Statistical Text Analysis for Social Science

Learning to Extract International Relations from the News

Brendan O'Connor

Machine Learning Department

Carnegie Mellon University

http://brenocon.com

UW iSchool, Feb 24, 2014

Computational Social Science

Official social data

Data collection

100 BCE

Data analysis

1829

1900

2000

Computational Social Science

Official social data

Data collection

100 BCE

Data analysis

1829

Newly available social data

Digitized behavior
Billions of users
Billions of messages/day

Digitized news
Thousands of articles/day

Digitized archives
Millions of books/century

HATHI TRUST Digital Library

Google books

1900 2000

Text as "data"?

Details Agreed on Nuclear Deal With Iran, Set to Start Jan. 20
PARIS — Iran and six world powers have agreed on how to put in place an accord that would temporarily freeze much of Iran's nuclear program, American and Iranian officials said on Sunday. That accord would go into effect on Jan. 20. International negotiators worked out an agreement in November to constrain much of Iran's program for six months so that diplomats would have time to pursue a more comprehensive follow-up accord. But before the temporary agreement could take effect, negotiators had to work out the technical procedures for carrying it out and resolve some of its ambiguities in concert with the International Atomic Energy Agency.
Antigovernment Protesters Try to Shut Down Bangkok
BANGKOK — Antigovernment protesters seeking to block next month's elections in Thailand took over major roads in Bangkok on Sunday as they began their campaign to shut down the city. In this vast metropolis of well over 10 million people, the protesters were unlikely to paralyze all movement and commerce. But they vowed that by Monday morning they would close busy intersections, make major government offices inaccessible and besiege the homes of top officials in the administration of Prime Minister Yingluck Shinawatra, whose party is most likely to win the general elections that are scheduled for Feb. 2. "We have to shut down Bangkok," said Ratchanee Saengarun, a protester who stood in the middle of an intersection in the city. "This is our last resort." By late Sunday, protesters had blocked several roads using double-decker buses and sandbags, and had diverted traffic.

Text as "data"?

46 183 3388 43 135 2727 35258 149 14001 69 24 225
37 57124 7 9641 176 252 15 2086 183 3388 218 14001 161 10830 97 2128 33
5268 1459 28 5 449 14210 6966 43 45564 360 9641 3 363 3734 3388 39465
5268 33 1459 165 570 90 3388 24 7097 261 11 48 611 2128 197 10830 42
14001 2 449 14210 16347 398 5338 176 442 499 5268 5 1459 2086 480
14001 26 12709 1251 23 1 27181 2248 338 30775 28 197 739 248 38678 11
1139 14001 257 611 30775 37 24 5338 20 3837 611 9641 17 1073 14210
2341 2 10830 3 2727 30775 261 1 85 88741
17877 10 70 14001 11 438 2
2 65417 59555 10 87 14001 40 427 43199 31 10830 3 152 560 367 7 10830 2
3388 19 2857 1639 129 1159 73 14001 11 438 30775 47956 10830 1529 15
75989 14210 260 560 327 2692 51472 30775 10 1177 23 14001 90351 717 30
9641 24040 2248 1639 9 5268 2811 135 39 1639 1459 199 20 13554 406 367
552 51 1 9641 35951 30775 37 14210 121 363 10830 30775 165 14210 57 59
90525 87723 108 78 4750 597 179 14001 60 30775 257 31 5268 2563 68
5338 14 15012 2679 2086 14001 11 438 14456 3734 16286 44733 12709 1
1031 14 10830 30775 25 14210 2128 49392 10830 30775 20260 738 4750
250 797 32407 2811 195 90338 10 1139 4 244 7 111 3 7 9641 75964 9641
1139 5 95973

Text as “data”?

46 183 3388 43 135 2727 35258 149 14001 69 24 225
37 57124 7 9641 176 252 15 2086 183 3388 218 14001 161 10830 97 2128 33
5268 1459 28 5 449 14210 6966 43 45564 360 9641 3 363 3734 3388 39465
097 261 11 48 611 2128 197 10830 42
176 442 499 5268 5 1459 2086 480
48 338 30775 28 197 739 248 38678 11
38 20 3837 611 9641 17 1073 14210
5 88741

43199 31 10830 3 152 560 367 7 10830 2
01 11 438 30775 47956 10830 1529 15
72 30775 10 1177 23 14001 90351 717 30
135 39 1639 1459 199 20 13554 406 367
552 51 1 9641 35951 30775 37 14210 121 363 10830 30775 165 14210 57 59
90525 87723 108 78 4750 597 179 14001 60 30775 257 31 5268 2563 68
5338 14 15012 2679 2086 14001 11 438 14456 3734 16286 44733 12709 1
1031 14 10830 30775 25 14210 2128 49392 10830 30775 20260 738 4750
250 797 32407 2811 195 90338 10 1139 4 244 7 111 3 7 9641 75964 9641
1139 5 95973

Data collection

Data analysis

Social discovery and measurement from text

Society

Writing

HATHI TRUST Digital Library

Text

Social discovery and measurement from text

Society

1. Infer attributes of society from text data: opinion, events...

Writing

HATHI TRUST Digital Library

Text

Social discovery and measurement from text

Society

1. Infer attributes of society from text data: opinion, events...

2. Learn about the text generation process: bias, influence, media...

Writing

HATHI TRUST Digital Library

Text

Discovery and measurement in social media

Data

Statistical text analysis

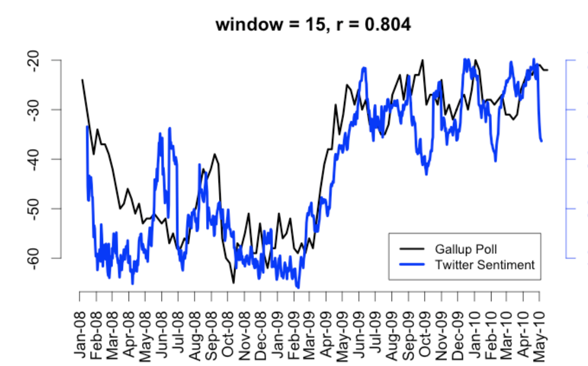
Linguistic analysis tools

[ACL 2011, NAACL 2013]

ikr	smh	he	asked	fir	yo	last	name
!	G	O	V	P	D	A	N

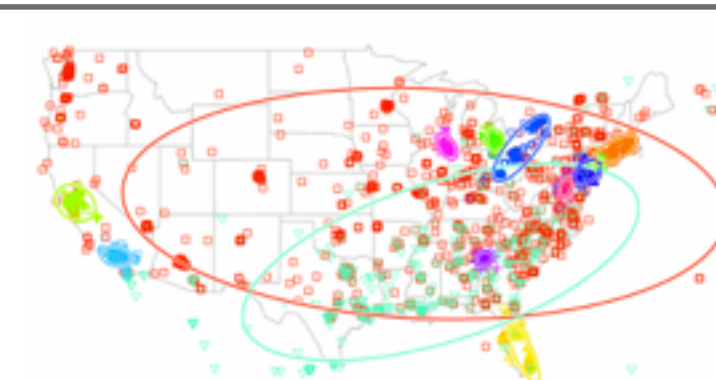
Opinion polls and sentiment analysis

[O'Connor, Balasub., Routledge, Smith 2010]



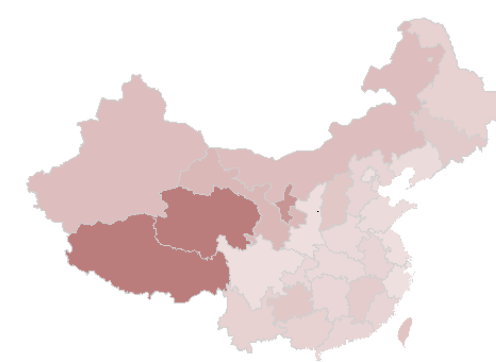
Geographic and demographic factors in slang and language change

[Eisenstein, O'Connor, Xing, Smith 2010, 2012]



Censorship in Chinese social media

[Bamman, O'Connor, Smith 2011]



Analysis methods for **Text** and **Social Context**

concepts, attitudes, events

community, author, time, space

Statistics

Computation

Linguistics

... motivated by analysis problems in the social sciences and humanities

Politics

Literature

Business

Economics

Sociology

Health

Topics

- Textual social data
- **Linguistic semantic learning**

- Examples
  - Sentiment and opinion polls
  - **International relations**
  - Geography and slang
  - Linguistic tools
  - Chinese censorship

International Relations

- Forecasting: When and where will future conflicts happen?
- Understanding: What causes war, peace, trade? How do conflicts resolve?
- Tools to acquire better data

Text as "data"?

Details Agreed on Nuclear Deal With Iran, Set to Start Jan. 20
PARIS — Iran and six world powers have agreed on how to put in place an accord that would temporarily freeze much of Iran's nuclear program, American and Iranian officials said on Sunday. That accord would go into effect on Jan. 20. International negotiators worked out an agreement in November to constrain much of Iran's program for six months so that diplomats would have time to pursue a more comprehensive follow-up accord. But before the temporary agreement could take effect, negotiators had to work out the technical procedures for carrying it out and resolve some of its ambiguities in concert with the International Atomic Energy Agency.
Antigovernment Protesters Try to Shut Down Bangkok
BANGKOK — Antigovernment protesters seeking to block next month's elections in Thailand took over major roads in Bangkok on Sunday as they began their campaign to shut down the city. In this vast metropolis of well over 10 million people, the protesters were unlikely to paralyze all movement and commerce. But they vowed that by Monday morning they would close busy intersections, make major government offices inaccessible and besiege the homes of top officials in the administration of Prime Minister Yingluck Shinawatra, whose party is most likely to win the general elections that are scheduled for Feb. 2. "We have to shut down Bangkok," said Ratchanee Saengarun, a protester who stood in the middle of an intersection in the city. "This is our last resort." By late Sunday, protesters had blocked several roads using double-decker buses and sandbags, and had diverted traffic.

Text as “data”?

Details Agreed on Nuclear Deal With Iran, Set to Start Jan. 20
PARIS — Iran and six world powers have agreed on how to put in place an accord that would temporarily freeze much of Iran’s nuclear program, American and Iranian officials said on Sunday. That accord would go into effect on Jan. 20. International negotiators worked out an agreement in November to constrain much of Iran’s program for six months so that diplomats would have time to pursue a more comprehensive follow-up accord. But before the temporary agreement could take effect, negotiators had to work out the technical procedures for carrying it out and resolve some of its ambiguities in concert with the International Atomic Energy Agency.
Antigovernment Protesters Try to Shut Down Bangkok
BANGKOK — Antigovernment protesters seeking to block next month’s elections in Thailand took over major roads in Bangkok on Sunday as they began their campaign to shut down the city. In this vast metropolis of well over 10 million people, the protesters were unlikely to paralyze all movement and commerce. But they vowed that by Monday morning they would close busy intersections, make major government offices inaccessible and besiege the homes of top officials in the administration of Prime Minister Yingluck Shinawatra, whose party is most likely to win the general elections that are scheduled for Feb. 2. “We have to shut down Bangkok,” said Ratchanee Saengarun, a protester who stood in the middle of an intersection in the city. “This is our last resort.” By late Sunday, protesters had blocked several roads using double-decker buses and sandbags, and had diverted traffic.

Text as "data"?

Details Agreed on Nuclear Deal With Iran, Set to Start Jan. 20

PARIS — Iran and six world powers have agreed on how to put in place an accord that would temporarily freeze much of Iran's nuclear program, American and Iranian officials said on Sunday. That accord would go into effect on Jan. 20. International negotiators worked out an agreement in November to constrain much of Iran's program for six months so that diplomats would have time to pursue a more comprehensive follow-up accord. But before the temporary agreement could take effect, negotiators had to work out the technical procedures for carrying it out and resolve some of its ambiguities in concert with the International Atomic Energy Agency.

Antigovernment Protesters Try to Shut Down Bangkok

BANGKOK — Antigovernment protesters seeking to block next month's elections in Thailand took over major roads in Bangkok on Sunday as they began their campaign to shut down the city. In this vast metropolis of well over 10 million people, the protesters were unlikely to paralyze all movement and commerce. But they vowed that by Monday morning they would close busy intersections, make major government offices inaccessible and besiege the homes of top officials in the administration of Prime Minister Yingluck Shinawatra, whose party is most likely to win the general elections that are scheduled for Feb. 2. "We have to shut down Bangkok," said Ratchanee Saengarun, a protester who stood in the middle of an intersection in the city. "This is our last resort." By late Sunday, protesters had blocked several roads using double-decker buses and sandbags, and had diverted traffic.

Text as “data”?

Details Agreed on Nuclear Deal With Iran, Set to Start Jan. 20

PARIS — Iran and six world powers have agreed on how to put in place an accord that would temporarily freeze much of Iran’s nuclear program, American and Iranian officials said on Sunday. That accord would go into effect on Jan. 20. International negotiators worked out an agreement in November to constrain much of Iran’s program for six months so that diplomats would have time to pursue a more comprehensive follow-up accord. But before the temporary agreement could take effect, negotiators had to work out the technical procedures for carrying it out and resolve some of its ambiguities in concert with the International Atomic Energy Agency.

