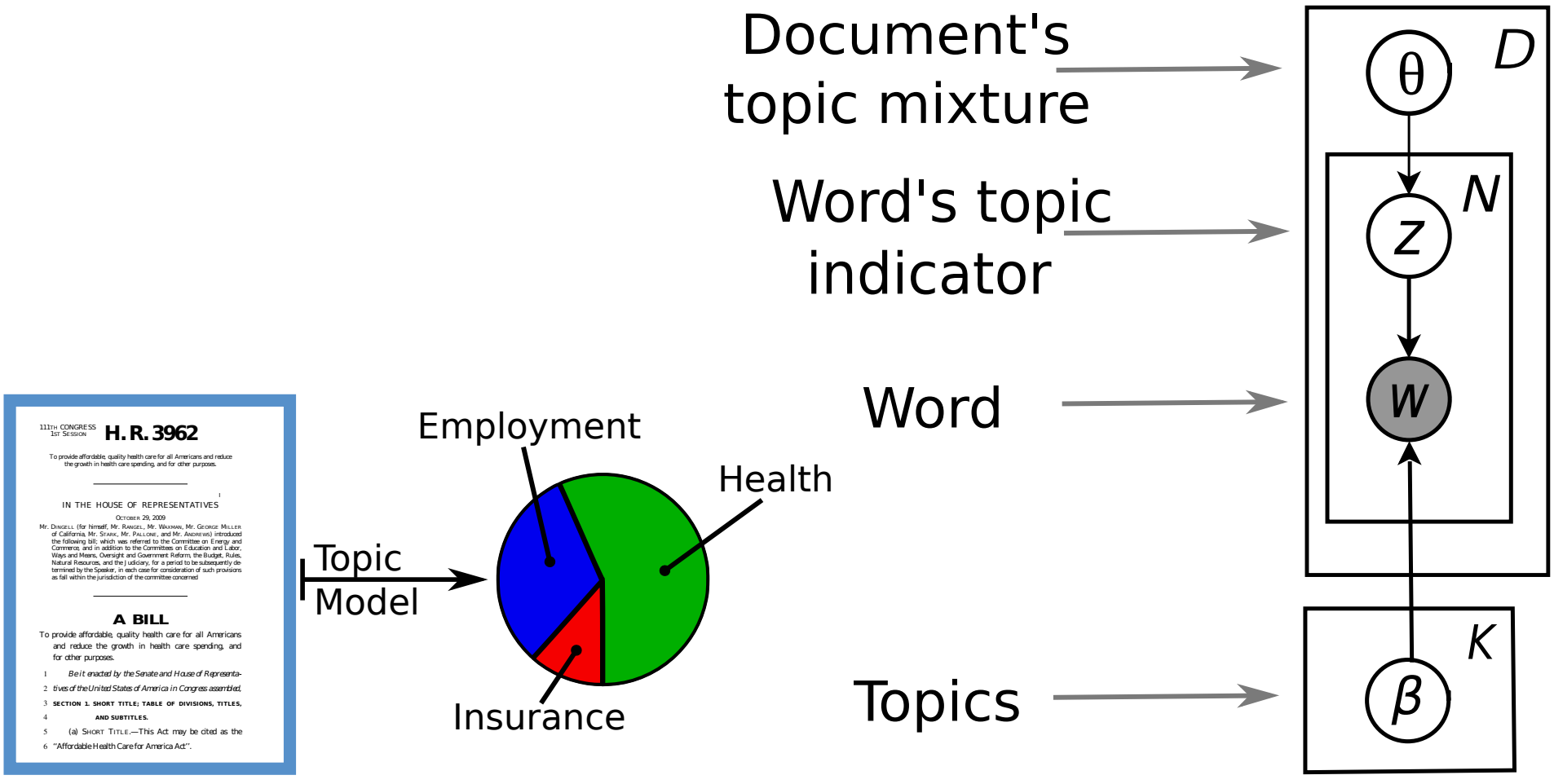


- We develop a probabilistic model to capture the influence of documents.
- The intuition: influential documents use language that becomes more popular in later years.
- With posterior inference, we retrospectively see which documents have been influential on the corpus.

Latent Dirichlet Allocation

[Blei et al., 2003]

Topic models decompose a corpus into a set of topics, i.e. distributions over terms.

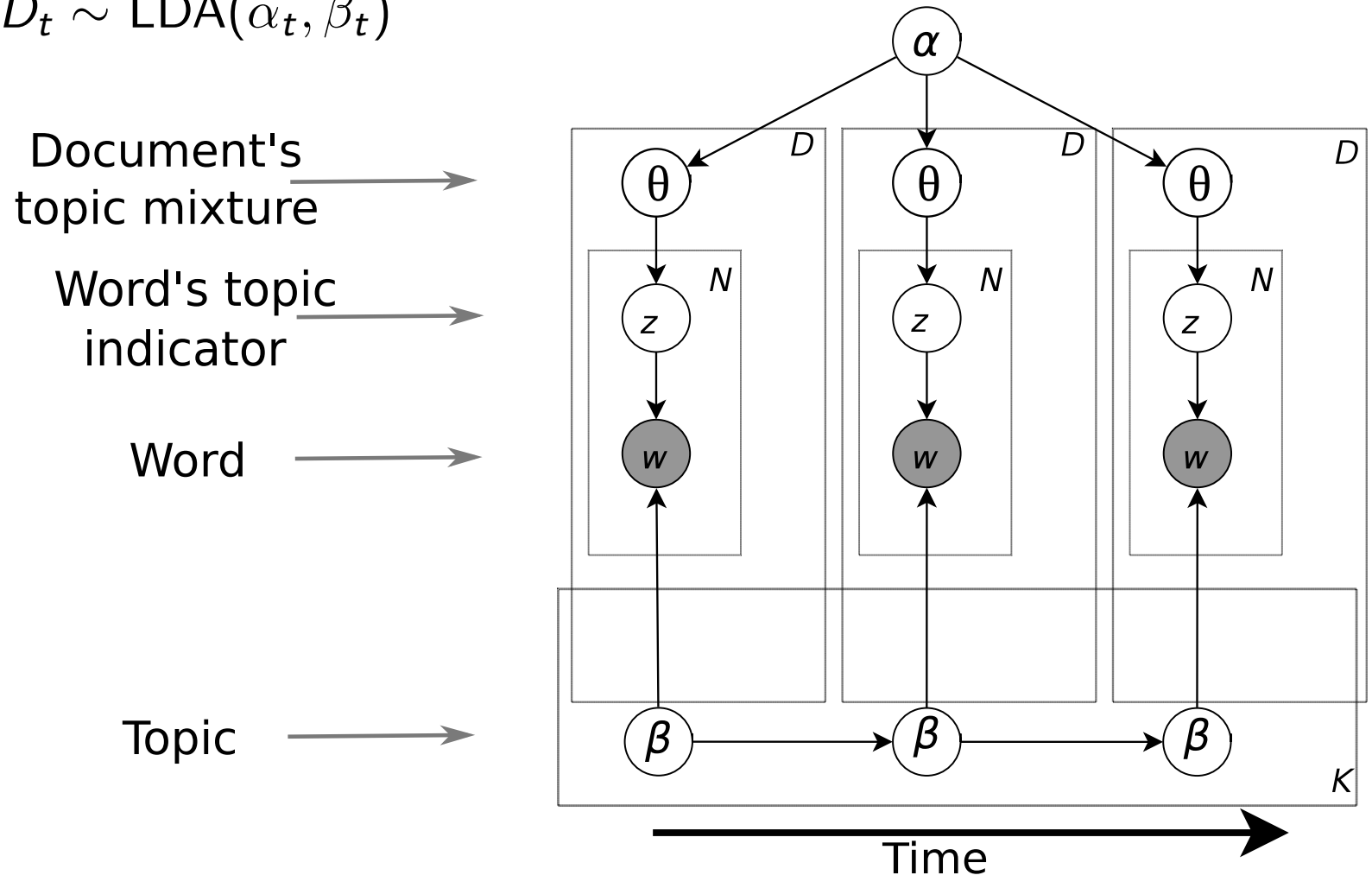


Example topics

Image adapted from Chang et al. [Chang et al., 2009]

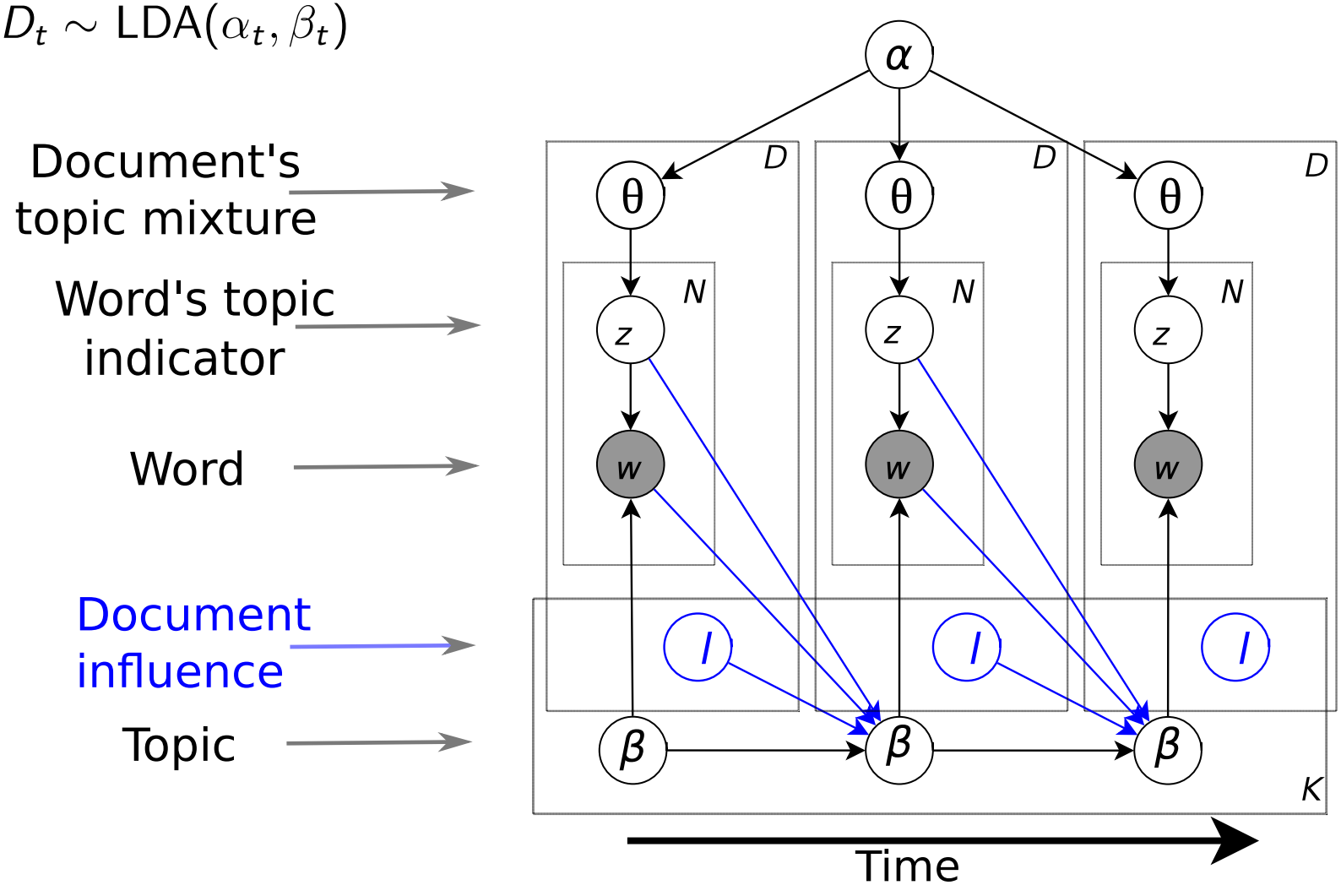
Dynamic topics [Blei and Lafferty, 2006]

- Topics drift in a Markov chain:
  $\beta_t \sim \mathcal{N}(\beta_{t-1}, \sigma^2)$
- Documents are generated from Latent Dirichlet Allocation
  $D_t \sim \text{LDA}(\alpha_t, \beta_t)$



The Document Influence Model

- Topics drift in a Markov chain:
  $\beta_t \sim \mathcal{N}(\beta_{t-1} + \text{Infl}(...), \sigma^2)$
- Documents are generated from Latent Dirichlet Allocation
  $D_t \sim \text{LDA}(\alpha_t, \beta_t)$



The influence function

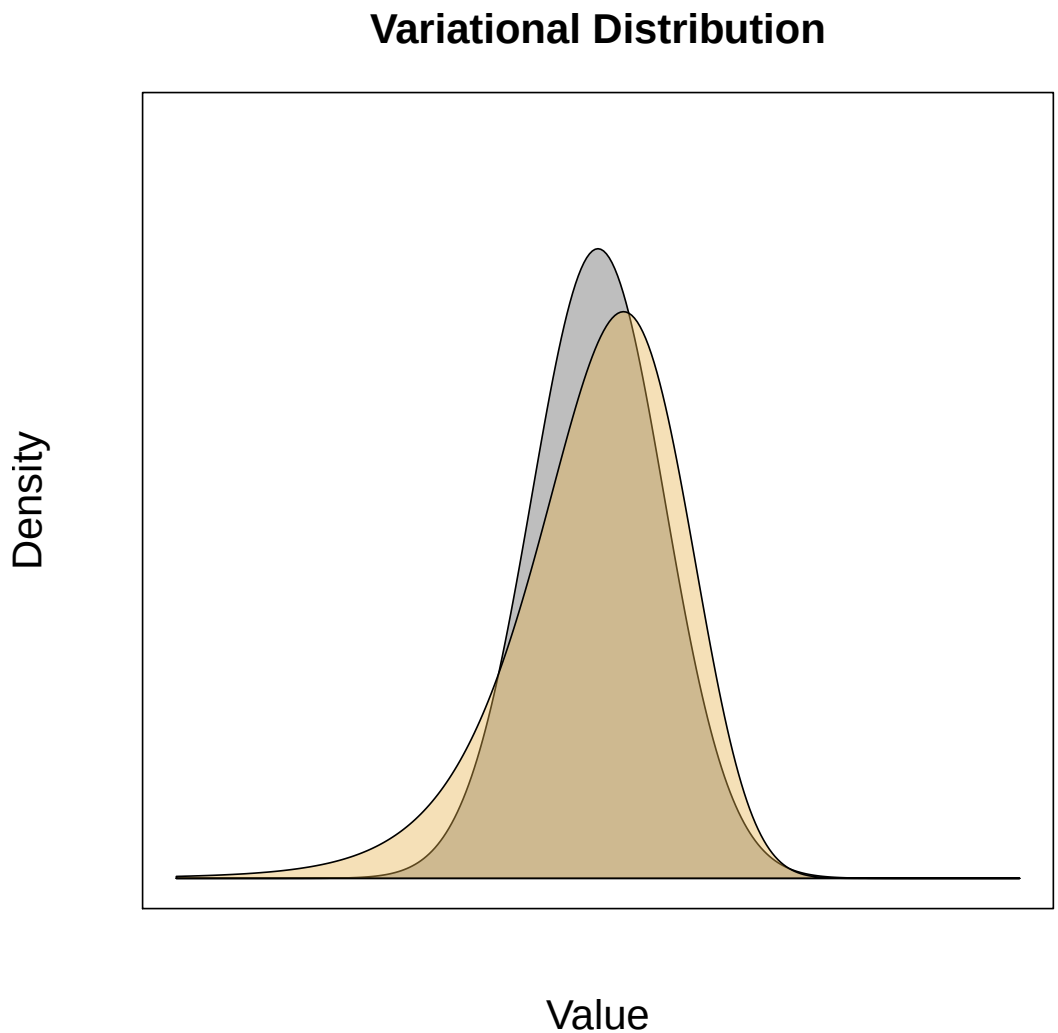
Markov step: $\beta_{t,k} \sim \mathcal{N}(\beta_{t-1,k} + \sum_D \mathsf{Infl}_{d,t,k}, \sigma^2 I)$,

We developed the model to have certain characteristics:

1. More recent documents have greater influence.
   - Progress is sequential
   - Empirically, we see that research has a short half-life
2. A document only influences relevant topics.
   - A document about mad cow disease could influence a medicine topic.
   - A document about mad cow disease *will not* influence language in the space/cosmology topic.

Posterior Inference

- We observe only the words $\mathbf{w}$.
- We want to find the posterior $p(l, \theta, z, \beta|\mathbf{w})$.
- We use variational methods [Jordan et al., 1999].

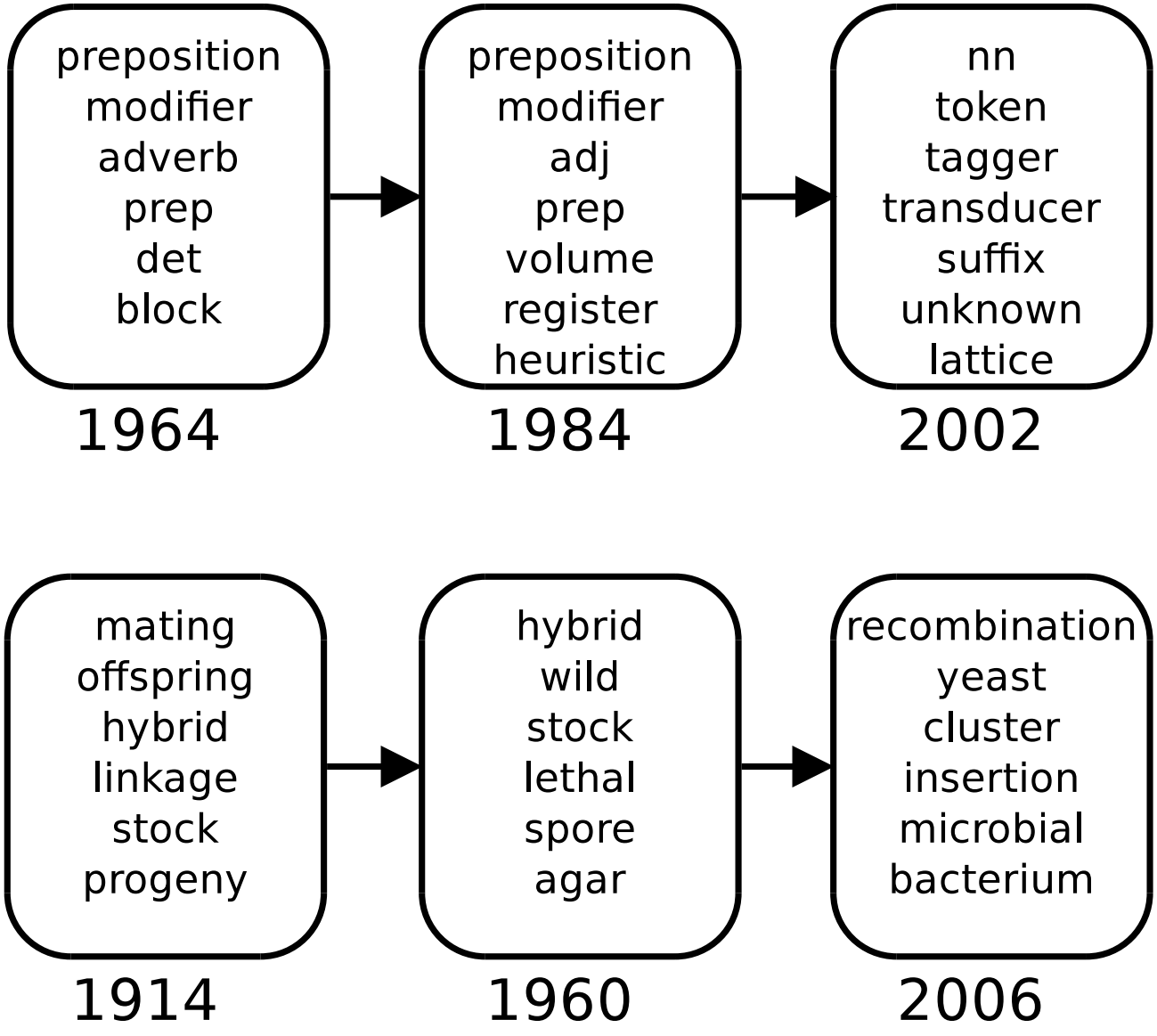


Experiments

- We analyzed three corpora with the DIM.
  - The *ACL Anthology*: 7561 documents
  - *Nature*: 34454 documents
  - *PNAS*: 11855 documents

- This provides estimates of the influence of each article.

