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

Talk at Wharton Statistics Colloquium, Jan 15 2014

slides: http://brenocon.com/wharton_2014-01-15.pdf

abstract: http://brenocon.com/wharton_seminar_announcement_201401.pdf

Outline

- **Introduction: Textual Social Data**
- Case Study: International Relations
- Examples: Social Media Analysis

Computational Social Science

Official social data

Data collection

Data analysis

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

Google books

1900 2000

Thursday, January 16, 14

Roman census ~ 100 BC
Guerry's crime map = 1829
U.S. Census punchcard counter = 1890

Computation = tools

ROMAN CENSUS
http://en.wikipedia.org/wiki/File:Altar-of-Domitius-Ahenobarb.jpg
FRANCE CRIME / ED
http://www.datavis.ca/papers/swiss/swiss2x2.pdf
http://arxiv.org/pdf/0801.4263.pdf
http://www.datavis.ca/milestones/index.php?group=1800%2B
CENSUS TABULATION
http://www.columbia.edu/cu/computinghistory/census-tabulator.html
Hollerith machine -- spinoff became IBM
CONTEMPORARY
http://www.businessinsider.com/a-global-social-media-census-2013-10

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.
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.

Discovery and measurement in social media

Data

Statistical text analysis

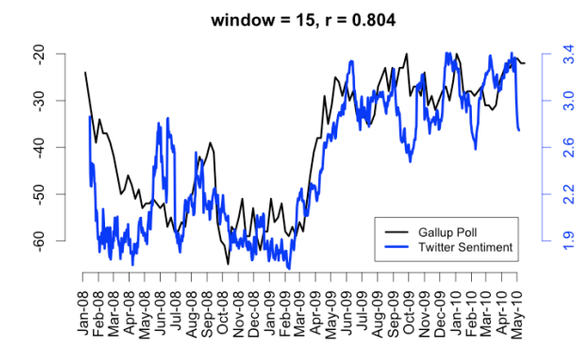
Linguistic analysis tools

[ACL 2011, NAACL 2013]

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

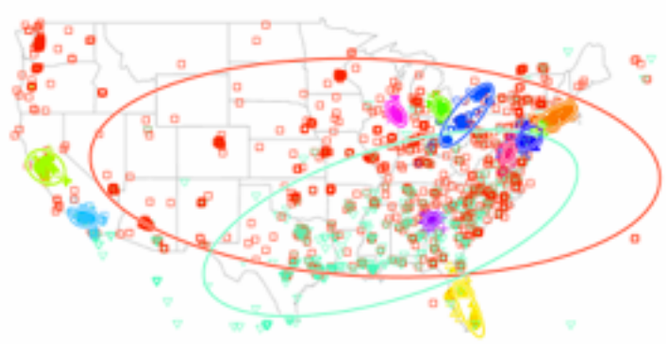
Opinion polls and sentiment analysis

[ICWSM 2010]



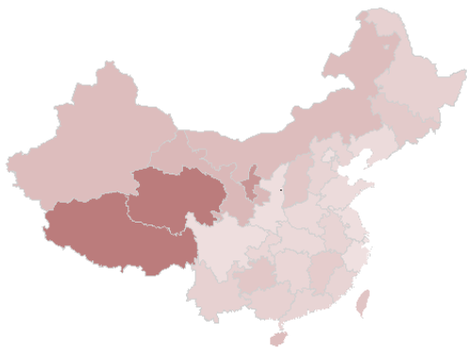
Geographic and demographic factors in slang and language change

[EMNLP 2010, work-under-review]



Censorship in Chinese social media

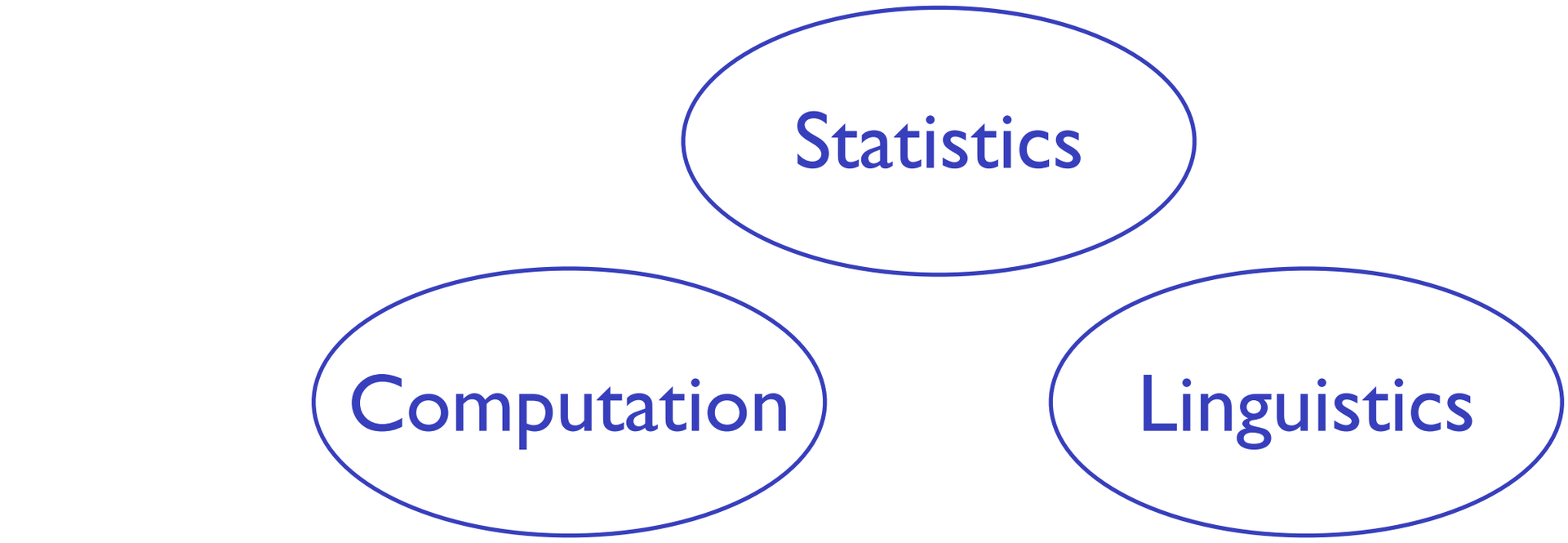
[FM 2011]



Advance methods in

Statistical Modeling of Text

Tools for discovery and measurement of concepts, attitudes, events



Politics
Literature
Business
Economics
Sociology
Health

... applied to the **social sciences** and humanities

first case study -- all three

Outline

- Introduction: Textual Social Data
- **Case Study: International Relations**
- Examples: Social Media Analysis

International Relations

- Forecasting: When and where will future conflicts happen?
- Understanding: What causes war? When do crises escalate?

Knowledge engineering approach

[Schrodt 1994, Leetaru and Schrodt 2013] http://gdelt.utdallas.edu

Event classes
(~200)

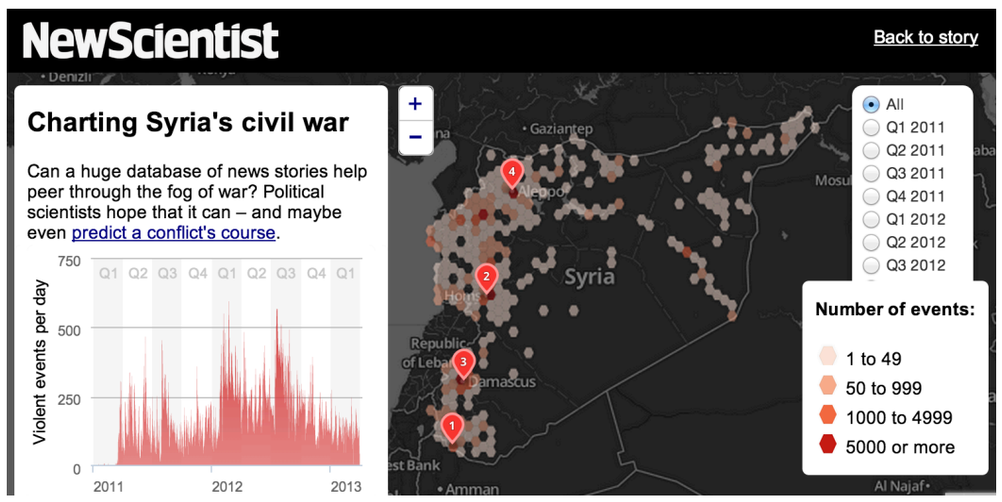
Dictionary:
Verb patterns per event class
(~15000)

Extract events from news text

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

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



Issue: Hard to maintain and adapt to new domains

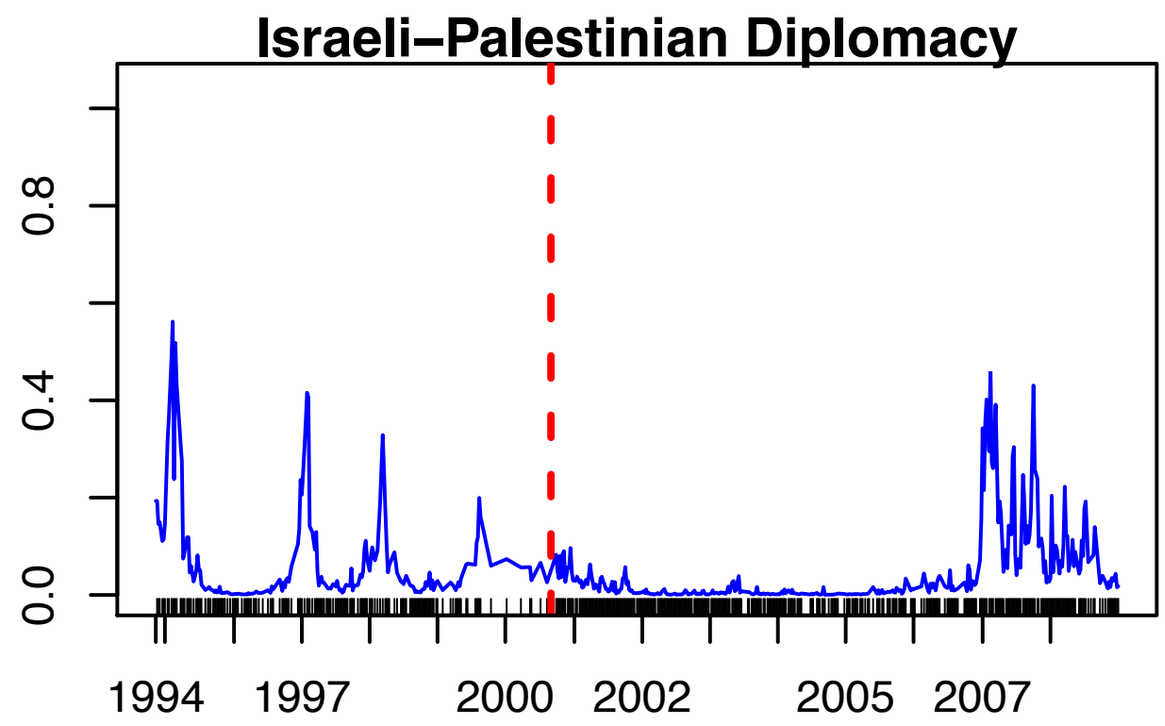
Our approach

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

Natural language processing

Event phrases

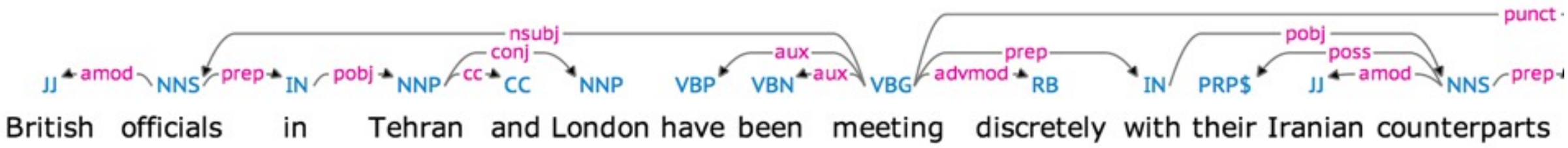
Hierarchical Model



Jointly learn

- Event class dictionaries
- Political dynamics

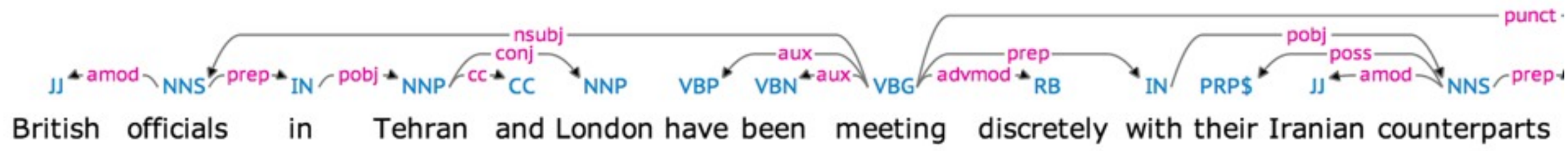
Event Extraction:
Who did what to whom?