Antigove

BANGKOK … s in
Thailand … ign to
shut dow … otesters
were unli … Monday
morning … accessible
and besie … gluck
Shinawat … duled for
Feb. 2. “ … r who
stood in … e Sunday,
protester … nd had
diverted …

Semantic parsing
a.k.a.
Information extraction

[e.g. MUC-3: *Lehnert, Williams, Cardie, Riloff, Fisher 1991*]

Event data through knowledge engineering

[Schrodt 1994, Leetaru and Schrodt 2013]

Event classes
(~200)

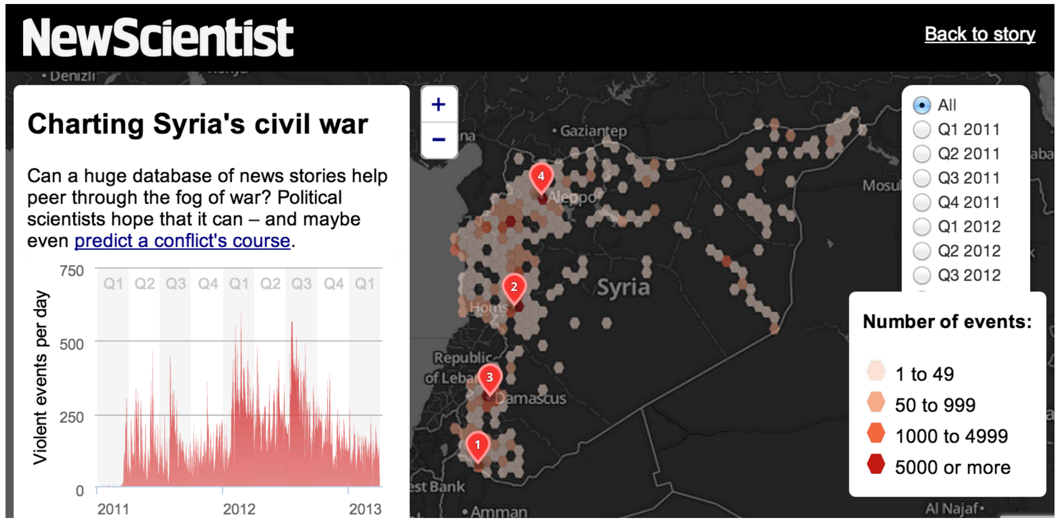
03 - EXPRESS INTENT TO COOPERATE
07 - PROVIDE AID
15 - EXHIBIT MILITARY POSTURE

Dictionary:
Verb patterns per event class
(~15000)

191 - Impose blockade, restrict movement

not_ allow to_ enter ;mj 02 aug 2006
barred travel
block traffic from ;ab 17 nov 2005
block road ;hux 1/7/98

Extract events from news text



Issue: Hard to maintain and adapt to new domains

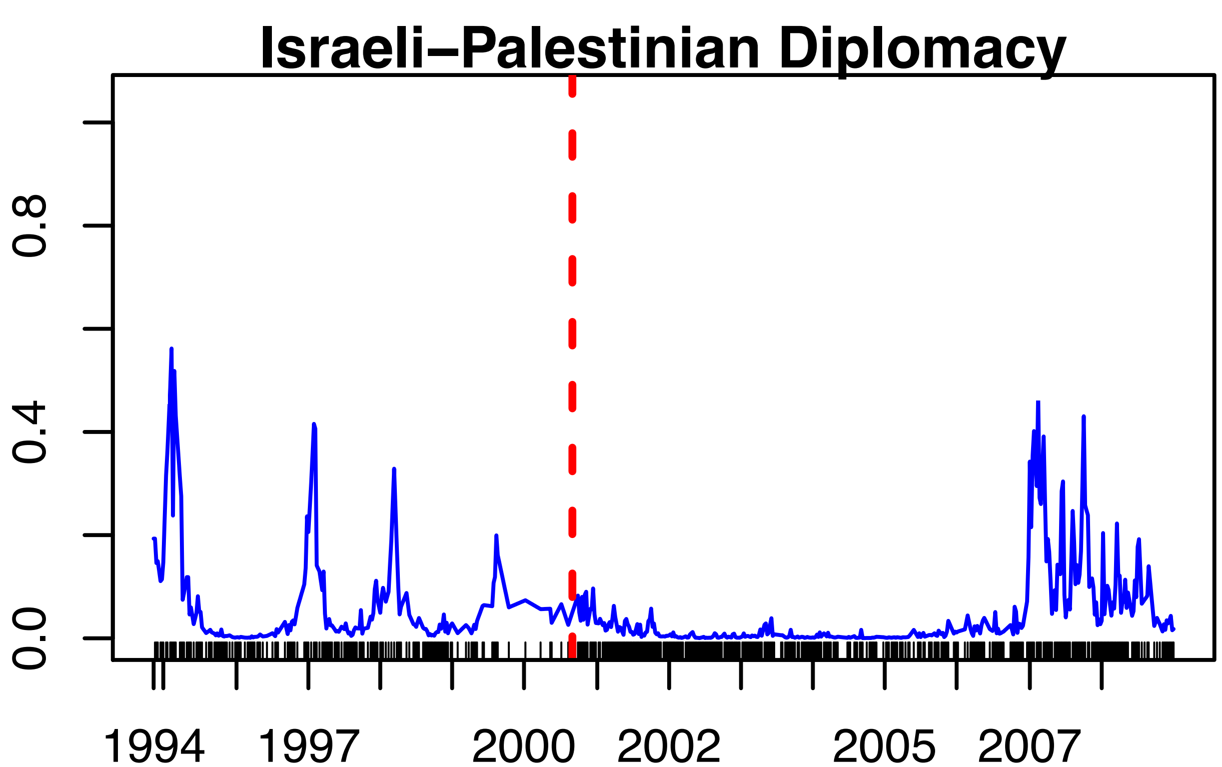
Our approach

[O'Connor, Stewart, Smith *Assoc. Comp. Ling. 2013*]

Natural Language Processing

Event phrases

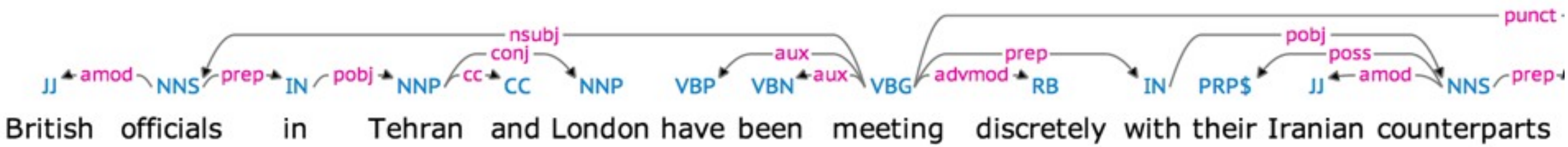
Probabilistic Graphical Model



Jointly learn

- Event class dictionaries
- Political dynamics

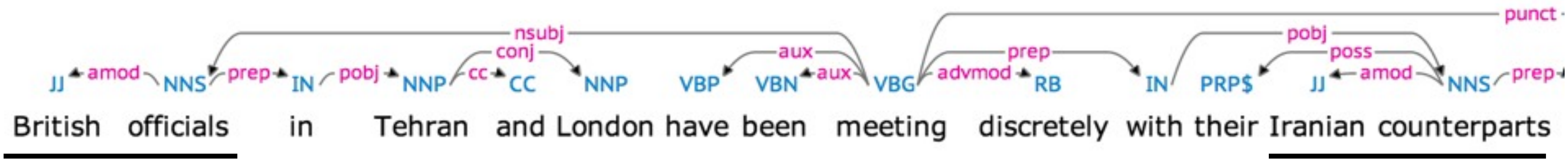
Event Extraction:
Who did what to whom?



[e.g. *Dowty 1991*]

Source *(s)*:
Recipient *(r)*:
Event phrase *(w)*:

Event Extraction:
Who did what to whom?



GBR

IRN

Match
country name list

Source *(s)*:
Recipient *(r)*:
Event phrase *(w)*:

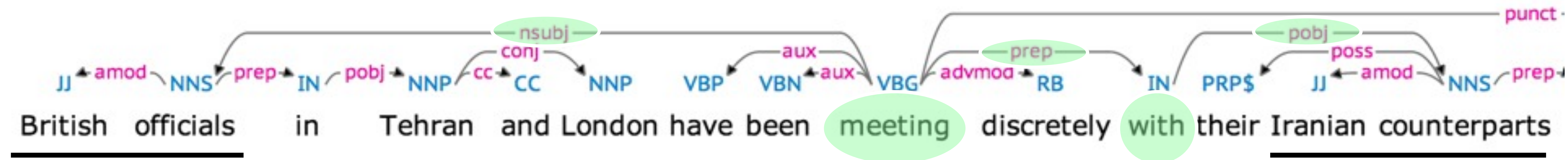
[e.g. *Dowty 1991*]

Event Extraction:
Who did what to whom?

Source *(s)*:
Recipient *(r)*:
Event phrase *(w)*:

[e.g. *Dowty 1991*]

Event Extraction:
Who did what to whom?



GBR

IRN

Match
country name list

Extract
event phrase

Source (*s*): GBR

Recipient (*r*): IRN

Event phrase (*w*): <--*nsubj*-- meet --*prep*--> with --*pobj*-->

[e.g. *Dowty 1991*]

"X meets with *Y*"

Event Extraction:
Who did what to whom?

- Structured linguistic analysis pipeline
  - Document classifier
  - Part-of-speech tagging
  - Syntactic parsing (rare in text-as-data) (*CoreNLP*)
  - POS and parse filtering rules
    - Factivity, verb paths, and parse quality

- Inputs
  1. 6.5 million news articles, 1987-2008 (Gigaword)
  2. Fixed list of country names
- Output:

time	**s**ender	**r**ecipient	**w**ords (event phrase)
1995-08-02	CHN	USA	say <-ccomp expel <-nsubjpass
1997-08-13	IGOUNO	IRQ	approve plan <-poss
2001-11-06	POL	IGONAT	campaign for
2002-09-04	PSE	ISR	fall with
2003-03-19	USA	IGOUNO	tell
2005-07-28	TUR	GRC	invade by supporter of union with
2006-08-07	IGOUNO	USA	debate
2007-05-18	CHN	RUS	host of talk <-rcmod involve
2008-06-05	MEX	USA	call upon
2008-12-02	IND	PAK	have

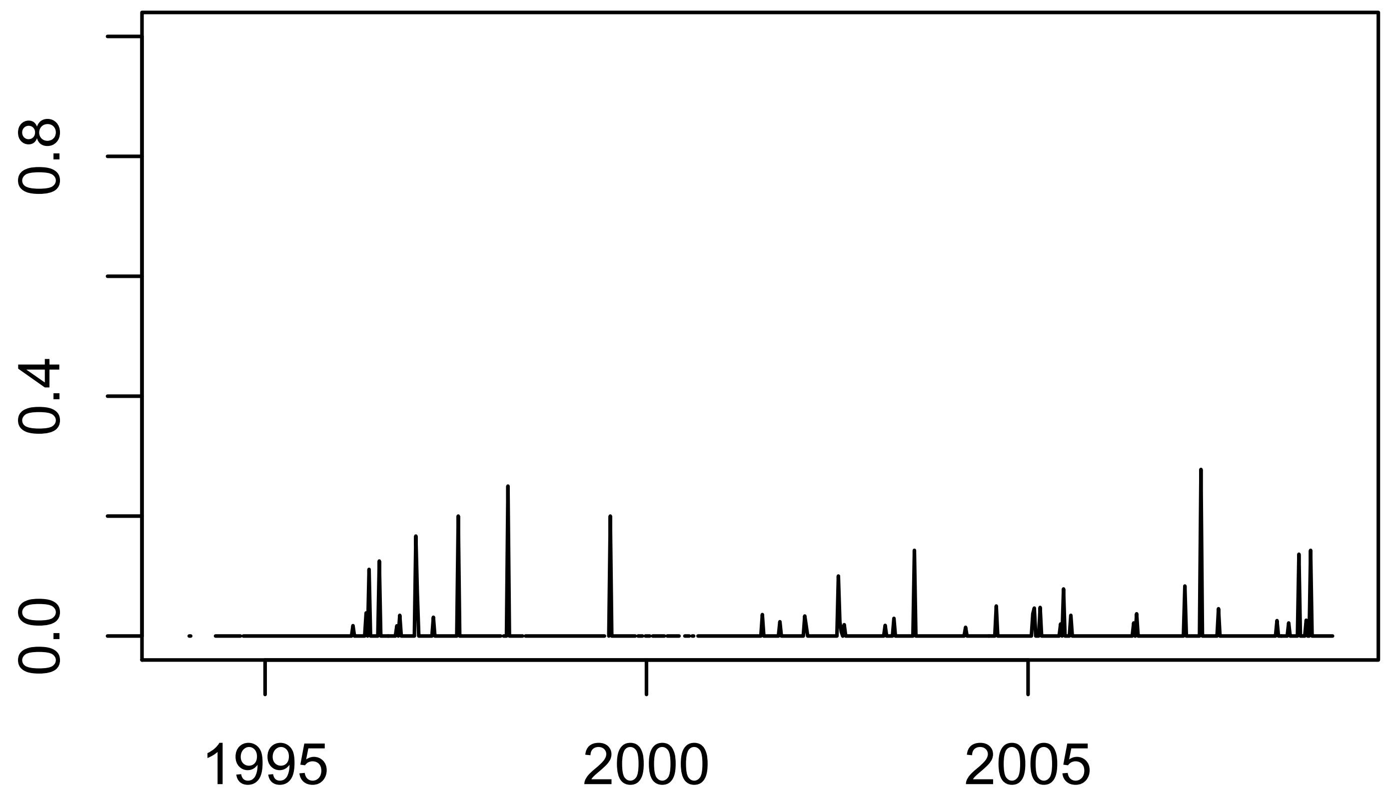
Filter to
- event phrases with count >= 10
- dyads with count >= 500

365,623 event tuples
421 directed dyads (s,r)
10,457 event phrases (w)
1,149 weeks (t)

Event phrases

"*ISR* meet with *PSE*"

P(w = "meet with" | t, s=*ISR*, r=*PSE*)



Too sparse for human interpretability

Do word semantics cluster on social context?

s=ISR, *r*=PSE

t= Jul 15-21, 2002

say <-ccomp be to
release to
take control of
occupy
wound in
scuffle with
be <-xcomp meet
meet with
meet with
arrest

t= Jul 3-9, 2006

commit to
strike
carry in
continue in
reject
fire at target in
start around
ratchet pressure on
shell
hit

s=USA, *r*=FRA

t= Feb 2-8, 1998

travel <-xcomp meet with
consider
meet with
meet with
meet with

t= Dec 22-28, 2003

release with
welcome
welcome by
win
agree with
indict
win from
concern over
win
indict

Do word semantics cluster on social context?

s=ISR, *r*=PSE

<u>*t*= Jul 15-21, 2002</u>

say <-ccomp be to
release to
take control of
occupy
wound in
scuffle with
be <-xcomp meet
meet with
meet with
arrest

<u>*t*= Jul 3-9, 2006</u>

commit to
strike
carry in
continue in
reject
fire at target in
start around
ratchet pressure on
shell
hit

s=USA, *r*=FRA

<u>*t*= Feb 2-8, 1998</u>

travel <-xcomp meet with
consider
meet with
meet with
meet with

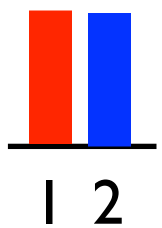
<u>*t*= Dec 22-28, 2003</u>

release with
welcome
welcome by
win
agree with
indict
win from
concern over
win
indict