Topics in ACL and PNAS



High influence and high citations

The Mathematics of Statistical Machine Translation: Parameter Estimation

Peter F. Brown*
IBM T.J. Watson Research Center

Stephen A. Della Pietra*
IBM T.J. Watson Research Center

Vincent J. Della Pietra*
IBM T.J. Watson Research Center

Robert L. Mercer*
IBM T.J. Watson Research Center

We describe a series of five statistical models of the translation process and give algorithms for estimating the parameters of these models given a set of pairs of sentences that are translations of one another. We define a concept of word-by-word alignment between such pairs of sentences. For any given pair of such sentences each of our models assigns a probability to each of the possible word-by-word alignments. We give an algorithm for seeking the most probable of these alignments. Although the algorithm is suboptimal, the alignment thus obtained accounts well for the word-by-word relationships in the pair of sentences. We have a great deal of data in French and English from the proceedings of the Canadian Parliament. Accordingly, we have restricted our work to these two languages; but we feel that because our algorithms have minimal linguistic content they would work well on other pairs of languages. We also feel, again because of the minimal linguistic content of our algorithms, that it is reasonable to argue that word-by-word alignments are inherent in any sufficiently large bilingual corpus.

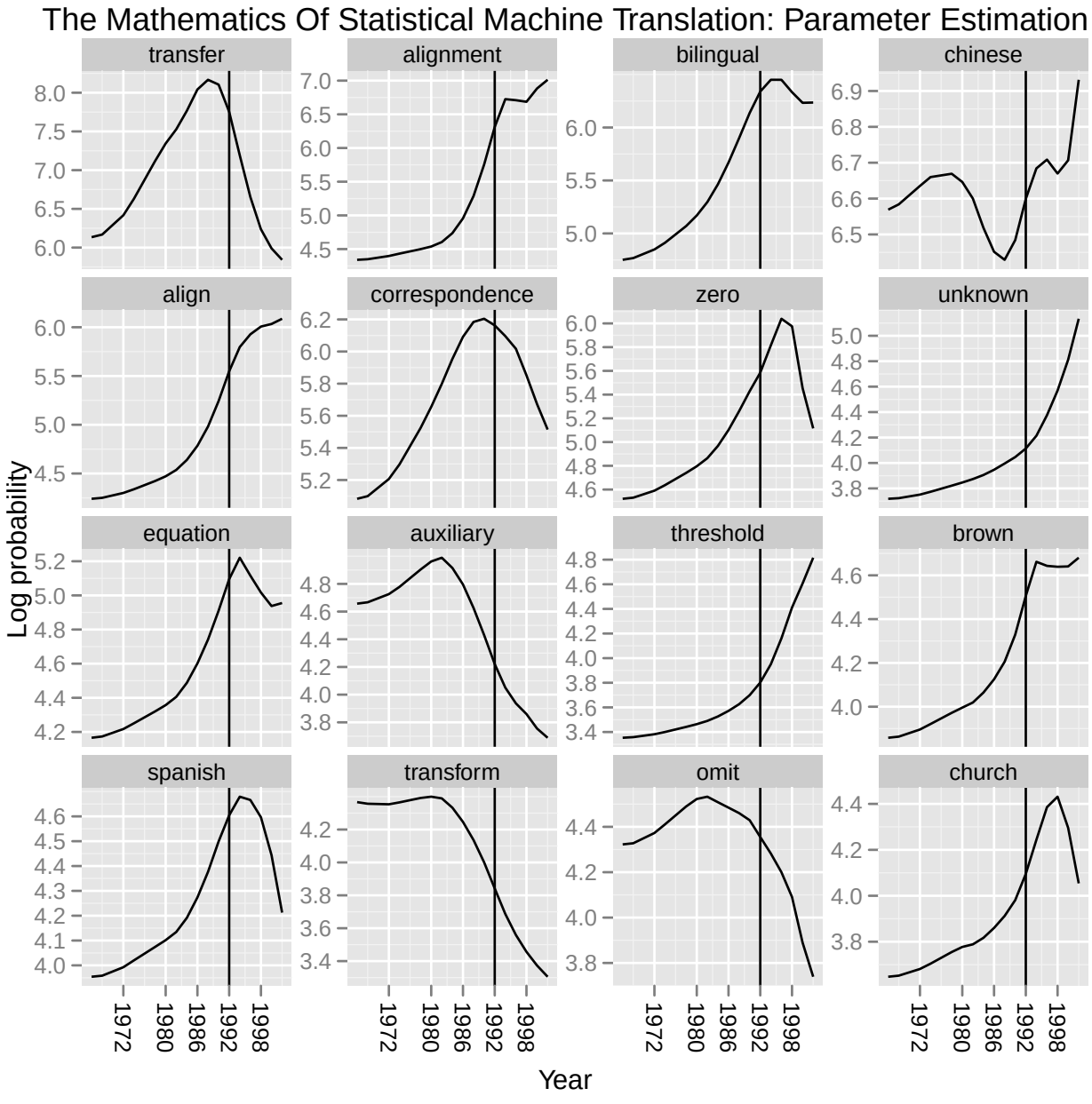
1. Introduction

The growing availability of bilingual, machine-readable texts has stimulated interest in methods for extracting linguistically valuable information from such texts. For example, a number of recent papers deal with the problem of automatically obtaining pairs of aligned sentences from parallel corpora (Warwick and Russell 1990; Brown, Lai, and Mercer 1991; Gale and Church 1991b; Kay 1991). Brown et al. (1990) assert, and Brown, Lai, and Mercer (1991) and Gale and Church (1991b) both show, that it is possible to obtain such aligned pairs of sentences without inspecting the words that the sentences contain. Brown, Lai, and Mercer base their algorithm on the number of words that the sentences contain, while Gale and Church base a similar algorithm on the number of characters that the sentences contain. The lesson to be learned from these two efforts is that simple, statistical methods can be surprisingly successful in achieving linguistically interesting goals. Here, we address a natural extension of that work: matching up the words within pairs of aligned sentences.

In recent papers, Brown et al. (1988, 1990) propose a statistical approach to machine translation from French to English. In the latter of these papers, they sketch an algorithm for estimating the probability that an English word will be translated into any particular French word and show that such probabilities, once estimated, can be used together with a statistical model of the translation process to align the words in an English sentence with the words in its French translation (see their Figure 3).

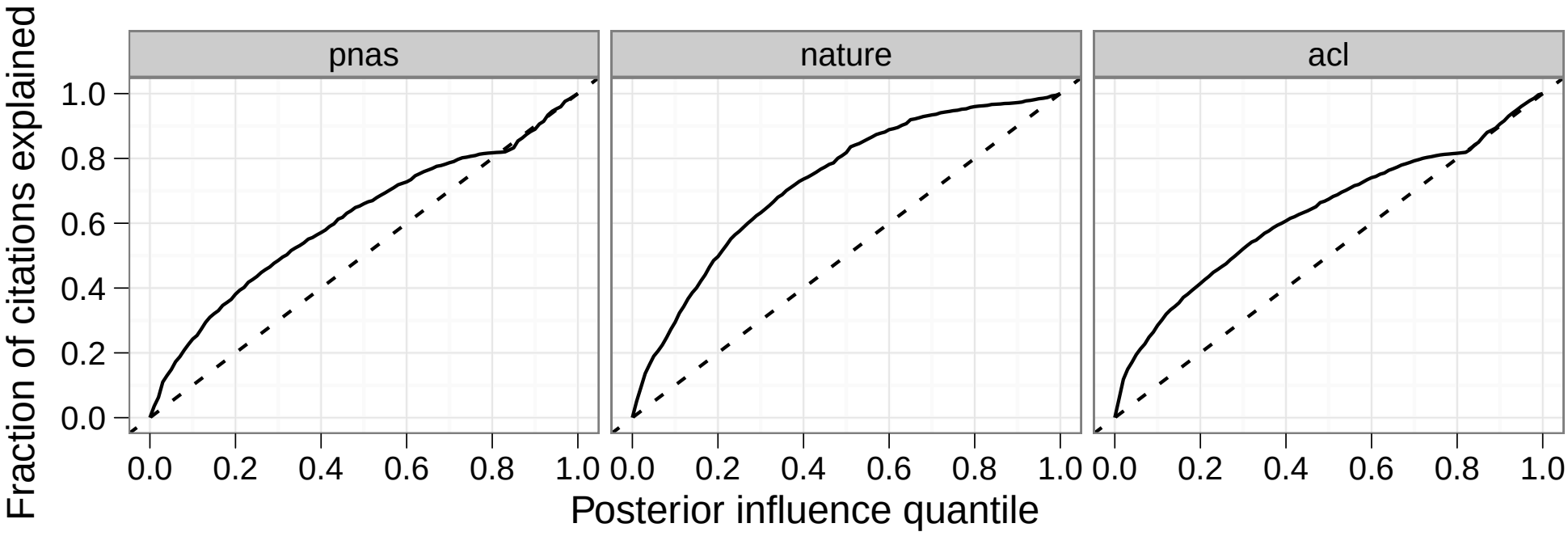
* IBM T.J. Watson Research Center, Yorktown Heights, NY 10598

ACL citations: 7561



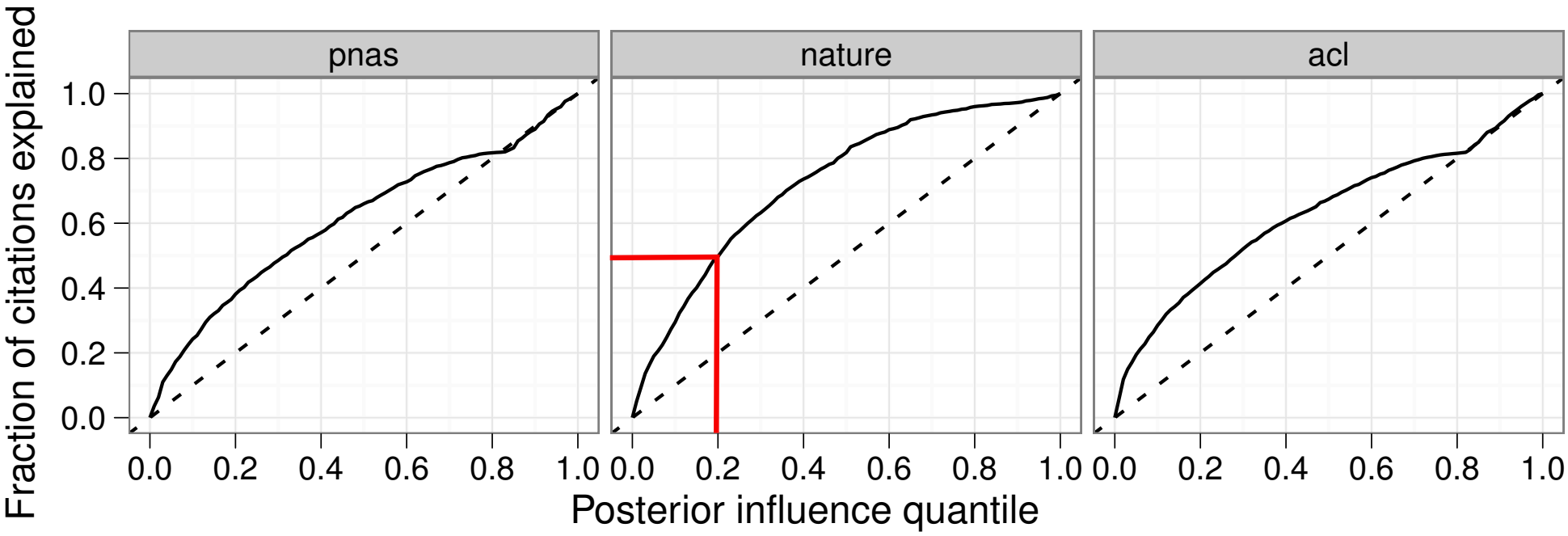
Comparison with citation counts

- We can sort articles by decreasing influence and inspect how many citations the top N have.
- For example, the top 20% of *Nature* articles (by the influence function) receive 50% of its citations.



Results

- We can sort articles by decreasing influence and inspect how many citations the top N have.
- For example, the top 20% of *Nature* articles (by the influence function) receive 50% of its citations.



"The most optimistic researchers believe that these storehouses of big data will for the first time reveal sociological laws of human behavior – enabling them to predict political crises, revolutions and other forms of social and economic instability, just as physicists and chemists can predict natural phenomena."

Government Aims to Build a Data Eye in the Sky. The New York Times. Feb 11, 2011

Libyans wary of NAT

by DAVID BELL

PERTH'S Libyan community joined a slew of activists

a decade ago, is distressed by
y. Photo by Alex Bainbridge
ould harm the

SATURDAY • MARCH 16 • 1991

Fort Worth Star-Telegram

U.S. advances in Iraq

Soviet ties survive, Baker says

Arms control talks remain deadlocked

Victories reported by rebels

The Philadelphia Inquirer

PUBLIC LEDGER

An Independent Newspaper for All the People

RUSSIA AT WAR WITH J

2,000,000 Red Troops to Be Hurled Into Action i

Virtually All Life Destroyed in Hiroshima, Tok

Today

Halsey's Planes Renew Assault, Batter Honshu

Honolulu Star-Bulletin 1st EXTRA

8 PAGES—HONOLULU, TERRITORY OF HAWAII, U. S. A., SUNDAY, DECEMBER 7, 1941—8 PAGES — PRICE FIVE CENTS

WAR!

(Associated Press by Transpacific Telephone)

SAN FRANCISCO, Dec. 7.—President Roosevelt announced this morning that Japanese planes had attacked Manila and Pearl Harbor.

OAHU BOMBED BY JAPANESE PLANES

Quad-City Times LATEST UPDATE!

WAR

'So far, so good'

Sunday, August 30, 2006

Martians invade earth

Incredible as it may seem, it has been confirmed that a large martian invasion fleet has landed on earth tonight.

First vessels were sighted over Great Britain, Denmark and Norway already in the late evening from where, as further

headed towards the North Pole and Santa Claus was taken hostage by the invaders.

Afterwards they split apart in order to approach most major cities around the earth. The streets filled as thousands fled their homes, many only wearing

Ren
follo
imp

The
that
rela
the
beh
of a
exp

ntonio Express

STREET

WILSON BEGINS TOUR FOR TREATY,
SEES VICTORY IN THE SENATE FIGHT;
REPUBLICANS FEAR PARTY SETBACK

SENATE TACTICS ASSAILED

Despite the savage cutbacks at hom
Britain tops world league in foreign

UK DOLES
OUT MORE
AID THAN
ANY OTHER
COUNTRY

Hugs for Life

The big relationship drought is connected to emotional and physical health

By John Sanderson

Mural dedicated

World reaches out to Haiti

The U.N. is releasing $10 million from its emerg
European Commission has approved $4.37 mill
after Tuesday's earthquake. Here is a selection

Canada: A military team to assess needed goods, sanitati packag

Irish telecommunications company donates $5 million

Spain: 100 tons of relief equipment

Netherlands: Rescue team

MAP
to illustrate the
TREATY OF BERLIN (1878)

Treaty of Paris).....
by Treaty of San Stefano .....
of Berlin

A model of foreign relations sentiment

Our goal is to automatically create a history of the relationship between countries from newspaper articles.

- Collect a bunch of newspaper articles.
- Ask Mechanical Turk workers to evaluate paragraphs that mention pairs of countries. [Chang and Blei, 2009]
- Use these labels to fit a *spatial* model of the sentiment between different countries.

[Chang and Blei, 2009, Hoff et al., 2002, Clinton et al., 2004, Martin and Quinn, 2002, Gartzke, 1998]

Labeling sentiment: typical task

although **israel** and neighboring **jordan** agreed with fanfare in late july to end their technical state of war and have since behaved in public like old and dear friends they have yet to sign a peace treaty and have no official links.

What is the relationship between **israel** and **jordan** as suggested by the text above?

- **There was no obvious relationship between these countries, or they were not discussed.**
- **Very Positive++** These states have a very good relationship.
- **Positive+** These states have a very good relationship.
- **Slightly Positive** These states are on decent terms.
- **Slightly Negative** There is a little tension between these states (tariffs might exist, for example.)
- **Negative-** These states have a bad relationship (e.g. the states are using negative, threatening remarks.)
- **Very Negative--** These states are mortal enemies.

