

## 12. Special Topics

DS-GA 3001, Text as Data  
Arthur Spirling

April 24, 2018

# Housekeeping

# Housekeeping

- 1 Homework coming in on [May 1](#) (to Leslie's mailbox, by 5pm). Note the rules about academic dishonesty: do not copy anyone's work (including code) and do not allow them to copy your work (including code).

# Housekeeping

- 1 Homework coming in on [May 1](#) (to Leslie's mailbox, by 5pm). Note the rules about academic dishonesty: do not copy anyone's work (including code) and do not allow them to copy your work (including code). Final papers come in to Leslie's mailbox on or before 5pm on [May 11](#). No extensions, no exceptions!

# Housekeeping

- 1 Homework coming in on [May 1](#) (to Leslie's mailbox, by 5pm). Note the rules about academic dishonesty: do not copy anyone's work (including code) and do not allow them to copy your work (including code). Final papers come in to Leslie's mailbox on or before 5pm on [May 11](#). No extensions, no exceptions!
- 2 Speaker series tomorrow: Sarah Bouchat.
- 3 Fill out class evals in lab Thursday (please!)

# Housekeeping

- 1 Homework coming in on **May 1** (to Leslie's mailbox, by 5pm). Note the rules about academic dishonesty: do not copy anyone's work (including code) and do not allow them to copy your work (including code). Final papers come in to Leslie's mailbox on or before 5pm on **May 11**. No extensions, no exceptions!
- 2 Speaker series tomorrow: Sarah Bouchat.
- 3 Fill out class evals in lab Thursday (please!)
- 4 Next week, May 1:

# Housekeeping

- 1 Homework coming in on [May 1](#) (to Leslie's mailbox, by 5pm). Note the rules about academic dishonesty: do not copy anyone's work (including code) and do not allow them to copy your work (including code). Final papers come in to Leslie's mailbox on or before 5pm on [May 11](#). No extensions, no exceptions!
- 2 Speaker series tomorrow: Sarah Bouchat.
- 3 Fill out class evals in lab Thursday (please!)
- 4 Next week, May 1: Work on your papers.

# Housekeeping

- 1 Homework coming in on [May 1](#) (to Leslie's mailbox, by 5pm). Note the rules about academic dishonesty: do not copy anyone's work (including code) and do not allow them to copy your work (including code). Final papers come in to Leslie's mailbox on or before 5pm on [May 11](#). No extensions, no exceptions!
- 2 Speaker series tomorrow: Sarah Bouchat.
- 3 Fill out class evals in lab Thursday (please!)
- 4 Next week, May 1: Work on your papers.



# Where Are We?

# Where Are We?



# Where Are We?



We've covered the main ideas of text analysis:  
representing text,

# Where Are We?



We've covered the main ideas of text analysis:  
representing text, supervised

# Where Are We?



We've covered the main ideas of text analysis:  
representing text, **supervised** and **unsupervised**  
learning.

# Where Are We?



We've covered the main ideas of text analysis: representing text, **supervised** and **unsupervised** learning.

**Now** look at some 'special topics' on

# Where Are We?



We've covered the main ideas of text analysis: representing text, **supervised** and **unsupervised** learning.

Now look at some 'special topics' on **debate**,

# Where Are We?



We've covered the main ideas of text analysis: representing text, **supervised** and **unsupervised** learning.

Now look at some 'special topics' on **debate**, **community** behavior,



# Where Are We?



We've covered the main ideas of text analysis: representing text, **supervised** and **unsupervised** learning.

Now look at some 'special topics' on **debate**, **community** behavior, **bursts** in streams,

# Where Are We?

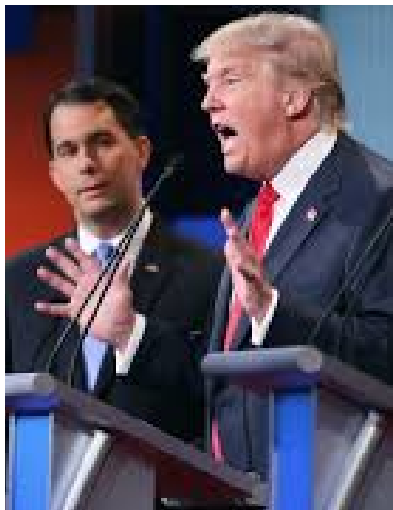


We've covered the main ideas of text analysis: representing text, **supervised** and **unsupervised** learning.

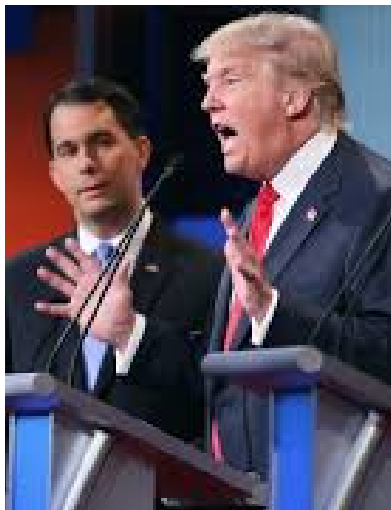
Now look at some 'special topics' on **debate**, **community** behavior, **bursts** in streams, **memes** and spreading of stories/information.

# Modeling Debate

# Modeling Debate

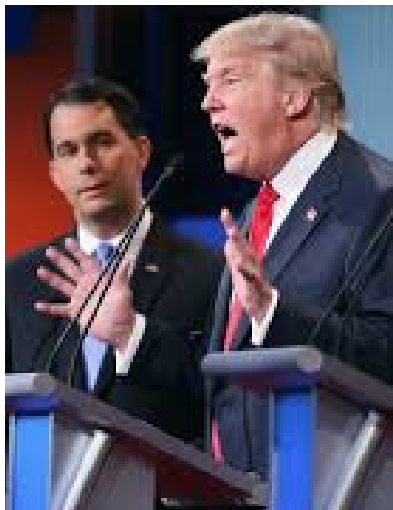


# Modeling Debate



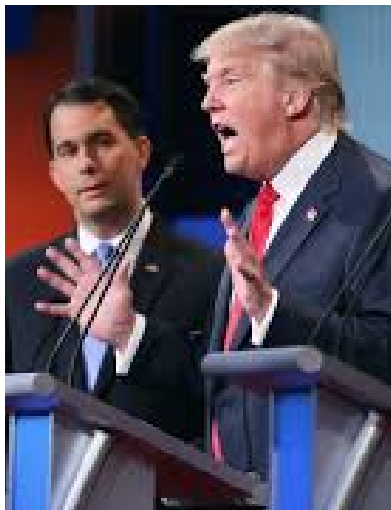
Most politics take place via *debate*

# Modeling Debate



Most politics take place via *debate*  
i.e. some kind of 'back and forth' between  
actors with particular *roles* and *views*.

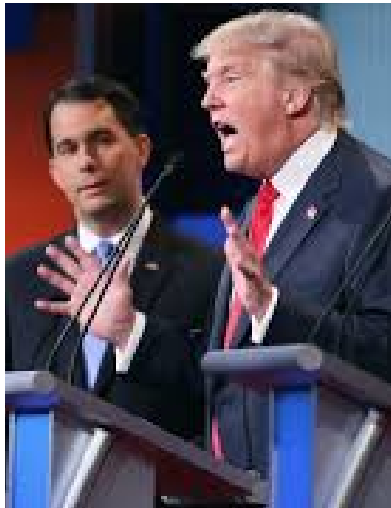
# Modeling Debate



Most politics take place via *debate*  
i.e. some kind of 'back and forth' between  
actors with particular *roles* and *views*.

This is especially common in  
*institutions* like legislatures and courts

# Modeling Debate

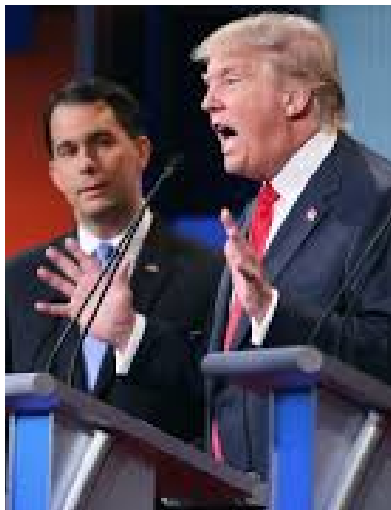


Most politics take place via *debate*  
i.e. some kind of 'back and forth' between  
actors with particular *roles* and *views*.

This is especially common in  
*institutions* like legislatures and courts  
but we also see it in 'new media' like  
twitter.



# Modeling Debate



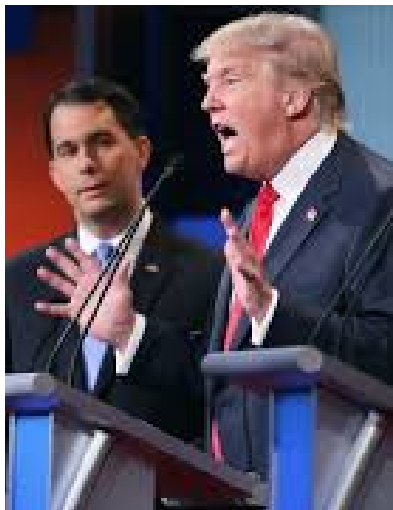
Most politics take place via *debate*

i.e. some kind of 'back and forth' between actors with particular *roles* and *views*.

This is especially common in *institutions* like legislatures and courts but we also see it in 'new media' like twitter.

e.g. Westminster/Parliamentary systems:

# Modeling Debate



Most politics take place via *debate*  
i.e. some kind of 'back and forth' between  
actors with particular *roles* and *views*.

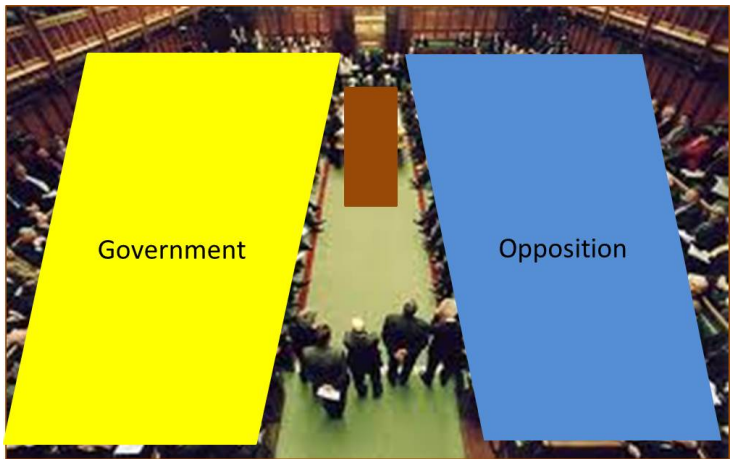
This is especially common in  
*institutions* like legislatures and courts  
but we also see it in 'new media' like  
twitter.

e.g. Westminster/Parliamentary systems:  
*government-vs-opposition dynamic* in  
which most debate takes place  
between one party vs the other(s).

# Modern Arrangement



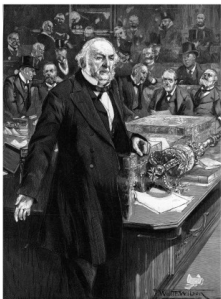
# Modern Arrangement



# British Political Development

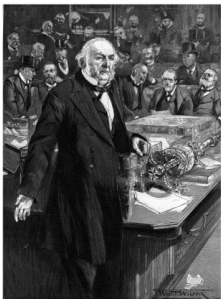
# British Political Development

Theory is that between 1880–1902,

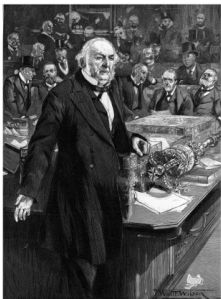


# British Political Development

Theory is that between 1880–1902, ministers became more 'responsive' to (non-cabinet) MPs.



# British Political Development

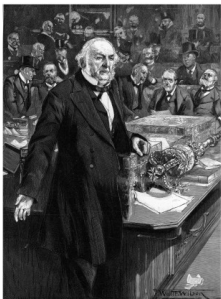


Theory is that between 1880–1902, ministers became more ‘responsive’ to (non-cabinet) MPs.

Increase in ‘responsiveness’ disproportionately accrued to opposition,



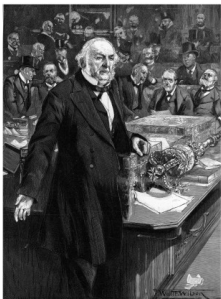
# British Political Development



Theory is that between 1880–1902, ministers became more ‘responsive’ to (non-cabinet) MPs.

Increase in ‘responsiveness’ disproportionately accrued to **opposition**, relative to government backbenchers,

# British Political Development

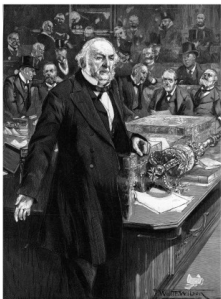


Theory is that between 1880–1902, ministers became more ‘responsive’ to (non-cabinet) MPs.