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

from APW_ENG_200704
source: subject or by-agent of verb
receiver: most any path, usually direct object or prep arg

Event Extraction:
Who did what to whom?



GBR

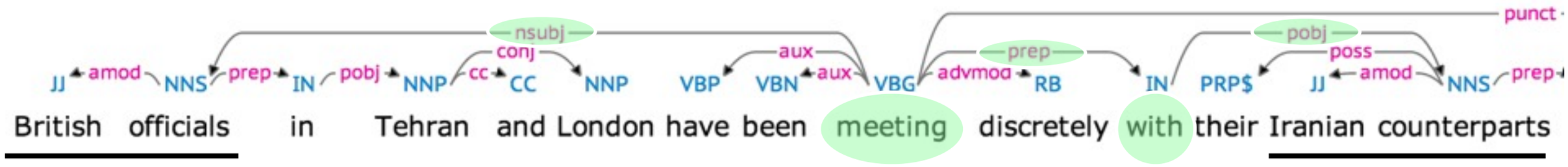
IRN

Match
country name list

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

from APW_ENG_200704
source: subject or by-agent of verb
receiver: most any path, usually direct object or prep arg

Event Extraction:
Who did what to whom?



GBR

IRN

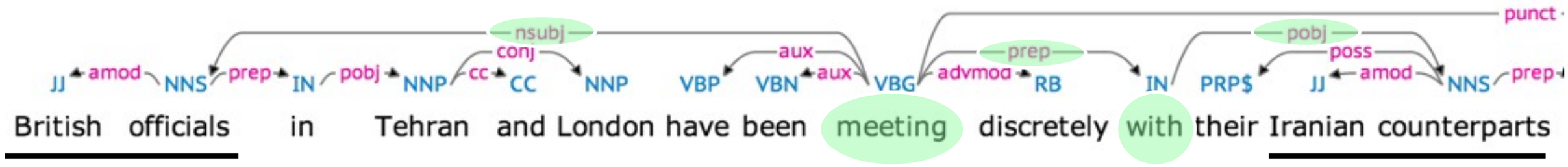
Match
country name list

Extract
event phrase

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

from APW_ENG_200704
source: subject or by-agent of verb
receiver: most any path, usually direct object or prep arg

Event Extraction:
Who did what to whom?



GBR

IRN

Match
country name list

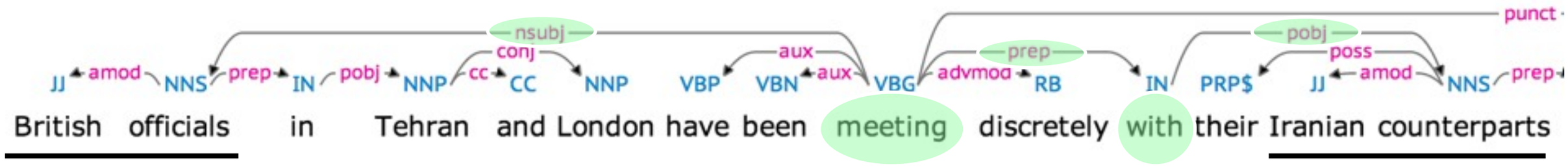
Extract
event phrase

Source *(s)*: GBR
Recipient *(r)*: IRN
Event phrase *(w)*: <--*nsubj*-- meet --*prep*--> with --*pobj*-->

"*X* meets with *Y*"

from APW_ENG_200704
source: subject or by-agent of verb
receiver: most any path, usually direct object or prep arg

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

from APW_ENG_200704
source: subject or by-agent of verb
receiver: most any path, usually direct object or prep arg

- Inputs
  1. 6.5 million news articles, 1987-2008
  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

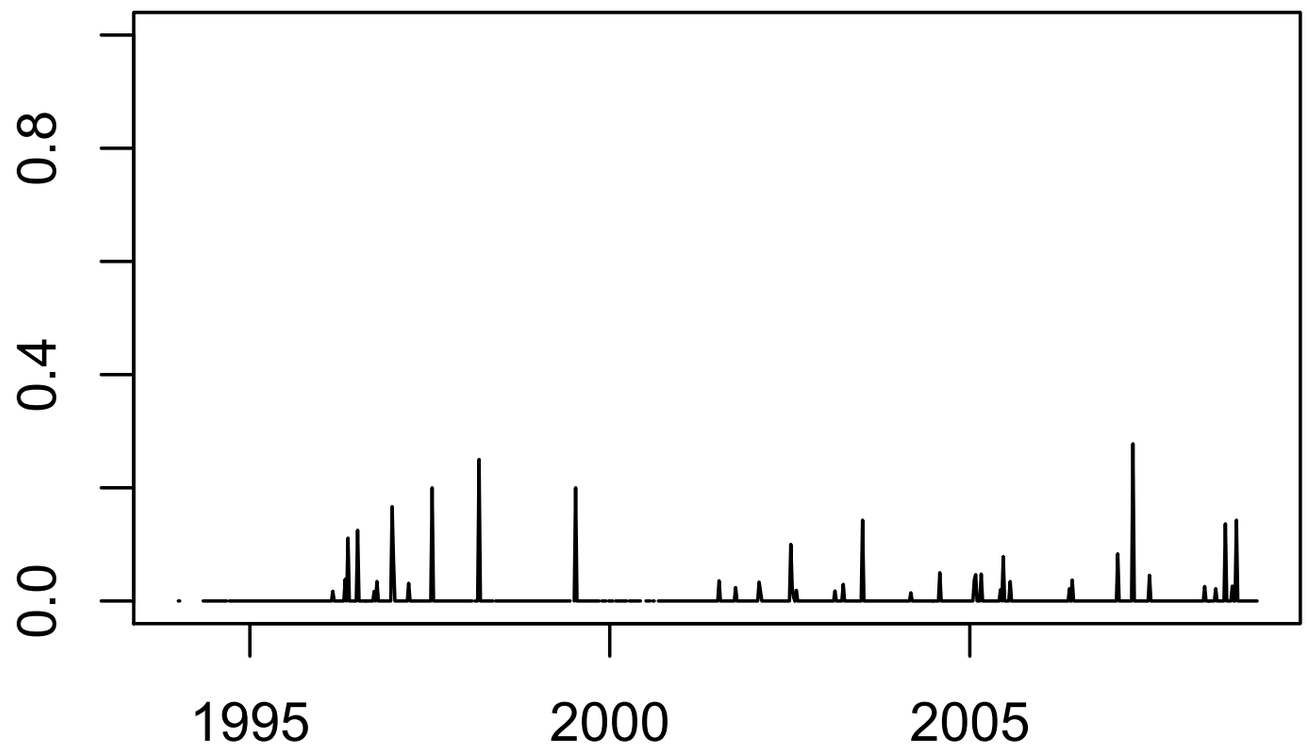
365,623 event tuples
421 directed dyads (s,r)
10,457 event phrases (w)
1,149 weeks (t)

Thursday, January 16, 14

84 actors (union of s,r)
“invade by supporter of union with”: error, historical (cyprus) (misparse? subord clause rule looksl ike it should work)

"*ISR* meet with *PSE*"

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



Too sparse for human interpretability

ISR-PSE is the most common dyad
max week count = 9

Do event classes reflect dyadic-temporal structure?

s=*ISR*, r=*PSE*

t=811	t=1018
say <-ccomp be to	commit to
release to	strike
take control of	carry in
occupy	continue in
wound in	reject
scuffle with	fire at target in
be <-xcomp meet	start around
meet with	ratchet pressure on
meet with	shell
arrest	hit

s=*USA*, r=*FRA*

t=579	t=886
travel <-xcomp meet with	release with
consider	welcome
meet with	welcome by
meet with	win
meet with	agree with
	indict
	win from
	concern over
	win
	indict

- Approach: Hierarchical Dirichlet and logistic normal models

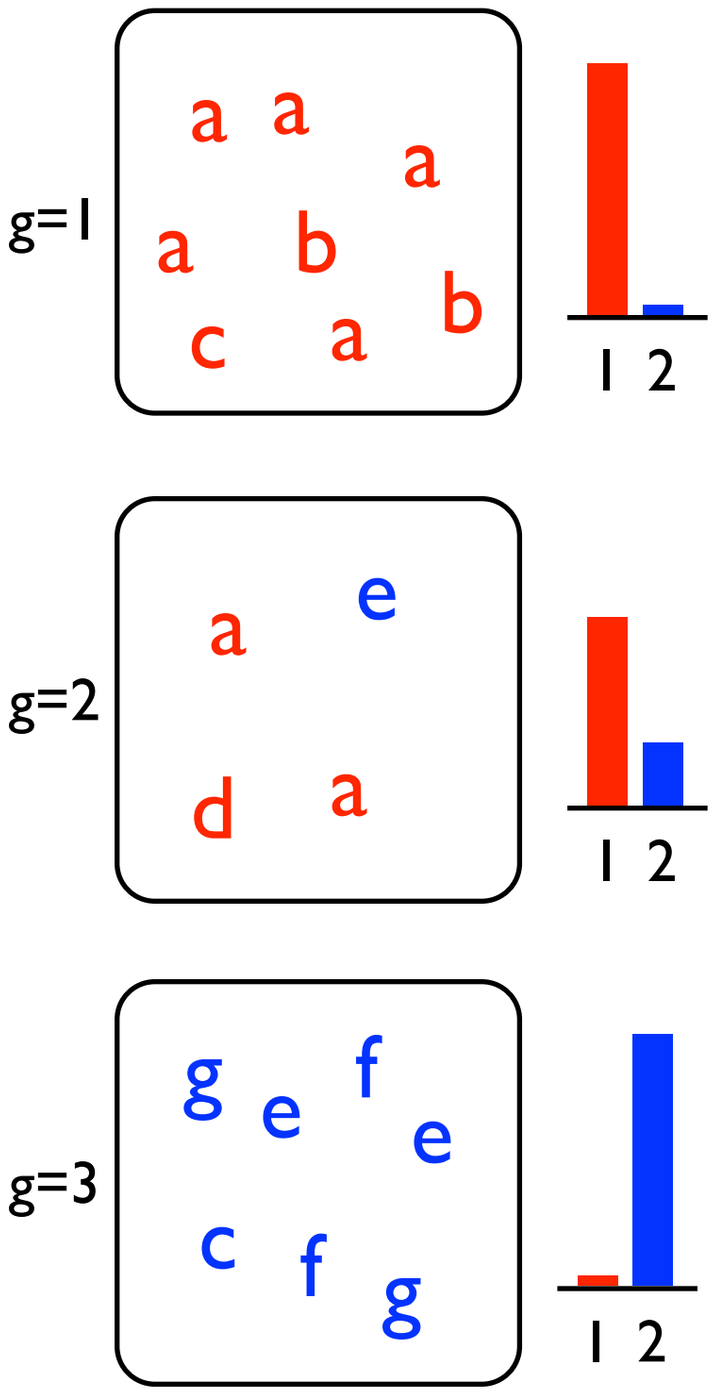
highest “meet with” count, vs highest “kill” count