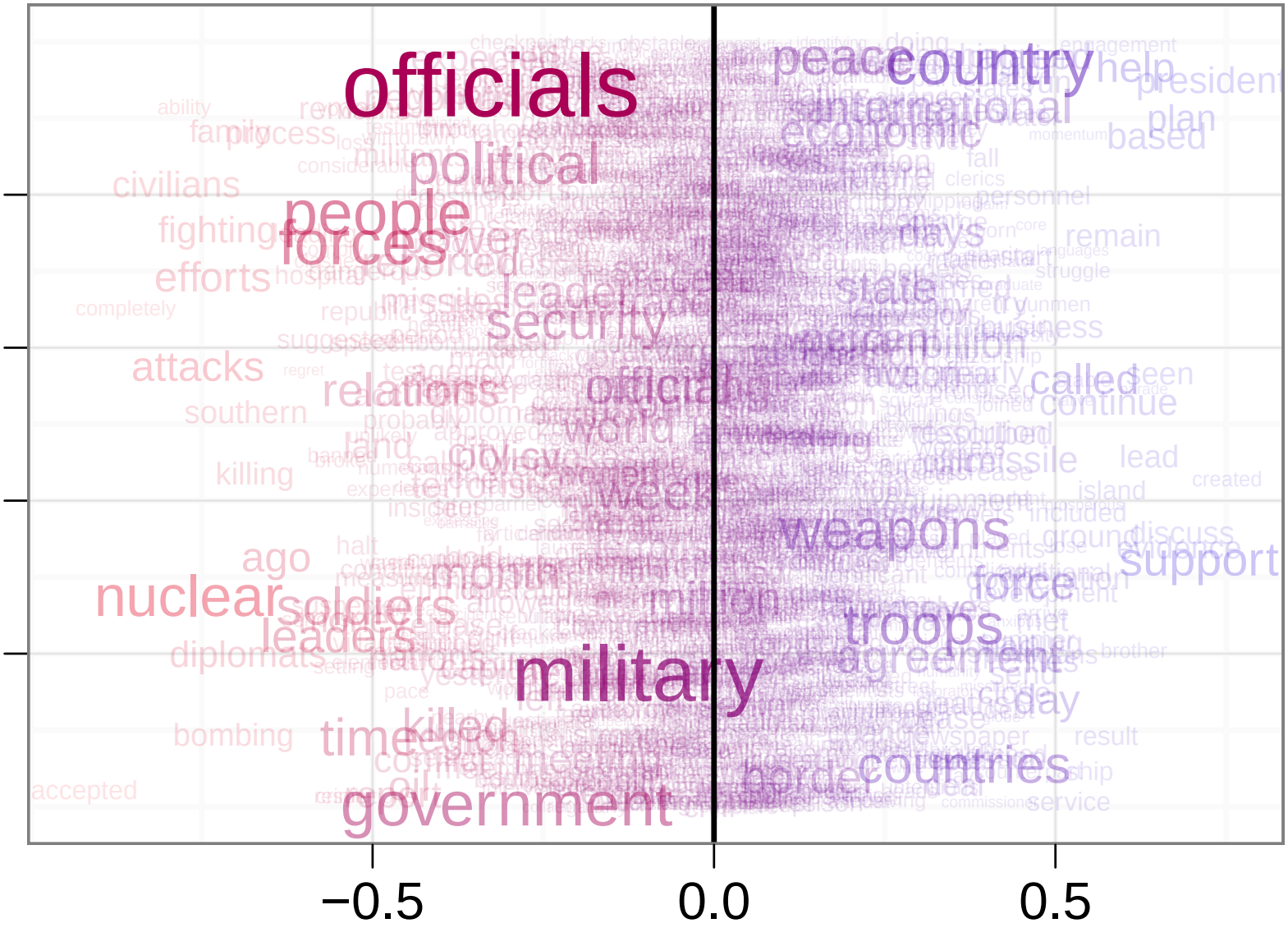
Sentiment and news articles: text regression

[Kogan et al., 2009]

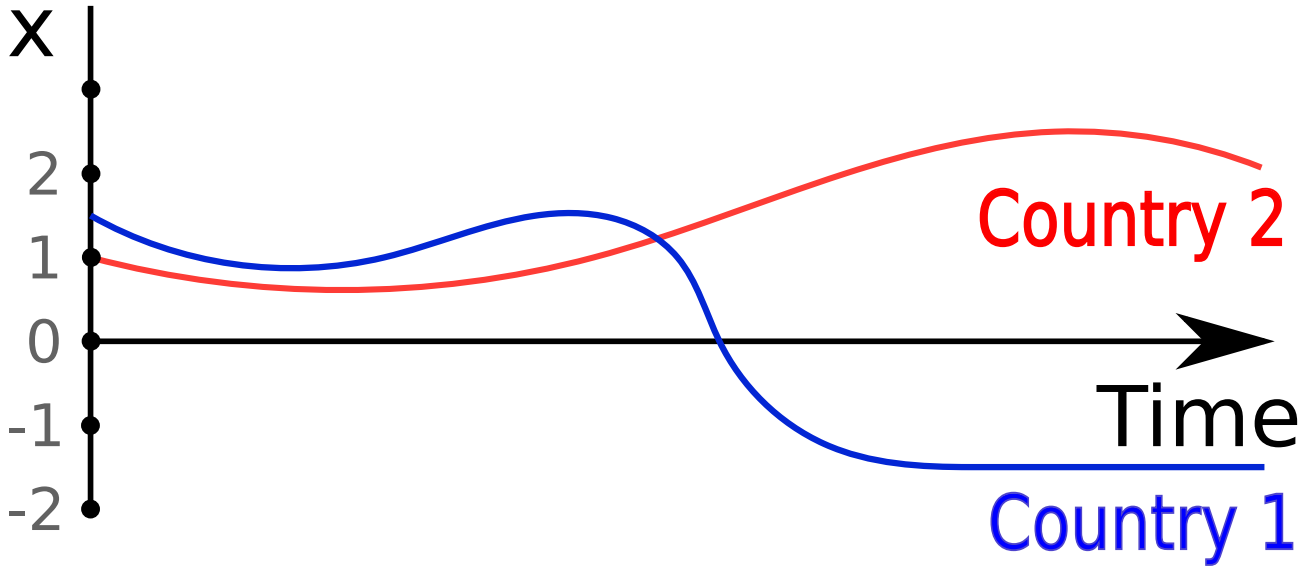
$$s_d = \mathbf{w}_d^T \boldsymbol{\beta} + \varepsilon$$

- $\mathbf{w}_d \in \mathbb{R}^V$ is the text of a news paragraph
- $s_d \in \mathbb{R}$ is the sentiment between two countries
- $\boldsymbol{\beta} \in \mathbb{R}^V$ is the "weight" of each word

Sentiment and news articles: text parameter β

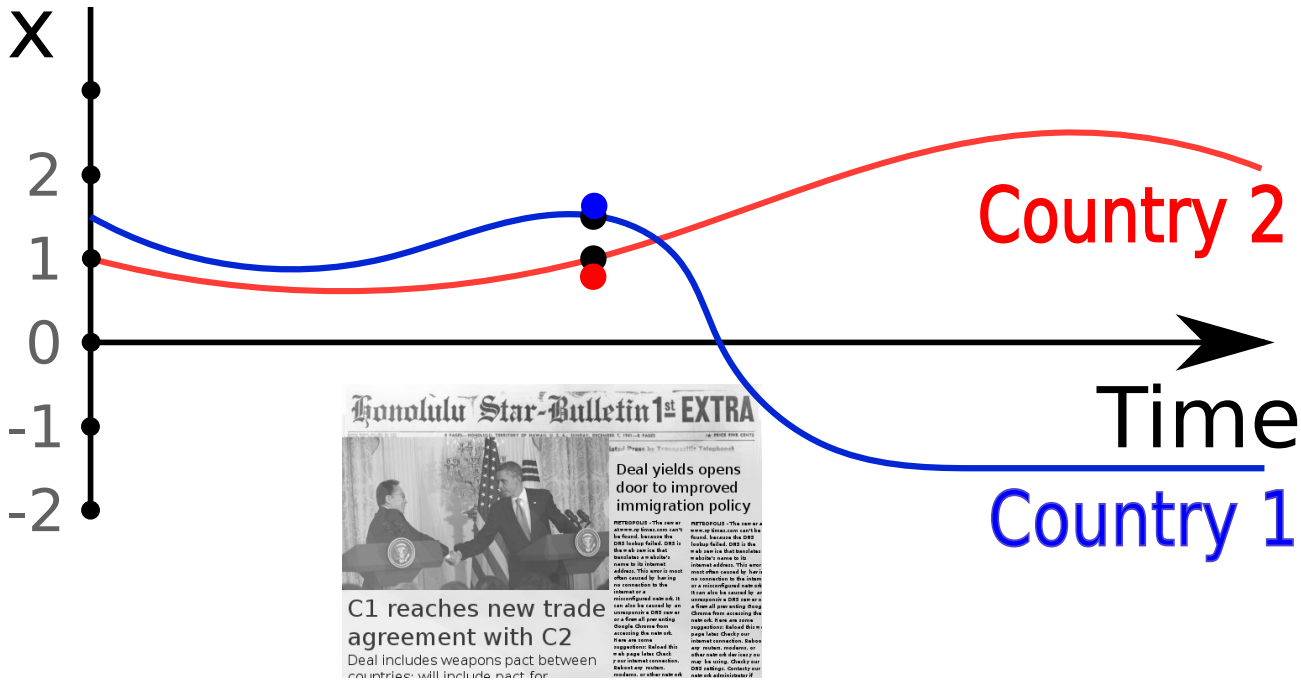


Countries take latent positions $\bar{x}_{ct}$ over time



$$\bar{x}_{c,t} | \bar{x}_{c,t-1} \sim N(\bar{x}_{c,t-1}, \sigma_K^2)$$

The relationship between countries is observed in the news.

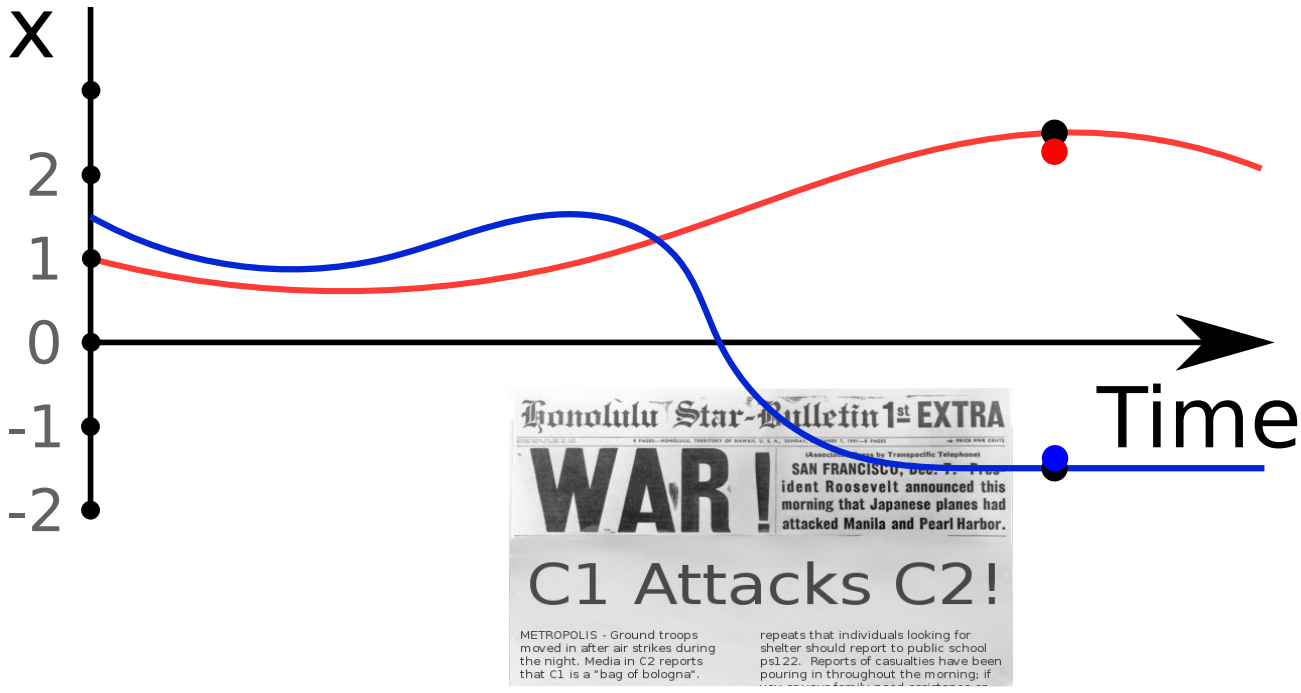


$$x_{c_1,d} \sim N(\bar{x}_{c_1,t}, \sigma_D^2)$$

$$x_{c_2,d} \sim N(\bar{x}_{c_2,t}, \sigma_D^2)$$

$$\text{Sentiment } s_d := x_{c_1,d}{}^T x_{c_2,d}$$

The relationship between countries is observed in the news.

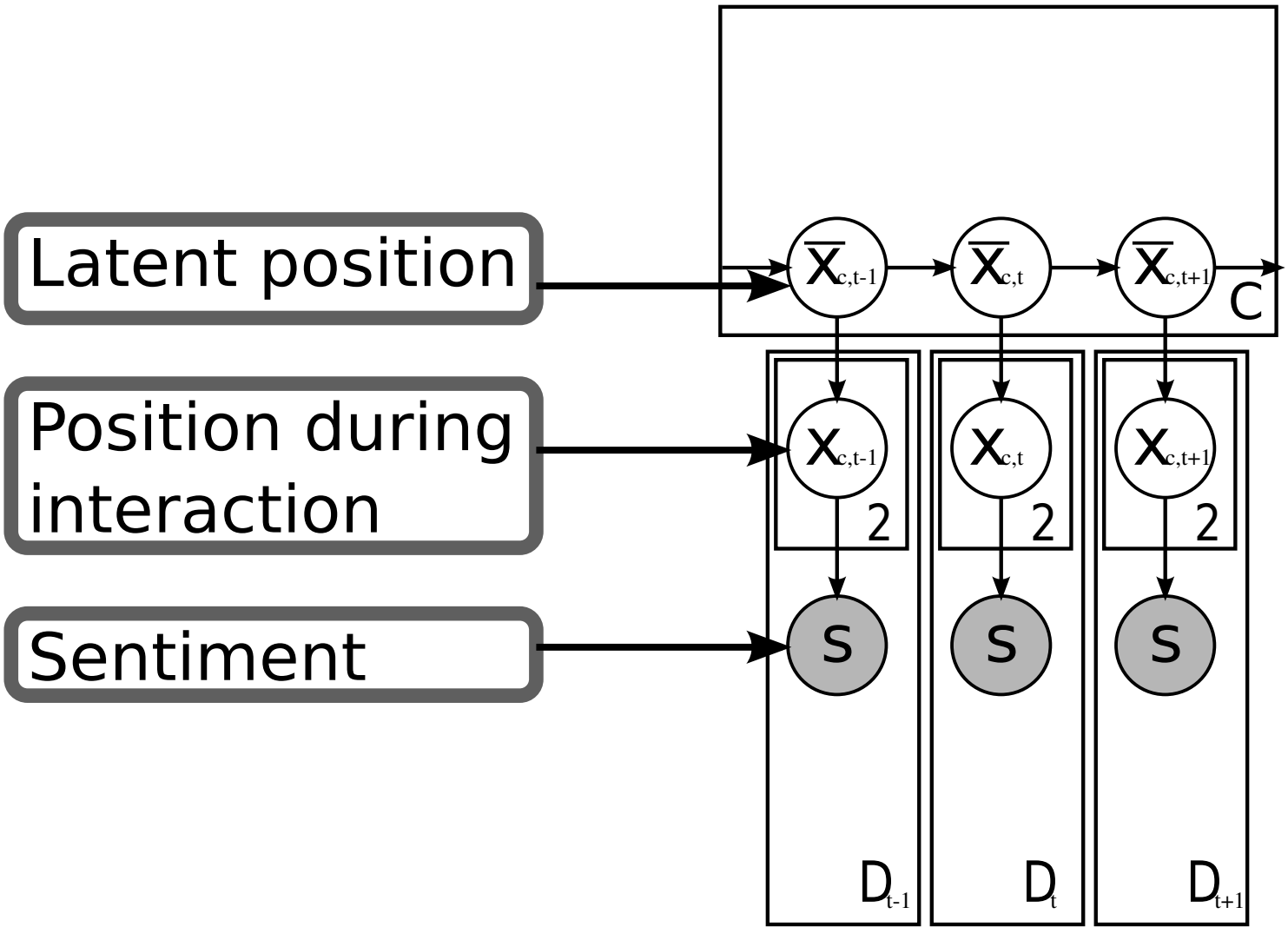


$$x_{c_1,d} \sim N(\bar{x}_{c_1,t}, \sigma_D^2)$$

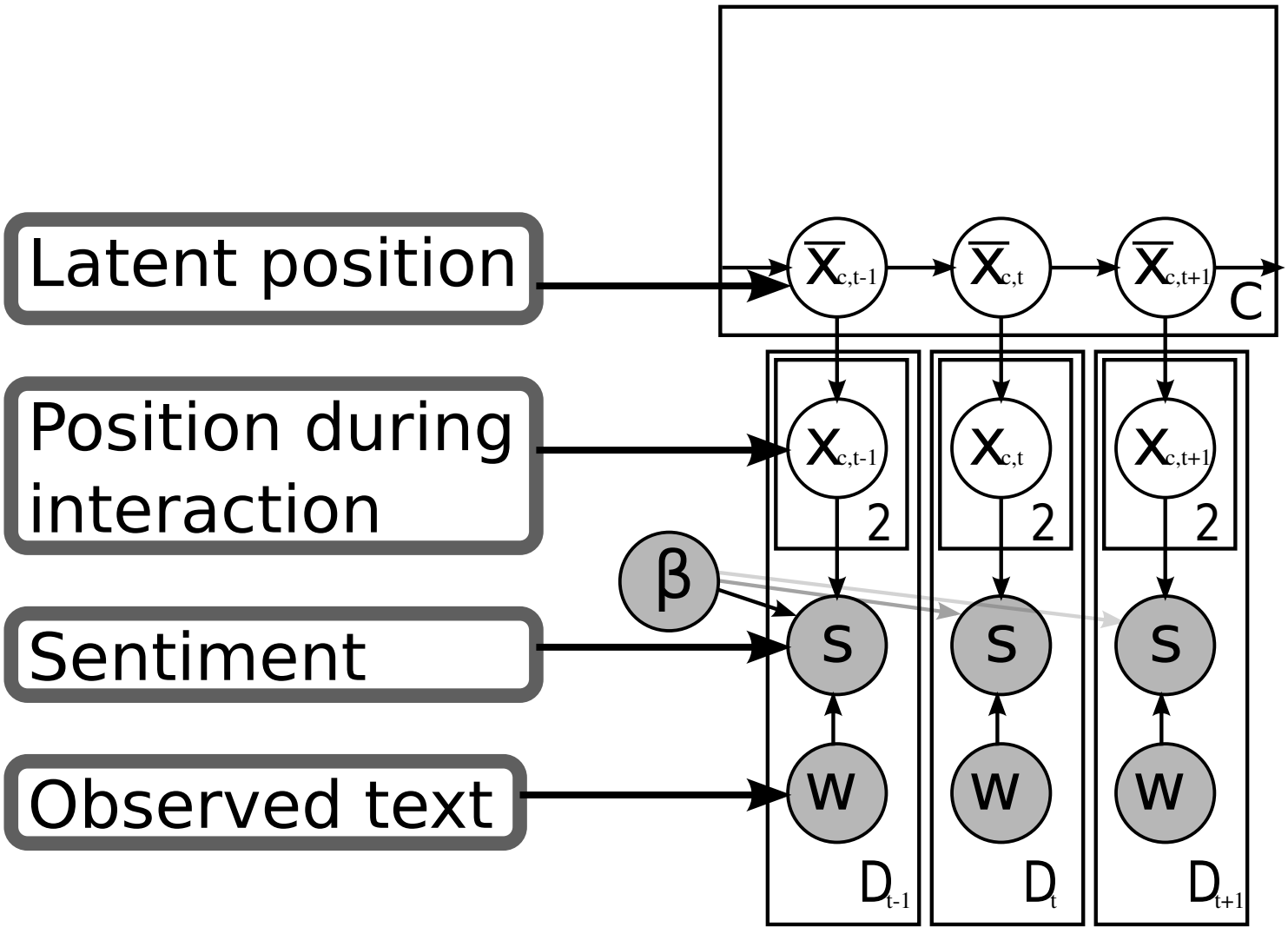
$$x_{c_2,d} \sim N(\bar{x}_{c_2,t}, \sigma_D^2)$$

$$\text{Sentiment } s_d := x_{c_1,d}{}^T x_{c_2,d}$$

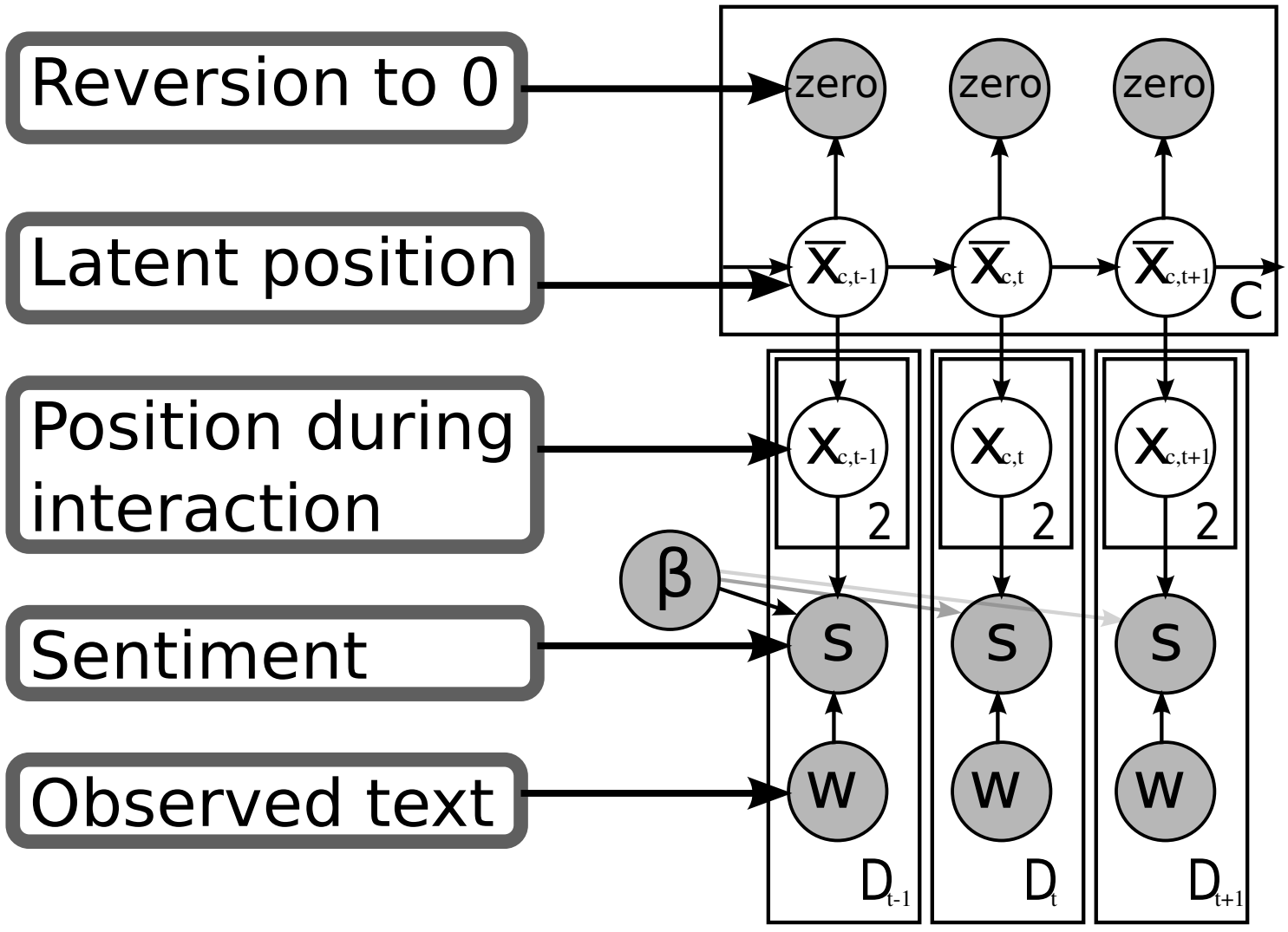
The relationship between countries over time