Increase in ‘responsiveness’ disproportionately accrued to **opposition**, relative to government backbenchers,

Increase is ‘**once-and-for-all**’, not response to special circumstances

# British Political Development



Theory is that between 1880–1902, ministers became more ‘responsive’ to (non-cabinet) MPs.

Increase in ‘responsiveness’ disproportionately accrued to **opposition**, relative to government backbenchers,

Increase is ‘**once-and-for-all**’, not response to special circumstances

→ need to measure **responsiveness**

# Three Types of Actors

# Three Types of Actors



# Three Types of Actors

$G_B$  government backbencher

# Three Types of Actors

$G_B$  government backbencher

$G_M$

# Three Types of Actors

$G_B$  government backbencher

$G_M$  government minister



# Three Types of Actors

$G_B$  government backbencher

$G_M$  government minister

$O$

# Three Types of Actors

$G_B$  government backbencher

$G_M$  government minister

$O$  opposition member

# Three Types of Actors



$G_B$  government backbencher

$G_M$  government minister

$O$  opposition member

# Three Types of Actors

$G_B$  government backbencher

$G_M$  government minister

$O$  opposition member



# Three Types of Actors

$G_B$  government backbencher

$G_M$  government minister

$O$  opposition member



# What we have

# What we have

Every speech made in parliament,  
1803–today

# What we have

Every speech made in parliament,  
1803–today

Focus on period between 1832  
(1st RA) and 1915 (4th RA)



# What we have

Every speech made in parliament,  
1803–today

Focus on period between 1832  
(1st RA) and 1915 (4th RA)

1M+ digitized speech records:

# What we have

Every speech made in parliament,  
1803–today

Focus on period between 1832  
(1st RA) and 1915 (4th RA)

1M+ digitized speech records:  
organized within 'debates'

# What we have

Every speech made in parliament,  
1803–today

Focus on period between 1832  
(1st RA) and 1915 (4th RA)

1M+ digitized speech records:  
organized within ‘debates’

Problem: **party** not recorded  
(needed for Gov/Oppn  
assignment),

# What we have

Every speech made in parliament,  
1803–today

Focus on period between 1832  
(1st RA) and 1915 (4th RA)

1M+ digitized speech records:  
organized within ‘debates’

Problem: party not recorded  
(needed for Gov/Oppn  
assignment), lots of  
missingness/errors in terms of MP  
ids

```
<p id="S3V0094P0-00140">
<member>SIR GEORGE GREY</member>
<membercontribution>
, in reply, stated that he had not as yet received any reply from the coroner of the
district, to whom, as well as to the magistrates, he had written; neither had he
received any communication from the magistrates tending to confirm the charges
made against the owners of the colliery. He had, in consequence of the statement
which had been made by the hon. Member for Finsbury respecting the accident,
addressed a communication to the magistrates and coroner of the district, offering
any assistance which could be given by the Home Office to forward the inquiry;
and he had directed the magistrates to inquire rigidly into the means adopted for
saving the lives of the persons who had been left in the pit, and to investigate the
substance of the charges made against the proprietors of the colliery. He had just
received a letter, dated the 6th of July, from the magistrates, in which they stated,
that in consequence of the letter from the Home Office, they had directed their
clerk to call a meeting of the magistrates, and that they had heard the statements
of several parties upon the subjects alluded to in the communication. The result of
the inquiry was, that they had come to an unanimous opinion as to cause of the
accident. As that question, however, was still under the consideration of the
coroner's inquest, he (Sir G. Grey) did not think it would be right for him to state
the nature of their opinion until the verdict of the coroner's jury should have been
ascertained. As to the question of the subsequent conduct of the owners of the
colliery in preventing persons from descending into the pit to rescue those who
might
<image src="S3V0094P0I0035" />
<col>49</col>
have been left alive in it, the magistrates were convinced that no man left in the pit
after the explosion could have been alive, and that every exertion that could have
been made was made to get them out. That letter was signed by five magistrates.
As he had before stated, he had received no letter from the coroner, whose
investigation was still proceeding; but he would observe, that the gentleman who
had been alluded to by the hon. Member for Finsbury had had every opportunity
during the inquest of examining and cross-examining any witnesses he chose.
</membercontribution>
</p>
<p id="S3V0094P0-00141">
<member>MR. DUNCOMBE</member>
<membercontribution>expressed his astonishment at the hon. Member for Berwick
denying the grounds for the statement which he had made. He had informed
Gentlemen who was his authority. The man himself had been in London, and might
have been examined in the lobby of the House by the hon. Member, had he chosen
to satisfy himself upon the subject. And now he (Mr. Duncombe) was prepared to
support the statement he had made. If the masters could have contradicted those
statements, they had had opportunities of going before the coroner, whose inquiry
had been adjourned from Thursday last to that very day. But he would state what
one of the owners, Mr. Robert Lankester, had himself stated. Mr. Robert Lankester
said the men were bricked up and could not escape.</membercontribution>
</p>
</section>
</section>
```







Use speech information—



Use speech information—‘to and fro’

Use speech information—‘to and fro’— to measure how responsive front bench is to legislature

Use speech information—‘to and fro’— to measure how responsive front bench is to legislature



Use speech information—‘to and fro’— to measure how responsive front bench is to legislature



Use speech information—‘to and fro’— to measure how responsive front bench is to legislature



Use speech information—‘to and fro’— to measure how responsive front bench is to legislature



Use speech information—‘to and fro’— to measure how responsive front bench is to legislature



# Debate

Use speech information—‘to and fro’— to measure how responsive front bench is to legislature





# Debate

Use speech information—‘to and fro’— to measure how responsive front bench is to legislature



# Debate

Use speech information—‘to and fro’— to measure how responsive front bench is to legislature



$O_{15}$

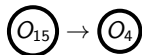
# Debate

Use speech information—‘to and fro’— to measure how responsive front bench is to legislature



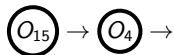
# Debate

Use speech information—‘to and fro’— to measure how responsive front bench is to legislature



# Debate

Use speech information—‘to and fro’— to measure how responsive front bench is to legislature



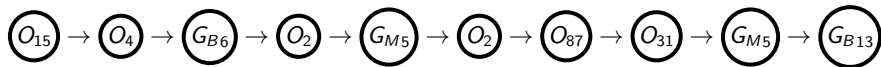
# Debate

Use speech information—‘to and fro’— to measure how responsive front bench is to legislature



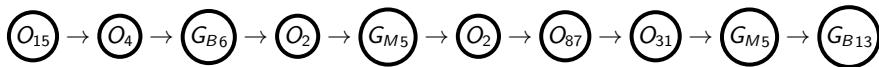
# Debate

Use speech information—‘to and fro’— to measure how responsive front bench is to legislature



# Debate

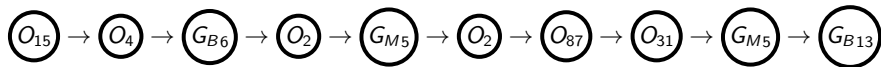
Use speech information—‘to and fro’— to measure how responsive front bench is to legislature





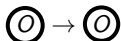
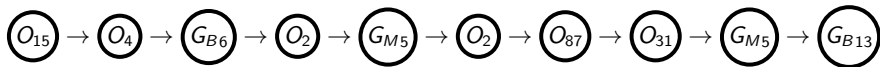
# Debate

Use speech information—‘to and fro’— to measure how responsive front bench is to legislature



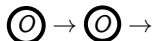
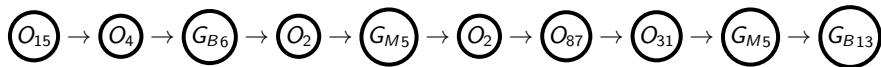
# Debate

Use speech information—‘to and fro’— to measure how responsive front bench is to legislature



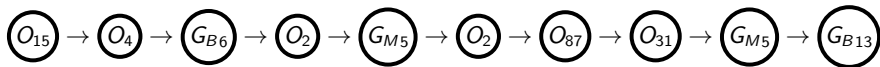
# Debate

Use speech information—‘to and fro’— to measure how responsive front bench is to legislature



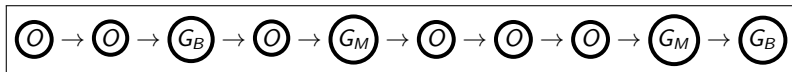
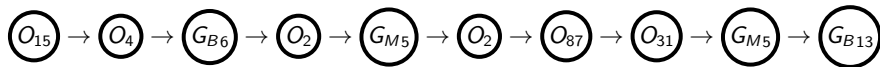
# Debate

Use speech information—‘to and fro’— to measure how responsive front bench is to legislature

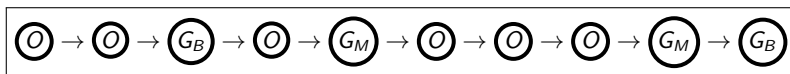


# Debate

Use speech information—‘to and fro’— to measure how responsive front bench is to legislature



Use speech information—‘to and fro’— to measure how responsive front bench is to legislature



If...

If...

For given debate...



# If...

For given debate...

Three **states**

# If...

For given debate...

Three **states**: ministers, government backbenchers, opposition

# If...

For given debate...

Three **states**: ministers, government backbenchers, opposition

Sequence of speaker (identities) has **Markov property**: identity at  $t + 1$  depends only on speaker at  $t$ .

# If...

For given debate...

Three **states**: ministers, government backbenchers, opposition

Sequence of speaker (identities) has **Markov property**: identity at  $t + 1$  depends only on speaker at  $t$ . That is,

# If...

For given debate...

Three **states**: ministers, government backbenchers, opposition

Sequence of speaker (identities) has **Markov property**: identity at  $t + 1$  depends only on speaker at  $t$ . That is, **conditional independence** holds.

# If...

For given debate...

Three **states**: ministers, government backbenchers, opposition

Sequence of speaker (identities) has **Markov property**: identity at  $t + 1$  depends only on speaker at  $t$ . That is, **conditional independence** holds.

→ speaker responds to man immediately proceeding him.

# If...

For given debate...

Three **states**: ministers, government backbenchers, opposition

Sequence of speaker (identities) has **Markov property**: identity at  $t + 1$  depends only on speaker at  $t$ . That is, **conditional independence** holds.

→ speaker responds to man immediately proceeding him. Surely true of interrogatives.

# If...

For given debate...

Three **states**: ministers, government backbenchers, opposition

Sequence of speaker (identities) has **Markov property**: identity at  $t + 1$  depends only on speaker at  $t$ . That is, **conditional independence** holds.

→ speaker responds to man immediately proceeding him. Surely true of interrogatives.

Resulting **Markov chain** is **discrete**:



# If...

For given debate...

Three **states**: ministers, government backbenchers, opposition

Sequence of speaker (identities) has **Markov property**: identity at  $t + 1$  depends only on speaker at  $t$ . That is, **conditional independence** holds.

→ speaker responds to man immediately proceeding him. Surely true of interrogatives.

Resulting **Markov chain** is **discrete**: i.e. only relevant information is identity of speaker (not 'how long' we have been in particular state)

# If...

For given debate...

Three **states**: ministers, government backbenchers, opposition

Sequence of speaker (identities) has **Markov property**: identity at  $t + 1$  depends only on speaker at  $t$ . That is, **conditional independence** holds.

→ speaker responds to man immediately proceeding him. Surely true of interrogatives.

Resulting **Markov chain** is **discrete**: i.e. only relevant information is identity of speaker (not 'how long' we have been in particular state)

Resulting **Markov chain** is **(time) homogenous**:

# If...

For given debate...

Three **states**: ministers, government backbenchers, opposition

Sequence of speaker (identities) has **Markov property**: identity at  $t + 1$  depends only on speaker at  $t$ . That is, **conditional independence** holds.

→ speaker responds to man immediately proceeding him. Surely true of interrogatives.

Resulting **Markov chain** is **discrete**: i.e. only relevant information is identity of speaker (not 'how long' we have been in particular state)

Resulting **Markov chain** is **(time) homogenous**: i.e. probability of moving from state  $i$  to  $j$  does not depend on  $t$ .

# Then...

# Then...

We can characterize chain with the set of **transition probabilities**

# Then...

We can characterize chain with the set of **transition probabilities**  
In particular:

Then...

We can characterize chain with the set of **transition probabilities**

In particular:

$$\begin{array}{c} G_M \\ G_B \\ O \end{array} \begin{array}{ccc} G_M & G_B & O \\ \left( \begin{array}{ccc} m_{MM} & m_{MB} & m_{MO} \\ \textcolor{red}{m_{BM}} & m_{BB} & m_{BO} \\ \textcolor{red}{m_{OM}} & m_{OB} & m_{OO} \end{array} \right) \end{array}$$

Then...