Clustering approach: Mixed-membership models
("topic models," "admixtures")

Contextual event class probabilities

s=ISR, r=PSE

$\theta_{s,r} =$ 

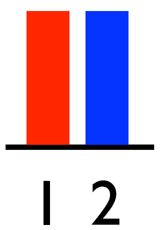
t= Jul 15-21, 2002

say <-ccomp be to
release to
take control of
occupy
wound in
scuffle with
be <-xcomp meet
meet with
meet with
arrest

t= Jul 3-9, 2006

commit to
strike
carry in
continue in
reject
fire at target in
start around
ratchet pressure on
shell
hit

s=USA, r=FRA

$\theta_{s,r} =$ 

t= Feb 2-8, 1998

travel <-xcomp meet with
consider
meet with
meet with
meet with

t= Dec 22-28, 2003

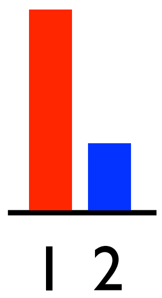
release with
welcome
welcome by
win
agree with
indict
win from
concern over
win
indict

Event class dictionaries ϕ_1 ϕ_2

agree with, arrest, be <-xcomp meet, carry in, commit to, concern over, consider, continue in, fire at target in, hit, indict, meet with, occupy, ratchet pressure on, reject, release to, release with, say <-ccomp be to, scuffle with, shell, start around, strike, take control of, travel <-xcomp meet with, welcome, welcome by, win, win from, wound in

Contextual event class probabilities

s=ISR, r=PSE

$\theta_{s,r} =$ 

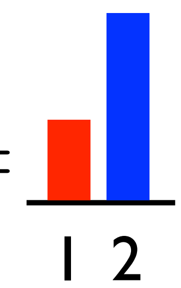
t= Jul 15-21, 2002

say <-ccomp be to
release to
take control of
occupy
wound in
scuffle with
be <-xcomp meet
meet with
meet with
arrest

t= Jul 3-9, 2006

commit to
strike
carry in
continue in
reject
fire at target in
start around
ratchet pressure on
shell
hit

s=USA, r=FRA

$\theta_{s,r} =$ 

t= Feb 2-8, 1998

travel <-xcomp meet with
consider
meet with
meet with
meet with

t= Dec 22-28, 2003

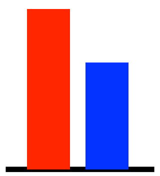
release with
welcome
welcome by
win
agree with
indict
win from
concern over
win
indict

Event class dictionaries ϕ_1 ϕ_2

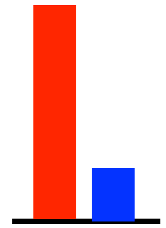
agree with, arrest, be <-xcomp meet, carry in, commit to, concern over, consider, continue in, fire at target in, hit, indict, meet with, occupy, ratchet pressure on, reject, release to, release with, say <-ccomp be to, scuffle with, shell, start around, strike, take control of, travel <-xcomp meet with, welcome, welcome by, win, win from, wound in

Contextual event class probabilities

s=ISR, *r*=PSE

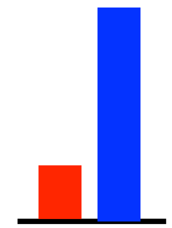
$\theta_{s,r,t} =$ 

t= Jul 15-21, 2002
say <-ccomp be to
release to
take control of
occupy
wound in
scuffle with
be <-xcomp meet
meet with
meet with
arrest

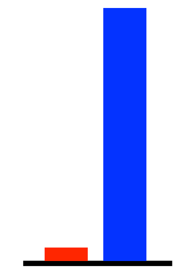
$\theta_{s,r,t} =$ 

t= Jul 3-9, 2006
commit to
strike
carry in
continue in
reject
fire at target in
start around
ratchet pressure on
shell
hit

s=USA, *r*=FRA

$\theta_{s,r,t} =$ 

t= Feb 2-8, 1998
travel <-xcomp meet with
consider
meet with
meet with
meet with

$\theta_{s,r,t} =$ 

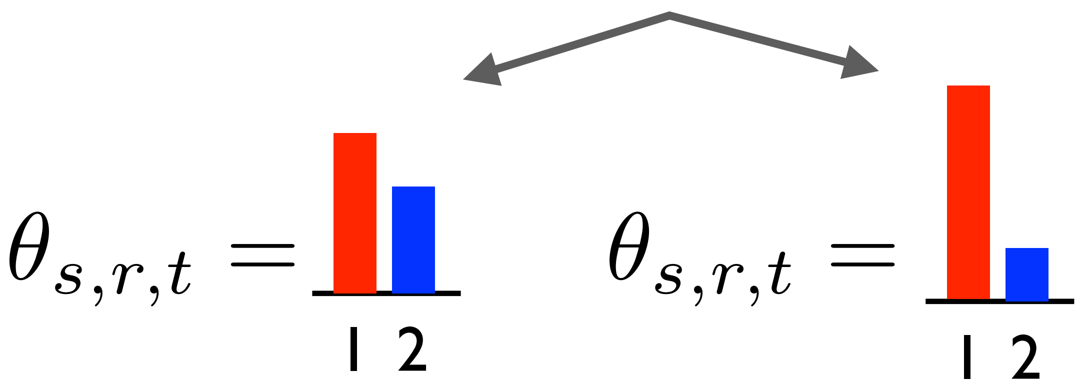
t= Dec 22-28, 2003
release with
welcome
welcome by
win
agree with
indict
win from
concern over
win
indict

Event class dictionaries ϕ_1 ϕ_2

agree with, arrest, be <-xcomp meet, carry in, commit to, concern over, consider, continue in, fire at target in, hit, indict, meet with, occupy, ratchet pressure on, reject, release to, release with, say <-ccomp be to, scuffle with, shell, start around, strike, take control of, travel <-xcomp meet with, welcome, welcome by, win, win from, wound in

Contextual event class probabilities

s=ISR, r=PSE



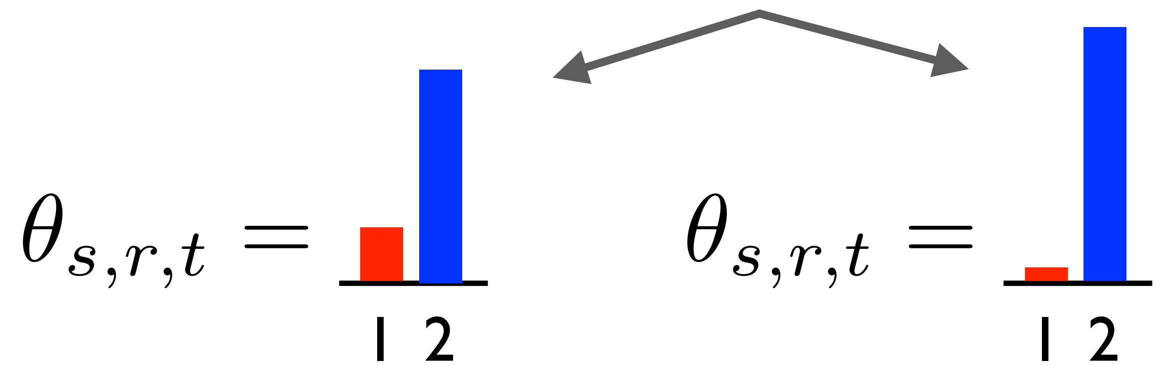
t= Jul 15-21, 2002

say <-ccomp be to
release to
take control of
occupy
wound in
scuffle with
be <-xcomp meet
meet with
meet with
arrest

t= Jul 3-9, 2006

commit to
strike
carry in
continue in
reject
fire at target in
start around
ratchet pressure on
shell
hit

s=USA, r=FRA



t= Feb 2-8, 1998

travel <-xcomp meet with
consider
meet with
meet with
meet with

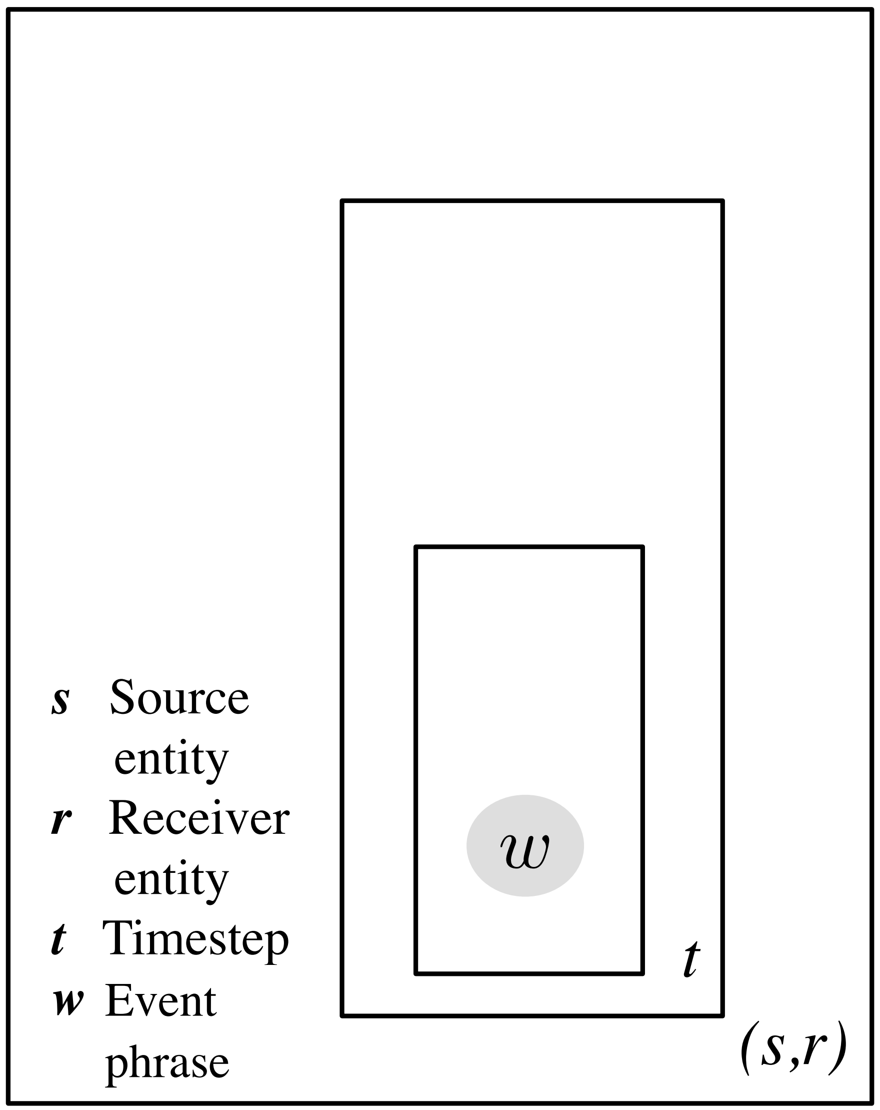
t= Dec 22-28, 2003

release with
welcome
welcome by
win
agree with
indict
win from
concern over
win
indict

Event class dictionaries ϕ_1 ϕ_2

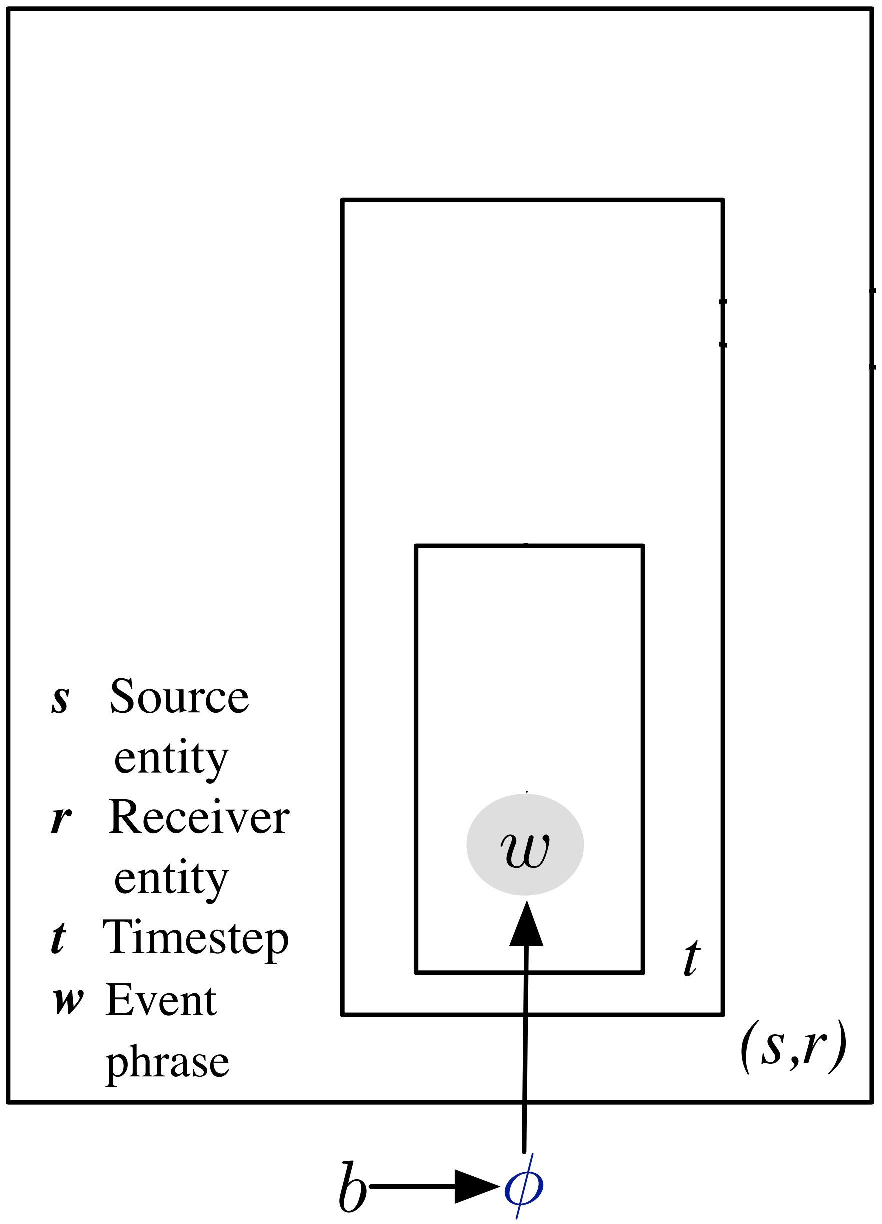
agree with, arrest, be <-xcomp meet, carry in, commit to, concern over, consider, continue in, fire at target in, hit, indict, meet with, occupy, ratchet pressure on, reject, release to, release with, say <-ccomp be to, scuffle with, shell, start around, strike, take control of, travel <-xcomp meet with, welcome, welcome by, win, win from, wound in