The relationship between countries over time



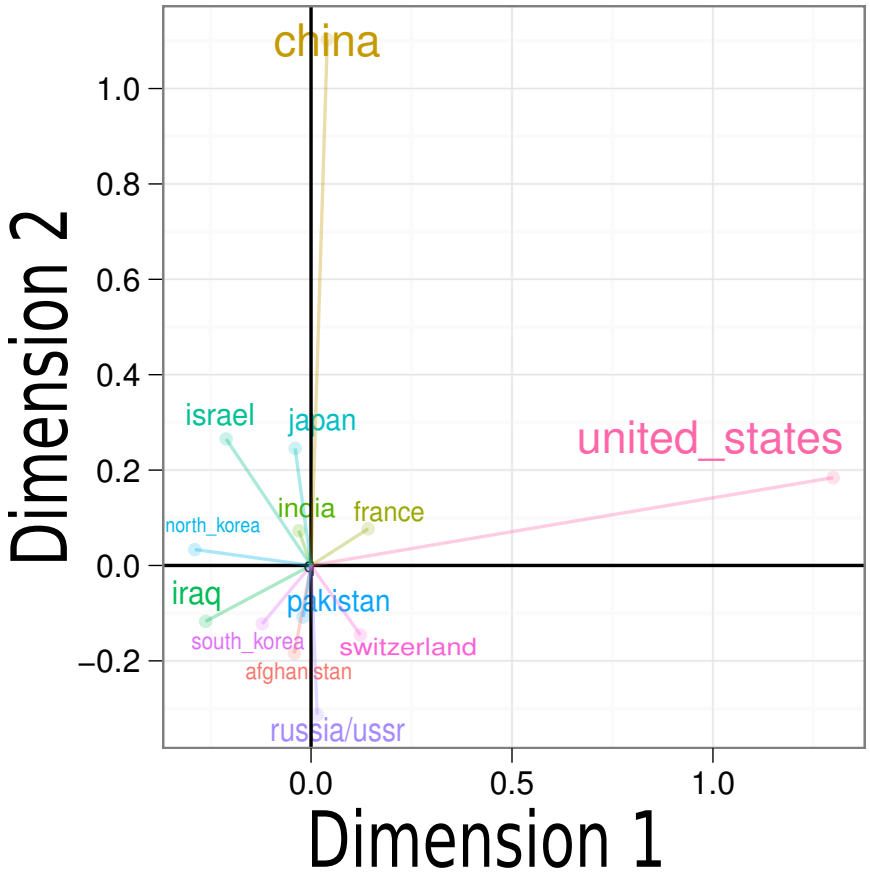
The relationship between countries over time



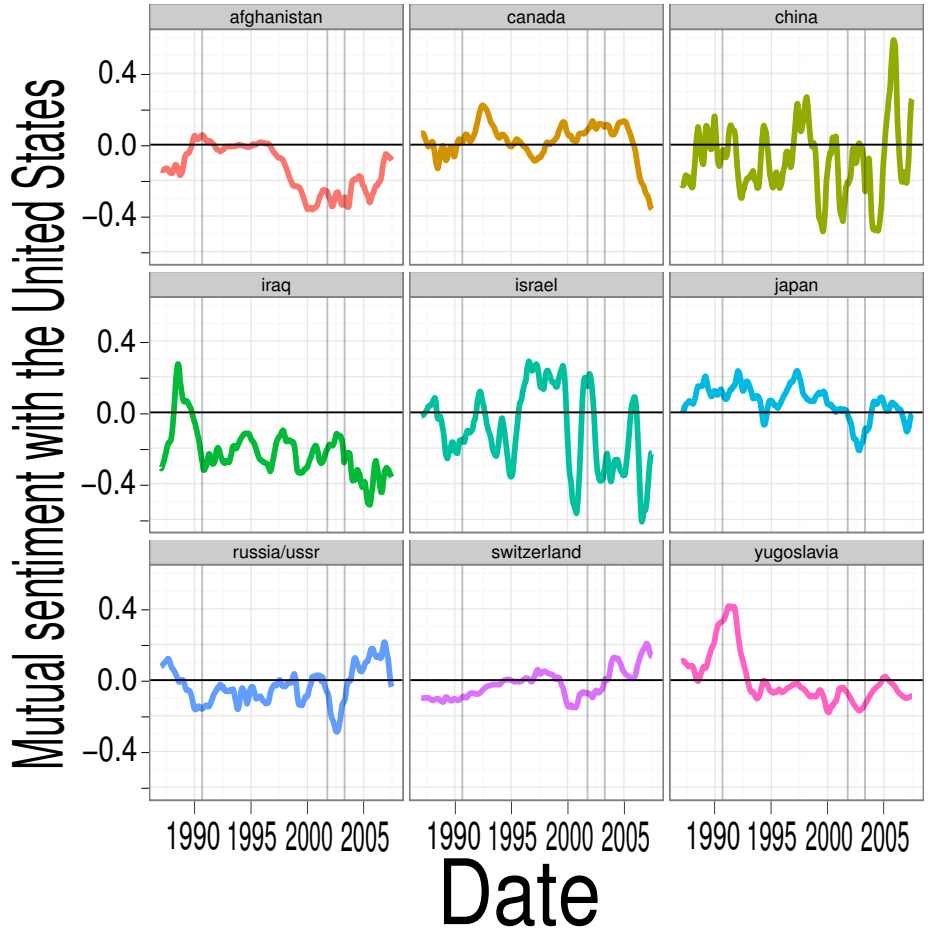
Experiments

- Randomly select 3607 paragraphs discussing pairs of 245 countries and territories.
- Label each of these paragraphs' sentiment with two ratings from Amazon Mechanical Turk.
- Hold out 42 random pairs (244 paragraphs) for testing.
- Fit sentiment model parameters β on training paragraphs.
- Fit the spatial sentiment model with these parameters on all 257,472 paragraphs from 1988 to 2008. (We used MAP.)

Results: countries' latent positions



Latent positions in 2007



Mutual sentiment with U.S. over time

Evaluation

The model does better than text regression and individual *Mechanical Turk* workers compared against one another.

Model	Mean Squared Error	Mean Absolute Error
Inter-rater agreement	1.77 (7.11)	1.037 (2.07)
Text regression	5.53	1.94
Reversion variance 0.1	2.36	1.09
Reversion variance 1	**2.32**	**1.07**
Reversion variance 10	2.32	1.08
Reversion variance 100	2.34	1.09
Reversion variance 1000	2.33	1.08

Next steps

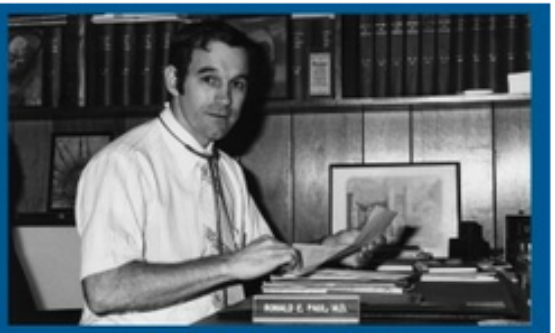
- Use this approach to understand how different news sources frame international relations
- Use better supervision than Mechanical Turk workers
- Fit the foreign relations model in an unsupervised way (c.f. [Chang and Blei, 2009])
- Add intercepts to countries

THE ISSUES

RON PAUL ON THE ISSUES

A PRO-LIFE CHAMPION

Abortion



THE RIGHT REMEDY

Health Care

RESTORE AMERICA'S PROSPERITY

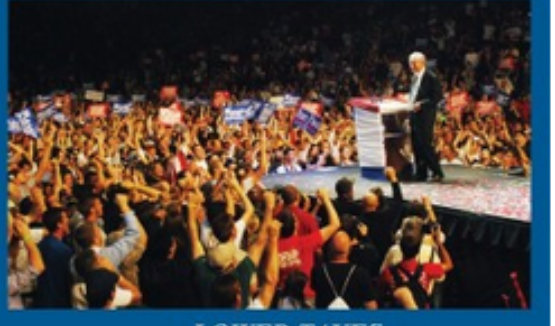
Economy

STRONG. SECURE. RESPECTED.

National Defense

END THE FED

End The Fed



LOWER TAXES

Taxes

Predicting Legislative Votes

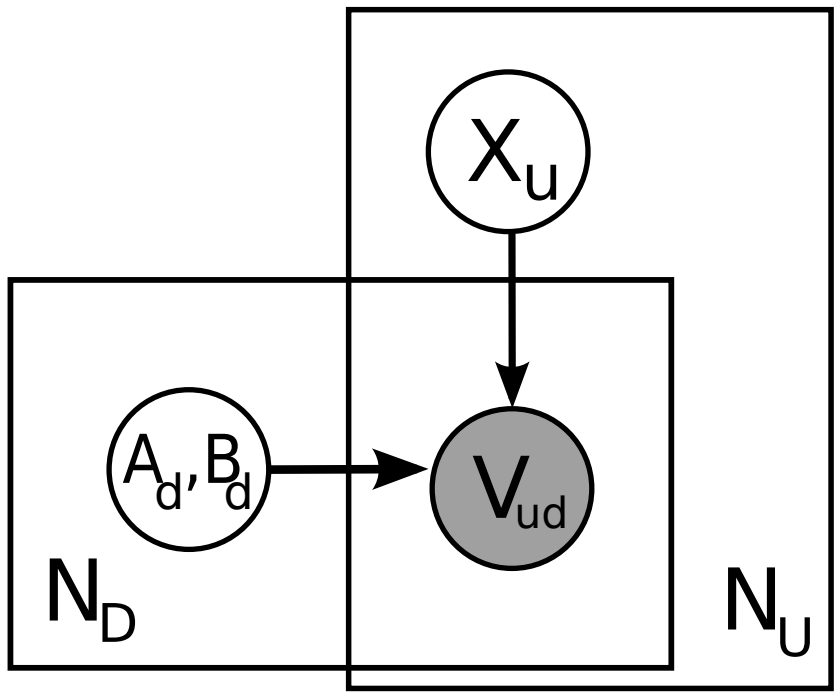
$$\begin{bmatrix} Y & & & N & Y & & & Y \\ & & & & & & N & \\ & & Y & & & & & \\ Y & & & & & N & & N \\ N & N & & & Y & & & Y \\ & & & & & & & N \\ & & & & Y & & Y & \\ Y & Y & & & & & & N \\ & & N & Y & & Y & & \\ & & & & & & & \end{bmatrix}$$

Images credit Susan Walsh (Associated Press) and EPA

Item Response Theory

Jackman, 2001
Jounson and Albert, 1999
Poole and Rosenthal, 1985
Clinton et al., 2004
Martin and Quinn, 2002

$$p(v_{ud} = \text{Yes} | x_u, a_d, b_d) = \text{logistic}(x_u \cdot a_d + b_d)$$

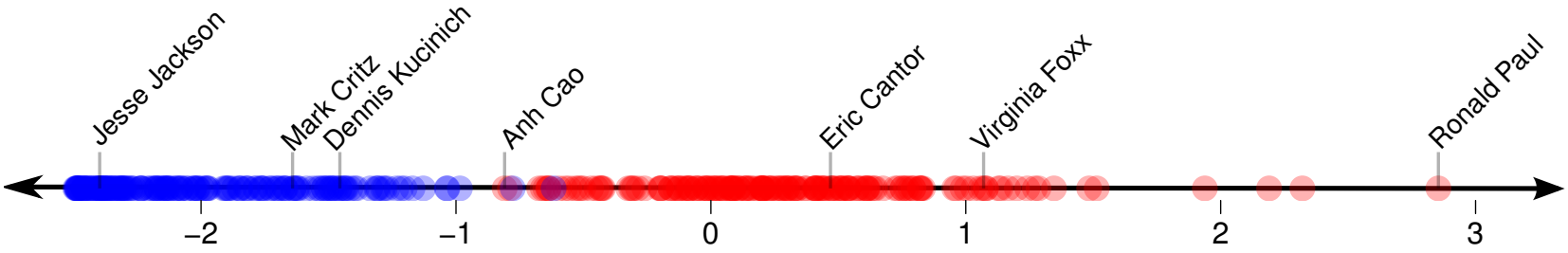


Ideal points

Ideal points x_u position lawmakers in a latent political space.

Jackman, 2001
Jounson and Albert, 1999
Poole and Rosenthal, 1985
Clinton et al., 2004
Martin and Quinn, 2002

$$p(v_{ud} = \text{Yes} | x_u, a_d, b_d) = \text{logistic}(x_u \cdot a_d + b_d)$$

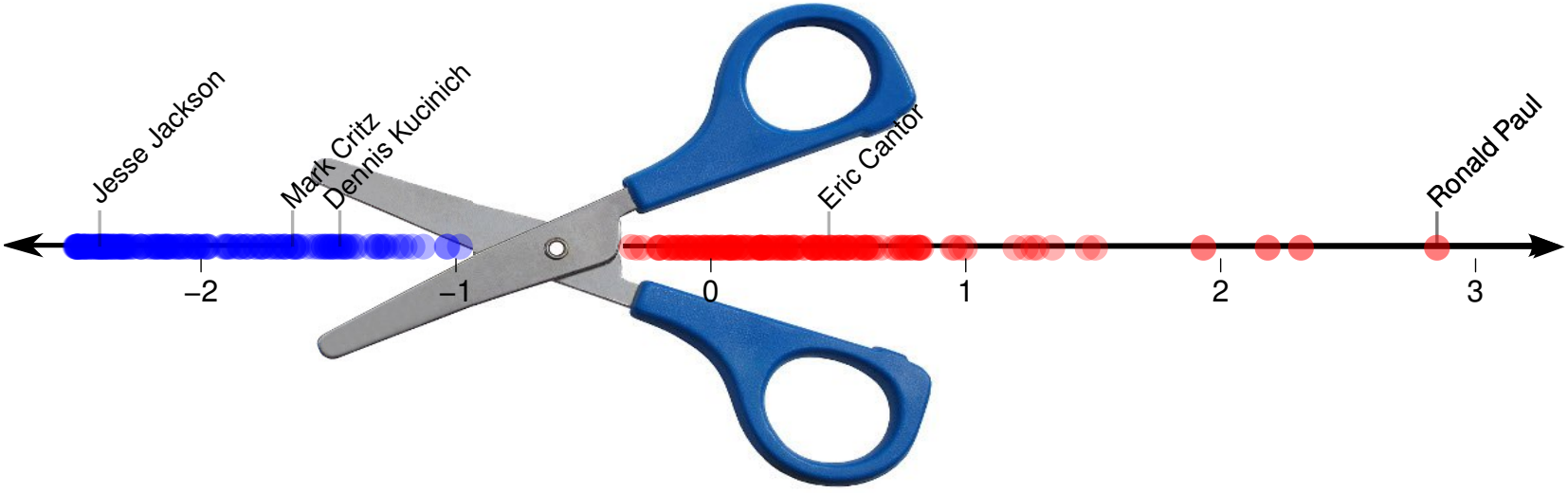


Ideal points

Ideal points x_u position lawmakers in a latent political space.

[Jackman, 2001] [Poole and Rosenthal, 1985] [Martin and Quinn, 2002]
[Johnson and Albert, 1999] [Clinton et al., 2004]

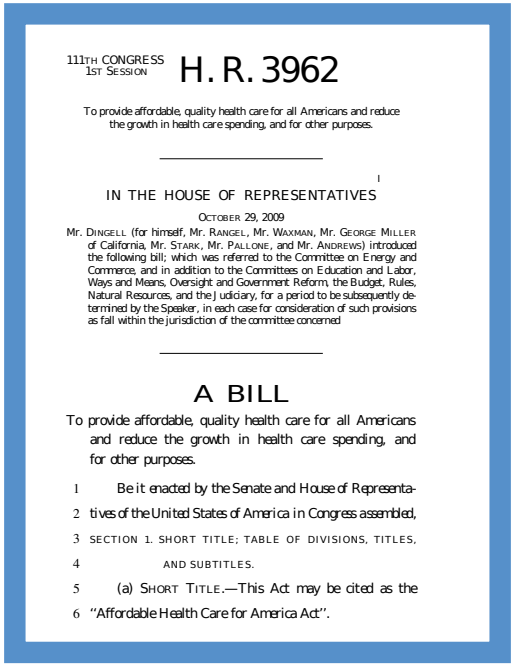
$$p(v_{ud} = \text{Yes} | x_u, a_d, b_d) = \text{logistic}(x_u \cdot a_d + b_d)$$



Each bill defines a *cut point* that splits lawmakers.

Labeled Latent Dirichlet Allocation

[Ramage et al., 2009]



111TH CONGRESS
1ST SESSION

H. R. 3962

To provide affordable, quality health care for all Americans and reduce the growth in health care spending, and for other purposes.