We can characterize chain with the set of **transition probabilities**

In particular:

$$\begin{array}{c} G_M \\ G_B \\ O \end{array} \begin{pmatrix} G_M & G_B & O \\ m_{MM} & m_{MB} & m_{MO} \\ \textcolor{red}{m_{BM}} & m_{BB} & m_{BO} \\ \textcolor{red}{m_{OM}} & m_{OB} & m_{OO} \end{pmatrix}$$

where  $m_{ij}$  is probability of a move from speaker of identity  $i$  to speaker of identity  $j$



# Partner Exercise

## Partner Exercise

- ① Suppose that the weather {good, bad} evolves according to a Markov chain. In particular:

$$\begin{array}{cc} & \begin{array}{cc} G & B \end{array} \\ \begin{array}{c} G \\ B \end{array} & \left( \begin{array}{cc} 0.8 & 0.2 \\ 0.6 & 0.4 \end{array} \right).$$

## Partner Exercise

- ① Suppose that the weather {good, bad} evolves according to a Markov chain. In particular:

$$\begin{array}{cc} & \begin{matrix} G & B \end{matrix} \\ \begin{matrix} G \\ B \end{matrix} & \begin{pmatrix} 0.8 & 0.2 \\ 0.6 & 0.4 \end{pmatrix} \end{array}.$$

If yesterday was a 'good' day, what is the probability that today is a 'bad' day?

## Partner Exercise

- ① Suppose that the weather {good, bad} evolves according to a Markov chain. In particular:

$$\begin{array}{cc} & \begin{matrix} G & B \end{matrix} \\ \begin{matrix} G \\ B \end{matrix} & \begin{pmatrix} 0.8 & 0.2 \\ 0.6 & 0.4 \end{pmatrix} \end{array}.$$

If yesterday was a 'good' day, what is the probability that today is a 'bad' day? How does your prediction for today change if you learn that the day before yesterday was a 'good' day?

## Partner Exercise

- 1 Suppose that the weather {good, bad} evolves according to a Markov chain. In particular:

$$\begin{array}{cc} & \begin{matrix} G & B \end{matrix} \\ \begin{matrix} G \\ B \end{matrix} & \begin{pmatrix} 0.8 & 0.2 \\ 0.6 & 0.4 \end{pmatrix} \end{array}.$$

If yesterday was a 'good' day, what is the probability that today is a 'bad' day? How does your prediction for today change if you learn that the day before yesterday was a 'good' day?

- 2 If today is a 'good' day, what is the probability that the weather two days from now will be good?

## Partner Exercise

- 1 Suppose that the weather {good, bad} evolves according to a Markov chain. In particular:

$$\begin{array}{cc} & \begin{matrix} G & B \end{matrix} \\ \begin{matrix} G \\ B \end{matrix} & \begin{pmatrix} 0.8 & 0.2 \\ 0.6 & 0.4 \end{pmatrix} \end{array}.$$

If yesterday was a 'good' day, what is the probability that today is a 'bad' day? How does your prediction for today change if you learn that the day before yesterday was a 'good' day?

- 2 If today is a 'good' day, what is the probability that the weather two days from now will be good?

# Task

# Task

Must estimate  $m_{ij}$



# Task

Must **estimate**  $m_{ij}$

MLE is well known, and available,

# Task

Must **estimate**  $m_{ij}$

MLE is well known, and available, but we want to include debate and ministry **random effects**.

# Task

Must **estimate**  $m_{ij}$

MLE is well known, and available, but we want to include debate and ministry **random effects**.

So use Markov Chain Monte Carlo approach to estimate **Generalized Linear Mixed Model**.

→ set up as **multinomial logit** with random effects

# Task

Must estimate  $m_{ij}$

MLE is well known, and available, but we want to include debate and ministry random effects.

So use Markov Chain Monte Carlo approach to estimate Generalized Linear Mixed Model.

→ set up as multinomial logit with random effects

$$\exp(\mathbf{X}\beta + \mathbf{Z}\mathbf{u} + \mathbf{e})$$

# Task

Must estimate  $m_{ij}$

MLE is well known, and available, but we want to include debate and ministry random effects.

So use Markov Chain Monte Carlo approach to estimate Generalized Linear Mixed Model.

→ set up as multinomial logit with random effects

$$\exp(\mathbf{X}\beta + \mathbf{Z}\mathbf{u} + \mathbf{e})$$

→ Predicted probabilities are then (estimates of) transition probabilities

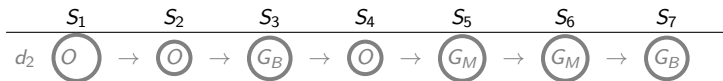
# Rearranging Data

# Rearranging Data

$s_1$        $s_2$        $s_3$        $s_4$        $s_5$        $s_6$        $s_7$

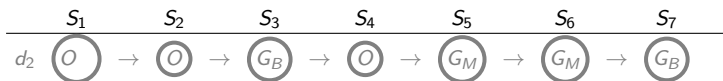
---

# Rearranging Data

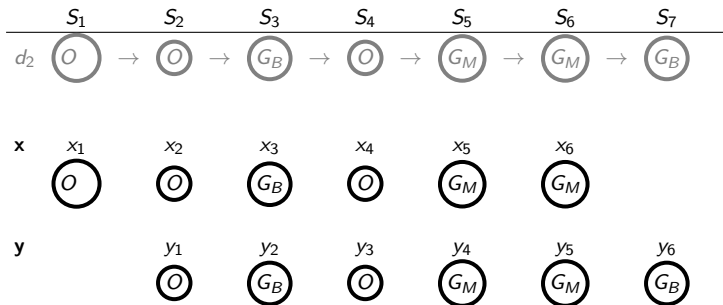




# Rearranging Data

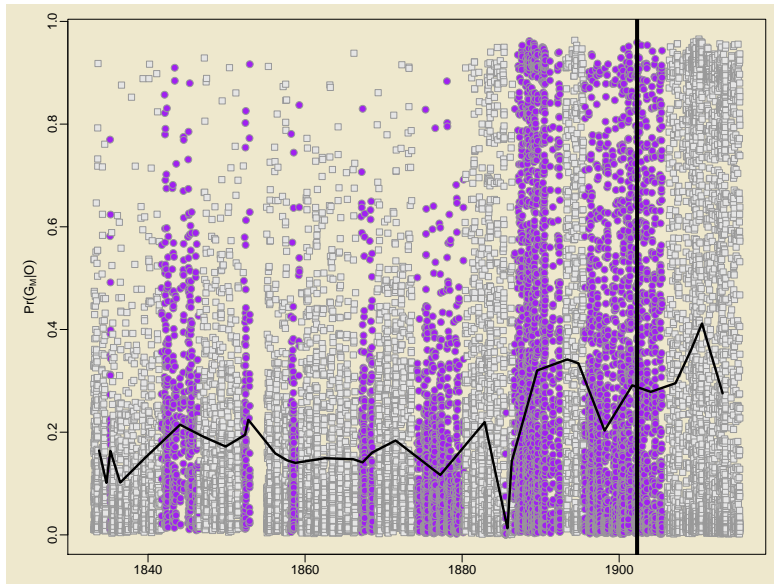


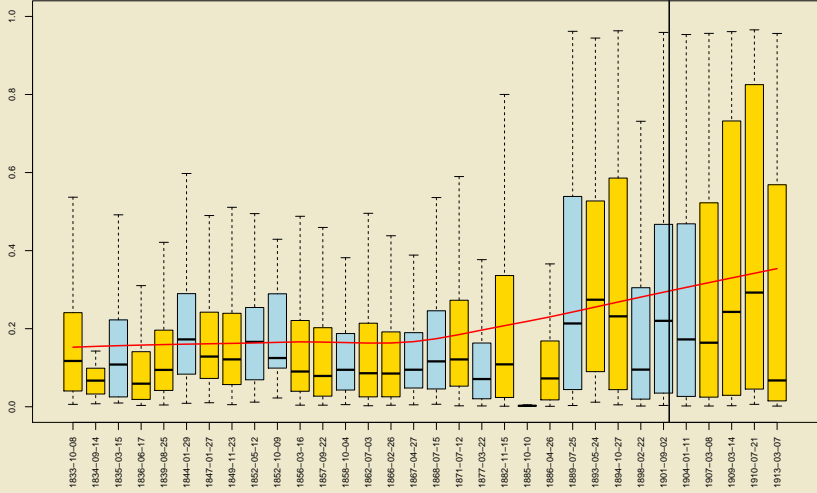
# Rearranging Data



# Results

# Results





# Linguistic Change in Online Communities

# Linguistic Change in Online Communities

Danescu-Niculescu-Mizil et al (2013) “No Country for Old Members”  
study linguistic innovation and adoption in RateBeer and  
BeerAdvocate

# Linguistic Change in Online Communities

Danescu-Niculescu-Mizil et al (2013) “No Country for Old Members”  
study linguistic innovation and adoption in RateBeer and  
BeerAdvocate

Look at users as they ‘age’ (from first post to last post):



# Linguistic Change in Online Communities

Danescu-Niculescu-Mizil et al (2013) “No Country for Old Members” study linguistic innovation and adoption in RateBeer and BeerAdvocate

Look at users as they ‘age’ (from first post to last post):

Well, early on, users adopt norms of community—until ‘**linguistic adolescence**’—after which they cease to respond to community changes.

# Linguistic Change in Online Communities

Danescu-Niculescu-Mizil et al (2013) “No Country for Old Members” study linguistic innovation and adoption in RateBeer and BeerAdvocate

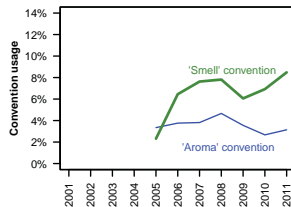
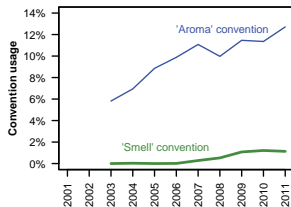
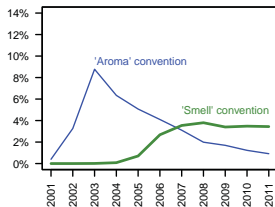
Look at users as they ‘age’ (from first post to last post):

Well, early on, users adopt norms of community—until ‘**linguistic adolescence**’—after which they cease to respond to community changes.

and results allow them to predict how long user will stay in community from early posts!

# Aroma v Smell: overall, 2003 joiners, 2007 joiners

# Aroma v Smell: overall, 2003 joiners, 2007 joiners



# Intuition

# Intuition

‘Snapshot language model’:

# Intuition

'Snapshot language model': use **cross-entropy** to compare language of a user's (held out) posts to see how different it was to rest of community at the time

# Intuition

'Snapshot language model': use **cross-entropy** to compare language of a user's (held out) posts to see how different it was to rest of community at the time—in terms of unusual **bigrams** it contained.



# Intuition

‘Snapshot language model’: use **cross-entropy** to compare language of a user’s (held out) posts to see how different it was to rest of community at the time—in terms of unusual **bigrams** it contained.

When cross entropy is **high**,

# Intuition

'Snapshot language model': use **cross-entropy** to compare language of a user's (held out) posts to see how different it was to rest of community at the time—in terms of unusual **bigrams** it contained.

When cross entropy is **high**, user is using different terms from community.

# Intuition

'Snapshot language model': use **cross-entropy** to compare language of a user's (held out) posts to see how different it was to rest of community at the time—in terms of unusual **bigrams** it contained.

When cross entropy is **high**, user is using different terms from community.

Specifically, find that members stop using **first person pronouns** and start using **beer specific vocab**.

# Intuition

'Snapshot language model': use **cross-entropy** to compare language of a user's (held out) posts to see how different it was to rest of community at the time—in terms of unusual **bigrams** it contained.

When cross entropy is **high**, user is using different terms from community.

Specifically, find that members stop using **first person pronouns** and start using **beer specific vocab**.

**and** all users die 'old' in sense that they stop adopting linguistic innovation proportionally (up to one third of lifespan) to how long they are in the community.

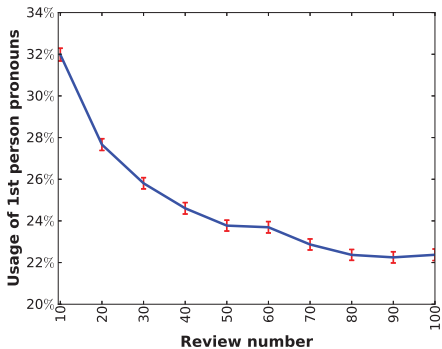
# Intuition

'Snapshot language model': use **cross-entropy** to compare language of a user's (held out) posts to see how different it was to rest of community at the time—in terms of unusual **bigrams** it contained.

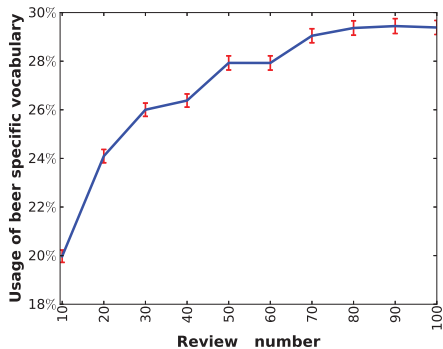
When cross entropy is **high**, user is using different terms from community.