WHY BAYES
- encode assumptions
- combine multiple sources of information
- pragmatic: rich toolset

Mixed-membership models for semantic learning

"admixture model," "topic model"

K latent word classes, to explain grouped discrete data



Distrib. over *K* classes

$$\theta_g \sim Dir$$

Distrib. over words

$$\phi_k \sim Dir$$

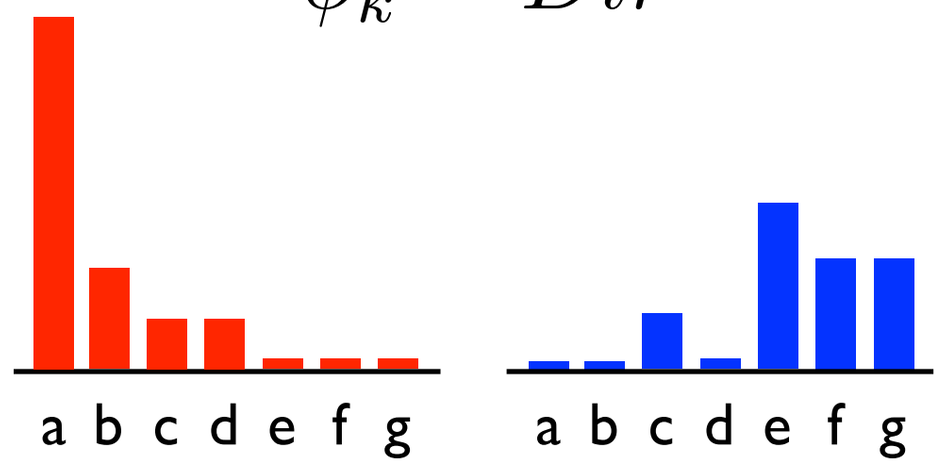
Word distrib. is mixture

$$w \sim \text{Mult}(\Phi\ \theta)$$

$$\updownarrow$$

$$z \sim \text{Mult}(\theta_g)$$

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



Latent Dirichlet allocation

[Pritchard et al. 2000, Blei et al. 2003]

Thursday, January 16, 14

```
CITE PRITCHARD, BLEI

w \sim \text{Mult}(\Phi\ \theta)
z \sim \text{Mult}(\theta)
w \sim \text{Mult}(\phi_z)

[1] 0.7272727
> 8/11 * 100
[1] 72.72727
> 8/11 * 200
[1] 145.4545
> 2/11 * 200
[1] 36.36364
> 1/8*200
[1] 25
```

Do event classes reflect dyadic-temporal structure?

s=*ISR*, r=*PSE*

t=811	t=1018
say <-ccomp be to	commit to
release to	strike
take control of	carry in
occupy	continue in
wound in	reject
scuffle with	fire at target in
be <-xcomp meet	start around
meet with	ratchet pressure on
meet with	shell
arrest	hit

s=*USA*, r=*FRA*

t=579	t=886
travel <-xcomp meet with	release with
consider	welcome
meet with	welcome by
meet with	win
meet with	agree with
	indict
	win from
	concern over
	win
	indict

K event classes (latent factors)
Distribution over event phrases

$$\phi_k \sim Dir$$

(M0) Dyad admixture: $p(w|s,r) = \Phi\ \theta_{s,r}$

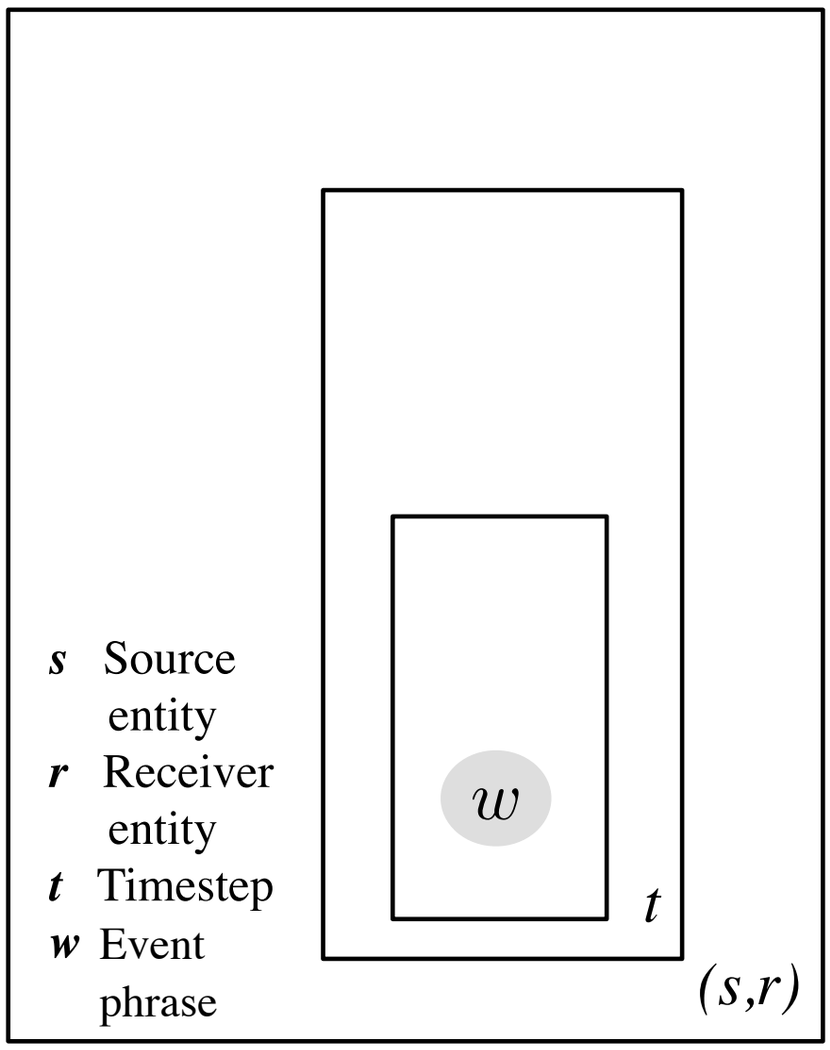
(M1) Temporal-dyad admixture: $p(w|s,r,t) = \Phi\ \theta_{s,r,t}$

Thursday, January 16, 14

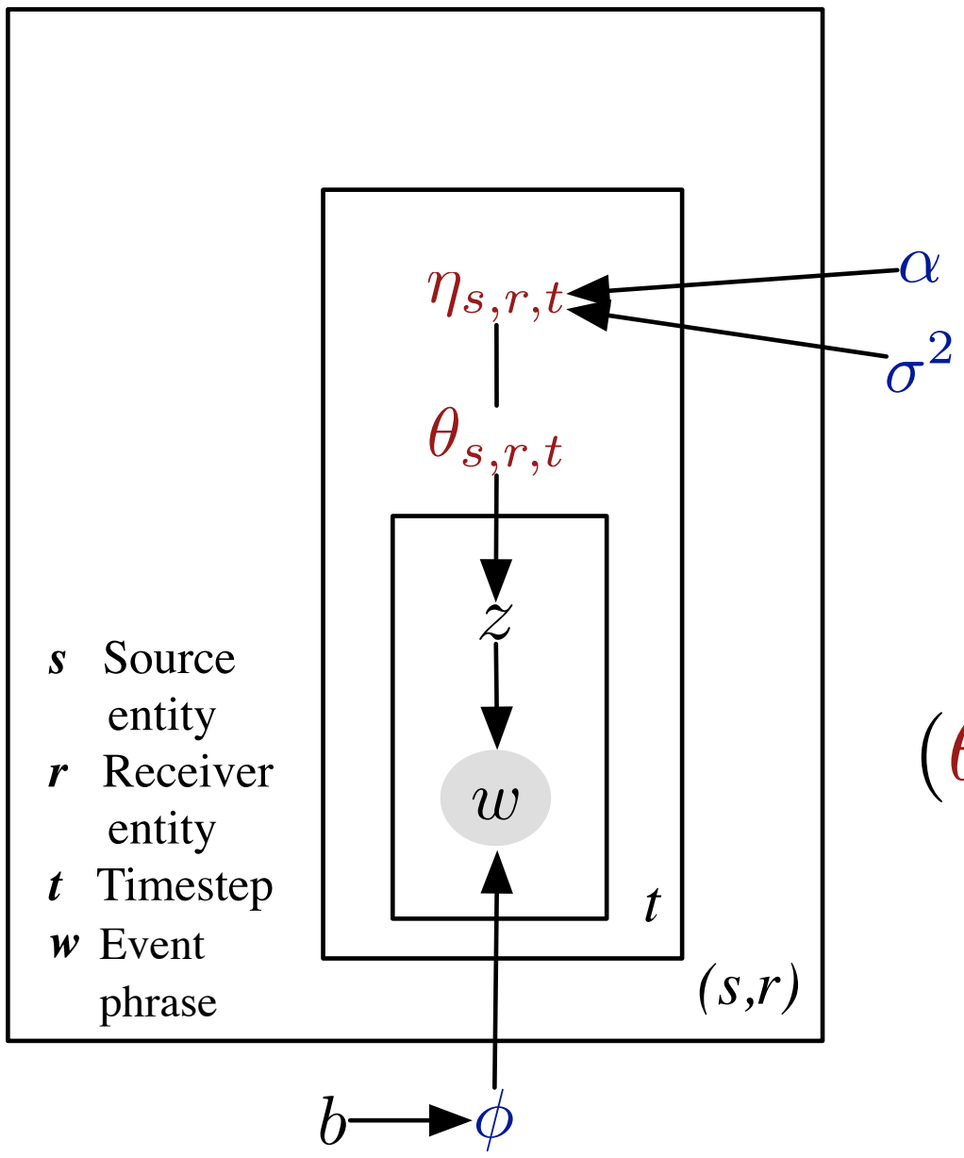
highest “meet with” count, vs highest “kill” count
dirichlet prior on theta => latent Dir allocation

```
p(w|s,r) = \Phi\ \theta_{s,r}
p(w|s,r,t) = \Phi\ \theta_{s,r,t}
```

Model


s Source entity
r Receiver entity
t Timestep
w Event phrase
w
t
(s,r)

Model



Logistic Normal prior

M1: independent contexts

$\in \mathbb{R}^K$: Event class prevalences

Per-context event class sparsity

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

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

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

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

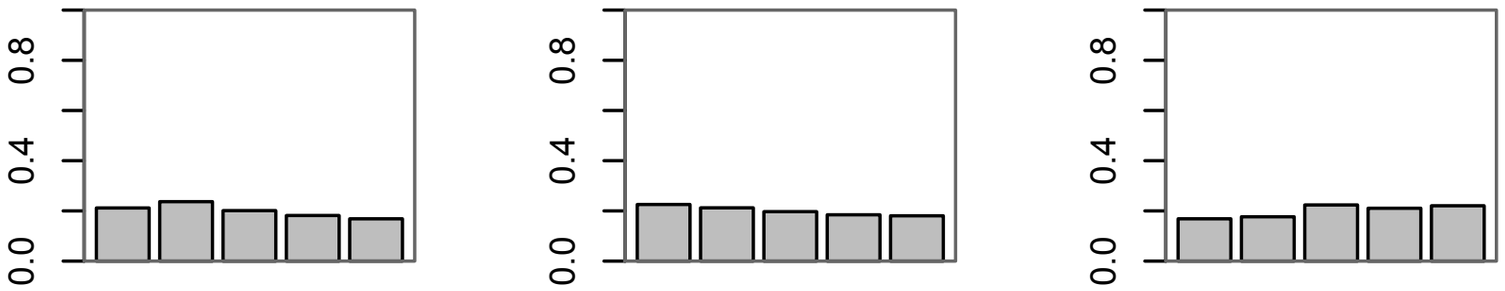
$$\phi_k \sim \text{Dir}(b)$$

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

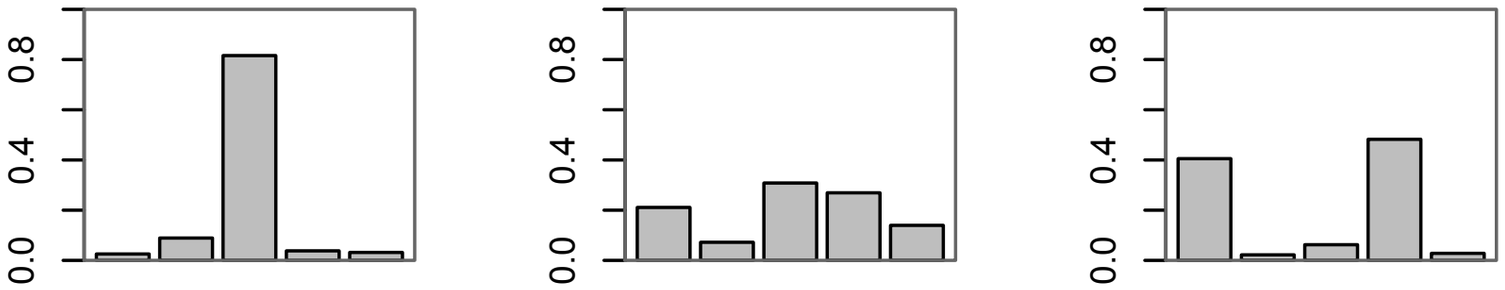
$$\eta_{s,r,t} \sim N(\alpha \quad , \mathrm{Diag}[\sigma_1^2..\sigma_K^2])$$