Model



Predicate-argument models: Pereira, Tishby, Lee 1993; Rooth et al. 1998

Model

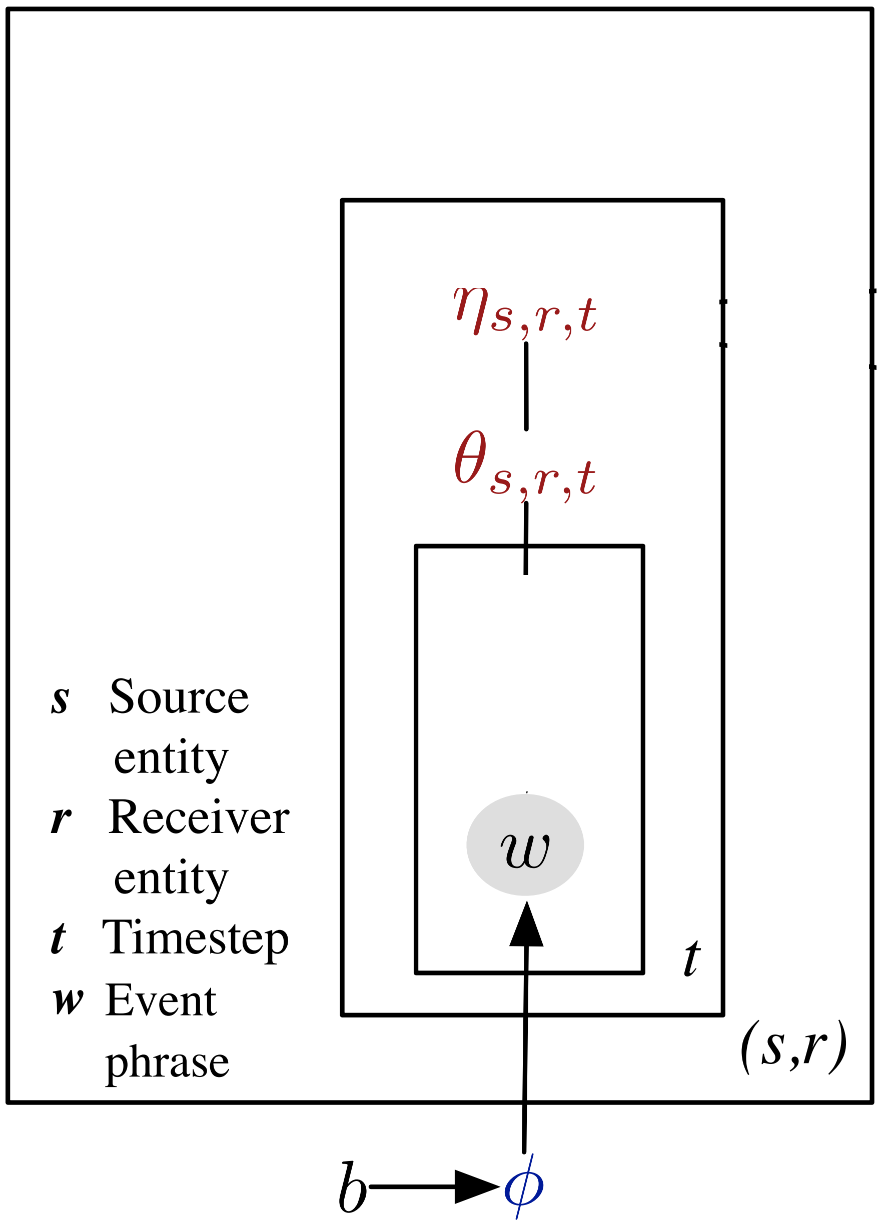


$$\phi_k \sim \mathrm{Dir}(b)$$

K phrase clusters (one per event class)

Linguistic definitions

Model



K = number of latent event classes

Event class prevalences per context

$$\eta_{s,r,t} \in \mathbb{R}^K$$

$$(\theta_{s,r,t})_k \propto \exp(\eta_{s,r,t,k})$$

Event class probabilities per context

Political context

$$\phi_k \sim \mathrm{Dir}(b)$$

K phrase clusters (one per event class)

Linguistic definitions

Logistic Normal [e.g. *Aitchison and Shen 1980*]

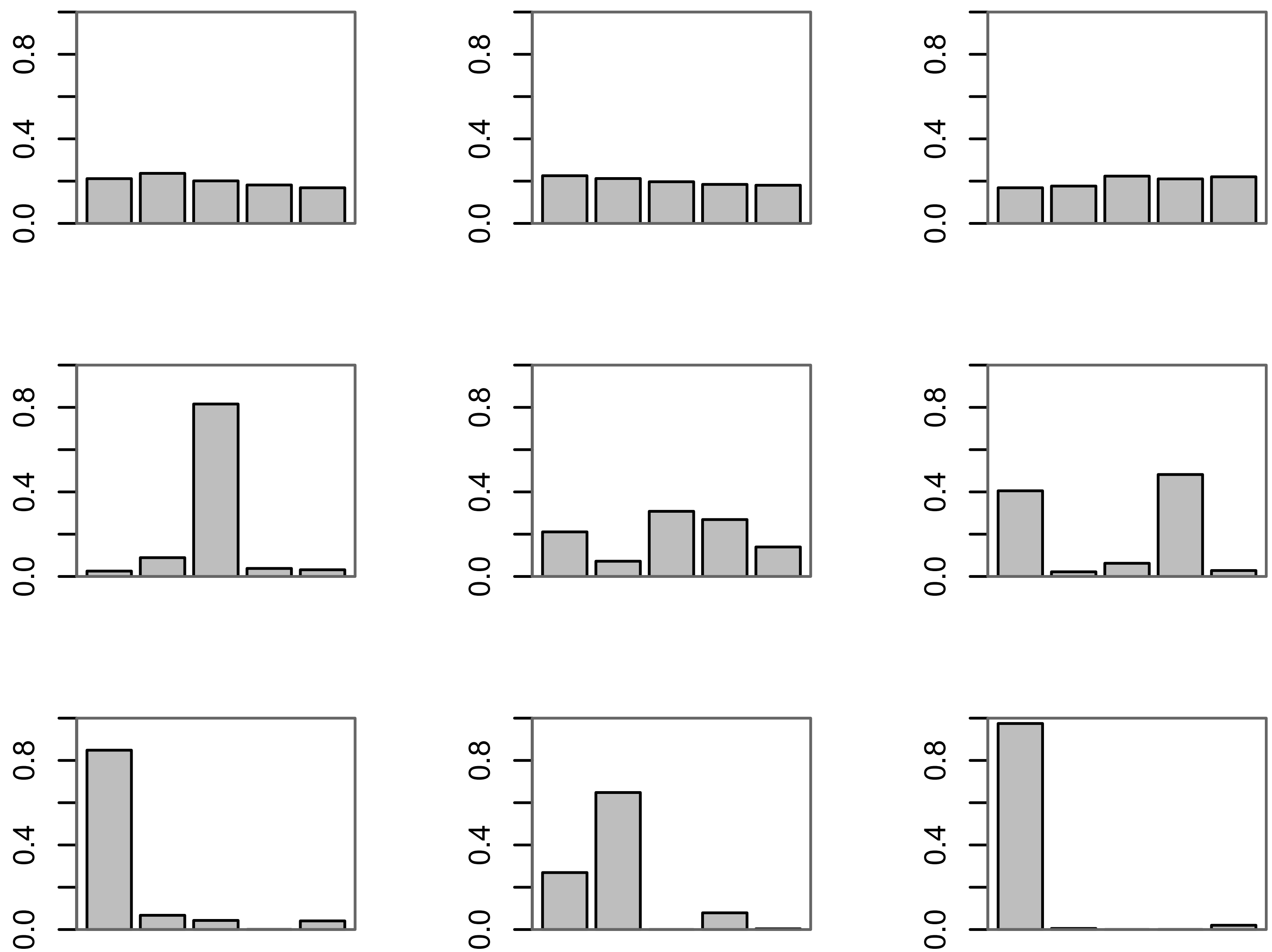
$$\eta_{s,r,t} \sim N(\alpha \qquad\qquad , \mathrm{Diag}[\sigma_1^2..\sigma_K^2])$$

$$(\theta_{s,r,t})_k \propto \exp(\eta_{s,r,t,k})$$

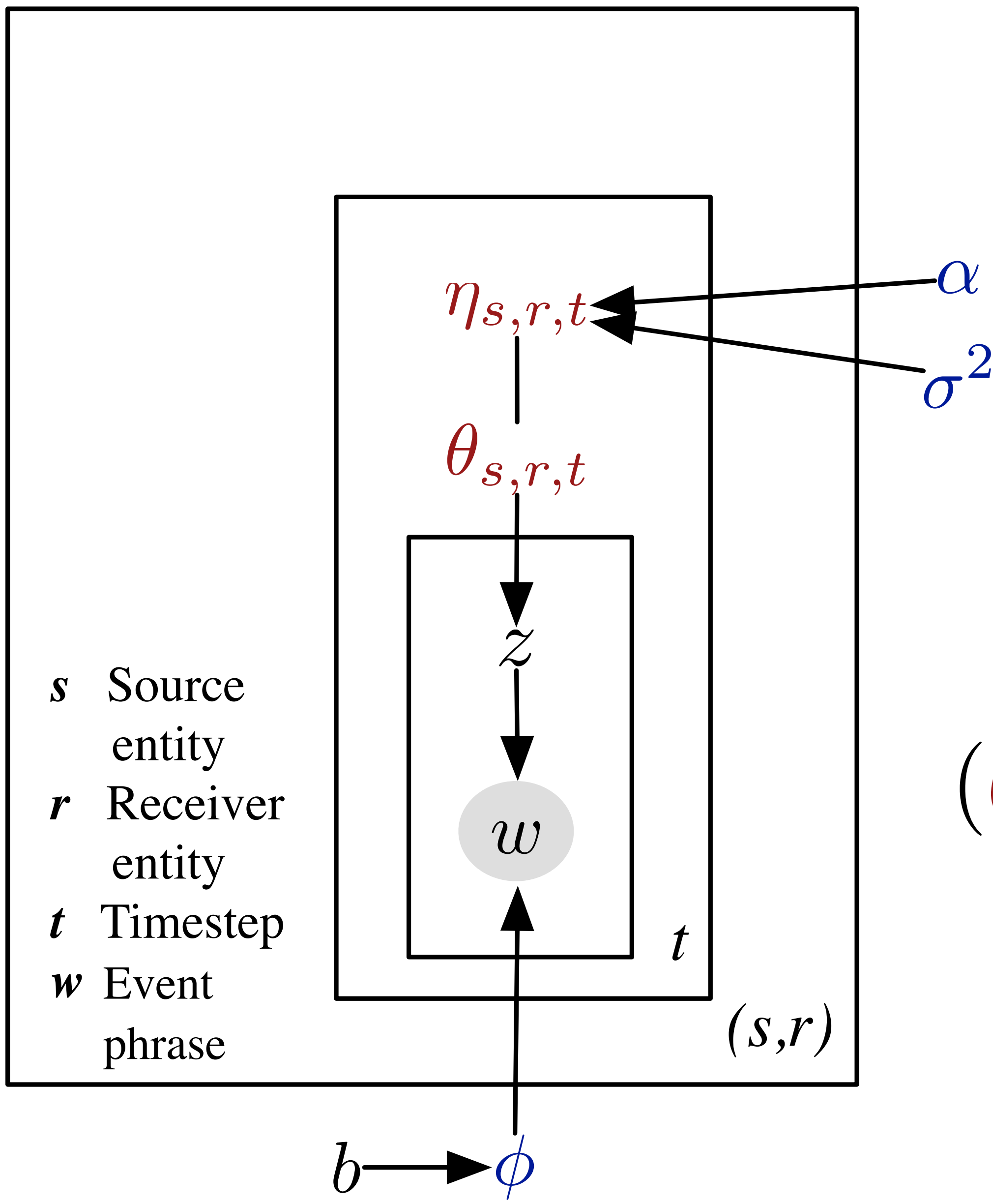
$\sigma = 0.1$

$\sigma = 1$

$\sigma = 5$



Model



<u>Event prior models</u>

M1: independent contexts

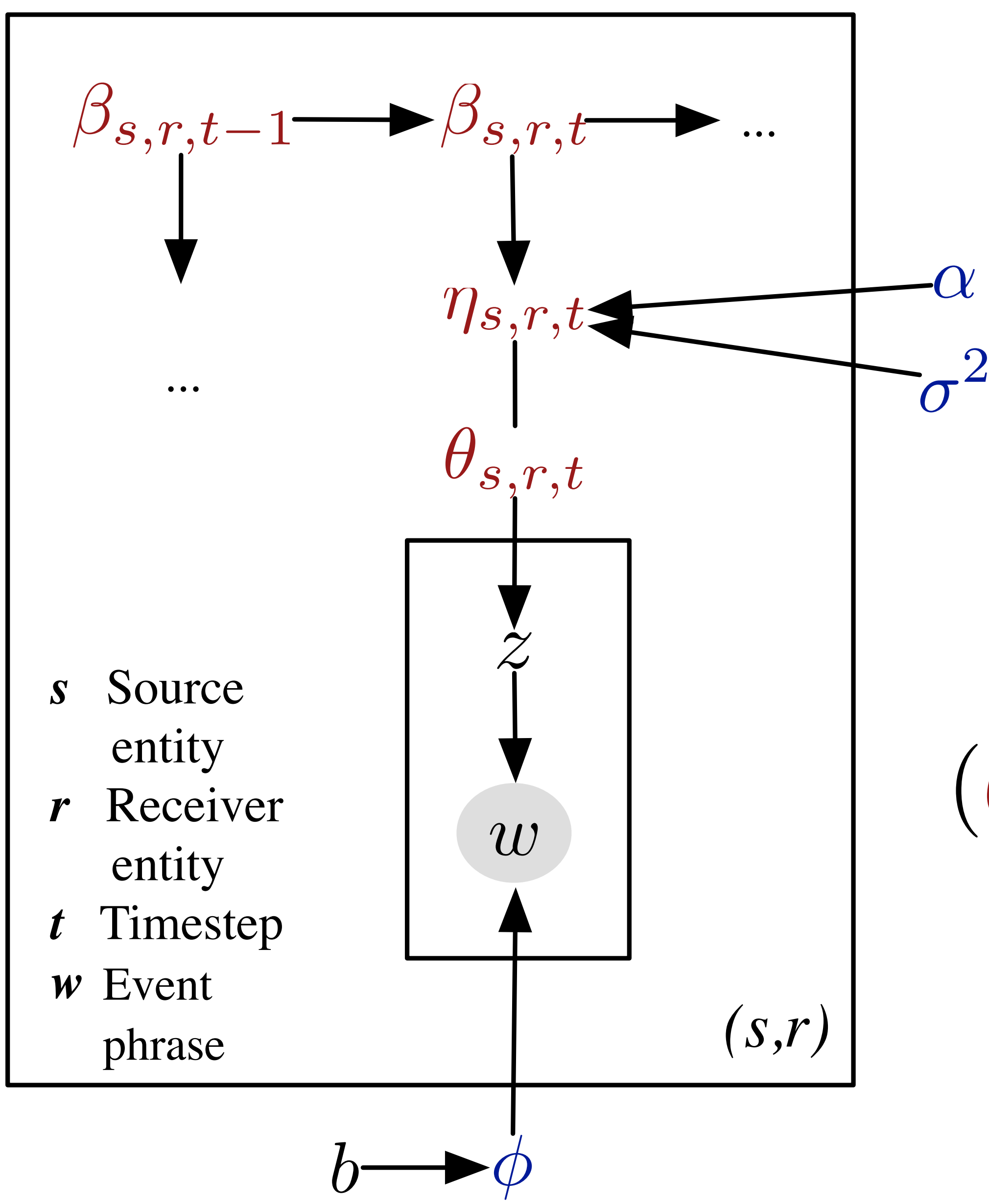
$$\eta_{s,r,t} \sim N(\alpha \qquad\qquad , \mathrm{Diag}[\sigma_1^2..\sigma_K^2])$$