Specifically, find that members stop using **first person pronouns** and start using **beer specific vocab**.

**and** all users die 'old' in sense that they stop adopting linguistic innovation proportionally (up to one third of lifespan) to how long they are in the community. Implies that '**adult language stability assumption**' is *relative* rather than absolute online.



(a) First person sing. pronouns



(b) Beer specific vocabulary

# Partner Exercise

## Partner Exercise

According to psychologists like Pennebaker, the use of first person singular pronouns ('I', 'me', 'my') is more common in some groups than others.



## Partner Exercise

According to psychologists like Pennebaker, the use of first person singular pronouns ('I', 'me', 'my') is more common in some groups than others. Who do you think uses them more (and why?) in blogs, essays etc:

# Partner Exercise

According to psychologists like Pennebaker, the use of first person singular pronouns ('I', 'me', 'my') is more common in some groups than others. Who do you think uses them more (and why?) in blogs, essays etc:

- ① men or women?
- ② people who are happy or people who are depressed?
- ③ extraverts or introverts?

# Burstiness

[Kleinberg](#) (2002) “Bursty and Hierarchical Structure in Streams” develops methods based on Markov models to detect interesting structure in [streams](#) of documents (such as email) arriving over time.

# Burstiness

[Kleinberg](#) (2002) “Bursty and Hierarchical Structure in Streams” develops methods based on Markov models to detect interesting structure in [streams](#) of documents (such as email) arriving over time.

Methods allow one to see which words and phrases have dramatic shifts in usage over time

[Kleinberg](#) (2002) “Bursty and Hierarchical Structure in Streams” develops methods based on Markov models to detect interesting structure in [streams](#) of documents (such as email) arriving over time.

Methods allow one to see which words and phrases have dramatic shifts in usage over time—oftentimes this corresponds to [substance](#) but sometimes [style](#).

[Kleinberg](#) (2002) “Bursty and Hierarchical Structure in Streams” develops methods based on Markov models to detect interesting structure in [streams](#) of documents (such as email) arriving over time.

Methods allow one to see which words and phrases have dramatic shifts in usage over time—oftentimes this corresponds to [substance](#) but sometimes [style](#).

Idea is to model [arrival times](#) of words.

# Burstiness

[Kleinberg](#) (2002) “Bursty and Hierarchical Structure in Streams” develops methods based on Markov models to detect interesting structure in [streams](#) of documents (such as email) arriving over time.

Methods allow one to see which words and phrases have dramatic shifts in usage over time—oftentimes this corresponds to [substance](#) but sometimes [style](#).

Idea is to model [arrival times](#) of words. As ‘gaps’ between use of same word become shorter,



# Burstiness

Kleinberg (2002) “Bursty and Hierarchical Structure in Streams” develops methods based on Markov models to detect interesting structure in [streams](#) of documents (such as email) arriving over time.

Methods allow one to see which words and phrases have dramatic shifts in usage over time—oftentimes this corresponds to [substance](#) but sometimes [style](#).

Idea is to model [arrival times](#) of words. As ‘gaps’ between use of same word become shorter, infer that term is ‘[bursty](#)’.

# Burstiness

[Kleinberg](#) (2002) “Bursty and Hierarchical Structure in Streams” develops methods based on Markov models to detect interesting structure in [streams](#) of documents (such as email) arriving over time.

Methods allow one to see which words and phrases have dramatic shifts in usage over time—oftentimes this corresponds to [substance](#) but sometimes [style](#).

Idea is to model [arrival times](#) of words. As ‘gaps’ between use of same word become shorter, infer that term is ‘[bursty](#)’.

Surges must be [long](#),

# Burstiness

Kleinberg (2002) “Bursty and Hierarchical Structure in Streams” develops methods based on Markov models to detect interesting structure in [streams](#) of documents (such as email) arriving over time.

Methods allow one to see which words and phrases have dramatic shifts in usage over time—oftentimes this corresponds to [substance](#) but sometimes [style](#).

Idea is to model [arrival times](#) of words. As ‘gaps’ between use of same word become shorter, infer that term is ‘[bursty](#)’.

Surges must be [long](#), and/or [intense](#),

# Burstiness

Kleinberg (2002) “Bursty and Hierarchical Structure in Streams” develops methods based on Markov models to detect interesting structure in [streams](#) of documents (such as email) arriving over time.

Methods allow one to see which words and phrases have dramatic shifts in usage over time—oftentimes this corresponds to [substance](#) but sometimes [style](#).

Idea is to model [arrival times](#) of words. As ‘gaps’ between use of same word become shorter, infer that term is ‘[bursty](#)’.

Surges must be [long](#), and/or [intense](#), depending on specification of model.

# Burstiness

Kleinberg (2002) “Bursty and Hierarchical Structure in Streams” develops methods based on Markov models to detect interesting structure in [streams](#) of documents (such as email) arriving over time.

Methods allow one to see which words and phrases have dramatic shifts in usage over time—oftentimes this corresponds to [substance](#) but sometimes [style](#).

Idea is to model [arrival times](#) of words. As ‘gaps’ between use of same word become shorter, infer that term is ‘[bursty](#)’.

Surges must be [long](#), and/or [intense](#), depending on specification of model.

# Applying to SOTU, 1790–2002

# Applying to SOTU, 1790–2002



# Applying to SOTU, 1790–2002



word	burst
gentlemen	1790–1800
british	1809–1814
slaves	1859–1863
japanese	1942–1945
health	1992–1994
help	1998–



# Burstiness in Parliament

# Burstiness in Parliament

Use `burstiness` to model MPs as  
`agenda-setters`

# Burstiness in Parliament

Use `burstiness` to model MPs as  
`agenda-setters`

# Burstiness in Parliament

Use **burstiness** to model MPs as  
**agenda-setters**

To ask the Secretary  
of State for the Home  
Department whether, having  
regard to the fact that  
the women suffragettes now  
imprisoned in Holloway Gaol  
are political rather than  
criminal offenders....

(Keir Hardie, Oct 31, 1906)

# Burstiness in Parliament

Use **burstiness** to model MPs as  
**agenda-setters**

Look specifically at changes to  
composition of opposition in terms of  
**concentration** of agenda-power.

To ask the Secretary  
of State for the Home  
Department whether, having  
regard to the fact that  
the women **suffragettes** now  
imprisoned in Holloway Gaol  
are political rather than  
criminal offenders....

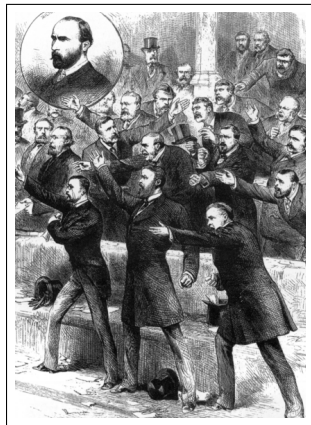
(Keir Hardie, Oct 31, 1906)

# Burstiness in Parliament

Use **burstiness** to model MPs as **agenda-setters**

Look specifically at changes to composition of opposition in terms of **concentration** of agenda-power.

→ timing of emergence of small(er) group of opposition agenda-setters is *prima facie* evidence of **shadow cabinet** (which is otherwise impossible to document)

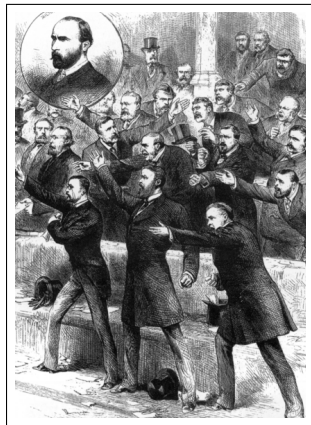


# Burstiness in Parliament

Use **burstiness** to model MPs as **agenda-setters**

Look specifically at changes to composition of opposition in terms of **concentration** of agenda-power.

→ timing of emergence of small(er) group of opposition agenda-setters is *prima facie* evidence of **shadow cabinet** (which is otherwise impossible to document)

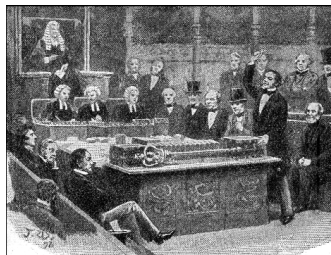


# Burstiness in Parliament

Use **burstiness** to model MPs as **agenda-setters**

Look specifically at changes to composition of opposition in terms of **concentration** of agenda-power.

- timing of emergence of small(er) group of opposition agenda-setters is *prima facie* evidence of **shadow cabinet** (which is otherwise impossible to document)





Idea...

# Idea...

Now, gaps are number of speeches between use of terms (*within* a speech, multiple uses count as one occurrence).

# Idea...

Now, gaps are number of speeches between use of terms (*within* a speech, multiple uses count as one occurrence). Levels of burstiness are **hierarchical**: burstiness of 3 is nested within burstiness of 2, is nested within burstiness of 1.

# Idea...

Now, gaps are number of speeches between use of terms (*within* a speech, multiple uses count as one occurrence). Levels of burstiness are **hierarchical**: burstiness of 3 is nested within burstiness of 2, is nested within burstiness of 1.

Innovation:

# Idea...

Now, gaps are number of speeches between use of terms (*within* a speech, multiple uses count as one occurrence). Levels of burstiness are **hierarchical**: burstiness of 3 is nested within burstiness of 2, is nested within burstiness of 1.

Innovation: score MPs as product of bursts they **started**:

# Idea...

Now, gaps are number of speeches between use of terms (*within* a speech, multiple uses count as one occurrence). Levels of burstiness are **hierarchical**: burstiness of 3 is nested within burstiness of 2, is nested within burstiness of 1.

Innovation: score MPs as product of bursts they **started**:

e.g. if MP made speech using term  $w_1$  that began burst of level 3 for 5 periods,

# Idea...

Now, gaps are number of speeches between use of terms (*within* a speech, multiple uses count as one occurrence). Levels of burstiness are **hierarchical**: burstiness of 3 is nested within burstiness of 2, is nested within burstiness of 1.

Innovation: score MPs as product of bursts they **started**:

e.g. if MP made speech using term  $w_1$  that began burst of level 3 for 5 periods, and speech using term  $w_2$  that began burst of level 2 for 7 periods,

# Idea...

Now, gaps are number of speeches between use of terms (*within* a speech, multiple uses count as one occurrence). Levels of burstiness are **hierarchical**: burstiness of 3 is nested within burstiness of 2, is nested within burstiness of 1.

Innovation: score MPs as product of bursts they **started**:

e.g. if MP made speech using term  $w_1$  that began burst of level 3 for 5 periods, and speech using term  $w_2$  that began burst of level 2 for 7 periods, his score for session is  $(3 \times 5) + (2 \times 7) = 29$ .



# Idea...

Now, gaps are number of speeches between use of terms (*within* a speech, multiple uses count as one occurrence). Levels of burstiness are **hierarchical**: burstiness of 3 is nested within burstiness of 2, is nested within burstiness of 1.

Innovation: score MPs as product of bursts they **started**:

e.g. if MP made speech using term  $w_1$  that began burst of level 3 for 5 periods, and speech using term  $w_2$  that began burst of level 2 for 7 periods, his score for session is  $(3 \times 5) + (2 \times 7) = 29$ .

Calculate for all words,

# Idea...

Now, gaps are number of speeches between use of terms (*within* a speech, multiple uses count as one occurrence). Levels of burstiness are **hierarchical**: burstiness of 3 is nested within burstiness of 2, is nested within burstiness of 1.

Innovation: score MPs as product of bursts they **started**:

e.g. if MP made speech using term  $w_1$  that began burst of level 3 for 5 periods, and speech using term  $w_2$  that began burst of level 2 for 7 periods, his score for session is  $(3 \times 5) + (2 \times 7) = 29$ .

Calculate for all words, all speeches,

## Idea...

Now, gaps are number of speeches between use of terms (*within* a speech, multiple uses count as one occurrence). Levels of burstiness are **hierarchical**: burstiness of 3 is nested within burstiness of 2, is nested within burstiness of 1.

Innovation: score MPs as product of bursts they **started**:

e.g. if MP made speech using term  $w_1$  that began burst of level 3 for 5 periods, and speech using term  $w_2$  that began burst of level 2 for 7 periods, his score for session is  $(3 \times 5) + (2 \times 7) = 29$ .

Calculate for all words, all speeches, all MPs,

# Idea...

Now, gaps are number of speeches between use of terms (*within* a speech, multiple uses count as one occurrence). Levels of burstiness are **hierarchical**: burstiness of 3 is nested within burstiness of 2, is nested within burstiness of 1.