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

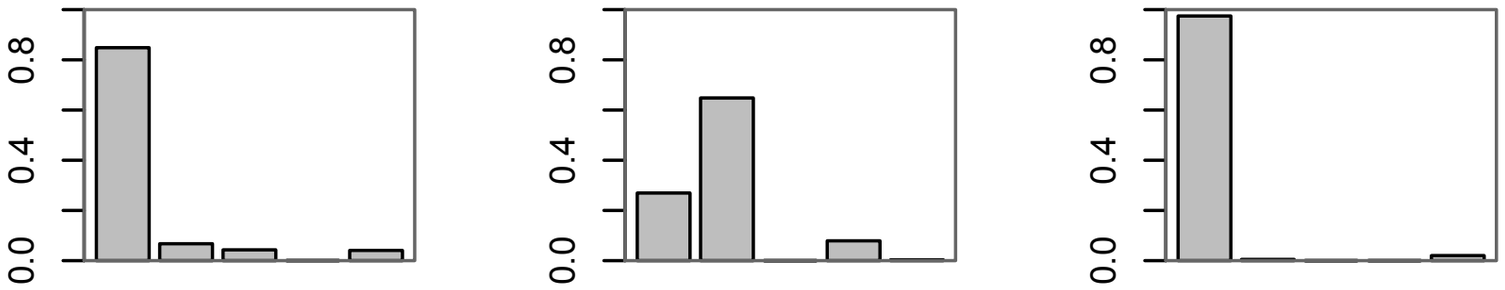
$\sigma = 0.1$



$\sigma = 1$

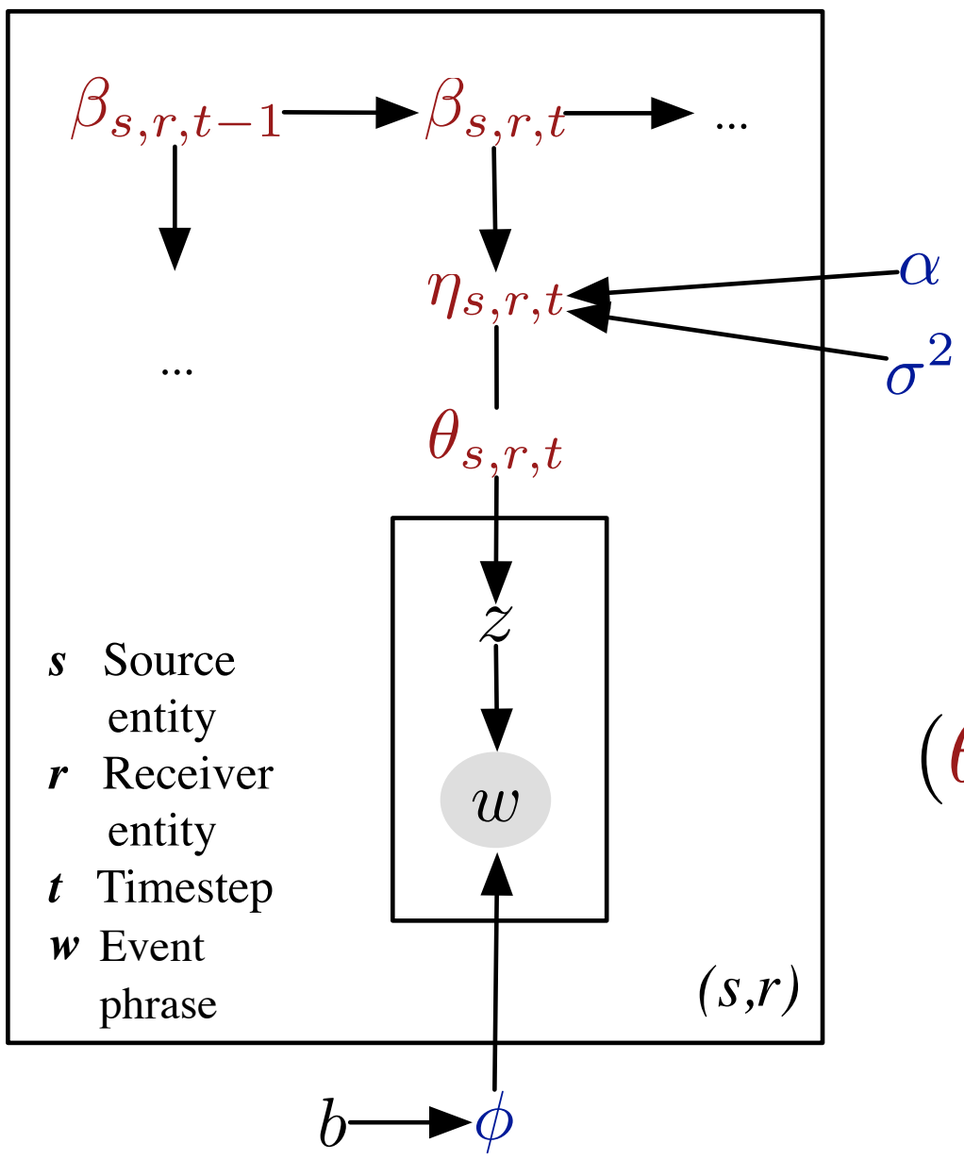


$\sigma = 5$



Dir = normalized gammas
LN = normalized exponentiated normals
sparsity control

Model



Logistic Normal prior

M1: independent contexts
M2: temporal smoothing
[Blei and Lafferty 2006, Quinn and Martin 2002]

Adjacent timestep similarity

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

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

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

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

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

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

K=100 ⟶ 80 million parameters

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

Thursday, January 16, 14

```
|s,r| = 421
|t| = 1149
|s,r|*|t| = 484k
|phi| = K*10k  =>  1mil at K=100

\cp{\beta_{s,r,t}} &\sim N(\cp{\beta_{s,r,t-1}}, \mathbb{I}\gp{\tau^2}) \\
\cp{\eta_{s,r,t}} &\sim N(\gp{\alpha} + \cp{\beta_{s,r,t}},
\text{Diag}[\gp{\sigma^2_1}..\gp{\sigma^2_K}]) \\
(\cp{\theta_{s,r,t}} )_k &\propto \exp(\cp{\eta_{s,r,t,k}}) \\
z &\sim \text{Mult}(\cp{\theta_{s,r,t}}) \\
w &\sim \text{Mult}(\gp{\phi_{z}}) \\
\gp{\phi_k} &\sim \text{Dir}(b)

\cp{\beta_{s,r,t}} &\sim N(\cp{\beta_{s,r,t-1}}, \mathbb{I}\gp{\tau^2}) \\
\cp{\eta_{s,r,t}} &\sim N(\gp{\alpha} + \cp{\beta_{s,r,t}},
\text{Diag}[\gp{\sigma^2_1}..\gp{\sigma^2_K}]) \\
(\cp{\theta_{s,r,t}} )_k &\propto \exp(\cp{\eta_{s,r,t,k}}) \\
z &\sim \text{Disc}(\cp{\theta_{s,r,t}}) \\
w &\sim \text{Disc}(\gp{\phi_{z}}) \\
\gp{\phi_k} &\sim \text{Dir}(b)
```

Inference: blocked Gibbs sampling

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

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

Conjugate normal

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

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

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

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

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

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

Dirichlet-multinomial
Collapsed sampling
[Griffiths and Steyvers 2005]

Slice sampling
[Neal 2003]

only tuning parameter is K

Laplace approx. to Logistic normal

$$\eta \sim N(\bar{\eta},\ \text{Diag}[\sigma_1^2..\sigma_K^2])$$
$$z \sim \text{Mult}(\theta(\eta))$$

$$\theta(\eta) = \exp(\eta)/\text{sum}(\exp(\eta))$$

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

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

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

Metropolis rejections correct approximation error

Alternative to variational inference for LN

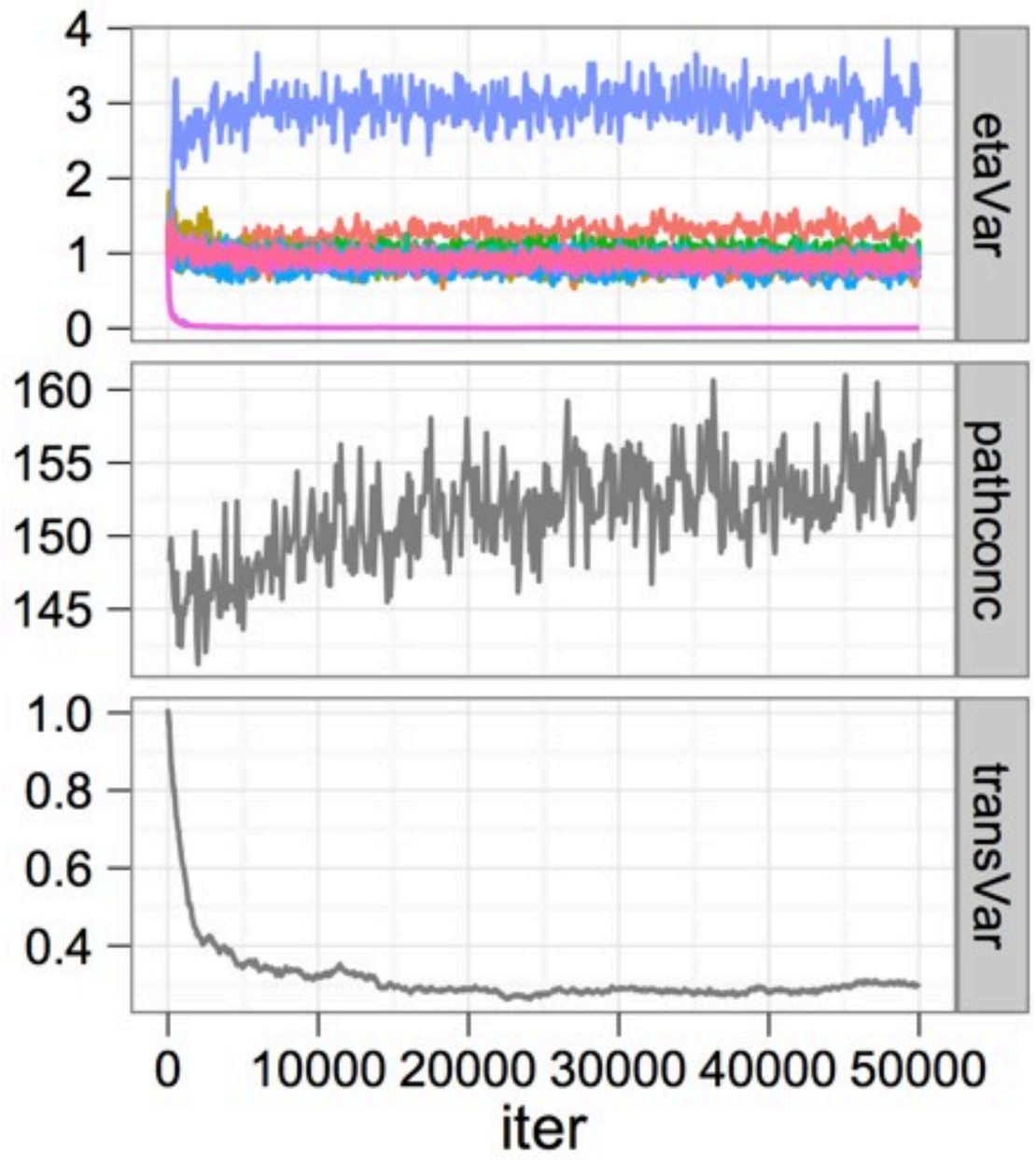
[e.g. Ahmed and Xing 2007, Blei and Lafferty 2006, Wang and Blei 2013]