$$(\theta_{s,r,t})_k \propto \exp(\eta_{s,r,t,k})$$

$$\left.\begin{aligned} z &\sim \mathrm{Mult}(\theta_{s,r,t}) \\ w &\sim \mathrm{Mult}(\phi_z) \end{aligned}\right] \; w \sim \mathrm{Mult}(\Phi\theta_{s,r,t})$$

$$\phi_k \sim \mathrm{Dir}(b)$$

Model



Event prior models

M1: independent contexts

M2: temporal smoothing

[Blei and Lafferty 2006, Quinn and Martin 2002]

$$\beta_{s,r,t} \sim N(\beta_{s,r,t-1}, \mathbb{I}\tau^2)$$

Adjacent timestep similarity

$$\eta_{s,r,t} \sim N(\alpha + \beta_{s,r,t}, \mathrm{Diag}[\sigma_1^2..\sigma_K^2])$$

$$(\theta_{s,r,t})_k \propto \exp(\eta_{s,r,t,k})$$

$$z \sim \mathrm{Mult}(\theta_{s,r,t})$$

$$w \sim \mathrm{Mult}(\phi_z)$$

$$w \sim \mathrm{Mult}(\Phi\theta_{s,r,t})$$

$$\phi_k \sim \mathrm{Dir}(b)$$

K=100 ⟶ 80 million parameters

Learning: blocked Gibbs sampling

$$p(\beta, (\eta, \theta), \sigma_1^2..\sigma_K^2, z, \phi, b \mid w)$$

$$\begin{aligned}
\beta_{s,r,t} &\sim N(\beta_{s,r,t-1}, \mathbb{I}\tau^2) \\
\eta_{s,r,t} &\sim N(\alpha + \beta_{s,r,t}, \mathrm{Diag}[\sigma_1^2..\sigma_K^2]) \\
(\theta_{s,r,t})_k &\propto \exp(\eta_{s,r,t,k}) \\
z &\sim \mathrm{Mult}(\theta_{s,r,t}) \\
w &\sim \mathrm{Mult}(\phi_z) \\
\phi_k &\sim \mathrm{Dir}(b)
\end{aligned}$$

Learning: blocked Gibbs sampling

$$p(\beta, (\eta, \theta), \sigma_1^2..\sigma_K^2, z, \phi, b \mid w)$$

Conjugate normal

Linear dynamical system
Forward filter backward sampler (FFBS)
[Carter and Kohn 1994, West and Harrison 1997]

$$\beta_{s,r,t} \sim N(\beta_{s,r,t-1}, \mathbb{I}\tau^2)$$

Logistic normal
Metropolis-within-Gibbs,
Laplace approximation proposal
[Hoff 2003]

$$\eta_{s,r,t} \sim N(\alpha + \beta_{s,r,t}, \mathrm{Diag}[\sigma_1^2..\sigma_K^2])$$

$$(\theta_{s,r,t})_k \propto \exp(\eta_{s,r,t,k})$$

$$z \sim \mathrm{Mult}(\theta_{s,r,t})$$

$$w \sim \mathrm{Mult}(\phi_z)$$

Dirichlet-multinomial
Collapsed sampling
[Griffiths and Steyvers 2005]

$$\phi_k \sim \mathrm{Dir}(b)$$

Slice sampling
[Neal 2003]

Laplace approx. to logistic normal

$$\eta \sim N(\bar{\eta},\ \mathrm{Diag}[\sigma_1^2..\sigma_K^2])$$
$$z \sim \mathrm{Mult}(\theta(\eta))$$
$$\theta(\eta) = \exp(\eta)/\mathrm{sum}(\exp(\eta))$$

$$p(\eta|\bar{\eta}, \Sigma, z) \propto N(\eta; \bar{\eta}, \Sigma)\ \mathrm{Mult}(\vec{z}; \theta(\eta))$$

1. Solve MAP $\hat{\eta} = \arg\max_{\eta} \sum_k \left(-\frac{1}{2\sigma_k^2}(\eta_k - \bar{\eta}_k)^2 + n_k \log \theta(\eta)_k \right)$

Newton's method with fast O(*K*) Sherman-Morrison steps (adapted from *Eisenstein et al. 2011*)

2. Proposal $\eta^* \sim N(\hat{\eta}, [H(-\ell(\hat{\eta})]^{-1})$

$$H_{kk} = n\theta_k(1-\theta_k) + 1/\sigma_k^2, \quad H_{jk} = -n\theta_j\theta_k$$

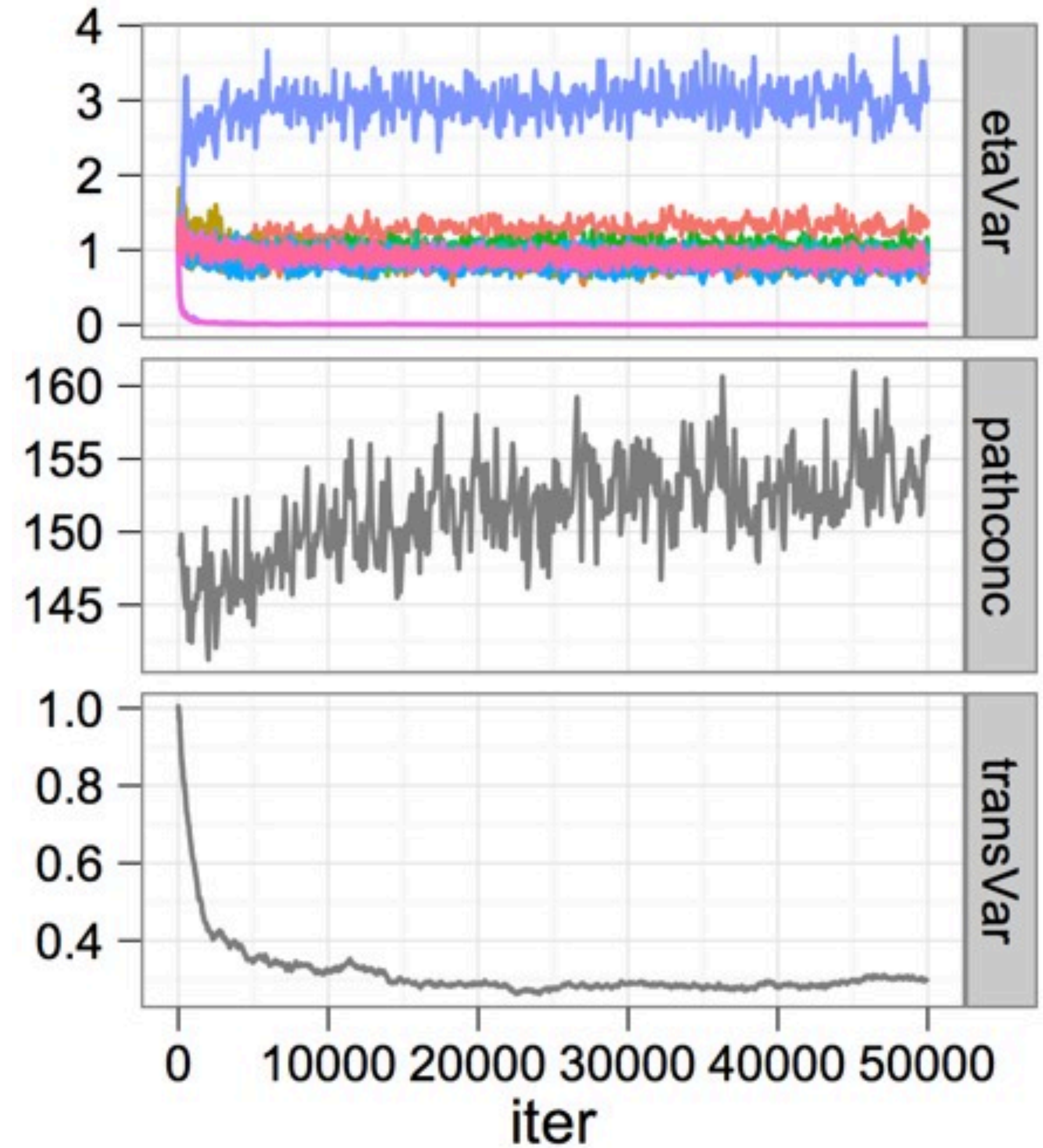
Metropolis rejections correct approximation error

Alternative to variational inference for LN

[Blei and Lafferty 2006, Ahmed and Xing 2007, Wang and Blei 2013 vs. *Mimno et al. 2008]*

Learning

- Markov Chain Monte Carlo
- Implementation
  - Parallelization
  - Few hours to few days
  - Thinning (600 MB/sample)
  - Java, Python, R



Event classes: word posteriors

Most probable phrases in ϕ_k

arrive in, visit, meet with, travel to, leave, hold with, meet, meet in, fly to, be in, arrive for talk with, say in, arrive with, head to, hold in, due in, leave for, make to, arrive to, praise

accuse, blame, say, break with, sever with, blame on, warn, call, attack, rule with, charge, say←ccomp come from, say ←ccomp, suspect, slam, accuse government ←poss, accuse agency ←poss, criticize, identify

kill in, have troops in, die in, be in, wound in, have soldier in, hold in, kill in attack in, remain in, detain in, have in, capture in, stay in, about ←pobj troops in, kill, have troops ←partmod station in, station in, injure in, invade, shoot in

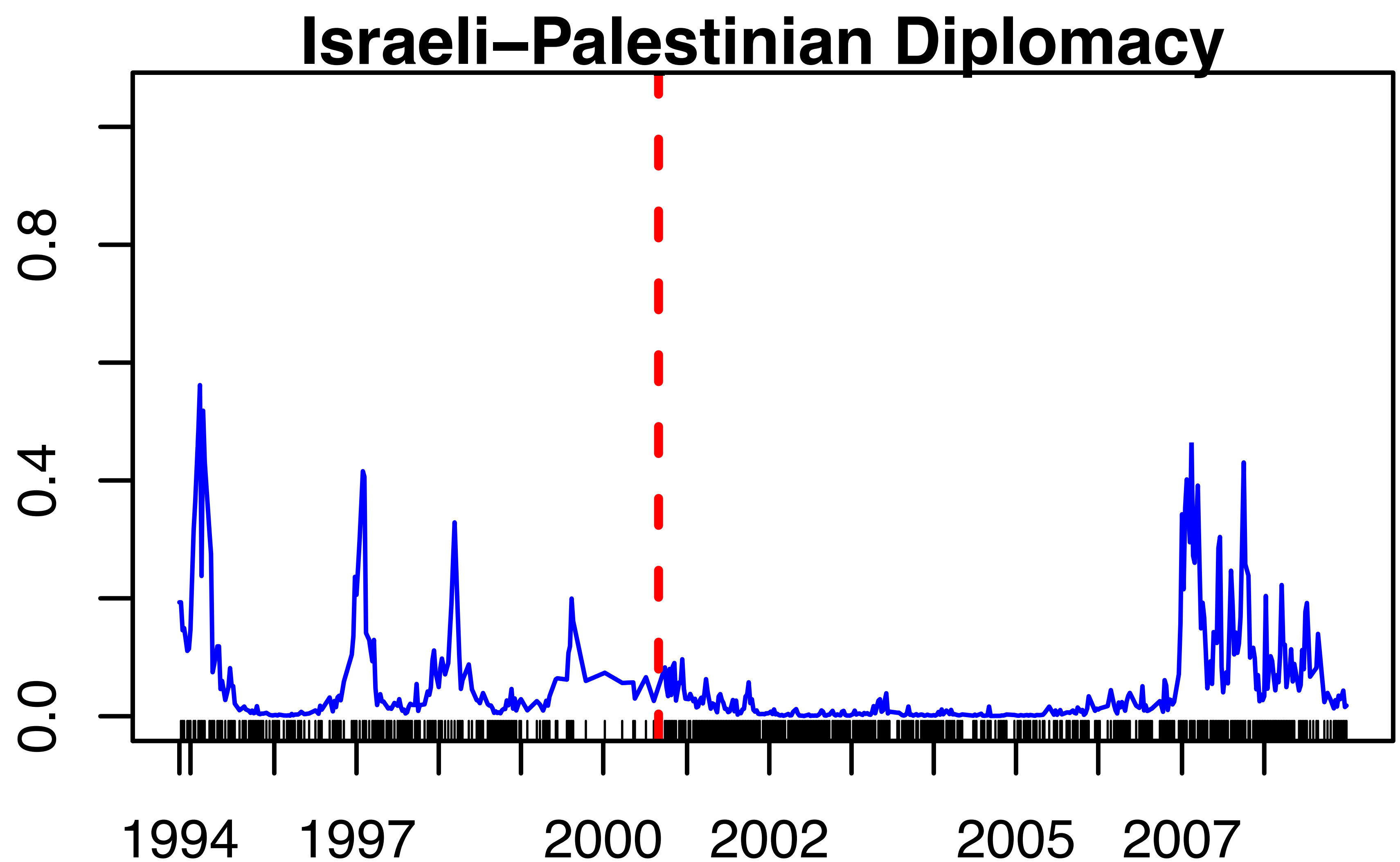
Event classes: word posteriors

Most probable phrases in ϕ_k

"diplomacy"	arrive in, visit, meet with, travel to, leave, hold with, meet, meet in, fly to, be in, arrive for talk with, say in, arrive with, head to, hold in, due in, leave for, make to, arrive to, praise
"verbal conflict"	accuse, blame, say, break with, sever with, blame on, warn, call, attack, rule with, charge, say←ccomp come from, say ←ccomp, suspect, slam, accuse government ←poss, accuse agency ←poss, criticize, identify
"material conflict"	kill in, have troops in, die in, be in, wound in, have soldier in, hold in, kill in attack in, remain in, detain in, have in, capture in, stay in, about ←pobj troops in, kill, have troops ←partmod station in, station in, injure in, invade, shoot in