Innovation: score MPs as product of bursts they **started**:

e.g. if MP made speech using term  $w_1$  that began burst of level 3 for 5 periods, and speech using term  $w_2$  that began burst of level 2 for 7 periods, his score for session is  $(3 \times 5) + (2 \times 7) = 29$ .

Calculate for all words, all speeches, all MPs, all sessions.



denote the gap 'times' as  $x$ .

denote the gap 'times' as  $x$ . Suppose that these gaps are exponentially distributed:

denote the gap 'times' as  $x$ . Suppose that these gaps are exponentially distributed:

$$f(x) = \alpha_i e^{-\alpha_i x}$$



denote the gap 'times' as  $x$ . Suppose that these gaps are exponentially distributed:

$$f(x) = \alpha_i e^{-\alpha_i x}$$

where  $\alpha_i$  is the [rate](#)

denote the gap 'times' as  $x$ . Suppose that these gaps are exponentially distributed:

$$f(x) = \alpha_i e^{-\alpha_i x}$$

where  $\alpha_i$  is the **rate**  
**wrt** model parameters,

denote the gap 'times' as  $x$ . Suppose that these gaps are exponentially distributed:

$$f(x) = \alpha_i e^{-\alpha_i x}$$

where  $\alpha_i$  is the [rate](#)

[wrt](#) model parameters,  $\alpha_i = \frac{n}{T} s^i$ , where  $n$  is the number of occurrences and  $T$  is total time.

## Details

denote the gap 'times' as  $x$ . Suppose that these gaps are exponentially distributed:

$$f(x) = \alpha_i e^{-\alpha_i x}$$

where  $\alpha_i$  is the [rate](#)

[wrt](#) model parameters,  $\alpha_i = \frac{n}{T} s^i$ , where  $n$  is the number of occurrences and  $T$  is total time.

[i.e.](#) when  $\alpha_i$  is large,

denote the gap 'times' as  $x$ . Suppose that these gaps are exponentially distributed:

$$f(x) = \alpha_i e^{-\alpha_i x}$$

where  $\alpha_i$  is the [rate](#)

[wrt](#) model parameters,  $\alpha_i = \frac{n}{T} s^i$ , where  $n$  is the number of occurrences and  $T$  is total time.

[i.e.](#) when  $\alpha_i$  is large, (mean) wait is short (soon see word again).

denote the gap 'times' as  $x$ . Suppose that these gaps are exponentially distributed:

$$f(x) = \alpha_i e^{-\alpha_i x}$$

where  $\alpha_i$  is the **rate**

**wrt** model parameters,  $\alpha_i = \frac{n}{T} s^i$ , where  $n$  is the number of occurrences and  $T$  is total time.

**i.e.** when  $\alpha_i$  is large, (mean) wait is short (soon see word again).

If estimated  $\alpha$  changes up (down),

denote the gap 'times' as  $x$ . Suppose that these gaps are exponentially distributed:

$$f(x) = \alpha_i e^{-\alpha_i x}$$

where  $\alpha_i$  is the **rate**

**wrt** model parameters,  $\alpha_i = \frac{n}{T} s^i$ , where  $n$  is the number of occurrences and  $T$  is total time.

**i.e.** when  $\alpha_i$  is large, (mean) wait is short (soon see word again).

If estimated  $\alpha$  changes up (down), we have evidence that a burst is occurring (ending).

denote the gap 'times' as  $x$ . Suppose that these gaps are exponentially distributed:

$$f(x) = \alpha_i e^{-\alpha_i x}$$

where  $\alpha_i$  is the **rate**

**wrt** model parameters,  $\alpha_i = \frac{n}{T} s^i$ , where  $n$  is the number of occurrences and  $T$  is total time.

**i.e.** when  $\alpha_i$  is large, (mean) wait is short (soon see word again).

If estimated  $\alpha$  changes up (down), we have evidence that a burst is occurring (ending).

In principle,



denote the gap 'times' as  $x$ . Suppose that these gaps are exponentially distributed:

$$f(x) = \alpha_i e^{-\alpha_i x}$$

where  $\alpha_i$  is the [rate](#)

[wrt](#) model parameters,  $\alpha_i = \frac{n}{T} s^i$ , where  $n$  is the number of occurrences and  $T$  is total time.

[i.e.](#) when  $\alpha_i$  is large, (mean) wait is short (soon see word again).

If estimated  $\alpha$  changes up ([down](#)), we have evidence that a burst is occurring ([ending](#)).

In principle,  $s$  could be estimated, but typically set to [2](#).

## Details II

## Details II

Here gaps are number of speeches between uses of a term.

## Details II

Here gaps are number of speeches between uses of a term. Obviously discrete, but approximates something continuous (time).

## Details II

Here gaps are number of speeches between uses of a term. Obviously discrete, but approximates something continuous (time).

The base rate of word usage is  $f_0(x) = \alpha_0 e^{-\alpha_0 x}$

## Details II

Here gaps are number of speeches between uses of a term. Obviously discrete, but approximates something continuous (time).

The base rate of word usage is  $f_0(x) = \alpha_0 e^{-\alpha_0 x}$  where  $\alpha_0 = \frac{n}{T}$ .

## Details II

Here gaps are number of speeches between uses of a term. Obviously discrete, but approximates something continuous (time).

The base rate of word usage is  $f_0(x) = \alpha_0 e^{-\alpha_0 x}$  where  $\alpha_0 = \frac{n}{T}$ .

e.g. there are 100 speeches,

## Details II

Here gaps are number of speeches between uses of a term. Obviously discrete, but approximates something continuous (time).

The base rate of word usage is  $f_0(x) = \alpha_0 e^{-\alpha_0 x}$  where  $\alpha_0 = \frac{n}{T}$ .

e.g. there are 100 speeches, 5 of them mention word,



## Details II

Here gaps are number of speeches between uses of a term. Obviously discrete, but approximates something continuous (time).

The base rate of word usage is  $f_0(x) = \alpha_0 e^{-\alpha_0 x}$  where  $\alpha_0 = \frac{n}{T}$ .

e.g. there are 100 speeches, 5 of them mention word,  $\alpha_0 = \frac{1}{20}$ .

## Details II

Here gaps are number of speeches between uses of a term. Obviously discrete, but approximates something continuous (time).

The base rate of word usage is  $f_0(x) = \alpha_0 e^{-\alpha_0 x}$  where  $\alpha_0 = \frac{n}{T}$ .

e.g. there are 100 speeches, 5 of them mention word,  $\alpha_0 = \frac{1}{20}$ .

For a fixed value of  $s$ , going from base level to  $i=1$  to  $i=2$  requires decreasing mean wait time by factor of  $\frac{1}{s^i}$  each time.

## Details II

Here gaps are number of speeches between uses of a term. Obviously discrete, but approximates something continuous (time).

The base rate of word usage is  $f_0(x) = \alpha_0 e^{-\alpha_0 x}$  where  $\alpha_0 = \frac{n}{T}$ .

e.g. there are 100 speeches, 5 of them mention word,  $\alpha_0 = \frac{1}{20}$ .

For a fixed value of  $s$ , going from base level to  $i=1$  to  $i=2$  requires decreasing mean wait time by factor of  $\frac{1}{s^i}$  each time.

i.e. if  $s = 2$ , arrival rate has to increase (wrt to  $\frac{1}{\alpha_0}$ ):  $1 \rightarrow \frac{1}{2} \rightarrow \frac{1}{4} \rightarrow \frac{1}{8}$

## Details II

Here gaps are number of speeches between uses of a term. Obviously discrete, but approximates something continuous (time).

The base rate of word usage is  $f_0(x) = \alpha_0 e^{-\alpha_0 x}$  where  $\alpha_0 = \frac{n}{T}$ .

e.g. there are 100 speeches, 5 of them mention word,  $\alpha_0 = \frac{1}{20}$ .

For a fixed value of  $s$ , going from base level to  $i=1$  to  $i=2$  requires decreasing mean wait time by factor of  $\frac{1}{s^i}$  each time.

i.e. if  $s = 2$ , arrival rate has to increase (wrt to  $\frac{1}{\alpha_0}$ ):  $1 \rightarrow \frac{1}{2} \rightarrow \frac{1}{4} \rightarrow \frac{1}{8}$

→ This is a geometric process: unless wait time halves (somewhere), we never leave level  $i = 0$ .

## Details II

Here gaps are number of speeches between uses of a term. Obviously discrete, but approximates something continuous (time).

The base rate of word usage is  $f_0(x) = \alpha_0 e^{-\alpha_0 x}$  where  $\alpha_0 = \frac{n}{T}$ .

e.g. there are 100 speeches, 5 of them mention word,  $\alpha_0 = \frac{1}{20}$ .

For a fixed value of  $s$ , going from base level to  $i=1$  to  $i=2$  requires decreasing mean wait time by factor of  $\frac{1}{s^i}$  each time.

i.e. if  $s = 2$ , arrival rate has to increase (wrt to  $\frac{1}{\alpha_0}$ ):  $1 \rightarrow \frac{1}{2} \rightarrow \frac{1}{4} \rightarrow \frac{1}{8}$

→ This is a geometric process: unless wait time halves (somewhere), we never leave level  $i = 0$ .

↔ this is an infinite state hidden Markov model.

# Partner Exercise

# Partner Exercise

- 1 Do we need to **remove stop words** when using calculating burstiness of given tokens? Why or why not?

# Partner Exercise

- 1 Do we need to **remove stop words** when using calculating burstiness of given tokens? Why or why not?
- 2 Should we **stem the words** in the texts?



# Partner Exercise

- 1 Do we need to **remove stop words** when using calculating burstiness of given tokens? Why or why not?
- 2 Should we **stem the words** in the texts?
- 3 How do models of the burstiness of words differ from '**topic** models'? Which would you use to study changing subjects of debate over time? Which would you use to study conceptual change?

# Cost Term

# Cost Term

Relying only on the  $\alpha$  term would tend to produce a 'lot' of state changes

# Cost Term

Relying only on the  $\alpha$  term would tend to produce a 'lot' of state changes

so use a second, `cost` parameter.

# Cost Term

Relying only on the  $\alpha$  term would tend to produce a 'lot' of state changes

so use a second, **cost** parameter. Cost of moving from state  $i$  to state  $j$  is:

## Cost Term

Relying only on the  $\alpha$  term would tend to produce a 'lot' of state changes

so use a second, **cost** parameter. Cost of moving from state  $i$  to state  $j$  is:

$$\tau(i,j) = \begin{cases} (j-i)\gamma \ln n & \text{if } j > i; \\ 0 & \text{if } j < i. \end{cases}$$

where  $n$  is the total number of mentions;

## Cost Term

Relying only on the  $\alpha$  term would tend to produce a 'lot' of state changes

so use a second, **cost** parameter. Cost of moving from state  $i$  to state  $j$  is:

$$\tau(i,j) = \begin{cases} (j-i)\gamma \ln n & \text{if } j > i; \\ 0 & \text{if } j < i. \end{cases}$$

where  $n$  is the total number of mentions;  $\gamma$  is the parameter (set to 1).

# Cost Term

Relying only on the  $\alpha$  term would tend to produce a 'lot' of state changes

- so use a second, **cost** parameter. Cost of moving from state  $i$  to state  $j$  is:

$$\tau(i,j) = \begin{cases} (j-i)\gamma \ln n & \text{if } j > i; \\ 0 & \text{if } j < i. \end{cases}$$

where  $n$  is the total number of mentions;  $\gamma$  is the parameter (set to 1).

- so large changes to state 'upwards' will be costly; changing down is 'free'



# Cost Term

Relying only on the  $\alpha$  term would tend to produce a 'lot' of state changes

so use a second, **cost** parameter. Cost of moving from state  $i$  to state  $j$  is:

$$\tau(i,j) = \begin{cases} (j-i)\gamma \ln n & \text{if } j > i; \\ 0 & \text{if } j < i. \end{cases}$$

where  $n$  is the total number of mentions;  $\gamma$  is the parameter (set to 1).

so large changes to state 'upwards' will be costly; changing down is 'free'  
and finding 'best' **state sequence**  $\mathbf{q} = (q_{i1}, \dots, q_{in})$  involves minimizing  
 $c(\mathbf{q}|\mathbf{x}) = [\sum \tau(i,j)] + [\sum -\ln f_{it}(x_t)]$

# Cost Term

Relying only on the  $\alpha$  term would tend to produce a 'lot' of state changes

so use a second, **cost** parameter. Cost of moving from state  $i$  to state  $j$  is:

$$\tau(i,j) = \begin{cases} (j-i)\gamma \ln n & \text{if } j > i; \\ 0 & \text{if } j < i. \end{cases}$$

where  $n$  is the total number of mentions;  $\gamma$  is the parameter (set to 1).

so large changes to state 'upwards' will be costly; changing down is 'free'  
and finding 'best' **state sequence**  $\mathbf{q} = (q_{i1}, \dots, q_{in})$  involves minimizing  
 $c(\mathbf{q}|\mathbf{x}) = [\sum \tau(i,j)] + [\sum -\ln f_{it}(x_t)]$

# More Intuition

# More Intuition

Consider a word like **the**

# More Intuition

Consider a word like **the**: used in (almost) every speech,

# More Intuition

Consider a word like **the**: used in (almost) every speech, so gaps are **always** small.