IN THE HOUSE OF REPRESENTATIVES

OCTOBER 29, 2009

Mr. DINGELL (for himself, Mr. RANGEL, Mr. WAXMAN, Mr. GEORGE MILLER of California, Mr. STARK, Mr. PALLONE, and Mr. ANDREWS) introduced the following bill; which was referred to the Committee on Energy and Commerce, and in addition to the Committees on Education and Labor, Ways and Means, Oversight and Government Reform, the Budget, Rules, Natural Resources, and the Judiciary, for a period to be subsequently determined by the Speaker, in each case for consideration of such provisions as fall within the jurisdiction of the committee concerned

A BILL

To provide affordable, quality health care for all Americans and reduce the growth in health care spending, and for other purposes.

Be it enacted by the Senate and House of Representatives of the United States of America in Congress assembled,

SECTION 1. SHORT TITLE; TABLE OF DIVISIONS, TITLES, AND SUBTITLES.

(a) SHORT TITLE.—This Act may be cited as the "Affordable Health Care for America Act".

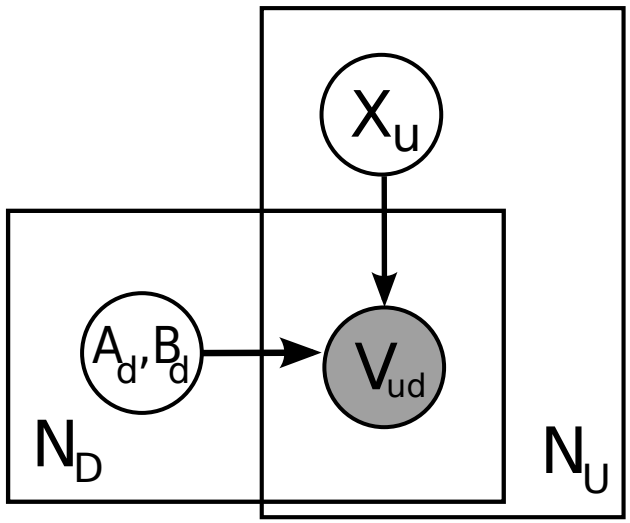
Labeled Topic model

0.01	Taxation
0.82	Health
0.02	Defense
0.14	Finance
0.01	Religion

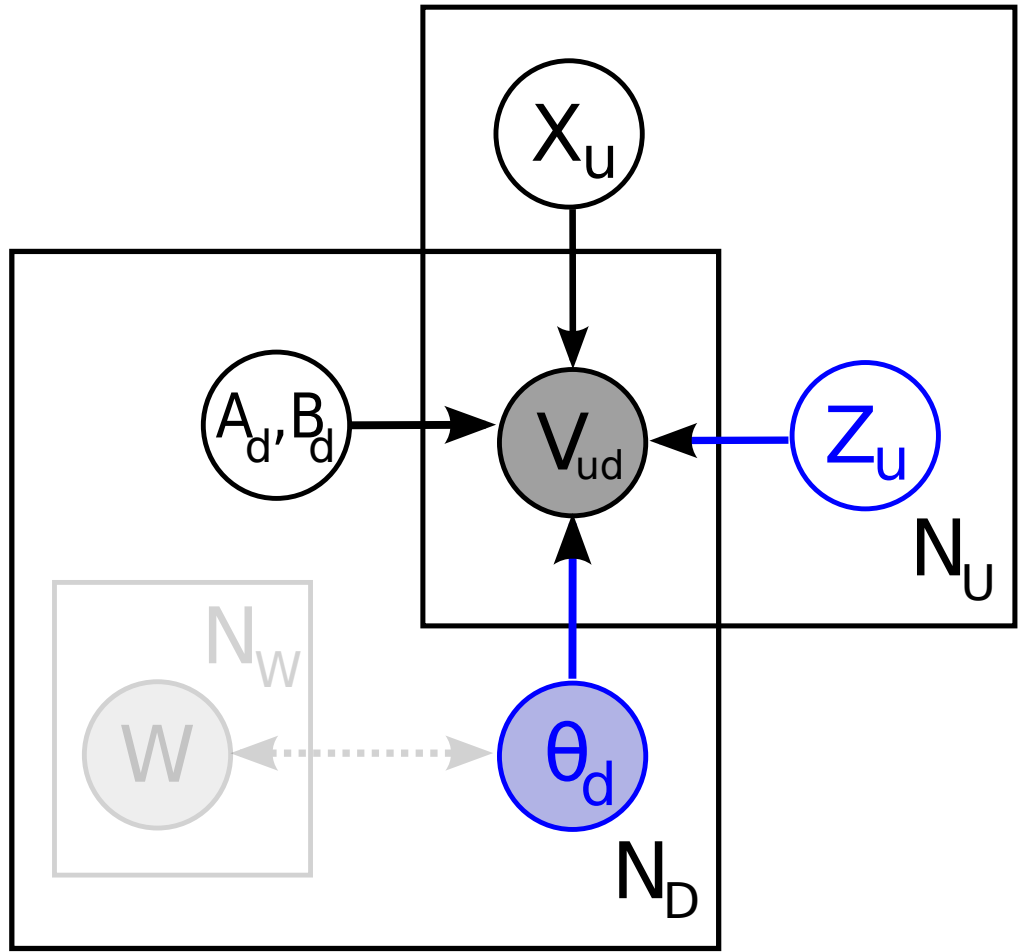
An issue-adjusted ideal point model

$$p(v_{ud} = \text{Yes} | \ldots) = \ldots$$

logistic$(x_u \cdot a_d + b_d)$



logistic$((x_u + \mathbf{z}_u^T \theta_d) \cdot a_d + b_d)$



The U.S. House and Senate, 1999-2010

Congress	**Years**	**Lawmakers**	**Bills**	**Votes** (Senate)
106	1999-2000	516	391	149,035 (7,612)
107	2001-2002	391	137	23,996 (5,547)
108	2003-2004	539	527	207,984 (7,830)
109	2005-2006	540	487	194,138 (7,071)
110	2007-2008	549	745	296,664 (9,019)
111	2009-2010	552	826	336,892 (5,936)

Heldout log-likelihood

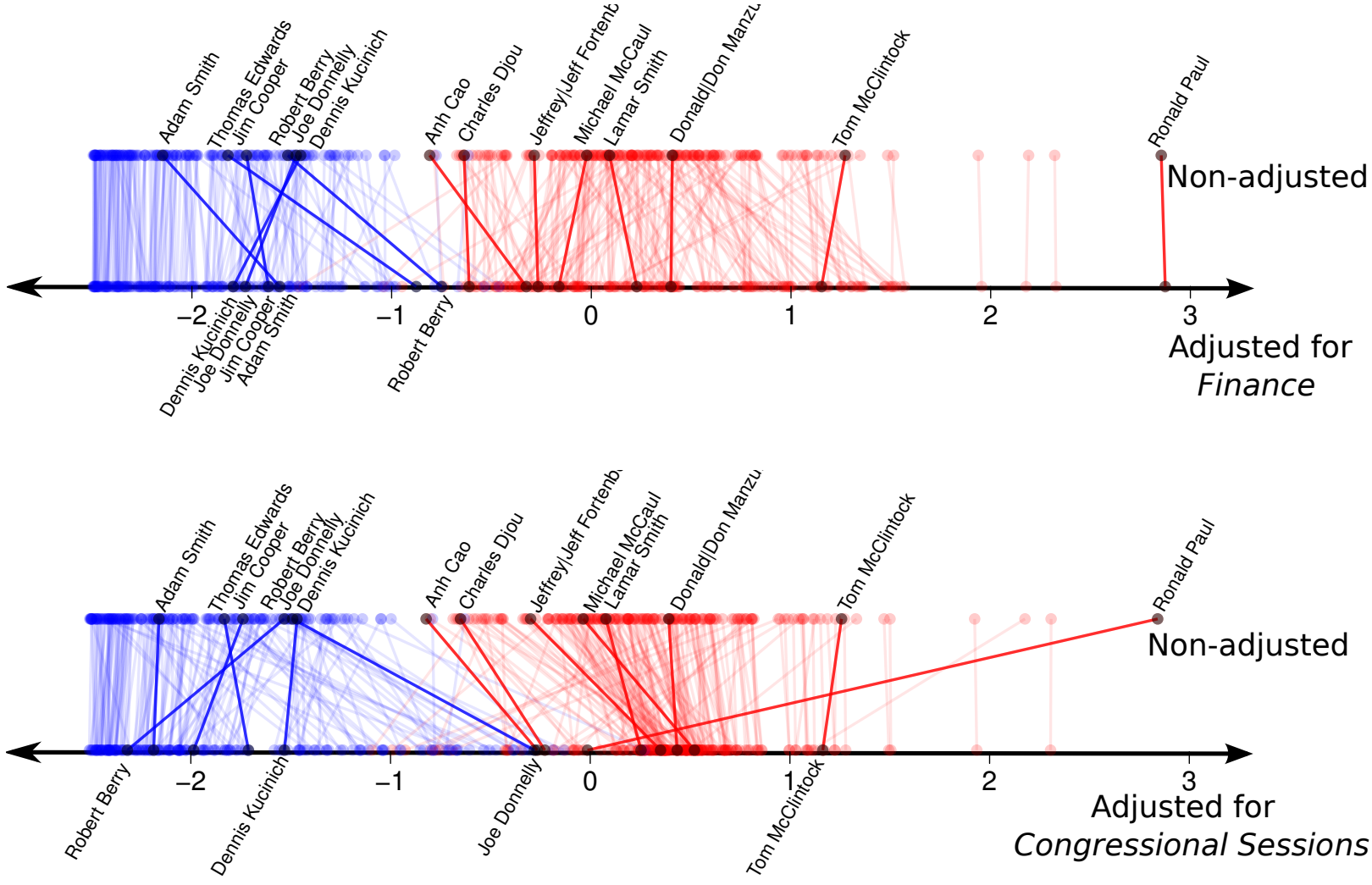
We used 6-fold cross-validation to assess the model.

Model	**Senate**					
Congress	106	107	108	109	110	111
Ideal	-0.209	-0.209	-0.182	-0.189	-0.206	-0.182
Issue	**-0.208**	**-0.209**	**-0.181**	**-0.188**	**-0.205**	**-0.180**
Perm. Issue	-0.210	-0.210	-0.183	-0.203	-0.211	-0.186
	House					
Ideal	-0.168	-0.154	-0.096	-0.120	-0.090	-0.182
Issue	**-0.166**	**-0.147**	**-0.093**	**-0.116**	**-0.087**	**-0.180**
Perm. Issue	-0.210	-0.211	-0.100	-0.123	-0.098	-0.187

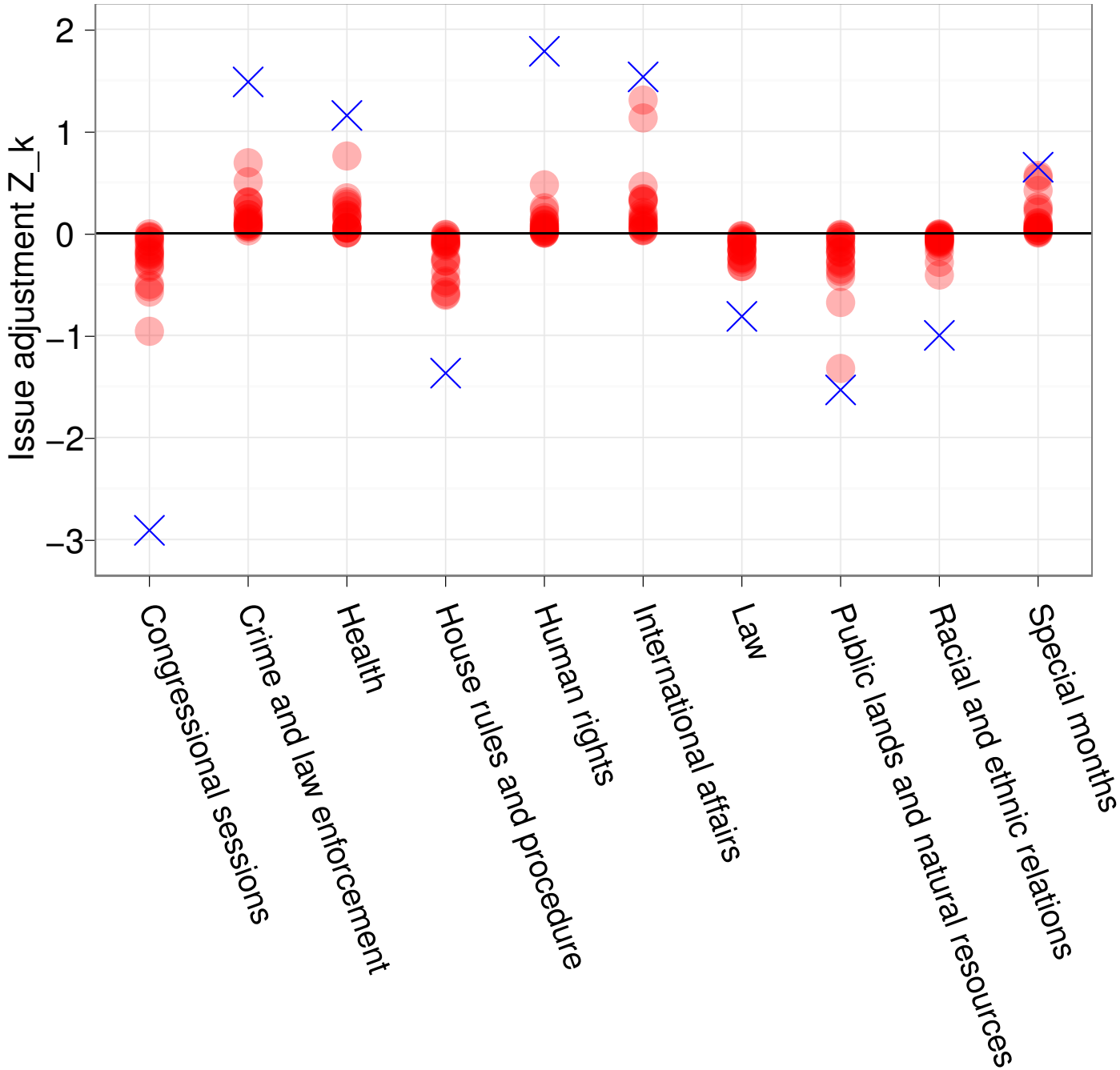
Issue-adjusted ideal points

Issue-adjusted ideal points x_u for specific issues.

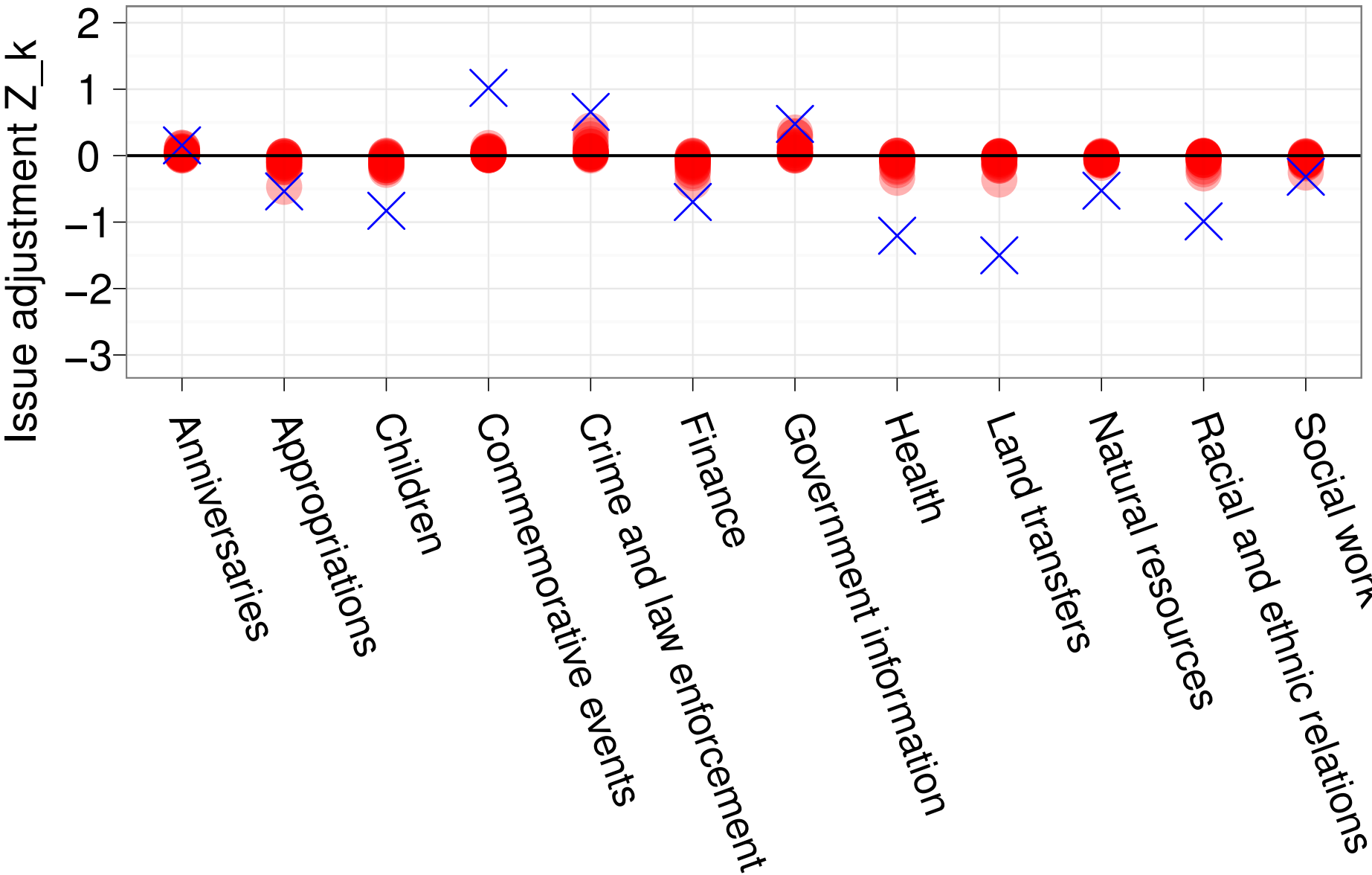
Recall, $\text{logistic}((x_u + \mathbf{z}_u^T \theta_d) \cdot a_d + b_d)$.



Example: Ronald Paul



Example: Donald Young



Summary

Two new models for finding patterns in collections of text data

- A method to find influence in text collections
- A method to map the interactions of countries through time

Automatically infer lawmakers' voting patterns on issues

Funding

Funding

- Google
- National Science Foundation
- Office of Naval Research
- Yahoo!

Bibliography I


Blei, D. and Lafferty, J. (2006).
Dynamic topic models.
Proc. of the 23rd ICML.


Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003).
Latent Dirichlet allocation.
Journal of Machine Learning Research, pages 993–1022.


Chang, J. and Blei, D. M. (2009).
Relational topic models for document networks.
Proceedings of the 12th International Conference on Artificial Intelligence and Statistics (AIStats) 2009, 5.


Chang, J., Boyd-Graber, J., Gerrish, S., Wang, C., and Blei, D. (2009).
Reading tea leaves: How humans interpret topic models.
In *Neural Information Processing Systems (NIPS)*.


Clinton, J., Jackman, S., and Rivers, D. (2004).
The statistical analysis of roll call data,.
American Political Science Review, 98(2):355–370.


Gartzke, E. (1998).
Kant we all just get along? opportunity, willingness, and the origins of the democratic peace.
American Journal of Political Science, 42(1):1–27.


Griffiths, T. L. and Steyvers, M. (2004).
Finding scientific topics.
Proceedings of the National Academy of Sciences, pages 5528–5235.