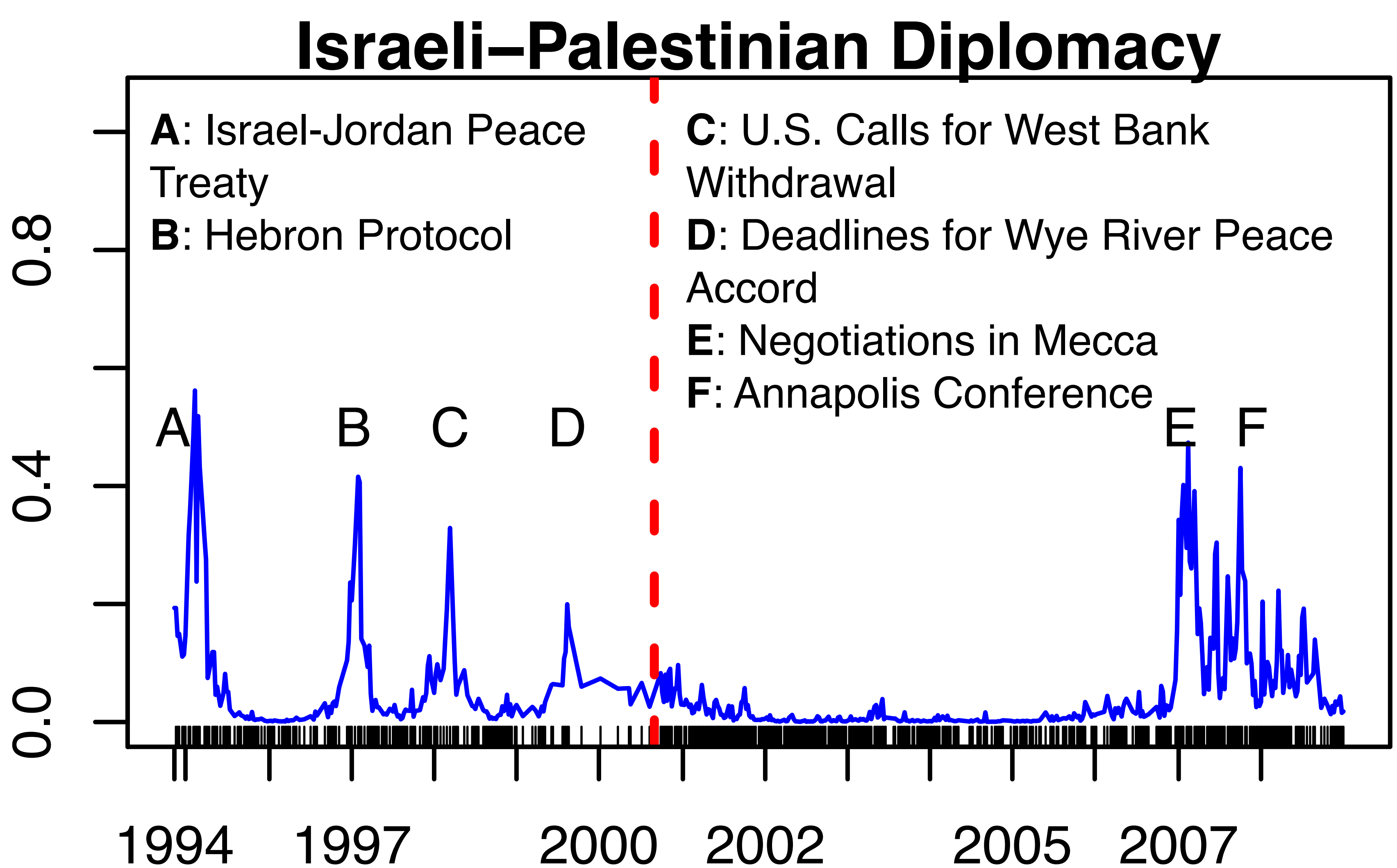
Case study

meet with, sign with, praise, say with, arrive in, host, tell, welcome, join, thank, meet, travel to, criticize, leave, take to, begin to, begin with, summon, reach with, hold with

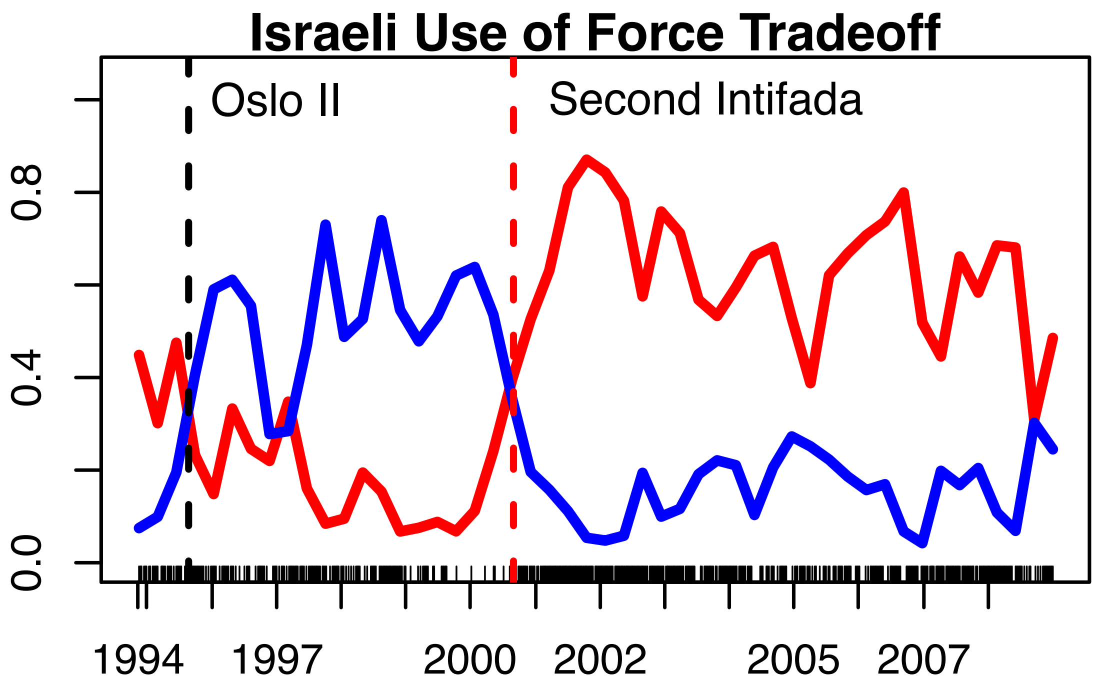


Case study

meet with, sign with, praise, say with, arrive in, host, tell, welcome, join, thank, meet, travel to, criticize, leave, take to, begin to, begin with, summon, reach with, hold with



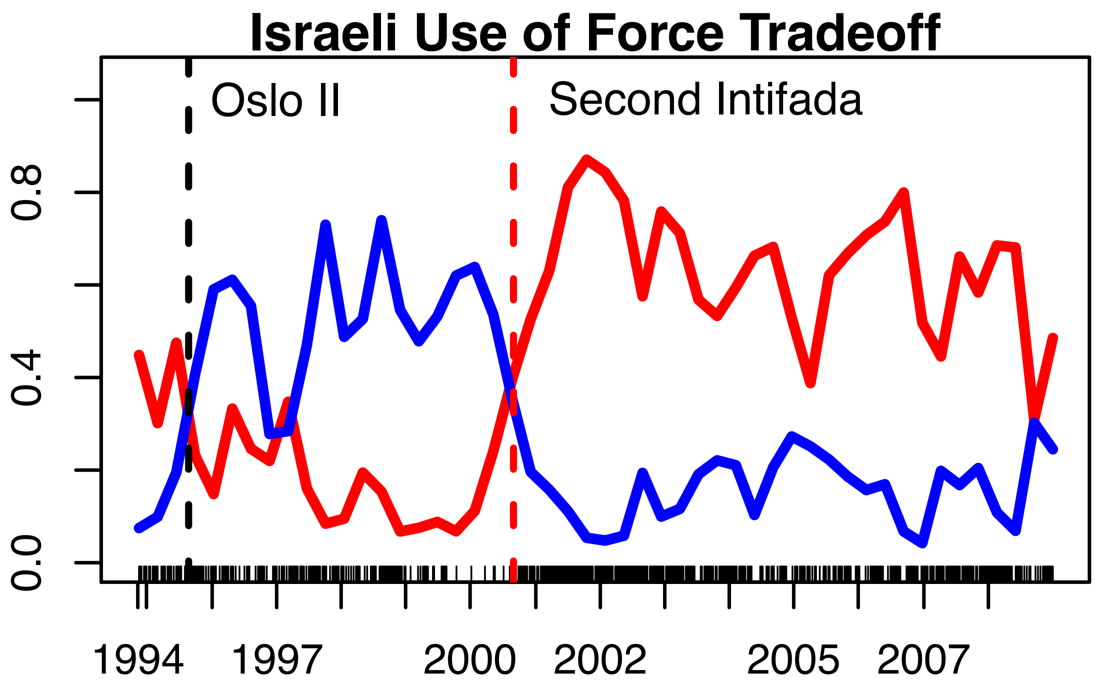
Validation of unsupervised models...



impose on, seal, capture from, seize from, arrest, ease closure of, close, deport, close with, release

kill, fire at, enter, kill in, attack, raid, strike in, move into, pound, bomb

Validation of unsupervised models...



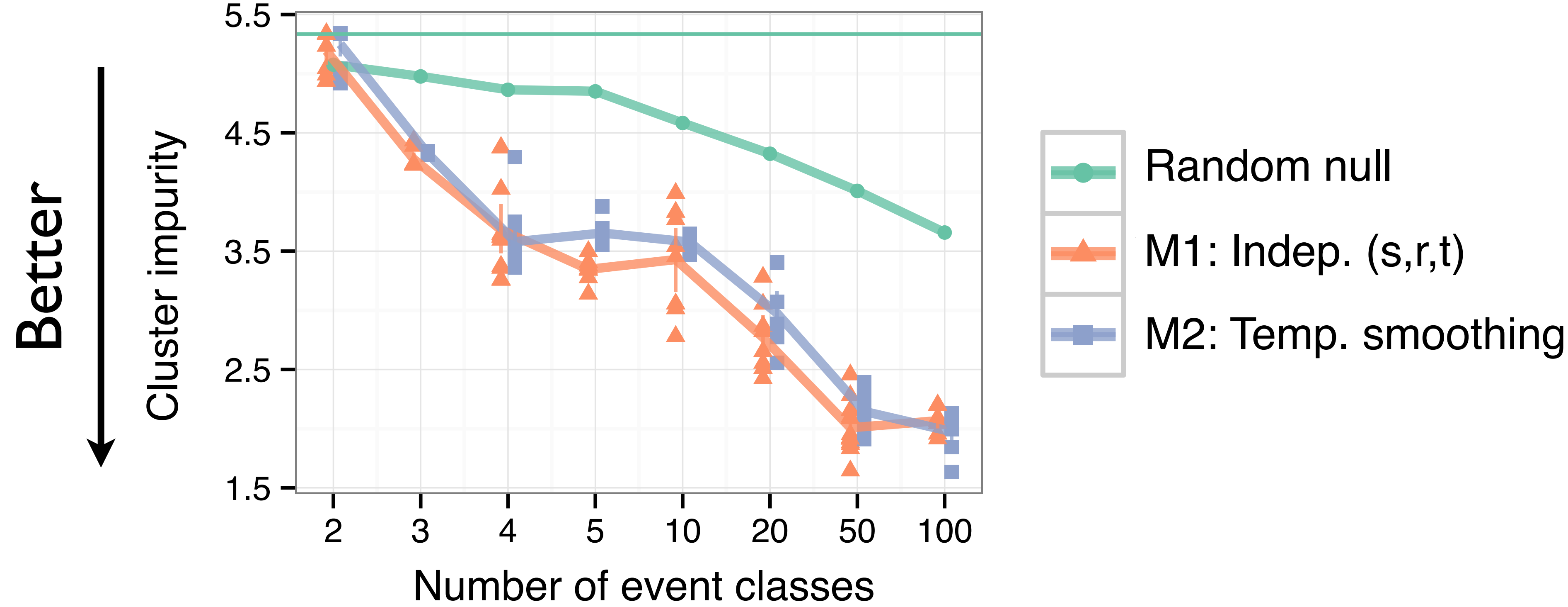
impose on, seal, capture from, seize from, arrest, ease closure of, close, deport, close with, release