# More Intuition

Consider a word like **the**: used in (almost) every speech, so gaps are **always** small. Would be almost impossible to make term bursty (given we have binary use metric).

## More Intuition

Consider a word like **the**: used in (almost) every speech, so gaps are **always** small. Would be almost impossible to make term bursty (given we have binary use metric).

Consider a word like **zither**



## More Intuition

Consider a word like **the**: used in (almost) every speech, so gaps are **always** small. Would be almost impossible to make term bursty (given we have binary use metric).

Consider a word like **zither**: very rarely used,

## More Intuition

Consider a word like **the**: used in (almost) every speech, so gaps are **always** small. Would be almost impossible to make term bursty (given we have binary use metric).

Consider a word like **zither**: very rarely used, so gaps are **always** long.

## More Intuition

Consider a word like **the**: used in (almost) every speech, so gaps are **always** small. Would be almost impossible to make term bursty (given we have binary use metric).

Consider a word like **zither**: very rarely used, so gaps are **always** long. Would be almost impossible to make term bursty (under any reasonable choice of  $s$ ).

# More Intuition

Consider a word like **the**: used in (almost) every speech, so gaps are **always** small. Would be almost impossible to make term bursty (given we have binary use metric).

Consider a word like **zither**: very rarely used, so gaps are **always** long. Would be almost impossible to make term bursty (under any reasonable choice of  $s$ ).

Consider a word like **boundary**

## More Intuition

Consider a word like **the**: used in (almost) every speech, so gaps are **always** small. Would be almost impossible to make term bursty (given we have binary use metric).

Consider a word like **zither**: very rarely used, so gaps are **always** long. Would be almost impossible to make term bursty (under any reasonable choice of  $s$ ).

Consider a word like **boundary**: **intense** bursts of usage in some years when redistribution of seats is discussed.

## More Intuition

Consider a word like **the**: used in (almost) every speech, so gaps are **always** small. Would be almost impossible to make term bursty (given we have binary use metric).

Consider a word like **zither**: very rarely used, so gaps are **always** long. Would be almost impossible to make term bursty (under any reasonable choice of  $s$ ).

Consider a word like **boundary**: **intense** bursts of usage in some years when redistribution of seats is discussed.

**btw** in keeping with original presentation, we treat base level as zero.

# More Intuition

Consider a word like **the**: used in (almost) every speech, so gaps are **always** small. Would be almost impossible to make term bursty (given we have binary use metric).

Consider a word like **zither**: very rarely used, so gaps are **always** long. Would be almost impossible to make term bursty (under any reasonable choice of  $s$ ).

Consider a word like **boundary**: **intense** bursts of usage in some years when redistribution of seats is discussed.

**btw** in keeping with original presentation, we treat base level as zero.

**NB** **no need** to stem or stop words:

## More Intuition

Consider a word like **the**: used in (almost) every speech, so gaps are **always** small. Would be almost impossible to make term bursty (given we have binary use metric).

Consider a word like **zither**: very rarely used, so gaps are **always** long. Would be almost impossible to make term bursty (under any reasonable choice of  $s$ ).

Consider a word like **boundary**: **intense** bursts of usage in some years when redistribution of seats is discussed.

**btw** in keeping with original presentation, we treat base level as zero.

**NB** **no need** to stem or stop words: stops should all be level 0 anyway.



# More Intuition

Consider a word like **the**: used in (almost) every speech, so gaps are **always** small. Would be almost impossible to make term bursty (given we have binary use metric).

Consider a word like **zither**: very rarely used, so gaps are **always** long. Would be almost impossible to make term bursty (under any reasonable choice of  $s$ ).

Consider a word like **boundary**: **intense** bursts of usage in some years when redistribution of seats is discussed.

**btw** in keeping with original presentation, we treat base level as zero.

**NB** **no need** to stem or stop words: stops should all be level 0 anyway. We do make everything lowercase.

## More Intuition

Consider a word like **the**: used in (almost) every speech, so gaps are **always** small. Would be almost impossible to make term bursty (given we have binary use metric).

Consider a word like **zither**: very rarely used, so gaps are **always** long. Would be almost impossible to make term bursty (under any reasonable choice of  $s$ ).

Consider a word like **boundary**: **intense** bursts of usage in some years when redistribution of seats is discussed.

**btw** in keeping with original presentation, we treat base level as zero.

**NB** **no need** to stem or stop words: stops should all be level 0 anyway. We do make everything lowercase.

**so** this is **not** tf-idf (or similar):

# More Intuition

Consider a word like **the**: used in (almost) every speech, so gaps are **always** small. Would be almost impossible to make term bursty (given we have binary use metric).

Consider a word like **zither**: very rarely used, so gaps are **always** long. Would be almost impossible to make term bursty (under any reasonable choice of  $s$ ).

Consider a word like **boundary**: **intense** bursts of usage in some years when redistribution of seats is discussed.

**btw** in keeping with original presentation, we treat base level as zero.

**NB** **no need** to stem or stop words: stops should all be level 0 anyway. We do make everything lowercase.

**so** this is **not** tf-idf (or similar): not based on which words typify documents,

## More Intuition

Consider a word like **the**: used in (almost) every speech, so gaps are **always** small. Would be almost impossible to make term bursty (given we have binary use metric).

Consider a word like **zither**: very rarely used, so gaps are **always** long. Would be almost impossible to make term bursty (under any reasonable choice of  $s$ ).

Consider a word like **boundary**: **intense** bursts of usage in some years when redistribution of seats is discussed.

**btw** in keeping with original presentation, we treat base level as zero.

**NB** **no need** to stem or stop words: stops should all be level 0 anyway. We do make everything lowercase.

**so** this is **not** tf-idf (or similar): not based on which words typify documents, but rather on relative intensity of use *within word* over time.

# Allocating Bursts to MPs

# Allocating Bursts to MPs

For a given word,

# Allocating Bursts to MPs

For a given word, we know **when** (i.e. in what speech) it went through a burst,

# Allocating Bursts to MPs

For a given word, we know **when** (i.e. in what speech) it went through a burst, and **how much** of a burst (i.e. what level, what duration).



# Allocating Bursts to MPs

For a given word, we know **when** (i.e. in what speech) it went through a burst, and **how much** of a burst (i.e. what level, what duration).

- for that word, obtain its burstiness *in that speech* by multiplying its level  $\times$  duration:

# Allocating Bursts to MPs

For a given word, we know **when** (i.e. in what speech) it went through a burst, and **how much** of a burst (i.e. what level, what duration).

- for that word, obtain its burstiness *in that speech* by multiplying its level  $\times$  duration: e.g. level 3 for 5 period = burst of 15.

# Allocating Bursts to MPs

For a given word, we know **when** (i.e. in what speech) it went through a burst, and **how much** of a burst (i.e. what level, what duration).

→ for that word, obtain its burstiness *in that speech* by multiplying its level  $\times$  duration: e.g. level 3 for 5 period = burst of 15.

Do this for every word, in every speech:

# Allocating Bursts to MPs

For a given word, we know **when** (i.e. in what speech) it went through a burst, and **how much** of a burst (i.e. what level, what duration).

→ for that word, obtain its burstiness *in that speech* by multiplying its level  $\times$  duration: e.g. level 3 for 5 period = burst of 15.

Do this for every word, in every speech: a speech may have multiple bursty words (or none).

# Allocating Bursts to MPs

For a given word, we know **when** (i.e. in what speech) it went through a burst, and **how much** of a burst (i.e. what level, what duration).

→ for that word, obtain its burstiness *in that speech* by multiplying its level  $\times$  duration: e.g. level 3 for 5 period = burst of 15.

Do this for every word, in every speech: a speech may have multiple bursty words (or none).

An MP's **score** is simply the **aggregate** of his speech scores.

# Allocating Bursts to MPs

For a given word, we know **when** (i.e. in what speech) it went through a burst, and **how much** of a burst (i.e. what level, what duration).

→ for that word, obtain its burstiness *in that speech* by multiplying its level  $\times$  duration: e.g. level 3 for 5 period = burst of 15.

Do this for every word, in every speech: a speech may have multiple bursty words (or none).

An MP's **score** is simply the **aggregate** of his speech scores.

(a) is burstiness just 'speechiness'?

# Allocating Bursts to MPs

For a given word, we know **when** (i.e. in what speech) it went through a burst, and **how much** of a burst (i.e. what level, what duration).

→ for that word, obtain its burstiness *in that speech* by multiplying its level  $\times$  duration: e.g. level 3 for 5 period = burst of 15.

Do this for every word, in every speech: a speech may have multiple bursty words (or none).

An MP's **score** is simply the **aggregate** of his speech scores.

(a) is burstiness just 'speechiness'? **no**.

# Allocating Bursts to MPs

For a given word, we know **when** (i.e. in what speech) it went through a burst, and **how much** of a burst (i.e. what level, what duration).

→ for that word, obtain its burstiness *in that speech* by multiplying its level  $\times$  duration: e.g. level 3 for 5 period = burst of 15.

Do this for every word, in every speech: a speech may have multiple bursty words (or none).

An MP's **score** is simply the **aggregate** of his speech scores.

- (a) is burstiness just 'speechiness'? **no**.
- (b) do certain MPs have procedural rights that enable them to speak first in response to events?



# Allocating Bursts to MPs

For a given word, we know **when** (i.e. in what speech) it went through a burst, and **how much** of a burst (i.e. what level, what duration).

→ for that word, obtain its burstiness *in that speech* by multiplying its level  $\times$  duration: e.g. level 3 for 5 period = burst of 15.

Do this for every word, in every speech: a speech may have multiple bursty words (or none).

An MP's **score** is simply the **aggregate** of his speech scores.

- (a) is burstiness just 'speechiness'? **no**.
- (b) do certain MPs have procedural rights that enable them to speak first in response to events? **yes**, and that's fine with us.

# Validation

# Validation

Ultimately, this is **measurement problem**,

# Validation

Ultimately, this is **measurement problem**, so **validation** matters:

# Validation

Ultimately, this is **measurement problem**, so **validation** matters:

**1** at a given **time**,

# Validation

Ultimately, this is **measurement problem**, so **validation** matters:

1 at a given **time**, when we know certain issues were at stake,

# Validation

Ultimately, this is **measurement problem**, so **validation** matters:

- 1 at a given **time**, when we know certain issues were at stake, they should be bursty

# Validation

Ultimately, this is **measurement problem**, so **validation** matters:

- 1 at a given **time**, when we know certain issues were at stake, they should be bursty
- 2 given **words** should be bursty at 'right' times



# Validation

Ultimately, this is **measurement problem**, so **validation** matters:

- 1 at a given **time**, when we know certain issues were at stake, they should be bursty
- 2 given **words** should be bursty at 'right' times (when we rescale all terms between 0 and 1)

# Validation

Ultimately, this is **measurement problem**, so **validation** matters:

- 1 at a given **time**, when we know certain issues were at stake, they should be bursty
- 2 given **words** should be bursty at 'right' times (when we rescale all terms between 0 and 1)
- 3 given **MPs** should be bursty at 'right' times

# Validation

Ultimately, this is **measurement problem**, so **validation** matters:

- 1 at a given **time**, when we know certain issues were at stake, they should be bursty
- 2 given **words** should be bursty at 'right' times (when we rescale all terms between 0 and 1)
- 3 given **MPs** should be bursty at 'right' times (when we rescale all terms between 0 and 1)

# Validation

Ultimately, this is **measurement problem**, so **validation** matters:



- 1 at a given **time**, when we know certain issues were at stake, they should be bursty
- 2 given **words** should be bursty at 'right' times (when we rescale all terms between 0 and 1)
- 3 given **MPs** should be bursty at 'right' times (when we rescale all terms between 0 and 1)

# Validating I: right words are bursty, given time


# Validating I: right words are bursty, given time

session	1846 (1841, 6)	1866 (1865, 1)	1885 (1885, 1)
---------	-------------------	-------------------	-------------------

# Validating I: right words are bursty, given time

session	1846 (1841, 6)	1866 (1865, 1)	1885 (1885, 1)
			

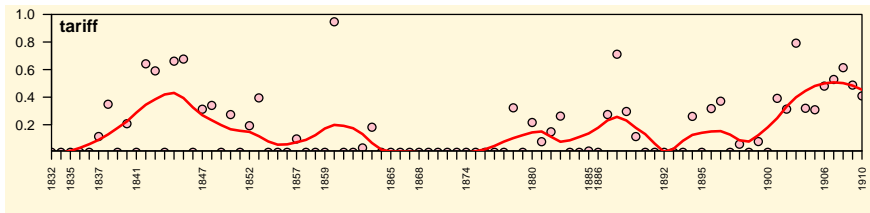
# Validating I: right words are bursty, given time

session	1846 (1841, 6)	1866 (1865, 1)	1885 (1885, 1)
			
terms (rank)	agriculturists (1) wheat (3) grain (5) farmer (6) prices (7)	suffrage (4) franchise (5) 1832 (7) redistribution (10) seats (11)	irishmen (2) 1782 (3) kingharmon (6) parnell (15) tenant (18)