Hoff, P., Raftery, A. E., and Handcock, M. S. (2002).
Latent space approaches to social network analysis.
Journal of the American Statistical Association, 97:1090–1098.

Bibliography II


Jackman, S. (2001).
Multidimensional analysis of roll call data via bayesian simulation: Identification, estimation, inference, and model checking.
Political Analysis, 9(3):227–241.


Johnson, V. E. and Albert, J. H. (1999).
Ordinal Data Modeling.
Springer-Verlag, New York.


Jordan, M. I., Ghahramani, Z., Jaakkola, T. S., and Saul, L. K. (1999).
An introduction to variational methods for graphical models.
Learning in Graphical Models, pages 183–233.


Kogan, S., Levin, D., Routledge, B., Sagi, J., and Smith, N. (2009).
Predicting risk from financial reports with regression.
In *ACL Human Language Technologies*, pages 272–280. Association for Computational Linguistics.


Martin, A. D. and Quinn, K. M. (2002).
Dynamic ideal point estimation via markov chain monte carlo for the u.s. supreme court, 1953-1999.
Political Analysis, 10:134–153.


Poole, K. T. and Rosenthal, H. (1985).
A spatial model for legislative roll call analysis.
American Journal of Political Science, pages 357–384.


Ramage, D., Hall, D., Nallapati, R., and Manning, C. D. (2009).
Labeled lda: A supervised topic model for credit attribution in multi-labeled corpora.
Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing.


Wells, G., Scott, A., Johnson, C., Gunning, R., Hancock, R., Jeffrey, M., Dawson, M., and Bradley, R. (1987).
A novel progressive spongiform encephalopathy in cattle.
Veterinary record, 121:419–420.

Bibliography III

Appendix

Low influence and high citations

Building a Large Annotated Corpus of English: The Penn Treebank

Mitchell P. Marcus*
University of Pennsylvania

Beatrice Santorini†
Northwestern University

Mary Ann Marcinkiewicz‡
University of Pennsylvania

1. Introduction

There is a growing consensus that significant, rapid progress can be made in both text understanding and spoken language understanding by investigating those phenomena that occur most centrally in naturally occurring unconstrained materials and by attempting to automatically extract information about language from very large corpora. Such corpora are beginning to serve as important research tools for investigators in natural language processing, speech recognition, and integrated spoken language systems, as well as in theoretical linguistics. Annotated corpora promise to be valuable for enterprises as diverse as the automatic construction of statistical models for the grammar of the written and the colloquial spoken language, the development of explicit formal theories of the differing grammars of writing and speech, the investigation of prosodic phenomena in speech, and the evaluation and comparison of the adequacy of parsing models.

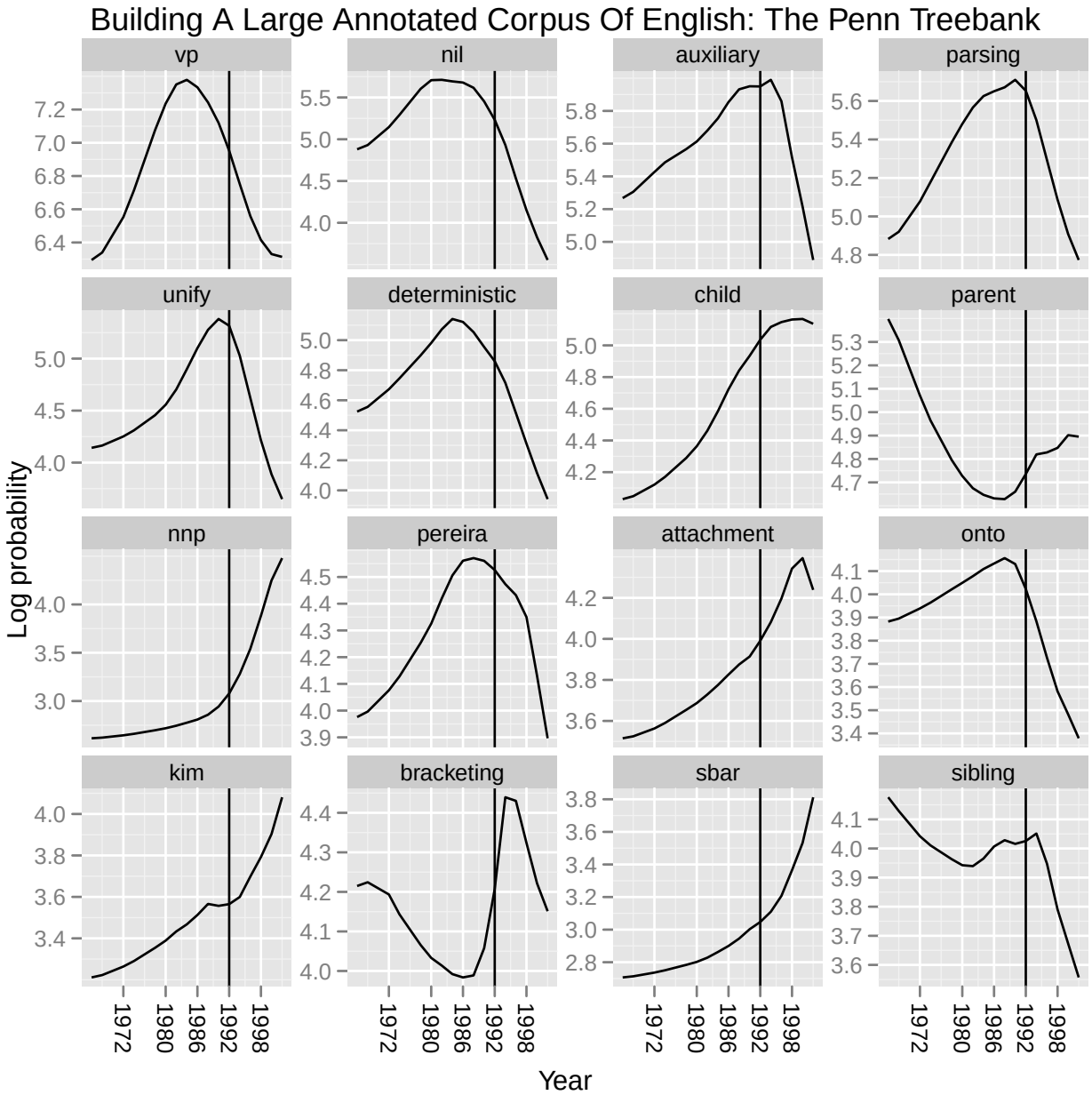
In this paper, we review our experience with constructing one such large annotated corpus—the Penn Treebank, a corpus[1] consisting of over 4.5 million words of American English. During the first three-year phase of the Penn Treebank Project (1989–1992), this corpus has been annotated for part-of-speech (POS) information. In addition, over half of it has been annotated for skeletal syntactic structure. These materials are available to members of the Linguistic Data Consortium; for details, see Section 5.1.

The paper is organized as follows. Section 2 discusses the POS tagging task. After outlining the considerations that informed the design of our POS tagset and presenting the tagset itself, we describe our two-stage tagging process, in which text is first assigned POS tags automatically and then corrected by human annotators. Section 3 briefly presents the results of a comparison between entirely manual and semi-automated tagging, with the latter being shown to be superior on three counts: speed, consistency, and accuracy. In Section 4, we turn to the bracketing task. Just as with the tagging task, we have partially automated the bracketing task: the output of

* Department of Computer and Information Sciences, University of Pennsylvania, Philadelphia, PA 19104.
† Department of Linguistics, Northwestern University, Evanston, IL 60208.
‡ Department of Computer and Information Sciences, University of Pennsylvania, Philadelphia, PA 19104.

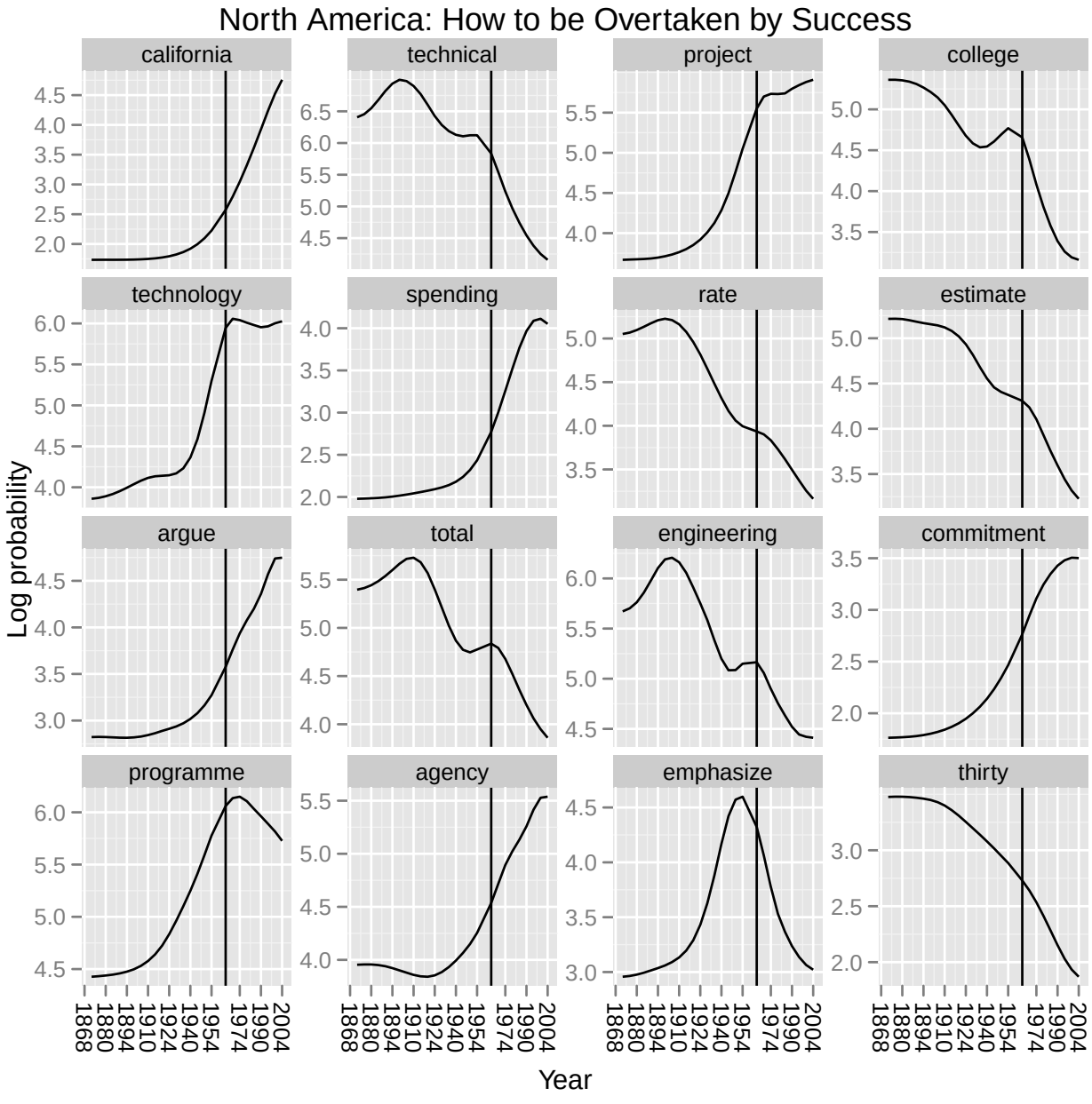
1 A distinction is sometimes made between a *corpus* as a carefully structured set of materials gathered together to jointly meet some design principles, and a *collection*, which may be much more opportunistic in construction. We acknowledge that from this point of view, the raw materials of the Penn Treebank form a collection.

ACL citations: 2180



High influence and low citations

Citations: NA



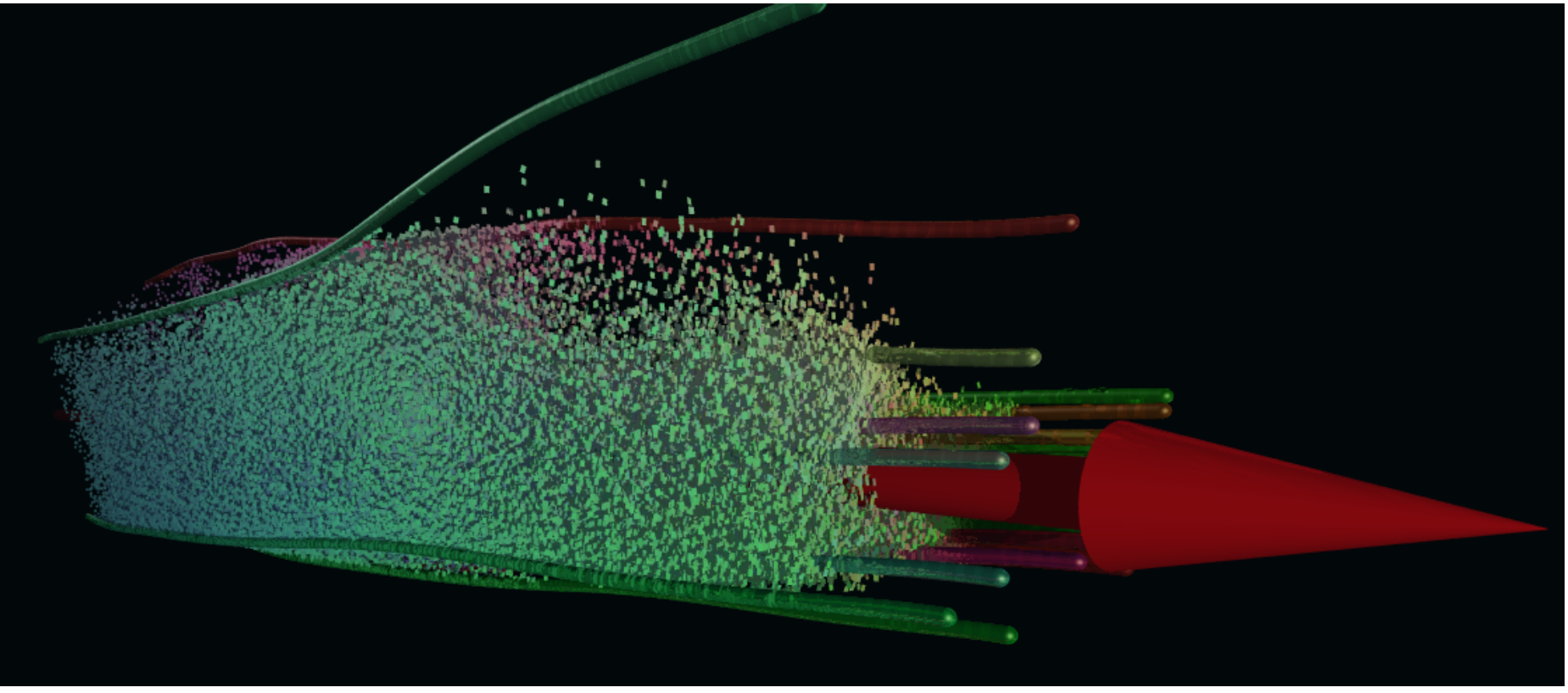
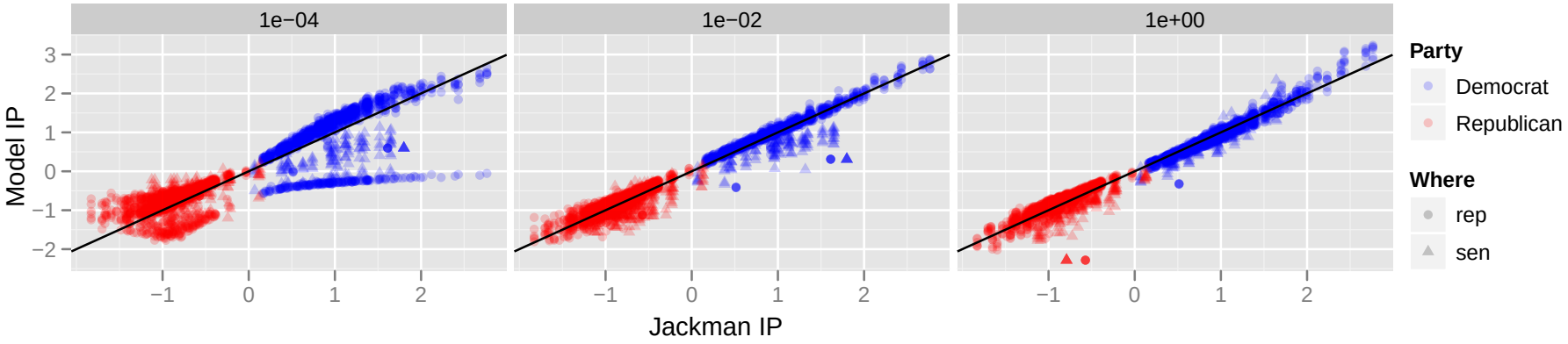
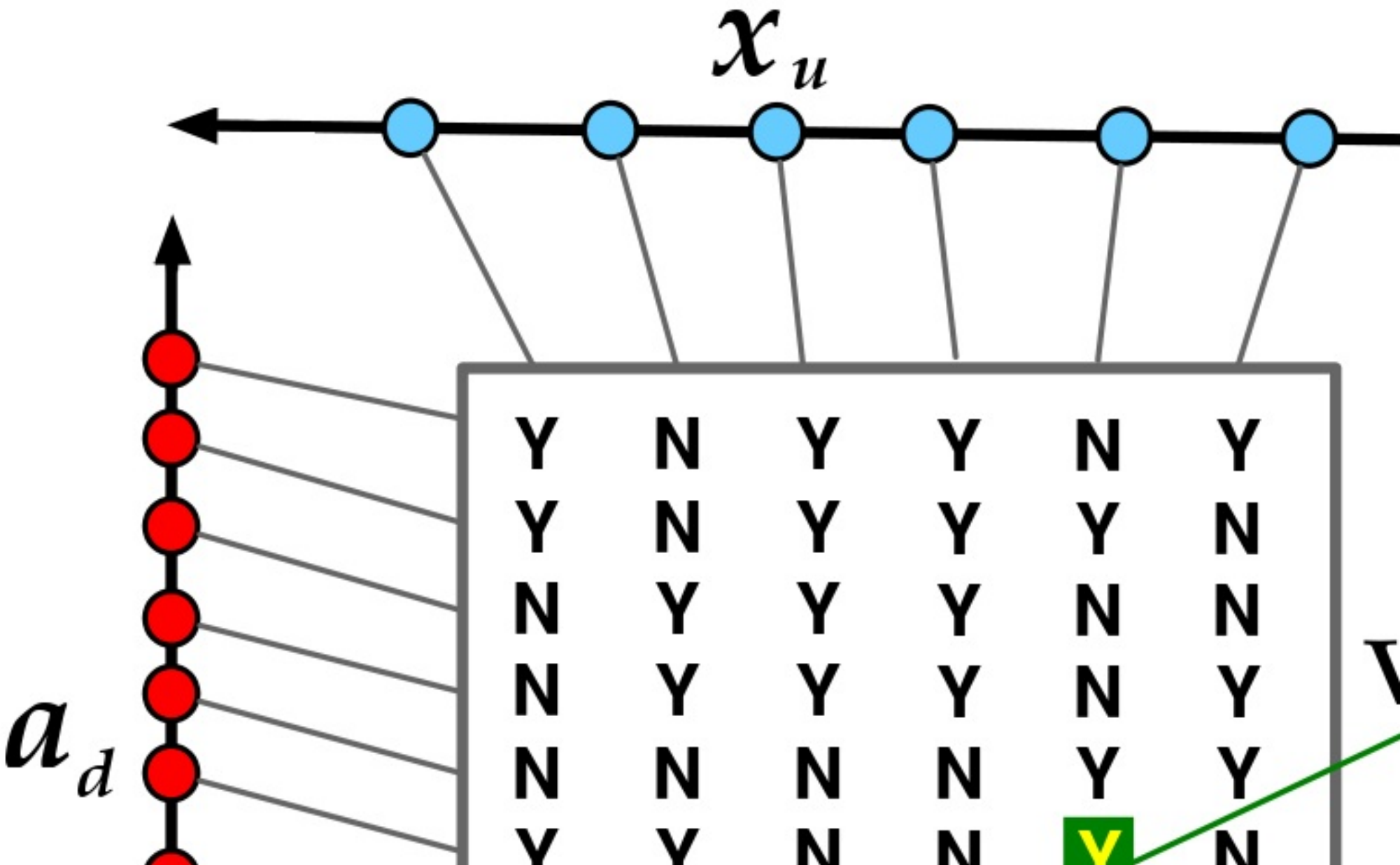