Thursday, January 16, 14

```
get MAP then propose mean-centered gaussian -- which is the taylor expansion
bayesian CLT


\cp{\eta} &\sim
N(\gp{\mu},\
\text{Diag}[\gp{\sigma^2_1}..\gp{\sigma^2_K}]) \\
\theta &= S(\eta) \equiv \exp(\eta)/\text{sum}(\exp(\eta)) \\
z &\sim \text{Mult}(\cp{\theta})

\theta(\eta) = \exp(\eta)/\text{sum}(\exp(\eta))


p(\eta|\mu,\Sigma,z) \propto N(\eta;\mu,\Sigma)
\ \text{Mult}(\vec{z}; S(\eta))

\hat{\eta}=\arg\max_\eta
\sum_k \left(
-\frac{1}{2 \sigma^2_k} (\eta_k-\bar{\eta}_k)^2 + n_k \log \theta(\eta)_k
\right)
```

Inference

- Implementation
  - Parallelization
    - Few hours to few days
    - Thinning
      (600 MB/sample)
    - Java (R, Python)
- Dispersion vs. mixture components



few days / 1000's of CPU hours
dozens to hundreds of CPU cores

Event Classes Posteriors

Most probable phrases in ϕ_k

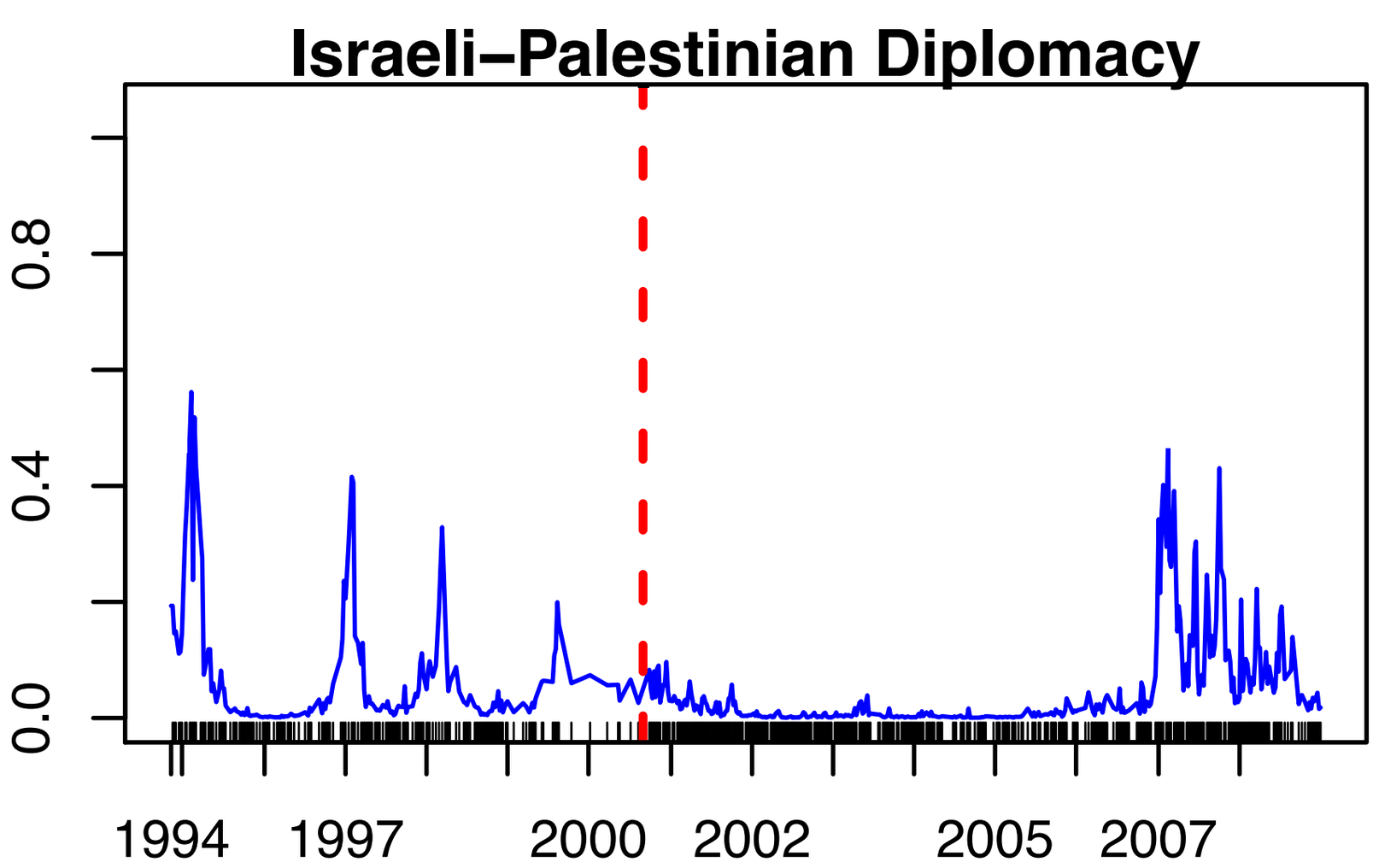
arrive in, visit, meet with, travel to, leave, hold with, meet, meet in, fly to, be in, arrive for talk with, say in, arrive with, head to, hold in, due in, leave for, make to, arrive to, praise
accuse, blame, say, break with, sever with, blame on, warn, call, attack, rule with, charge, say←ccomp come 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 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



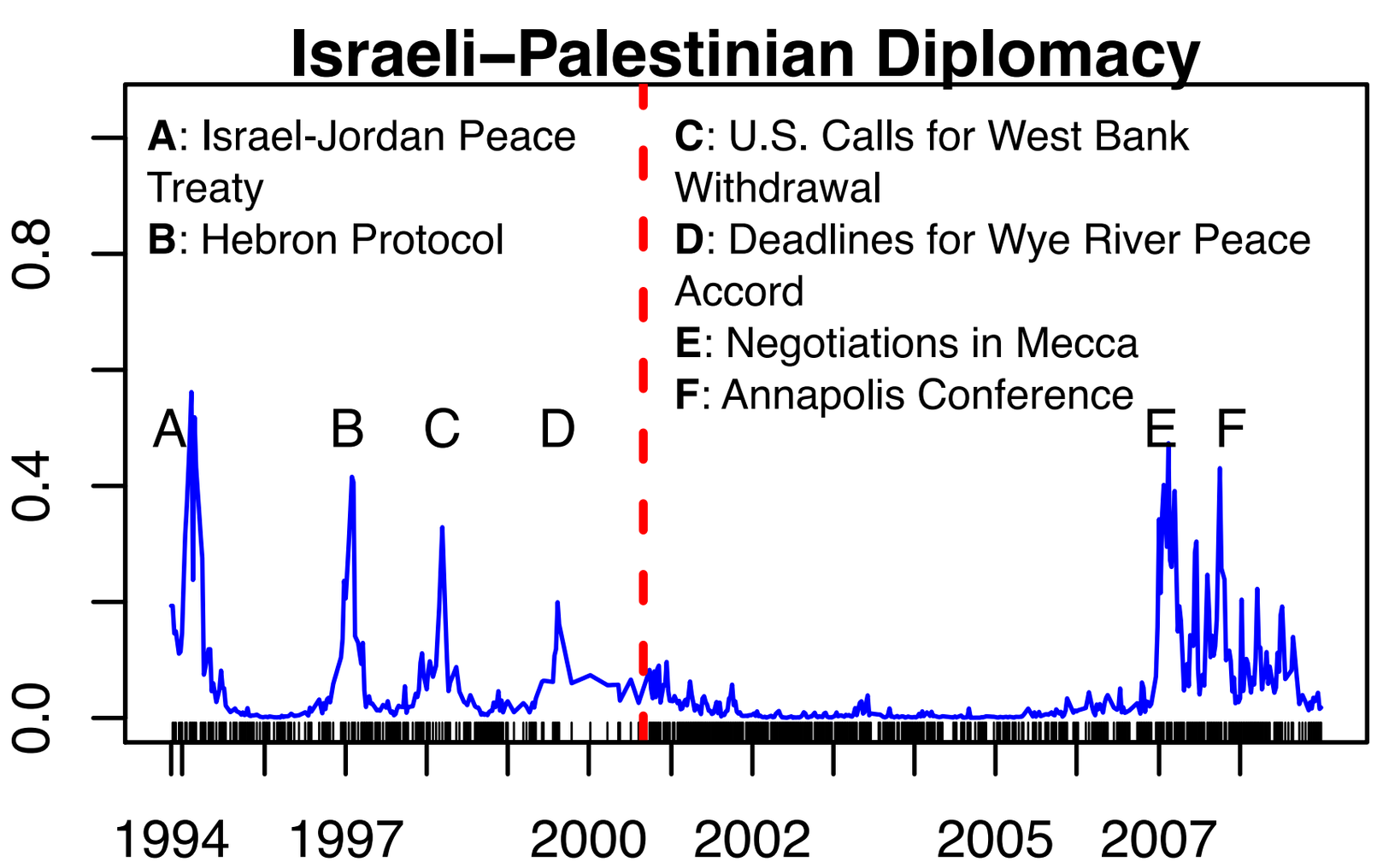
Thursday, January 16, 14

BLUE - Our model is also able to pick up positive diplomatic events as evidenced by the frame in Figure \ref{fig:case}(c), describing Israeli diplomatic actions towards Palestine (`meet with,sign with, praise,say with, arrive in'). Not only do the spikes coincide with major peace treaties and other negotiations, but the model correctly characterizes the relative lack of positively valenced action from the beginning of the Second Intafada until its end around 2005-2006.

GREEN -- In Figure \ref{fig:case}(b) we show that our model produces a frame which captures the legal aftermath of particular events (`accuse, criticize,' but also `detain, release, extradite, charge'). Each of the major spikes in the data coincides with a particular event which either involves the investigation of a particular attack or series of attacks (as in A,B,E) or a discussion about prisoner swaps or mass arrests (as in events D, F, J).