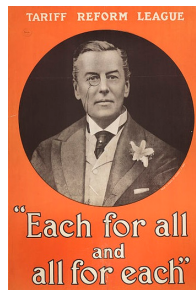
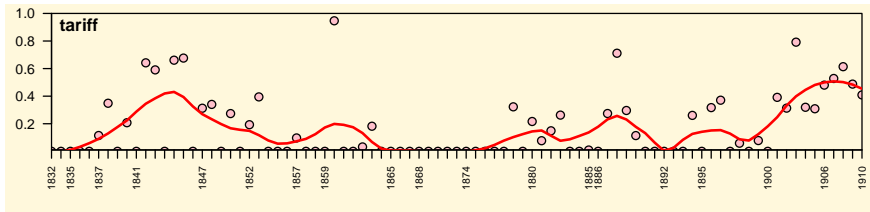


## Validating II: given words are bursty, at right time

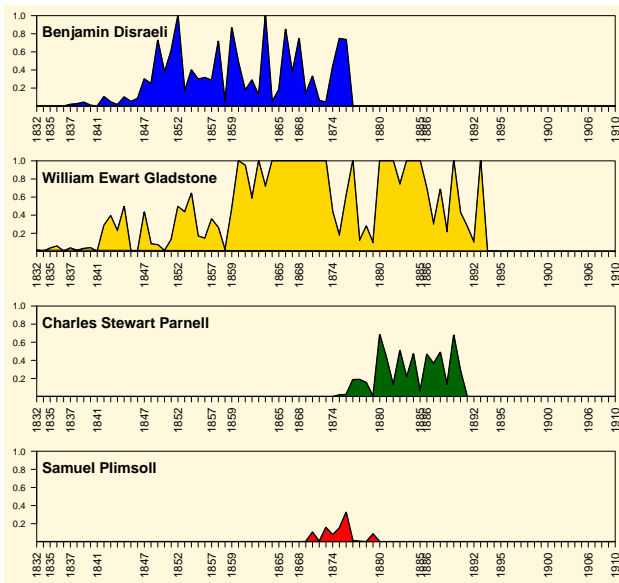
# Validating II: given words are bursty, at right time



# Validating II: given words are bursty, at right time



# Validating II: right MPs are bursty, at right time



# Shadow Cabinet Emerges

# Shadow Cabinet Emerges

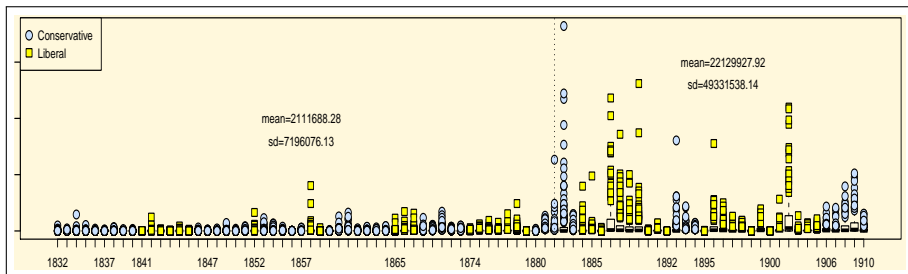
from 1880s,

# Shadow Cabinet Emerges

from 1880s, opposition **outliers** (defined in usual way wrt IQR) are noticeably more bursty. . .

# Shadow Cabinet Emerges

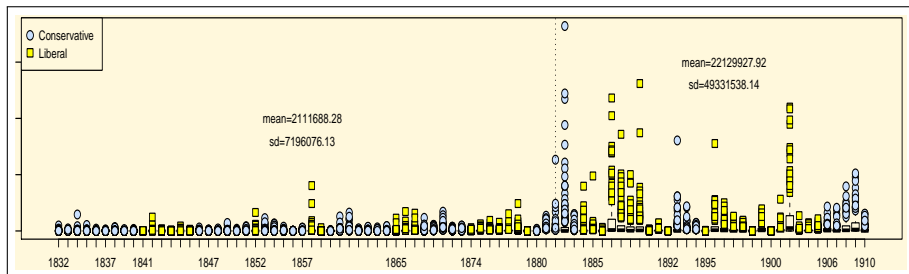
from 1880s, opposition **outliers** (defined in usual way wrt IQR) are noticeably more bursty. . .





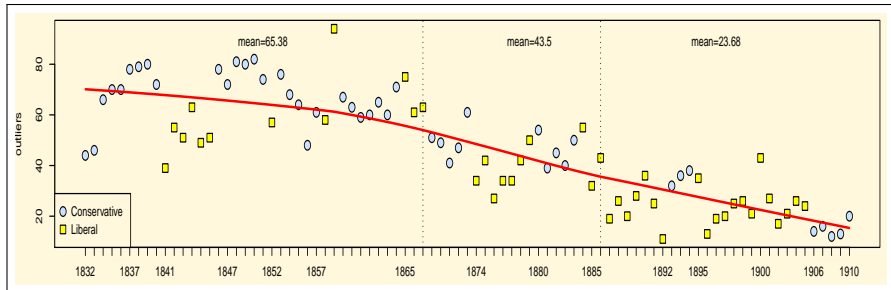
# Shadow Cabinet Emerges

from 1880s, opposition **outliers** (defined in usual way wrt IQR) are noticeably more bursty. . .

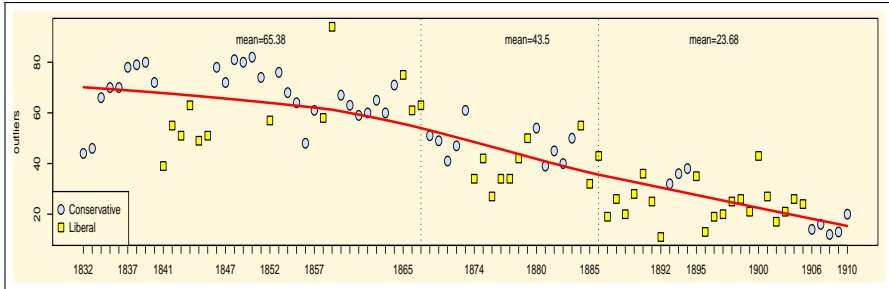


and fewer in number. . .

and fewer in number. . .



and fewer in number. . .



⇒ emergence of small group of opposition leaders

# “Meme-tracking and the Dynamics of the News Cycle”, Leskovec, Backstrom, and Kleinberg (2008)

# “Meme-tracking and the Dynamics of the News Cycle”, Leskovec, Backstrom, and Kleinberg (2008)

Want to understand how short, signature phrases—**memes**—diffuse and propagate over web:

# “Meme-tracking and the Dynamics of the News Cycle”, Leskovec, Backstrom, and Kleinberg (2008)

Want to understand how short, signature phrases—**memes**—diffuse and propagate over web: from news, to blogs etc.

# “Meme-tracking and the Dynamics of the News Cycle”, Leskovec, Backstrom, and Kleinberg (2008)

Want to understand how short, signature phrases—**memes**—diffuse and propagate over web: from news, to blogs etc.

Memes are ‘genetic signatures’: mutate over time,



# “Meme-tracking and the Dynamics of the News Cycle”, Leskovec, Backstrom, and Kleinberg (2008)

Want to understand how short, signature phrases—**memes**—diffuse and propagate over web: from news, to blogs etc.

Mememes are ‘genetic signatures’: mutate over time, which means very hard to track!

# “Meme-tracking and the Dynamics of the News Cycle”, Leskovec, Backstrom, and Kleinberg (2008)

Want to understand how short, signature phrases—**memes**—diffuse and propagate over web: from news, to blogs etc.

Mememes are ‘genetic signatures’: mutate over time, which means very hard to track! Key components:

# “Meme-tracking and the Dynamics of the News Cycle”, Leskovec, Backstrom, and Kleinberg (2008)

Want to understand how short, signature phrases—**memes**—diffuse and propagate over web: from news, to blogs etc.

Mememes are ‘genetic signatures’: mutate over time, which means very hard to track! Key components:

## 1 **imitation**:

# “Meme-tracking and the Dynamics of the News Cycle”, Leskovec, Backstrom, and Kleinberg (2008)

Want to understand how short, signature phrases—**memes**—diffuse and propagate over web: from news, to blogs etc.

Mememes are ‘genetic signatures’: mutate over time, which means very hard to track! Key components:

- 1 **imitation**: news sources copy each others’ decisions on what to cover,

# “Meme-tracking and the Dynamics of the News Cycle”, Leskovec, Backstrom, and Kleinberg (2008)

Want to understand how short, signature phrases—**memes**—diffuse and propagate over web: from news, to blogs etc.

Memes are ‘genetic signatures’: mutate over time, which means very hard to track! Key components:

- 1 **imitation**: news sources copy each others’ decisions on what to cover, subject to
- 2 **recency**,

# “Meme-tracking and the Dynamics of the News Cycle”, Leskovec, Backstrom, and Kleinberg (2008)

Want to understand how short, signature phrases—**memes**—diffuse and propagate over web: from news, to blogs etc.

Memes are ‘genetic signatures’: mutate over time, which means very hard to track! Key components:

- 1 **imitation**: news sources copy each others’ decisions on what to cover, subject to
- 2 **recency**, in sense that they want ‘fresh’ content.

# “Meme-tracking and the Dynamics of the News Cycle”, Leskovec, Backstrom, and Kleinberg (2008)

Want to understand how short, signature phrases—**memes**—diffuse and propagate over web: from news, to blogs etc.

Memes are ‘genetic signatures’: mutate over time, which means very hard to track! Key components:

- 1 **imitation**: news sources copy each others’ decisions on what to cover, subject to
- 2 **recency**, in sense that they want ‘fresh’ content.

How should we think about what counts as a ‘meme’?

# “Meme-tracking and the Dynamics of the News Cycle”, Leskovec, Backstrom, and Kleinberg (2008)

Want to understand how short, signature phrases—**memes**—diffuse and propagate over web: from news, to blogs etc.

Memes are ‘genetic signatures’: mutate over time, which means very hard to track! Key components:

- 1 **imitation**: news sources copy each others’ decisions on what to cover, subject to
- 2 **recency**, in sense that they want ‘fresh’ content.

How should we think about what counts as a ‘meme’? How does attention peak and decay?



# Data and Examples

# Data and Examples



# Data and Examples



“You can put lipstick on a pig,”

# Data and Examples



“You can put lipstick on a pig,” he said as the crowd cheered. “It’s still a pig.”

# Data and Examples



Gathered 1 million news articles in run up to 2008 election,

“You can put lipstick on a pig,” he said as the crowd cheered. “It’s still a pig.”

# Data and Examples



Gathered 1 million news articles in run up to 2008 election, plus 1.6 million blog posts

“You can put lipstick on a pig,” he said as the crowd cheered. “It’s still a pig.”

# Data and Examples



Gathered 1 million news articles in run up to 2008 election, plus 1.6 million blog posts

From that,

“You can put lipstick on a pig,” he said as the crowd cheered. “It’s still a pig.”

# Data and Examples



Gathered 1 million news articles in run up to 2008 election, plus 1.6 million blog posts

From that, gathered 112 million quoted strings (phrases),

“You can put lipstick on a pig,” he said as the crowd cheered. “It’s still a pig.”



# Data and Examples



Gathered 1 million news articles in run up to 2008 election, plus 1.6 million blog posts

From that, gathered 112 million quoted strings (phrases), which they have to parse down to **phrase clusters** to observe over time.

“You can put **lipstick on a pig**,” he said as the crowd cheered. “It’s still a pig.”

# Concept

# Concept

A **phrase cluster** is collection of phrases/quotes that are **close textual variants** of each other.

# Concept

A **phrase cluster** is collection of phrases/quotes that are **close textual variants** of each other.

Impose lower bound on how short phrases can be,

# Concept

A **phrase cluster** is collection of phrases/quotes that are **close textual variants** of each other.

Impose lower bound on how short phrases can be, and how commonly they must appear

# Concept

A **phrase cluster** is collection of phrases/quotes that are **close textual variants** of each other.

Impose lower bound on how short phrases can be, and how commonly they must appear

Build a **graph** of a quote in which each node is one of the phrases actually reported,

# Concept

A **phrase cluster** is collection of phrases/quotes that are **close textual variants** of each other.

Impose lower bound on how short phrases can be, and how commonly they must appear

Build a **graph** of a quote in which each node is one of the phrases actually reported, which contain some **contiguous subsequence** of the words in the original quote

# Concept

A **phrase cluster** is collection of phrases/quotes that are **close textual variants** of each other.