Figure: *Nature* articles and their dynamic topics

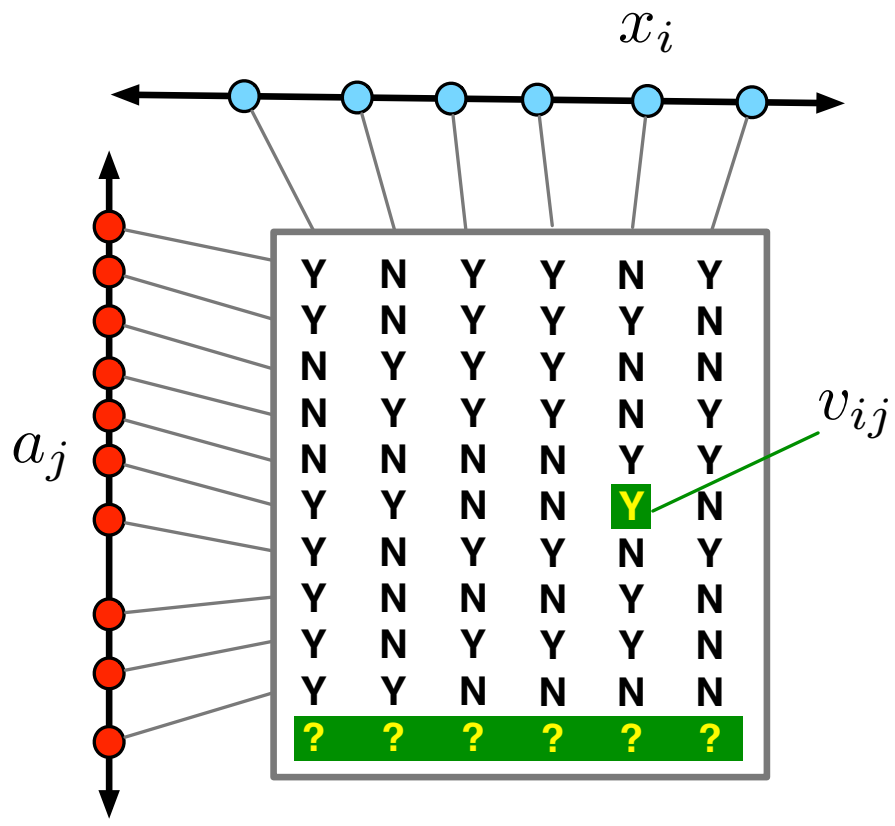

1e−04
1e−02
1e+00
Model IP
Jackman IP
3
2
1
0
−1
−2
−1
0
1
2
Party
Democrat
Republican
Where
rep
sen

Ideal Point Topic Models

The ideal point model



The ideal point model is limited for prediction



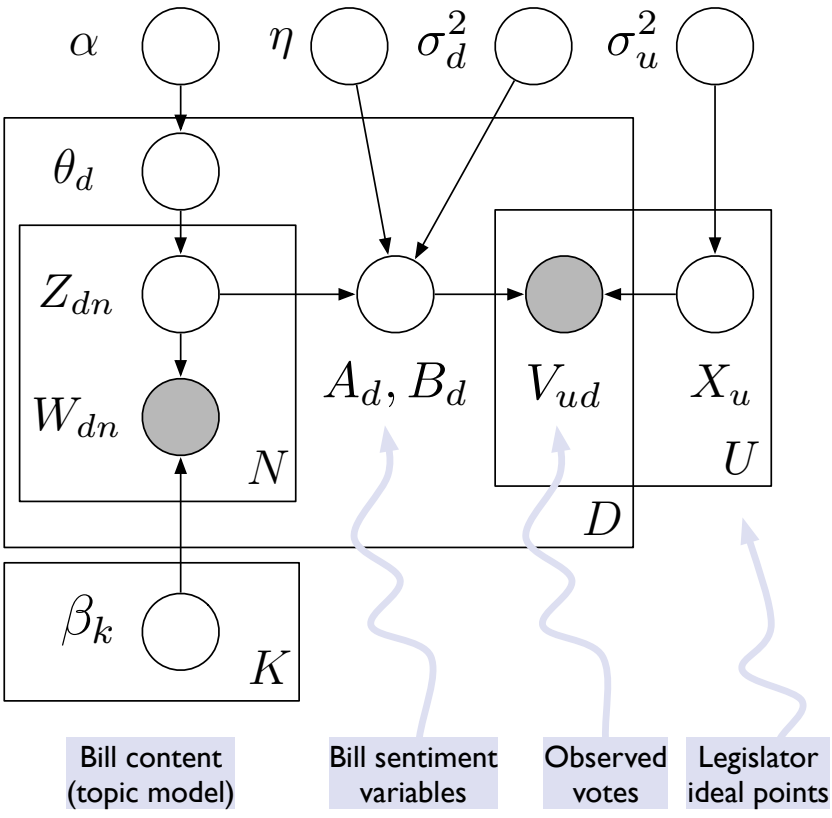
$$p(v_{ij}) = f(d(x_i, a_j))$$

- We can predict a missing vote.
- But we cannot predict all the missing votes from a bill.
- Cf. the limitations of collaborative filtering

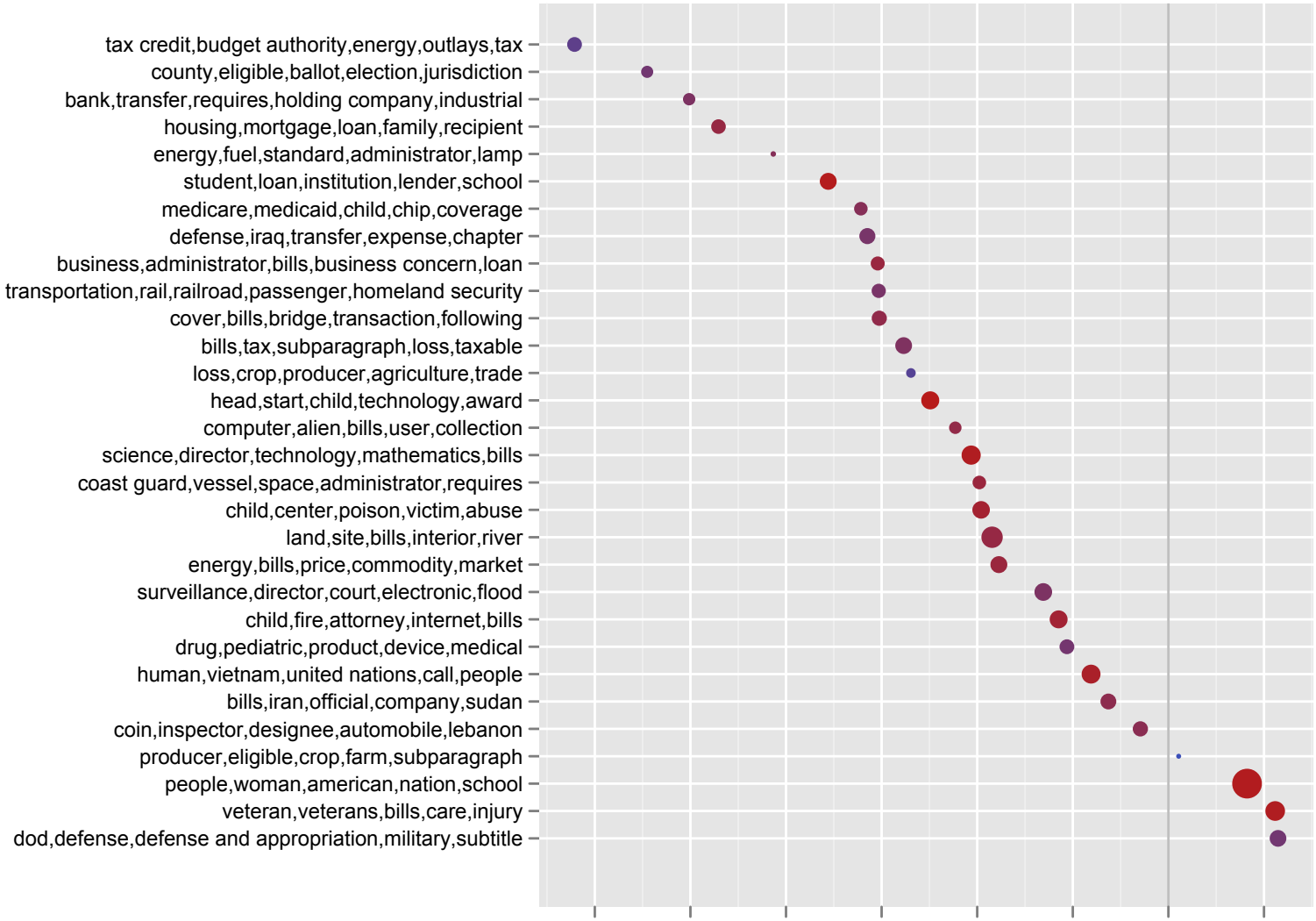
Ideal point topic models

Use supervised topic modeling assumptions as a predictive mechanism from bill texts to bill discrimination.

Ideal point topic models

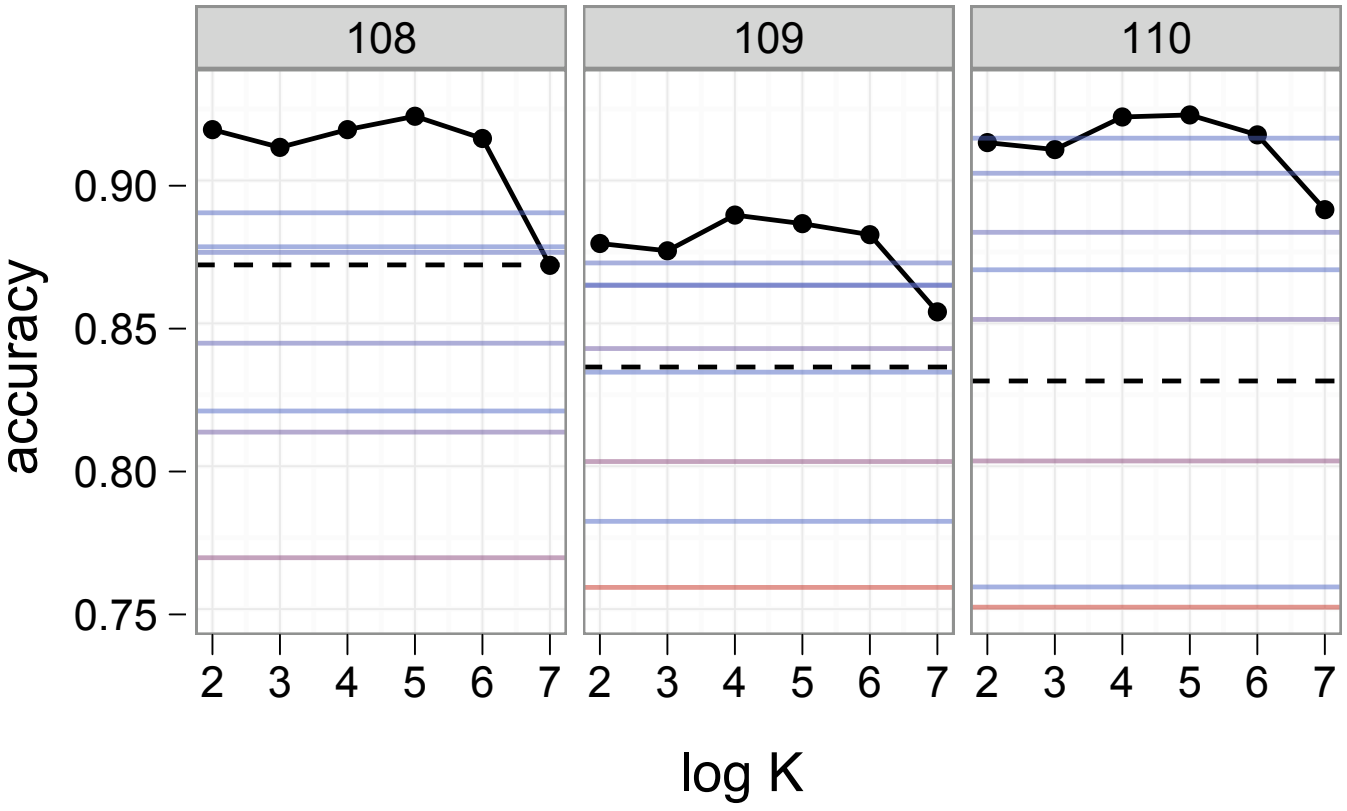


Ideal point topics



In addition to senators and bills, IPTM places **topics** on the spectrum.

Prediction on completely held-out votes



Versus the LASSO, the IPTM correctly predicted 126,000 more votes.

Ideal point topic models

- Ideal point topic model illustrates
  - Topic modeling embedded in a complex model
  - Topic modeling used to solve a real-world problem with text
- More generally, consider collaborative filtering.
  - Senators are *users*.
  - Bills are *items*.
- Existing collaborative filtering is akin to classical ideal point.
- Our model lets us predict preferences on *completely new items*.

Thank you

- Sean Gerrish (sgerrish@cs.princeton.edu)

Baseline

One possible heuristic is simple:

- Define a word's weight at time t as:

$$w_t := \frac{\text{Frequency of } w \text{ in } [t, t+f]}{\text{Frequency of } w \text{ in [t-b, t]}}$$

- Document **D**'s score is the weighted average of these:

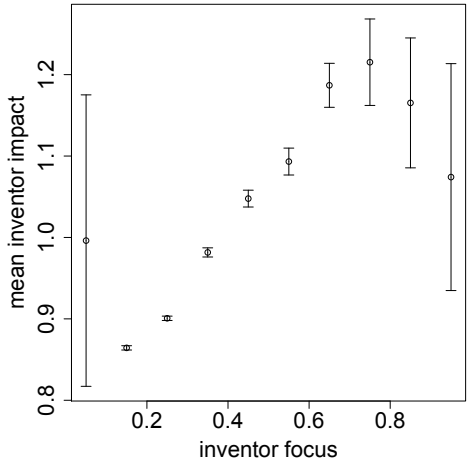
$$\mathcal{I}(\mathbf{D}) := \frac{\sum_{w \in \mathbf{D}} \text{Count}(w) w_t}{\sum_{w \in \mathbf{D}} \text{Count}(w)}$$

Heuristic solution

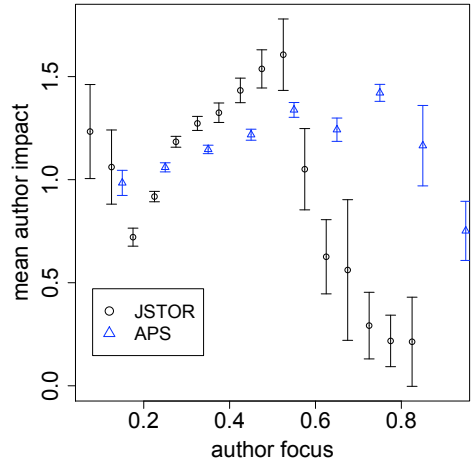
- Fast
- Easy to implement
- Not "optimal" in any obvious sense
- Does not incorporate information about documents' semantics
- Not focused at the level of research contributions
  - Too large a hammer
  - E.g.: cows and health policy around 1986

Focus and knowledge contribution

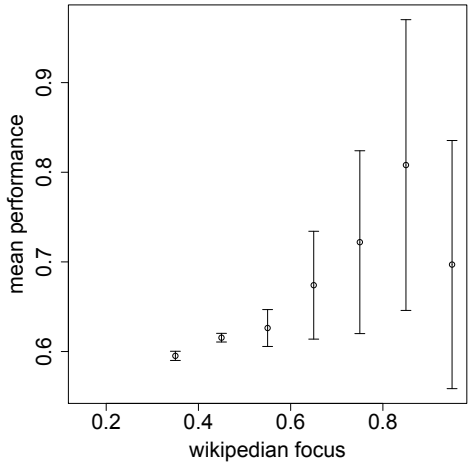
The quality of researchers' contributions increases with focus.



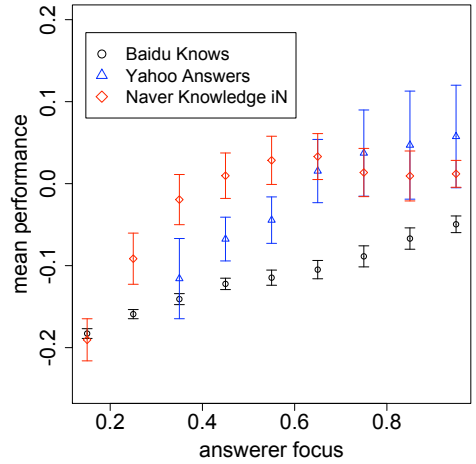
Patents



Research articles



Wikipedia



Q&A forums

Motivation for $\exp(-\beta)$ coefficient in $Infl(t, l, z, w)$

Markov step: $\beta_{t,k} \sim \mathcal{N}(\beta_{t-1,k} + \sum_D \sum_{s<t} \exp(-\beta_k)(W_d \circ Z_{dk}) I_d, \sigma^2 I)$,

$$
\begin{aligned}
& \exp(\beta_t) = \exp(\beta_{t-1}) + Infl_t \\
\iff \quad & 1 = \exp(\beta_{t-1} - \beta_t) + \exp(-\beta_t) Infl_t \\
\iff \quad & 1 - \exp(-\beta_t) Infl_t = \exp(\beta_{t-1} - \beta_t) \\
\iff \quad & \log(1 - \exp(-\beta_t) Infl_t) = \beta_{t-1} - \beta_t \\
\iff \quad & \beta_t = \beta_{t-1} - \log(1 - \exp(-\beta_t) Infl_t)
\end{aligned}
\tag{1}
$$

Note that when $\exp(-\beta_t) Infl_t$ is small, we have
$\beta_t \approx \beta_{t-1} + \exp(-\beta_t) Infl_t$.