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

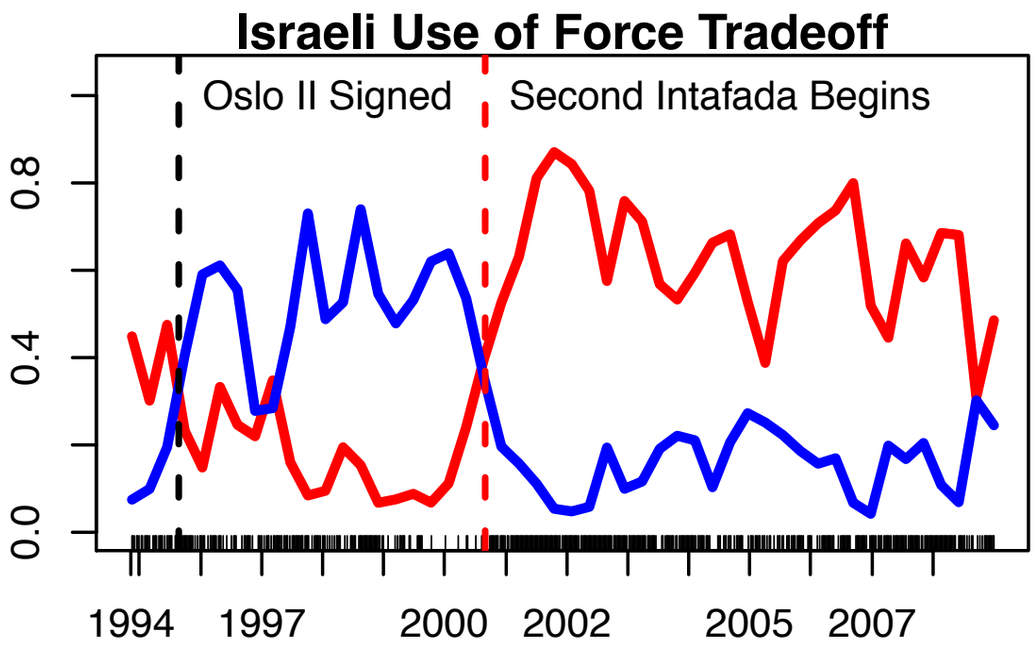


Thursday, January 16, 14

BLUE - Our model is also able to pick up positive diplomatic events as evidenced by the frame in Figure \ref{fig:case}(c), describing Israeli diplomatic actions towards Palestine (`meet with,sign with, praise,say with, arrive in'). Not only do the spikes coincide with major peace treaties and other negotiations, but the model correctly characterizes the relative lack of positively valenced action from the beginning of the Second Intafada until its end around 2005-2006.

GREEN -- In Figure \ref{fig:case}(b) we show that our model produces a frame which captures the legal aftermath of particular events (`accuse, criticize,' but also `detain, release, extradite, charge'). Each of the major spikes in the data coincides with a particular event which either involves the investigation of a particular attack or series of attacks (as in A,B,E) or a discussion about prisoner swaps or mass arrests (as in events D, F, J).

Case study



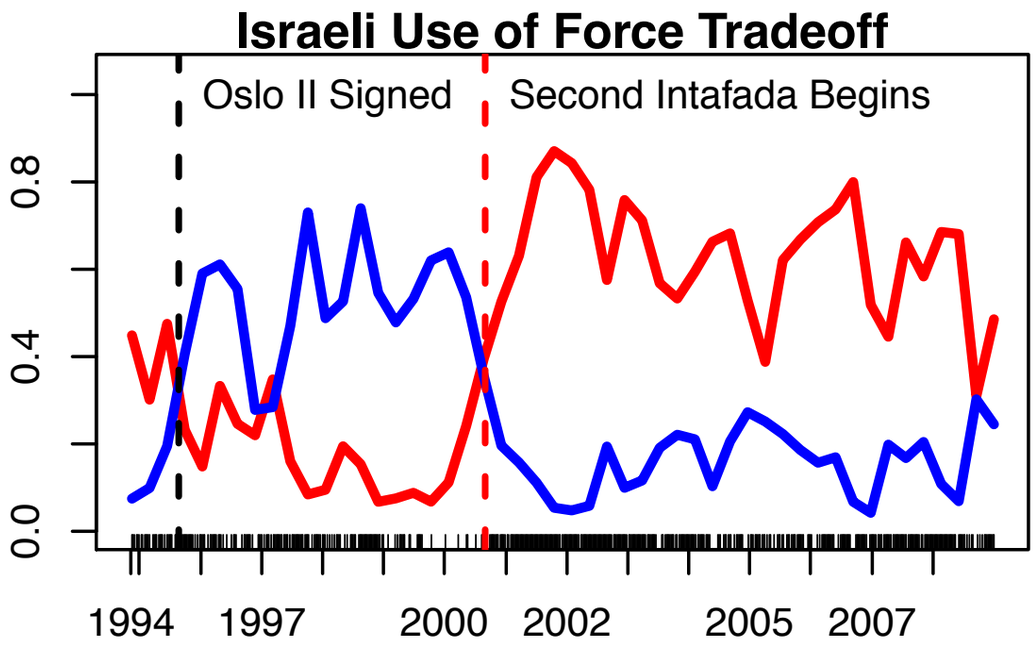
impose on, seal, capture from, seize from, arrest, ease closure of, close, deport, close with, release

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

Red line: collapse of Camp David negotiations
“include”: “Pales figure includes”
lowess

This balance persists until the abrupt breakdown in relations that followed the unsuccessful Camp David Summit in July of 2000, which generally marks the starting point of the wave of violence known as the Second Intifada.

Case study



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?

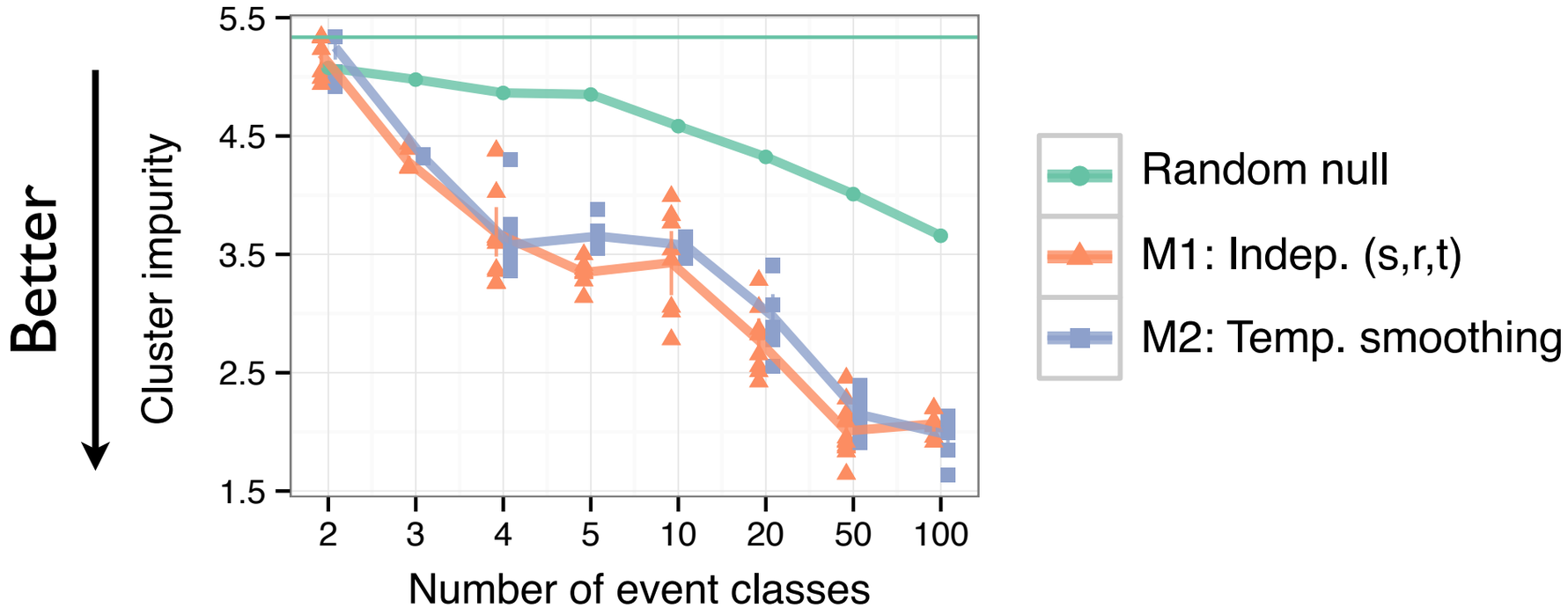
Semantic coherence?

Thursday, January 16, 14

Red line: collapse of Camp David negotiations
“include”: “Pales figure includes”
lowess

This balance persists until the abrupt breakdown in relations that followed the unsuccessful Camp David Summit in July of 2000, which generally marks the starting point of the wave of violence known as the Second Intifada.

Evaluation: lexicon/ontology



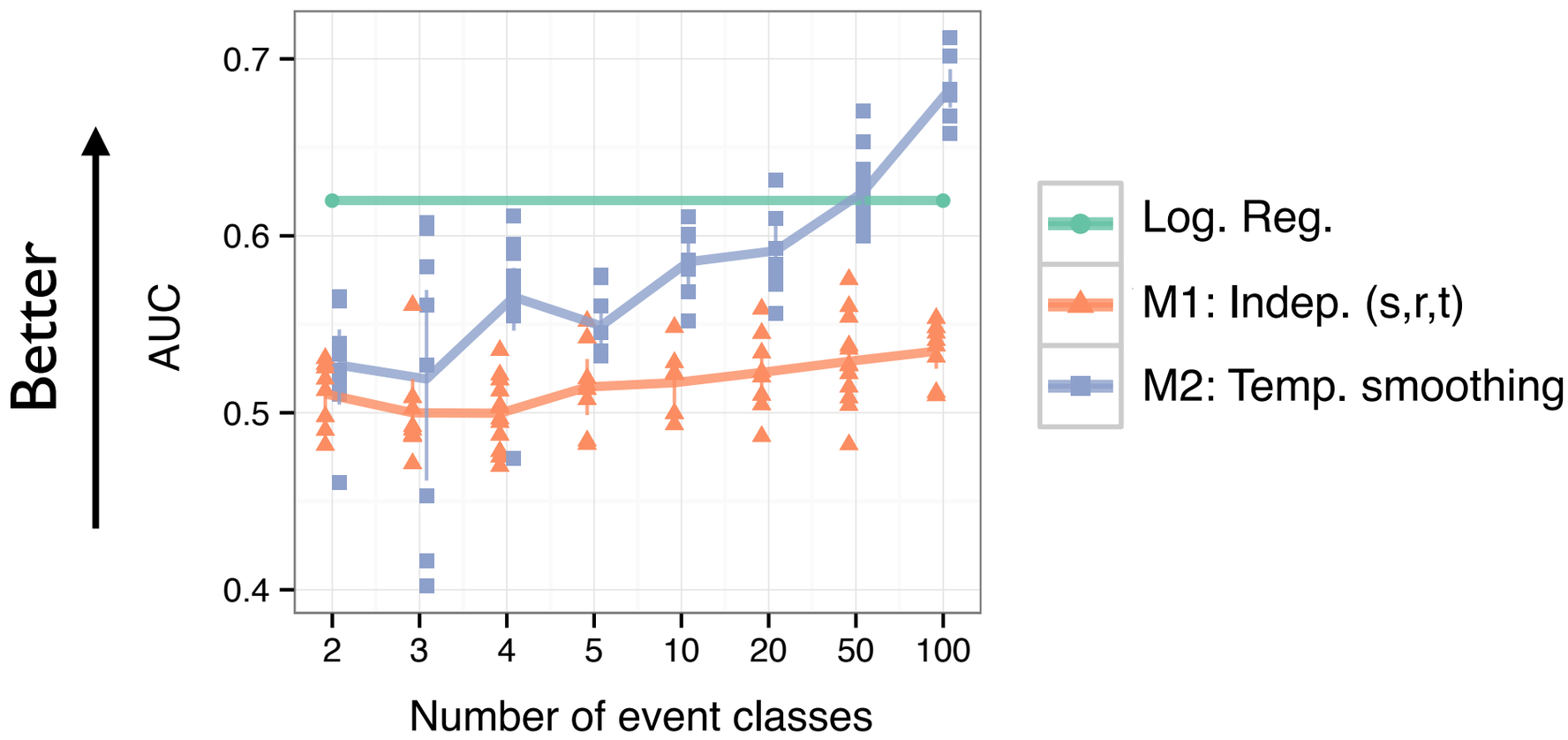
Do our event types (verb clusters) match the manually defined ontology?