Impose lower bound on how short phrases can be, and how commonly they must appear

Build a **graph** of a quote in which each node is one of the phrases actually reported, which contain some **contiguous subsequence** of the words in the original quote

then **partition** the graph back to a 'meme' of closely related (defined technically) phrases that can be followed through the news media as a **thread**



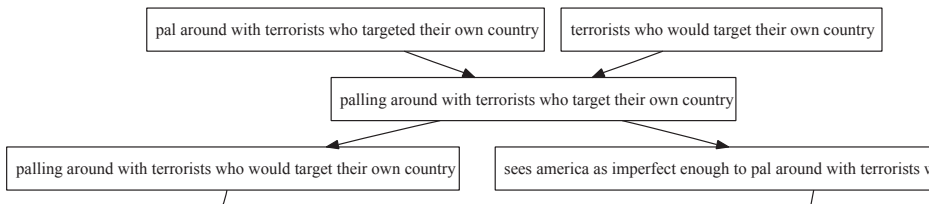
# Example

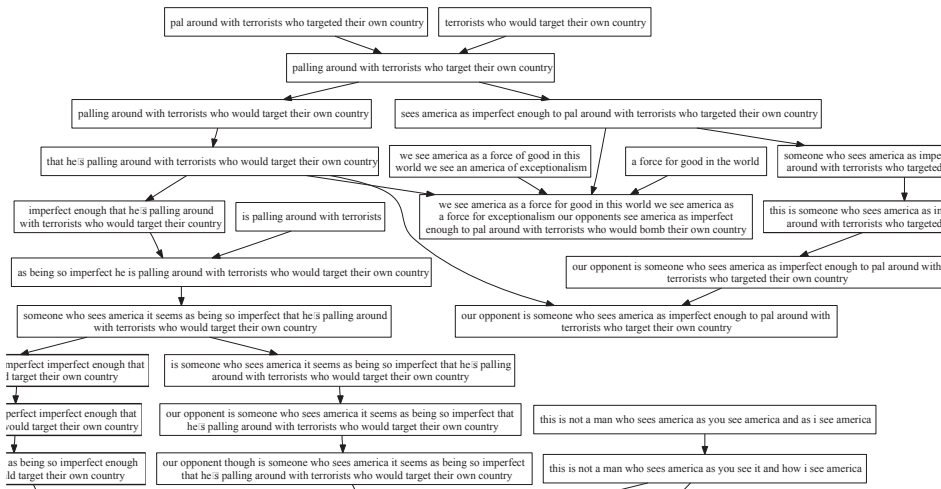


# Example



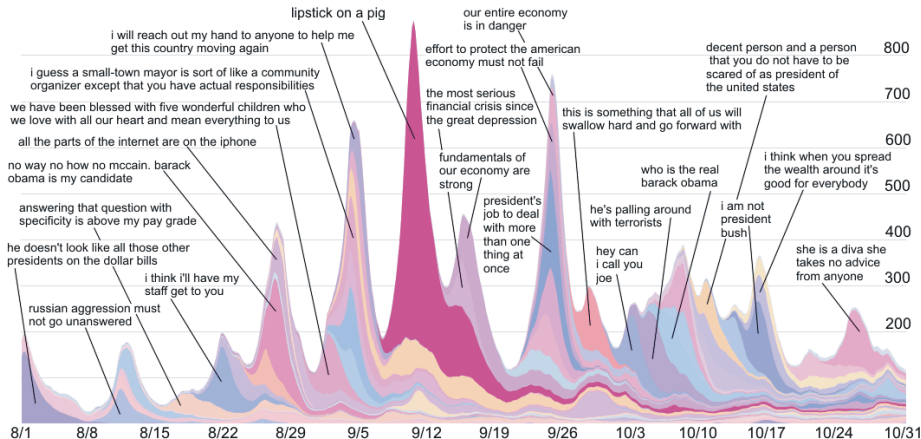
“Our opponent is someone who sees America, it seems, as being so imperfect, imperfect enough that hes palling around with terrorists who would target their own country.”





# Top 50 threads in 2008/9

# Top 50 threads in 2008/9



# Other Findings

## Other Findings

News cycle is a product of **imitation** and **recency**.



# Other Findings

News cycle is a product of **imitation** and **recency**.

Peak and decay is **symmetric**: not fast build up, and slow decay.

# Other Findings

News cycle is a product of **imitation** and **recency**.

Peak and decay is **symmetric**: not fast build up, and slow decay.

Peak blog intensity is around **2.5 hours** after news peak ('heart beat' pattern of decline in news, then picked up by blogs).

# Other Findings

News cycle is a product of **imitation** and **recency**.

Peak and decay is **symmetric**: not fast build up, and slow decay.

Peak blog intensity is around **2.5 hours** after news peak ('heart beat' pattern of decline in news, then picked up by blogs).

Generally movement is news → blogs,

# Other Findings

News cycle is a product of **imitation** and **recency**.

Peak and decay is **symmetric**: not fast build up, and slow decay.

Peak blog intensity is around **2.5 hours** after news peak ('heart beat' pattern of decline in news, then picked up by blogs).

Generally movement is news → blogs, but some phrases move the other way.

# *Quotus* (Niculae et al, 2015)

There is *prima facie* evidence that outlets of different hues embrace different quotes from Obama's speeches.

There is *prima facie* evidence that outlets of different hues embrace different quotes from Obama's speeches.

Niculae et al set out to examine whether there is a (ideological) *bias* in reporting.

There is *prima facie* evidence that outlets of different hues embrace different quotes from Obama's speeches.

Niculae et al set out to examine whether there is a (ideological) bias in reporting.

Data is six billion news articles,



There is *prima facie* evidence that outlets of different hues embrace different quotes from Obama's speeches.

Niculae et al set out to examine whether there is a (ideological) bias in reporting.

Data is six billion news articles, 2274 Obama public speeches.

There is *prima facie* evidence that outlets of different hues embrace different quotes from Obama's speeches.

Niculae et al set out to examine whether there is a (ideological) **bias** in reporting.

Data is six billion news articles, 2274 Obama public speeches. Idea is that what Obama says is **fixed** and prominent,

There is *prima facie* evidence that outlets of different hues embrace different quotes from Obama's speeches.

Niculae et al set out to examine whether there is a (ideological) **bias** in reporting.

Data is six billion news articles, 2274 Obama public speeches. Idea is that what Obama says is **fixed** and prominent, so question is how outlets choose to **quote** (may) reflect bias.

There is *prima facie* evidence that outlets of different hues embrace different quotes from Obama's speeches.

Niculae et al set out to examine whether there is a (ideological) **bias** in reporting.

Data is six billion news articles, 2274 Obama public speeches. Idea is that what Obama says is **fixed** and prominent, so question is how outlets choose to **quote** (may) reflect bias.

Main model matrix/**graph** has 275 media outlets,

There is *prima facie* evidence that outlets of different hues embrace different quotes from Obama's speeches.

Niculae et al set out to examine whether there is a (ideological) **bias** in reporting.

Data is six billion news articles, 2274 Obama public speeches. Idea is that what Obama says is **fixed** and prominent, so question is how outlets choose to **quote** (may) reflect bias.

Main model matrix/**graph** has 275 media outlets, and their selection of the 267,000 quotes they could use (binary).

There is *prima facie* evidence that outlets of different hues embrace different quotes from Obama's speeches.

Niculae et al set out to examine whether there is a (ideological) **bias** in reporting.

Data is six billion news articles, 2274 Obama public speeches. Idea is that what Obama says is **fixed** and prominent, so question is how outlets choose to **quote** (may) reflect bias.

Main model matrix/**graph** has 275 media outlets, and their selection of the 267,000 quotes they could use (binary).

# Partner Exercise

# Partner Exercise

- 1 Why don't the authors simply study how media outlets differ in terms of the stories they report?



# Partner Exercise

- 1 Why don't the authors simply study how media outlets differ in terms of the stories they report? Wouldn't that tell us something about ideological bias?

# Partner Exercise

- 1 Why don't the authors simply study how media outlets differ in terms of the stories they report? Wouldn't that tell us something about ideological bias?
- 2 In US politics, how do you expect ideological bias to manifest itself in terms of reporting Obama's speeches? (assuming he is not quoted out of context)

# Partner Exercise

- 1 Why don't the authors simply study how media outlets differ in terms of the stories they report? Wouldn't that tell us something about ideological bias?
- 2 In US politics, how do you expect ideological bias to manifest itself in terms of reporting Obama's speeches? (assuming he is not quoted out of context)

## Prima Facie evidence of bias

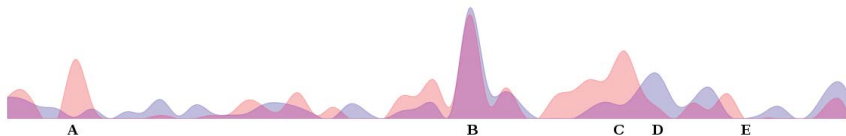


Figure 1: Volume of quotations for each word from a fragment of the 2010 State of the Union Address split by political leaning: conservative outlets shown in red and liberal outlets shown in blue. Quotes from the marked positions are reproduced in Table 1 and shown in the QUOTUS visualization in Figure 2.

Position	Quote from the 2010 State of the Union Address
A	And in the last year, hundreds of al Qaeda’s fighters and affiliates, including many senior leaders, have been captured or killed—far more than in 2008.
B	I will work with Congress and our military to finally repeal the law that denies gay Americans the right to serve the country they love because of who they are. It’s the right thing to do.
C	Each time lobbyists game the system or politicians tear each other down instead of lifting this country up, we lose faith. The more that TV pundits reduce serious debates to silly arguments, big issues into sound bites, our citizens turn away.
D	Democracy in a nation of 300 million people can be noisy and messy and complicated. And when you try to do big things and make big changes, it stirs passions and controversy. That’s just how it is.
E	But I wake up every day knowing that they are nothing compared to the setbacks that families all across this country have faced this year.

# Findings

Authors (basically) reduce the main matrix down to two dimensions,

# Findings

Authors (basically) reduce the main matrix down to two dimensions, and discover:

# Findings

Authors (basically) reduce the main matrix down to two dimensions, and discover:

- 1 first dimension is **independent-mainstream** while



# Findings

Authors (basically) reduce the main matrix down to two dimensions, and discover:

- 1 first dimension is **independent-mainstream** while
- 2 second dimension is **liberal-conservative**.

# Findings

Authors (basically) reduce the main matrix down to two dimensions, and discover:

- 1 first dimension is **independent-mainstream** while
- 2 second dimension is **liberal-conservative**.

Find that more conservative outlets tend to favor quotes that display **negative sentiment** (depressing!),

# Findings

Authors (basically) reduce the main matrix down to two dimensions, and discover:

- 1 first dimension is **independent-mainstream** while
- 2 second dimension is **liberal-conservative**.

Find that more conservative outlets tend to favor quotes that display **negative sentiment** (depressing!), more **negation** (controversial topics),

# Findings

Authors (basically) reduce the main matrix down to two dimensions, and discover:

- 1 first dimension is **independent-mainstream** while
- 2 second dimension is **liberal-conservative**.

Find that more conservative outlets tend to favor quotes that display **negative sentiment** (depressing!), more **negation** (controversial topics), more conservative **topics** of interest (e.g. troops rather than health care)

# Quote Results

# Quote Results

## First dimension of bias

<b>High</b>	<p>The principle that people of all faiths are welcome in this country, and will not be treated differently by their government, is essential to who we are.</p> <p>The United States is not, and will never be, at war with Islam. In fact, our partnership with the Muslim world is critical. At a time when our discourse has become so sharply polarized [...] it's important for us to pause for a moment and make sure that we are talking with each other in a way that heals, not a way that wounds.</p>
<b>Low</b>	<p>Tonight, we are turning the east room into a bona fide country music hall.</p> <p>You guys get two presidents for one, which is a pretty good deal.</p> <p>Now, nothing wrong with an art history degree—I love art history.</p>

## Second dimension of bias

<b>High</b>	<p>Those of you who are watching certain news channels, on which I'm not very popular, and you see folks waving tea bags around...</p> <p>If we don't work even harder than we did in 2008, then we're going to have a government that tells the American people, "you're on your own."</p> <p>By the way, if you've got health insurance, you're not getting hit by a tax.</p>
<b>Middle</b>	<p>Congress passed a temporary fix. A band-aid. But these cuts are scheduled to keep falling across other parts of the government that provide vital services for the American people.</p> <p>Keep in mind, nobody is asking them to raise income tax rates. All we're asking is for them to consider closing tax loopholes and deductions.</p> <p>The truth is, you could figure out on the back of an envelope how to get this done. The question is one of political will.</p>
<b>Low</b>	<p>By the end of the next year, all U.S. troops will be out of Iraq.</p> <p>We come together here in Copenhagen because climate change poses a grave and growing danger to our people.</p> <p>Wow, we must come together to end this war successfully.</p>