kill, fire at, enter, kill in, attack, raid, strike in, move into, pound, bomb

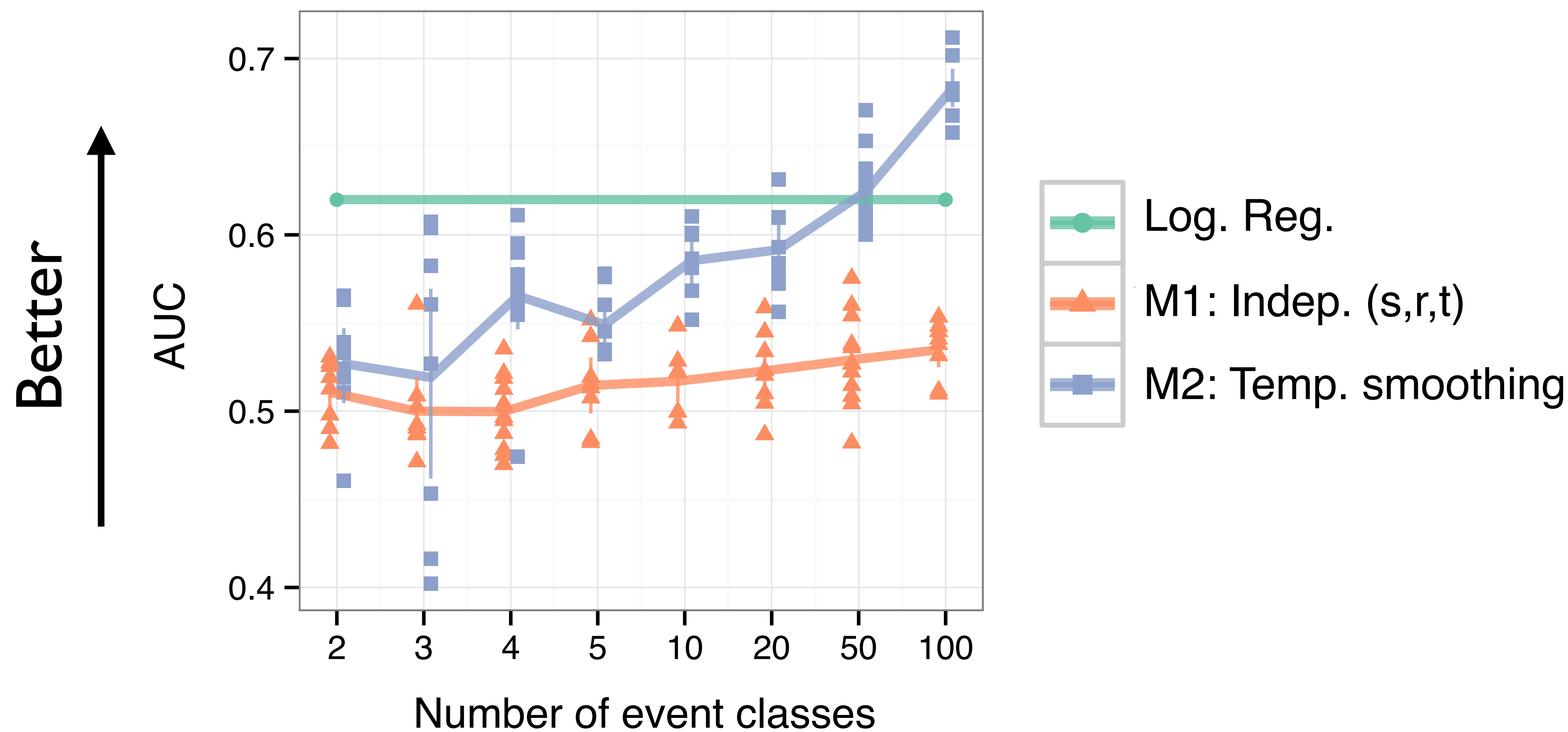
Correlates to conflict?

Semantic coherence?

Evaluations



Lexicon / Ontology reconstruction



Real-world conflict reconstruction

Applications of actor-event hierarchical models

[also e.g. *Chambers 2013, Cheung et al 2013...*]

- International events. From news, model:
  - Linguistic event classes
  - Event probabilities, through time
- Fictional narratives. From movie plot summaries, model:
  - Character types of attributes and actions
  - Conditioned on actors, genres, etc.

[Bamman, O'Connor, Smith *Assoc. Comp. Ling. 2013*]

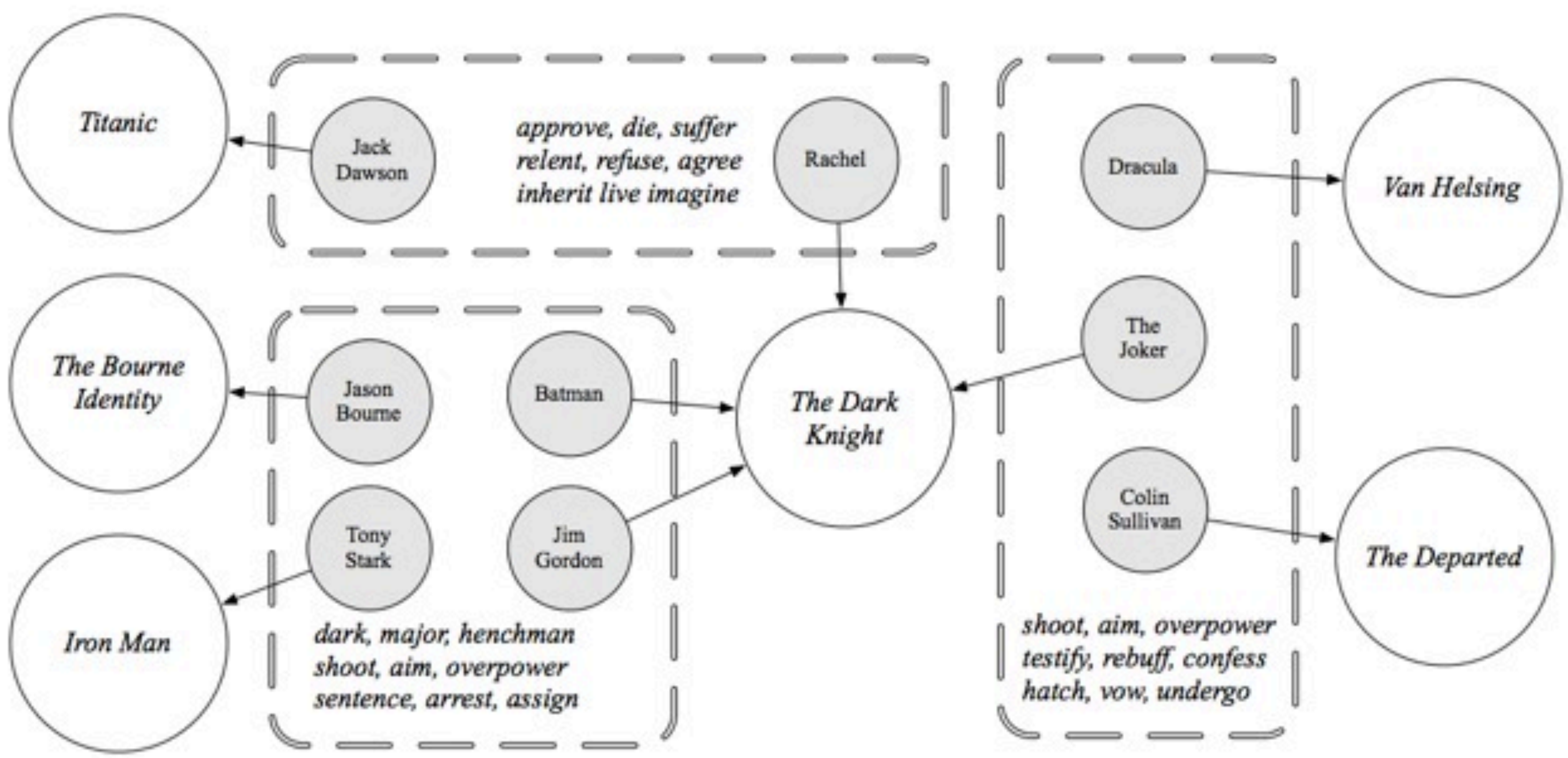
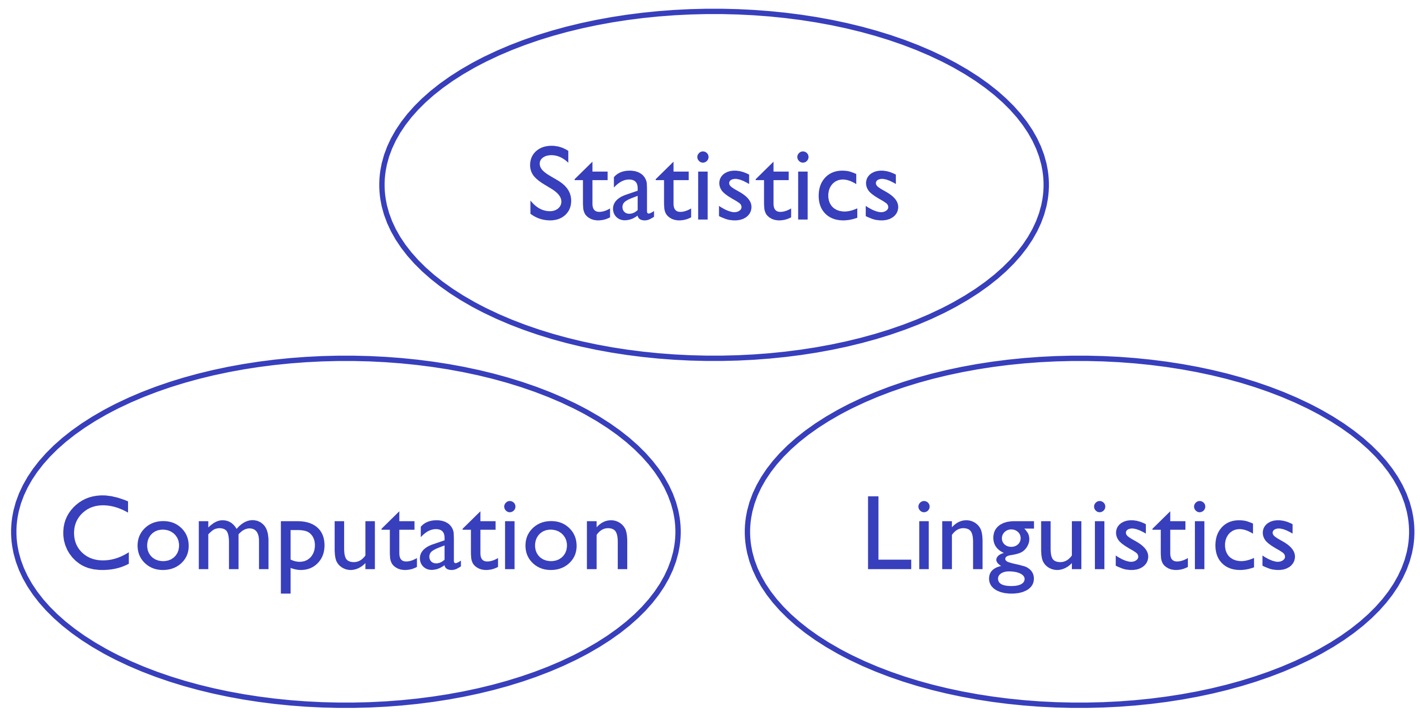


Figure 3: Dramatis personae of *The Dark Knight* (2008), illustrating 3 of the 100 character types learned by the persona regression model, along with links from other characters in those latent classes to other movies. Each character type is listed with the top three latent topics with which it is associated.

Analysis methods for **Text** and **Social Context**

concepts, attitudes, events

community, author, time, space



... motivated by analysis problems
in the social sciences and humanities

Politics
Literature
Business
Economics
Sociology
Health

Topics

- **Textual social data**
- Linguistic semantic learning

- Examples
  - Sentiment and opinion polls
  - International relations
  - **Geography and slang**
  - **Linguistic tools**
  - **Chinese censorship**

Geographic lexical variation in Twitter

[*Eisenstein, O'Connor, Smith, Xing 2010*]