Regularized linear regression for $\tilde{l}$ updates

$$g(s,q) := \Lambda_{\exp(-\tilde{m}_{q,k}+\tilde{V}_{q,k}/2)}(\mathbf{W}_{s,k} \circ \phi_{s,k}) \tag{2}$$

$$h(s,q) := ((\mathbf{W}_{s,k} \circ \phi_{s,k})^T \Lambda_{\exp(-2\tilde{m}_q+2\tilde{V}_q)+\exp(-2\tilde{m}_q+\tilde{V}_q)}(\mathbf{W}_{s,k} \circ \phi_{s,k}) \tag{3}$$

$$+\Lambda_{(\mathbf{W}_{s,k}\circ\mathbf{W}_{s,k}\circ(\phi_{s,k}-\phi_{s,k}\circ\phi_{s,k}))^T(\exp(-2\tilde{m}_q+2\tilde{V}_q)+\exp(-2\tilde{m}_q+\tilde{V}_q))}) \tag{4}$$

$$\tilde{l}_{t,k} \leftarrow \left(\frac{\sigma^2}{\sigma_d^2} I + \left(\sum_{i=t}^{T-1} r(i-t)^2 h(t,i)\right)\right)^{-1}$$
$$\left(\sum_{i=t}^{T-1} r(i-t) g(t,i)^T (\tilde{m}_{i+1,k} - \tilde{m}_{i,k} + \tilde{V}_{i,k} - \sum_{j=0\ldots i, j\neq t} r(i-j) g(j,i) \tilde{l}_{j,k})\right) \tag{5}$$

Dimensionality reduction

Bag-of-words model

- Only worry about the word counts in each document
- So a document is basically a sparse list of word counts:

$$\text{“the cat in the hat”} \rightarrow (0, 0, 1, 0, 2, \ldots, 0)$$

Doing anything with these huge lists is hard

- Popular statistics tools like Principal Component Analysis help us to reduce this to a smaller number
- Topic models accomplish a similar thing:

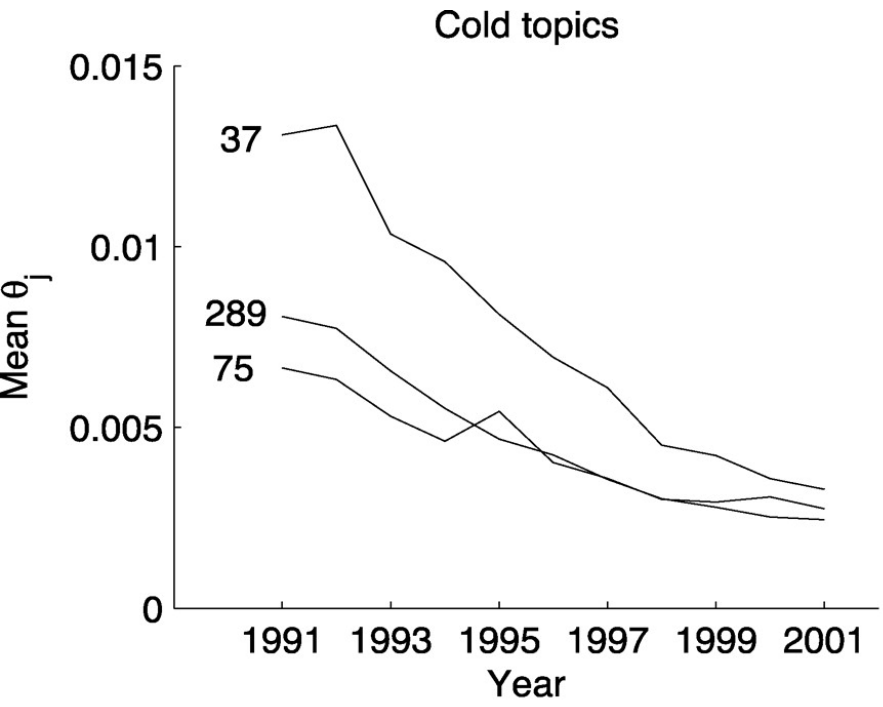
$$10^4 \text{ words} \rightarrow 50 \text{ topics}$$

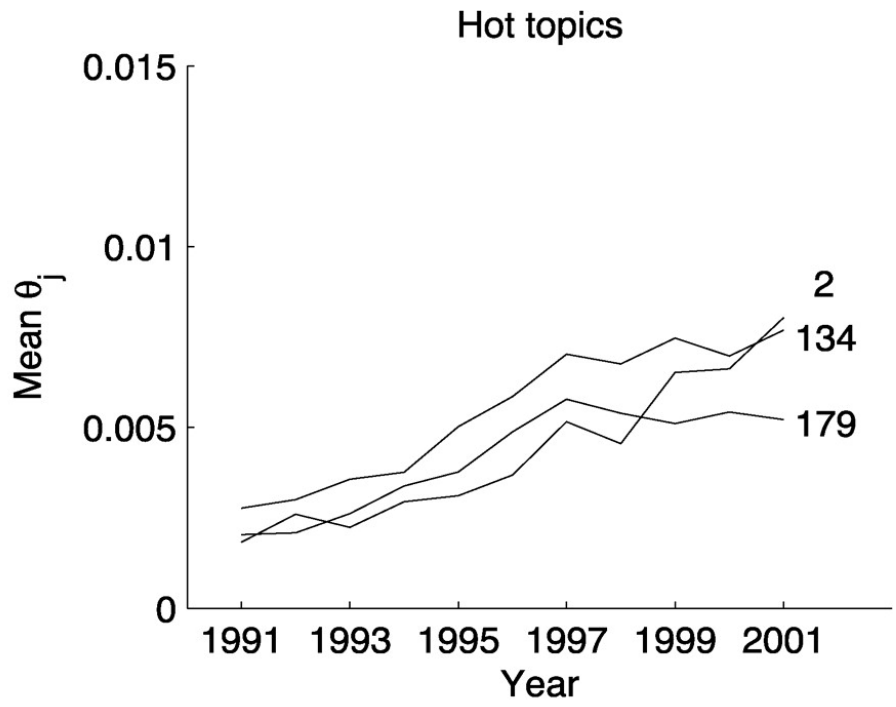
The DIM generative model

For time $t = 1, \ldots, T$:

- For topic $k = 1, \ldots, K$:
  Draw natural parameters
  $\beta_{t,k} | \beta_{t-1,k}, \mathbf{z}_{s<t}, l_{s<t} \sim \mathcal{N}(\beta_{t-1,k} + \mathit{Infl}(t,k), \sigma^2 I)$
- For each document d_t:
  - Generate document d_t using traditional LDA with parameters α_t and β_t.
  - For topic $k = 1, \ldots, K$, draw document weight $l_{d,k} \sim \mathcal{N}(\mathbf{0}, \sigma_d^2 I)$;

Topic models - applications





37	289	75
CDNA	KDA	ANTIBODY
AMINO	PROTEIN	ANTIBODIES
SEQUENCE	PURIFIED	MONOCLONAL
ACID	MOLECULAR	ANTIGEN
PROTEIN	MASS	IGG
ISOLATED	CHROMATOGRAPHY	MAB
ENCODING	POLYPEPTIDE	SPECIFIC
CLONED	GEL	EPITOPE
ACIDS	SDS	HUMAN
IDENTITY	BAND	MABS
CLONE	APPARENT	RECOGNIZED
EXPRESSED	LABELED	SERA

2	134	179
SPECIES	MICE	APOPTOSIS
GLOBAL	DEFICIENT	DEATH
CLIMATE	NORMAL	CELL
CO2	GENE	INDUCED
WATER	NULL	BCL
ENVIRONMENTAL	MOUSE	CELLS
YEARS	TYPE	APOPTOTIC
MARINE	HOMOZYGOUS	CASPASE
CARBON	ROLE	FAS
DIVERSITY	KNOCKOUT	SURVIVAL
OCEAN	DEVELOPMENT	PROGRAMMED
EXTINCTION	GENERATED	MEDIATED

Image source: [Griffiths and Steyvers, 2004]; Available 2010-01-15

http://www.pnas.org/content/101/suppl.1/5228/F5.large.jpg

Posterior Inference

Recall that we only observe words $\mathbf{W}$ and votes V.
We are interested in the posterior

$$p(\lambda_d, \kappa_d, x_u, \eta | V, \mathbf{W})$$

We derived a mean-field variational inference algorithm.

- This involves positing a family of fully factorized posterior distributions and finding the distribution from this family which is "closest" in K-L divergence to the true posterior.
- The resulting ideal points are correlated at over 0.98 with MCMC ideal points (the standard in this field).
- Variational methods like this are amenable to inference in large-scale datasets. Hoffman et al. 2010

Derivation and implementation of variational inference

Finding the variational posterior involves optimizing a lower bound on the model evidence $p(V, \mathbf{W})$.

This objective is optimized via gradient ascent, and we can use a few tricks to find the objective and converge better, e.g.:

- Use a second-order delta approximation for the bound Bickel et al. 2007
- Update subsets of lawmakers and legislation in a round-robin fashion to avoid cycles

Sterling similarity

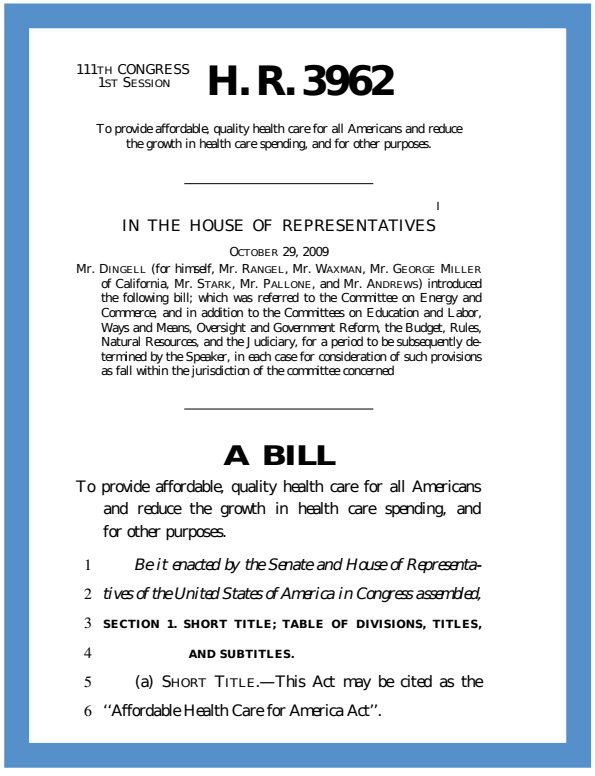
We aimed to use a metric that captures three qualities: *variety*, or how many different areas an individual contributes to; *balance*, or how evenly their efforts are distributed among these areas; and *similarity*, or how related those areas are. We use the Stirling measure $\mathcal{F}$, which captures all three aspects:

$$\mathcal{F} := \sum_{i,j} s_{ij} p_i p_j,$$

where p_i is the proportion of the individual's contributions in category i and $s_{ij} = n_{ij}/n_j$ is a measure of similarity between categories i and j, inferred from the number of joint contributors n_{ij} between two categories i and j.

Exploration

Understand the themes in a collection of documents and how they relate to one another



111TH CONGRESS
1ST SESSION

H. R. 3962

To provide affordable, quality health care for all Americans and reduce the growth in health care spending, and for other purposes.

IN THE HOUSE OF REPRESENTATIVES

OCTOBER 29, 2009

Mr. DINGELL (for himself, Mr. RANGEL, Mr. WAXMAN, Mr. GEORGE MILLER of California, Mr. STARK, Mr. PALLONE, and Mr. ANDREWS) introduced the following bill; which was referred to the Committee on Energy and Commerce, and in addition to the Committees on Education and Labor, Ways and Means, Oversight and Government Reform, the Budget, Rules, Natural Resources, and the Judiciary, for a period to be subsequently determined by the Speaker, in each case for consideration of such provisions as fall within the jurisdiction of the committee concerned

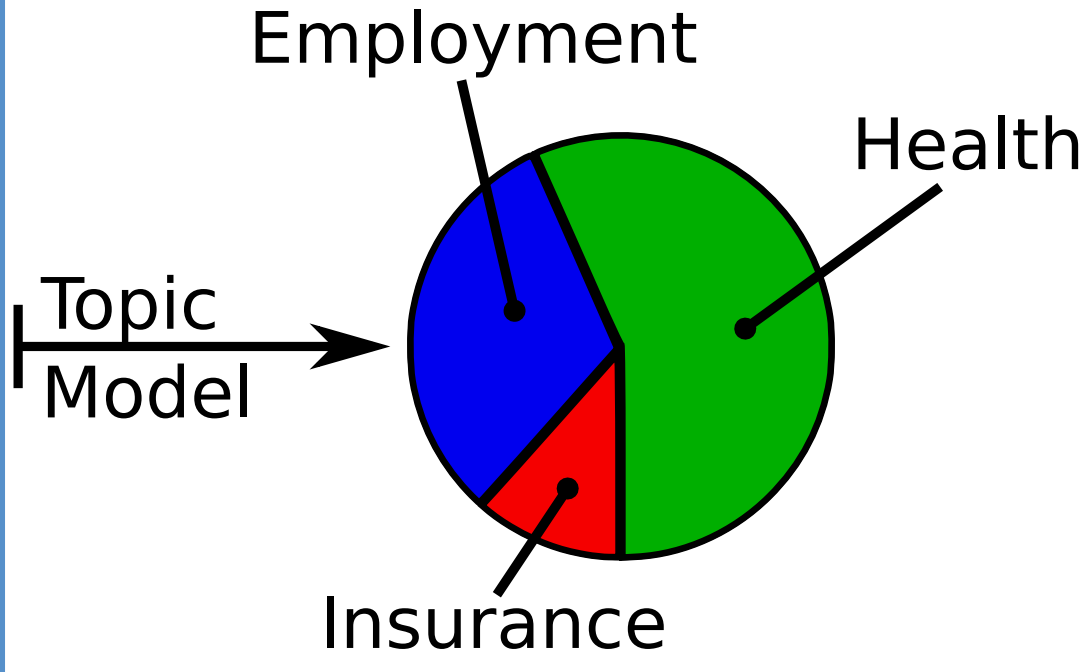
A BILL

To provide affordable, quality health care for all Americans and reduce the growth in health care spending, and for other purposes.

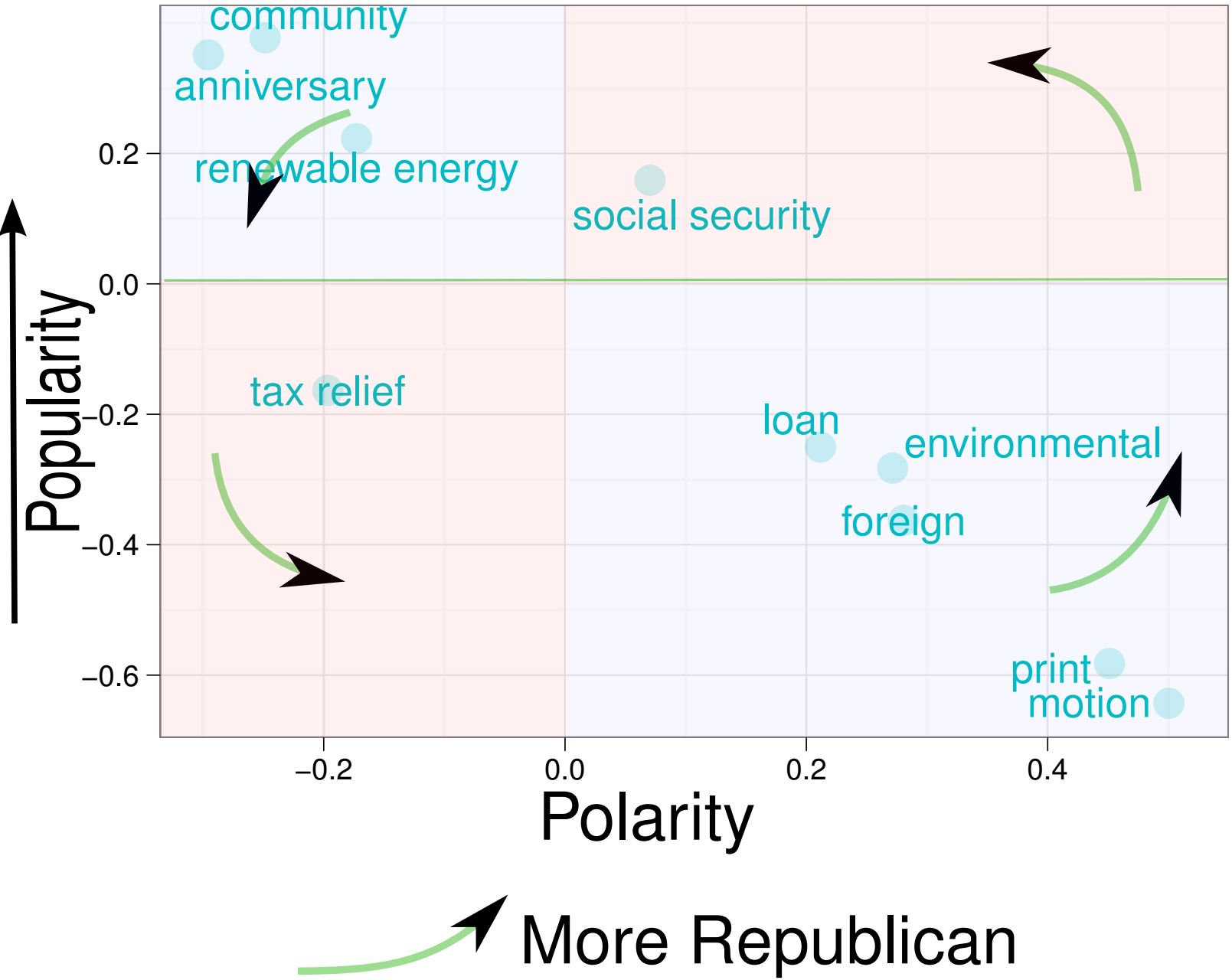
Be it enacted by the Senate and House of Representatives of the United States of America in Congress assembled,

SECTION 1. SHORT TITLE; TABLE OF DIVISIONS, TITLES, AND SUBTITLES.

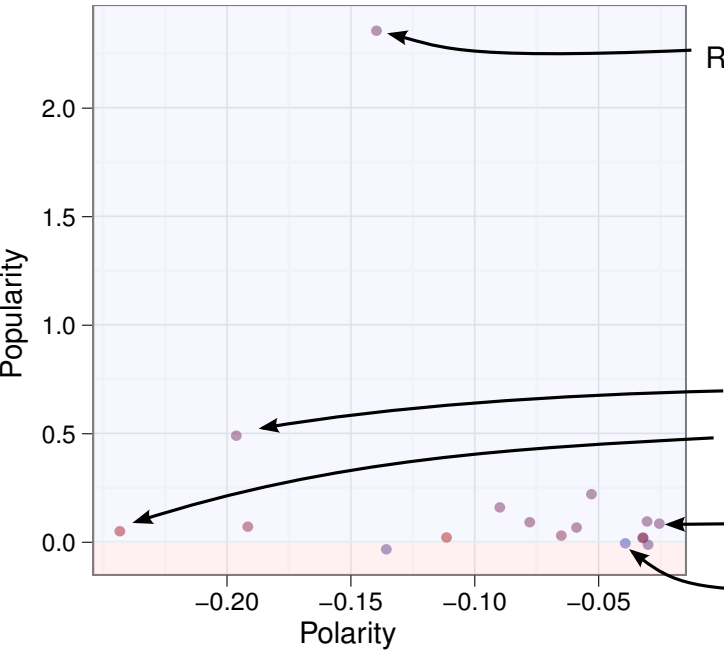
(a) SHORT TITLE.—This Act may be cited as the "Affordable Health Care for America Act".



Example - Ridge Regression parameters $\eta_\lambda, \eta_\kappa$



Results - Topics



Recognition: "people, month, recognize, history week, woman, awareness"

Position of topic in ideal space

−10
−5
0
5
10

Science and Industry: "research, technology, science, director, manufacturing

Finance and Lending: "mortgage, financial, compensation, bills, financial institution"

Land Management: "land, subtitle, wilderness, river, map, water, description"

Environmental Policy: "energy, emission, epa, administrator, allowance, standard"