Thursday, January 16, 14

Match dependency paths against TABARI patterns (536 / 10k)
Granularity invariance: use expert-assigned scale score (-10 to 10) [controversial?]
Lexical scale impurity: average difference between randomly chosen words
Random clusters baseline

Evaluation: conflict correlation

Do our event classes probabilities correspond to real-world conflict?



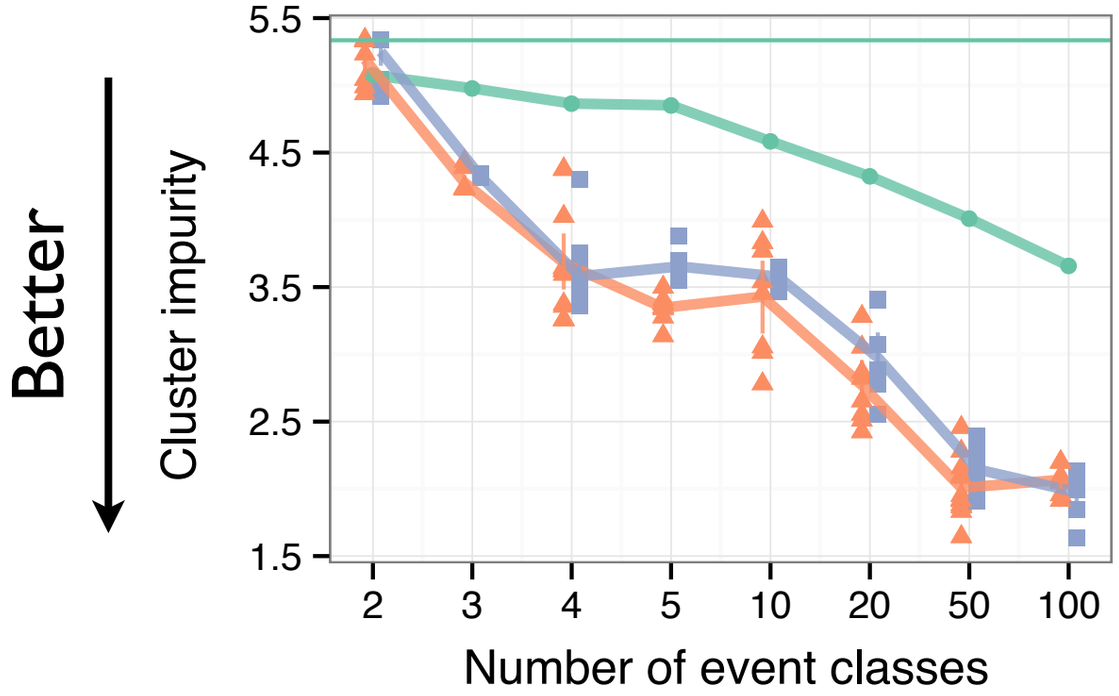
Thursday, January 16, 14

1. baseline
2. variability!

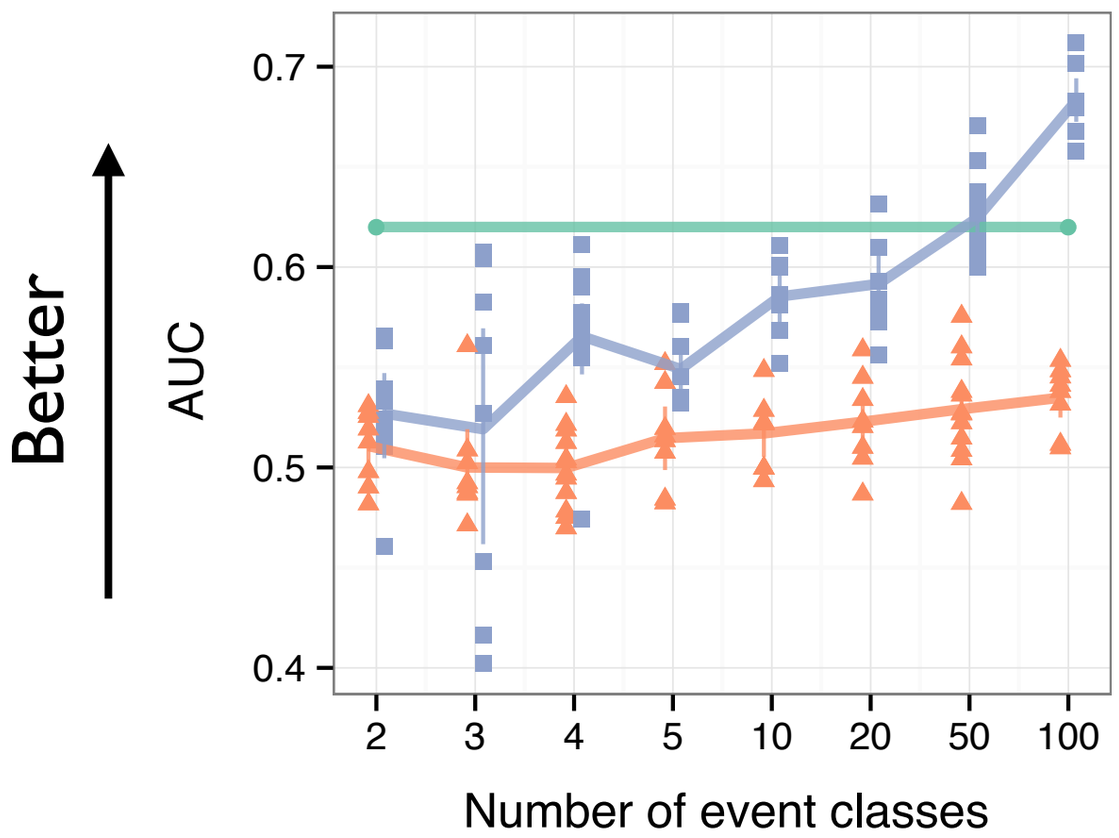
==

Do our event types correspond to real-world conflict?
“Gold” standard: Militarized Interstate Dispute dataset
(from Correlates of War project)
Regularized logistic regression from theta (event probs per dyad-time slice)
Baseline: regularized logistic regression from path counts

Evaluations



Lexicon / Ontology reconstruction



Real-world conflict reconstruction

Thursday, January 16, 14

tension b/w interp vs prediction
conflict reconstruction needs higher K
K=50, lexicon learning stalls out
Unsup model eval: multiple checks of reasonableness. Compare models.

is the ontology bad?

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*]

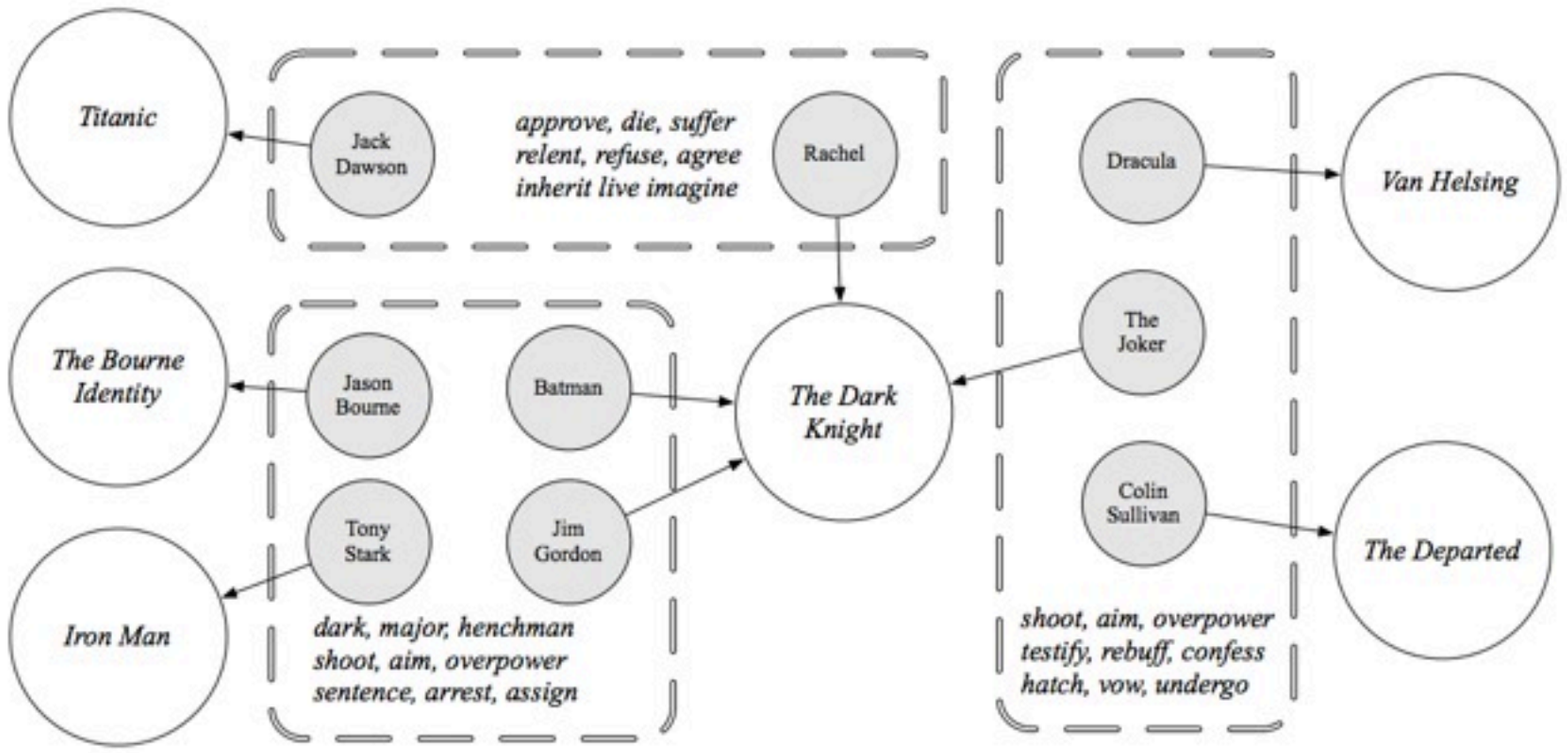


Figure 3: Dramatis personae of *The Dark Knight* (2008), illustrating 3 of the 100 character types learned by the persona regression model, along with links from other characters in those latent classes to other movies. Each character type is listed with the top three latent topics with which it is associated.