Geographic topic model

$$r \sim \vec{\pi}$$

$$(lat,\ lon) \sim N(\vec{\mu}_r, \mathbf{\Sigma}_r)$$

User's locations from DPMM Gaussian mixture

$$\theta \sim Dir(\vec{\alpha})$$

$$z \sim \vec{\theta}$$

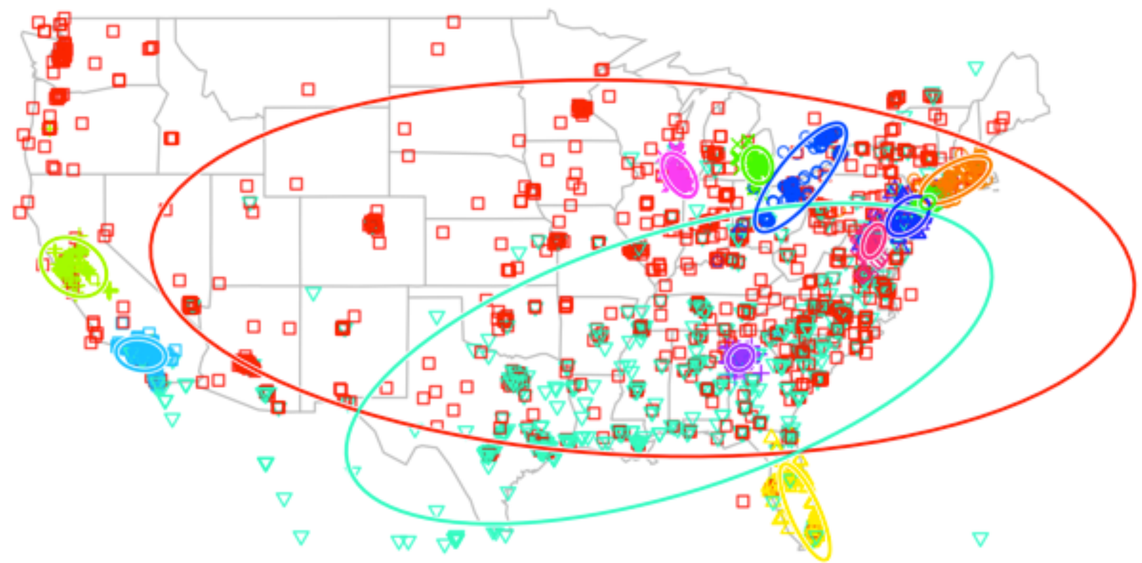
$$w \sim \exp(\vec{\eta}_{zr})$$

User's topics

$$\vec{\phi}_k \sim N(\vec{a}, b^2\mathbf{I})$$

$$\vec{\eta}_{kj} \sim N(\vec{\phi}_k, s_k^2\mathbf{I})$$

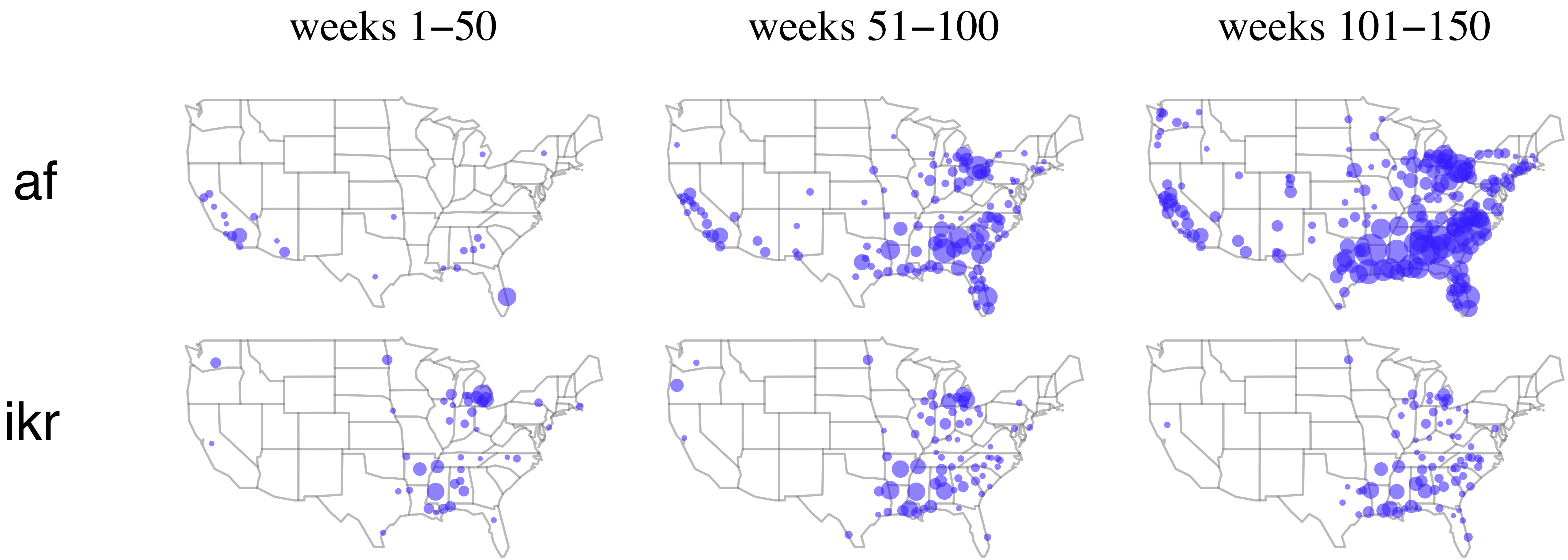
have regional variants



	"basketball"	"popular music"	"daily life"	"emoticons"	"chit chat"
	PISTONS KOBE LAKERS game DUKE NBA CAVS STUCKEY JETS KNICKS	album music beats artist video #LAKERS ITUNES tour produced vol	tonight shop weekend getting going chilling ready discount waiting iam	:) haha :d :(;) :p xd :/ hahaha hahah	lol smh jk yea wyd coo ima wassup somethin jp
Boston 	CELTICS victory BOSTON CHARLOTTE	playing daughter PEARL alive war comp	BOSTON	;p gna loveee	*ese* exam suttin sippin
N. California 	THUNDER KINGS GIANTS pimp trees clap	SIMON dl mountain seee	6am OAKLAND	*pues* hella koo SAN fckn	hella flirt hut iono OAKLAND

Social determinants of language change

[*Eisenstein, O'Connor, Smith, Xing 2012 and in review*]



Test sociolinguistic theories of how linguistic innovations diffuse

U.S. Census data

7 TB data, 200 regions, 2600 words, 165 timesteps = 85M parameters

$$n_{w,r,t} \sim \mathrm{Binom}(N_{r,t},\ \sigma(\nu_w + \tau_{r,t} + \eta_{w,*,t} + \eta_{w,r,t})$$

$$\boldsymbol{\eta}_{w,t} \sim \mathrm{Normal}(\mathbf{A}\boldsymbol{\eta}_{w,t-1},\ \boldsymbol{\Gamma})$$

$\mathbf{A}$ autoregressive coefficients (size $R \times R$)

Social Media NLP
Part-of-speech tagger for Twitter

Example

ikr	smh	he	asked	fir	yo	last	name
!	G	O	V	P	D	A	N

HMM word cluster (features for CRF tagger)

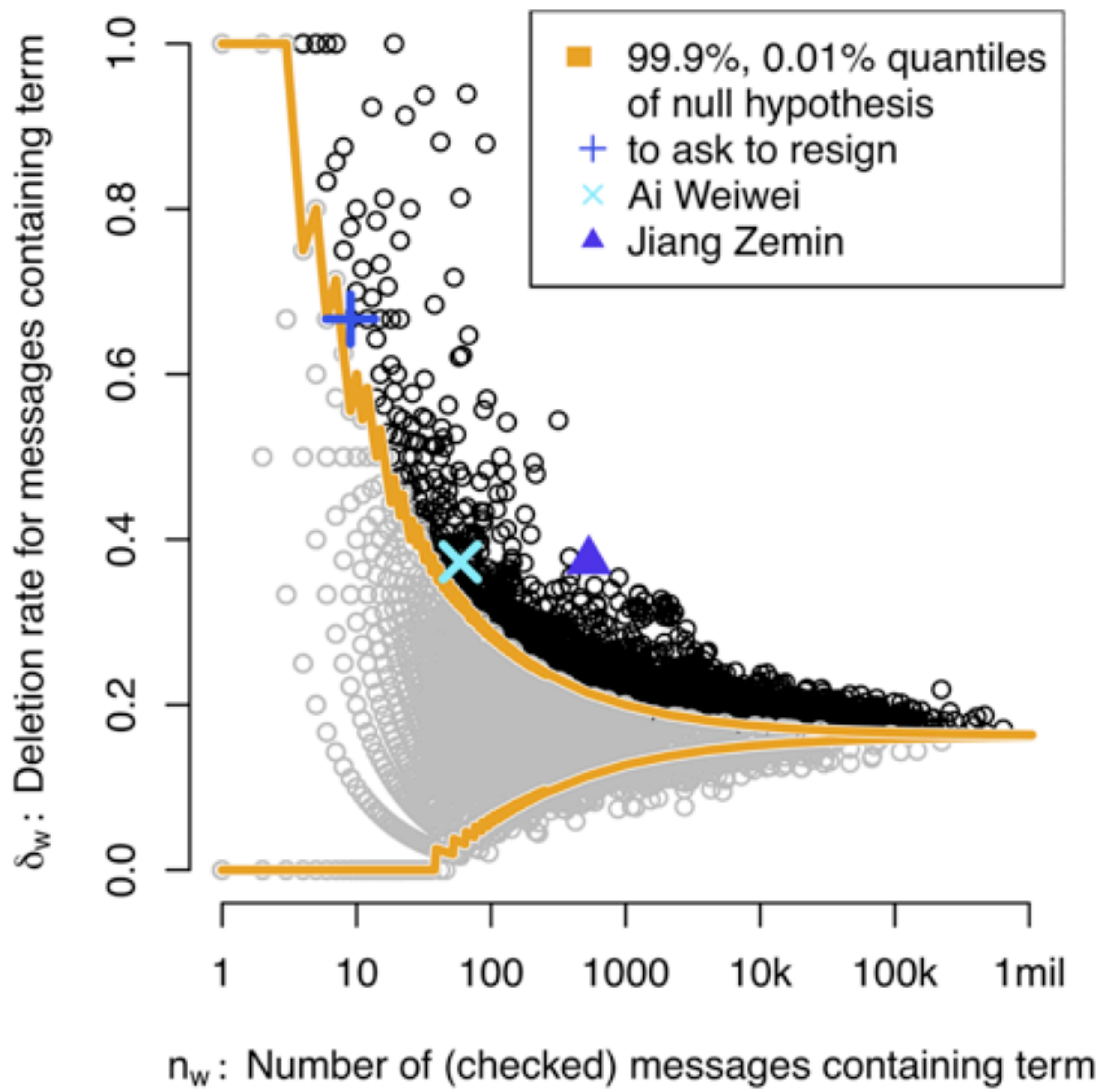
yeah yea nah naw yeahh nooo yeh noo noooo yeaa **ikr** nvm yeahhh nahh nooooo yh yeaaa yeaah yupp naa yeahhhh yeaaahiknow werd noes nahhh naww yeaaaa shucks yeaaaah yeahhhhh naaa naah nawl nawww yehh ino yeaaaaa yeeah yeeeah wordd yeaahh nahhhh naaah yeahhhhhh yeaaaaah naaaa yeeeeah nall yeaaaaaa

http://www.ark.cs.cmu.edu/TweetNLP/

[*Gimpel, Schneider, O'Connor, Das, Mills, Eisenstein, Heilman, Yogatama, Smith, 2011*]
[*Owoputi, O'Connor, Dyer, Gimpel, Schneider, Smith, 2013*]

Not just hierarchical models: Multiple hypothesis testing

Censorship in Chinese microblogs *[Bamman, O'Connor, Smith 2011]*



Benjamini-Hochberg
False discovery rate
calculation

Not just text:
Interests (online choice modeling)

LDA

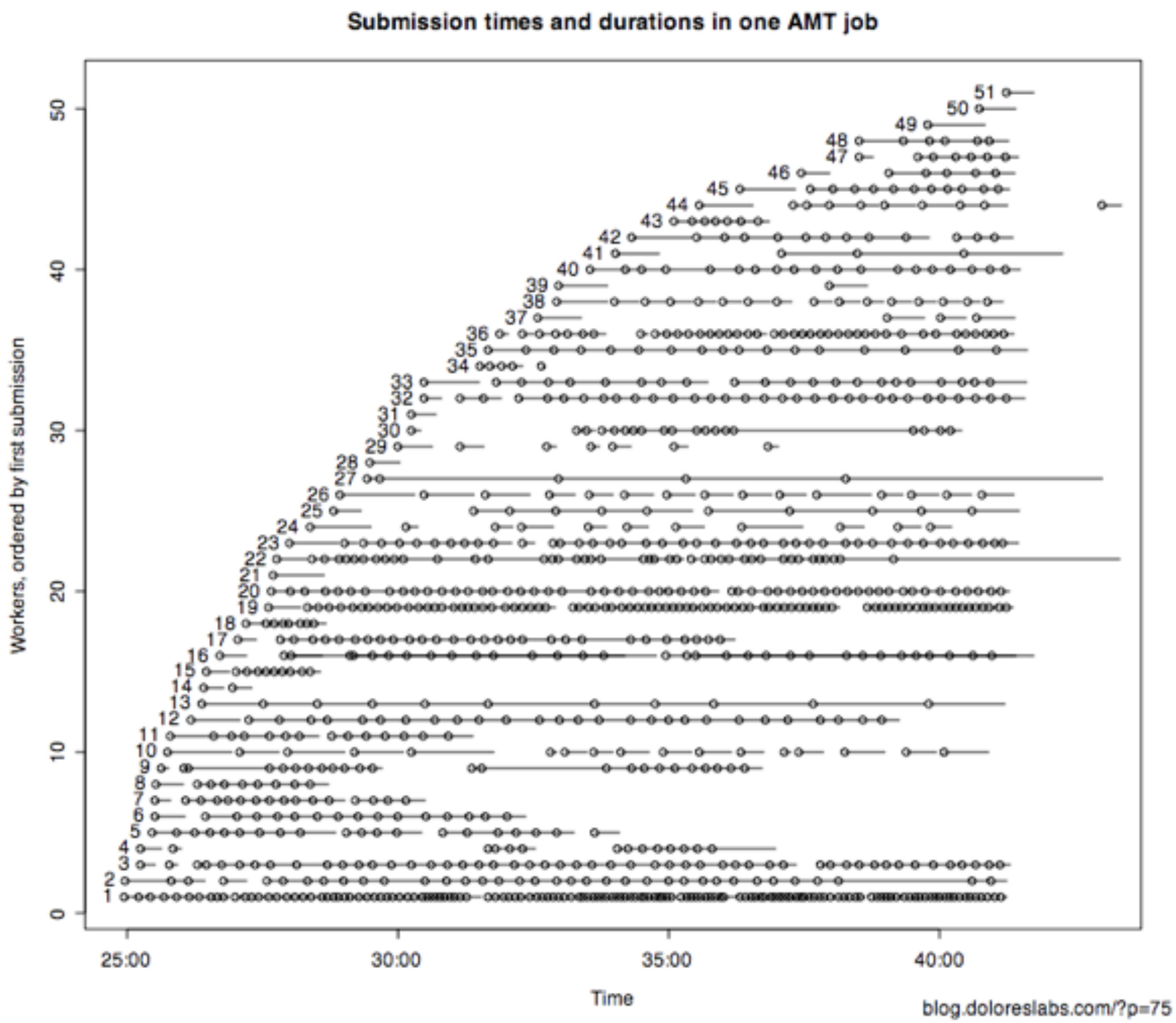
FreedomWorks, Sean Hannity, Conservative, Michelle Malkin, John Boehner, The Heritage Foundation, Mark Levin, Tea Party Patriots, Governor Jan Brewer, Americans for Prosperity, Tim Pawlenty, Marco Rubio

Ira Glass, NPR, This American Life, MoveOn.org, The Rachel Maddow Show, Can this poodle wearing a tinfoil hat get more fans than Glenn Beck?, Keith Olbermann, Telling Pat Robertson to STFU, Democracy Now!, Rachel Maddow, Al Franken

Friendship, Cross Country, Acting, Swimming, Listening to Music, Having fun, Talking, Singing, Volleyball, Pictures, Hanging Out, Action movies, Laughing, Writing Songs, Watching TV, Eating and Sleeping, Talking to Friends, Boys

[O'Connor 2010]

Not just analysis: Crowdsourced annotations



[Snow, O'Connor, Jurafsky, Ng 2008]

Text Analysis for Social Science

- Tools for discovery and measurement
  - Social, spatial, temporal context
  - Probabilistic models
  - Linguistic tools
- Future work
  - Semantics: belief structures from text
  - Incorporate a-priori knowledge
  - Information retrieval and text visualization / exploration tools

Thanks

- All papers available at: http://brenocon.com