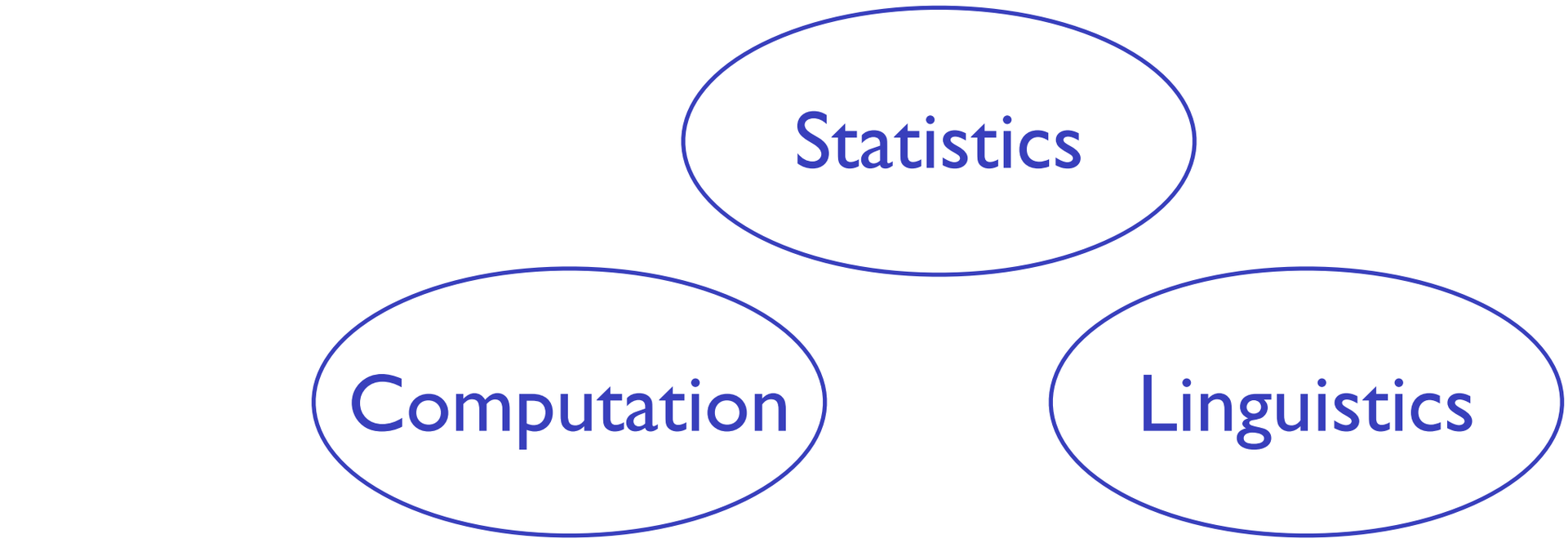
Thursday, January 16, 14

IR OTHER STUFF
token-level eval: Boschee et al. 2013, King and Lowe 2003, but no commonly available dataset
Linguistic learning based on all dyads, but inference for one dyad is much sparser (10's to 1000's of events)
We want good performance on rare dyads; e.g. the Congo is important but has very little media attention

Advance methods in

Statistical Modeling of Text

Tools for discovery and measurement of concepts, attitudes, events



Politics
Literature
Business
Economics
Sociology
Health

... applied to the **social sciences** and humanities

first case study -- all three

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

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

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

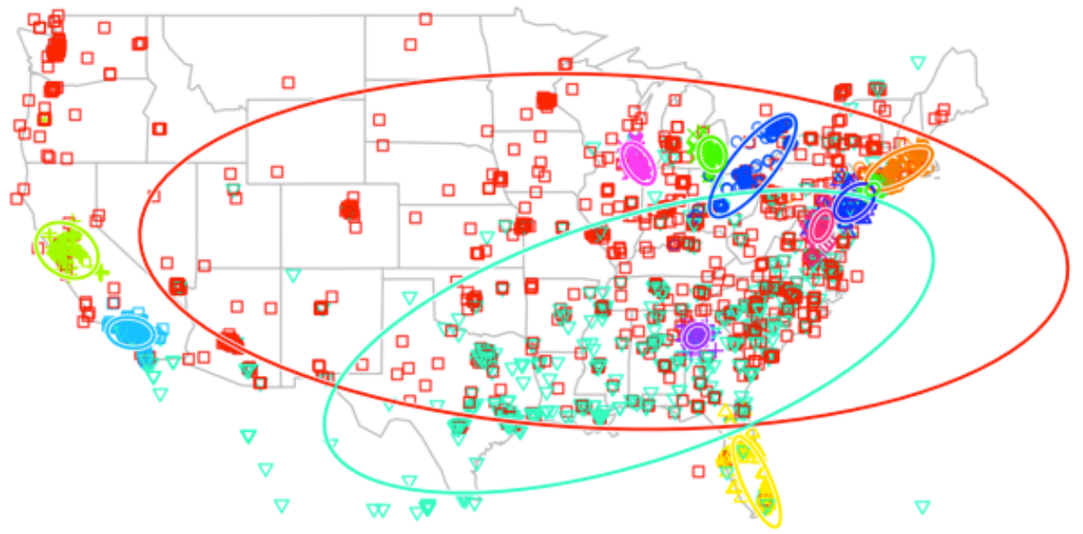
User's topics

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

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

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

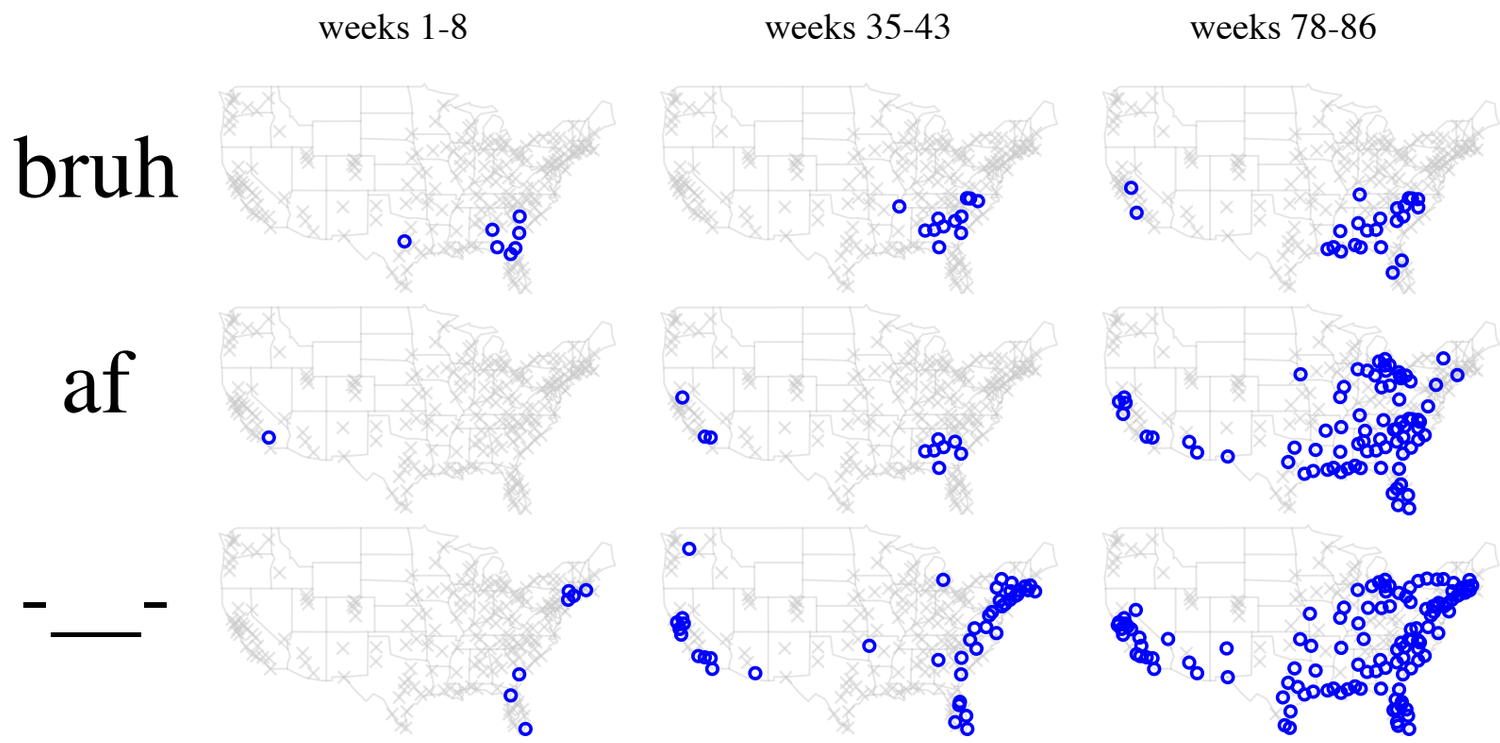
have regional variants



	"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 2011 and in review]



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

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

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

$\boldsymbol{\Gamma}$ variance of the autoregressive process (size $R \times R$)

7 TB data

200 regions, 2600 words, 165 timesteps = 85M parameters

Distributed implementations

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

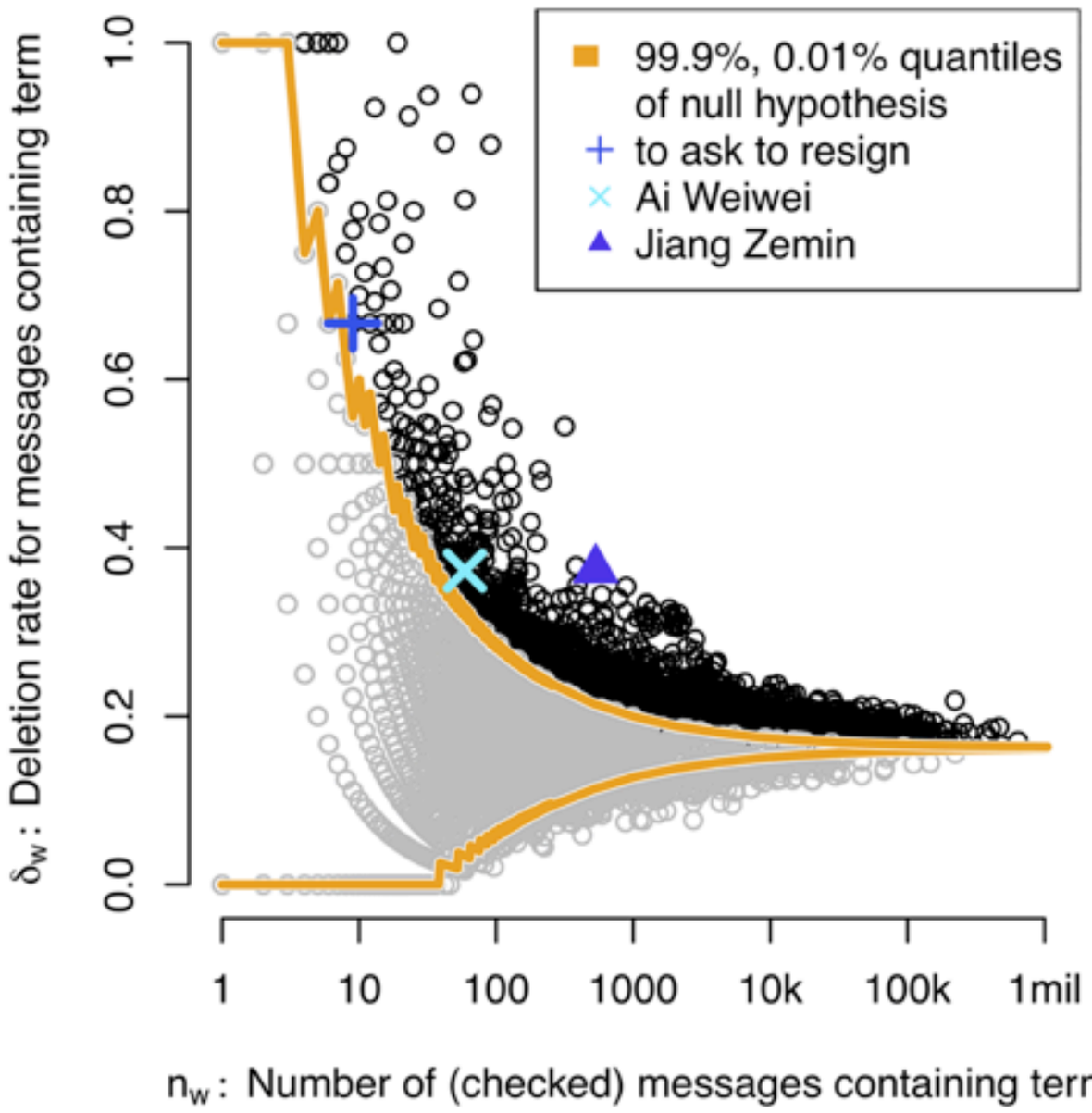
http://www.ark.cs.cmu.edu/TweetNLP/

[*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

Text Analysis for Social Science

- Tools for discovery and measurement of concepts, attitudes, events
  - Social, spatial, temporal context
- Future work
  - Incorporate a-priori knowledge
  - Easy-to-use analysis methods for practitioners

Thursday, January 16, 14

prev re-outline
 - socmed
 - IR
 - literature
Collaborations
Expert knowledge

social context: WHO and WHEN
ONTOLOGIES: evaluations, hypotheses to check OR expand
Applications to social science
Social contextual factors can drive linguistic learning
Using expert ontologies,
Test social scientific theories

Thanks

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