

# **Measuring Polarization in High-Dimensional Data:**

Method and Application to Congressional Speech

---

Matthew Gentzkow, Stanford and NBER  
Jesse M. Shapiro, Brown and NBER  
Matt Taddy, Microsoft and Chicago Booth

A word cloud featuring various political and social terms. The words are arranged in a non-uniform, overlapping manner. The colors used are red and blue, with some words appearing in both colors. The terms include:

- Wealthiest
- Pro life
- Tax Freedom
- entrepreneurs
- Tax Relief
- Tax Breaks
- 1 percent
- fair labor
- Pro choice
- ESTATE TAX
- War on Terror
- freedom fighters
- DEATH TAX
- terrorists
- Right to life
- equality
- Welfare Queens
- undocumented worker
- illegal alien
- Big Government
- living wage
- African American
- capitalist
- Washington takeover

## Origins

## THE LANGUAGE OF HEALTHCARE 2009

- THE 10 RULES FOR STOPPING THE "WASHINGTON TAKEOVER" OF HEALTHCARE**
- Abandon and exile ALL references to the "healthcare system." From now on, healthcare is about people. Before you speak, think of the three components of tone that matter most: Individualize. Personalize. Humanize.
- Humanize your approach. If you say there is no healthcare crisis, you give your listener permission to ignore everything else you say. It is a credibility killer for most Americans. A better approach is to define the crisis in your terms. "If you're one of the millions who can't afford healthcare, it is a crisis." Better yet, "If some bureaucrats put himself between you and your doctor, denying you exactly what you need, that's a crisis." And the best: "If you have to wait weeks for tests and months for treatment, that's a healthcare crisis."
  - "Time is the government healthcare killer. As Mick Jagger once sang, "Time is on Your Side." Nothing else turns people against the government takeover of healthcare than the realistic expectation that it will result in delayed and potentially even denied treatment, procedures and/or medications. "Waiting to buy a car or even a house kill you. But waiting for the healthcare you need - could. Delayed care is denied care."
  - The arguments against the Democrats' healthcare plan must center around "politics," "bureaucrats," and "Washington" ... not the free market, tax or competition. Stop talking economic theory and start personalizing the government takeover of healthcare. They don't want to hear that the gov't lower costs will be embraced) or because it's anti-competitive (they don't care about current limits to competition). But they are extremely receptive to Washington approach. It's not an economic issue. It's a bureaucratic takeover. ... horror stories from Canada & Co. do resonate. We recommend the phrase "government controlled" because the system like

## TAX RELIEF & SIMPLIFICATION

## OVERVIEW

- “ You may be tempted to talk about tax policy in terms of *reform*. Don’t. When Americans hear the word *reform*, they fear that they will end up paying more. Far better for you to talk about *simplification* – which everyone supports and sees a benefit.
- “ You may be tempted to talk about making the tax cuts from 2001 and 2003 “permanent.” Don’t. It is far more effective to talk about “the largest tax increase in American history if these tax cuts are *reversed*.” Remember, the American public dislikes a tax hike more than they like a tax cut.
- “ You may be tempted to talk about how Americans are overtaxed overall. Do, but also emphasize that Washington spends too much as well. The more you link high taxes to high spending, the greater the support for tax relief.
- If there is one debate

If there is one debate where framing the issue is as important as the policy itself, this is it. So here's what needs to be said to set the context and begin the tax relief and tax simplification effort:

- 1) ***Explain tax relief. Don't talk in numbers.*** Talk in terms of day-to-day life, and Americans at the end of every week. Don't talk about the overall size of the cut. Focus instead on the marriage penalty, death taxes, rate reductions, and so on.
- 2) ***"The only way to stop wasteful Washington spending now and forever is to keep the money with those who earned it."*** Americans are actually willing to pay their current tax bill, but what makes them angry is how Washington spends that money. So tell them: "If Washington doesn't have your money, Washington can't spend your money."
- 3) ***"It's the economy, stupid."*** The second strongest argument for maintaining the tax relief is to provide "economic security." Every day, more Americans are concerned about their personal job security and their individual financial situation. The economy isn't recovering as well under and jobs are coming back in record numbers. The economy is still with us. ***"A tax hike will only hike uncertainty and anxiety."***
- 4) ***The IRS is still the most hated institution of government. You cannot override it when it comes to attacking the IRS.*** The single greatest public relations success of the Republican Congressional majority was the 1997 public hearings on the IRS. For about 11 days you were the face of the country — true political heroes. History may repeat itself, but why not try it? The more you focus on confronting IRS abuses and the need for IRS reform, the better.

## Example: Social Security

- Luntz (2006):
  - “Never say ‘**privatization / private accounts.**’ Instead say ‘**personalization / personal accounts.**’ Two-thirds of America want to personalize security while only one third would privatize it. Why? [Personalization] suggests ownership and control... while [privatization] suggests a profit motive and winners and losers.”

## Example: Social Security

- 2005 Congress

	Rep	Dem
“personal account”	184	48
“private account”	5	542

## Example: Social Security

- 2005 Congress

	Rep	Dem
“personal account”	184	48
“private account”	5	542

- Media coverage, 6/23/05
  - “House GOP offers plan for Social Security; Bush’s **private accounts** would be scaled back” (*Washington Post*)
  - “GOP backs use of Social Security surplus; Finds funding for **personal accounts**” (*Washington Times*)

**Is partisan speech a new phenomenon?**

# This Paper

- Goal: Measure trends in partisanship of political speech
- Data: US Congressional Record, 1873-2009
- Challenge: Speech is high-dimensional choice data
  - Potential for severe finite-sample bias
  - Computation can be difficult
- Solution: Structural estimation with machine-learning methods
  - Approach exportable to other contexts (e.g. web browsing, residential segregation)



# Literature

- Polarization in Congress
  - E.g., Poole & Rosenthal (1984, 1997); McCarty et al. (2006)
- Polarization more broadly
  - E.g., Fiorina et al. (2006); Fiorina & Abrams (2006); Abramowitz & Saunders (2008)
- Congressional speech
  - E.g., Grimmer (2010, 2013); Quinn et al (2010)
  - Jensen et al (2012)

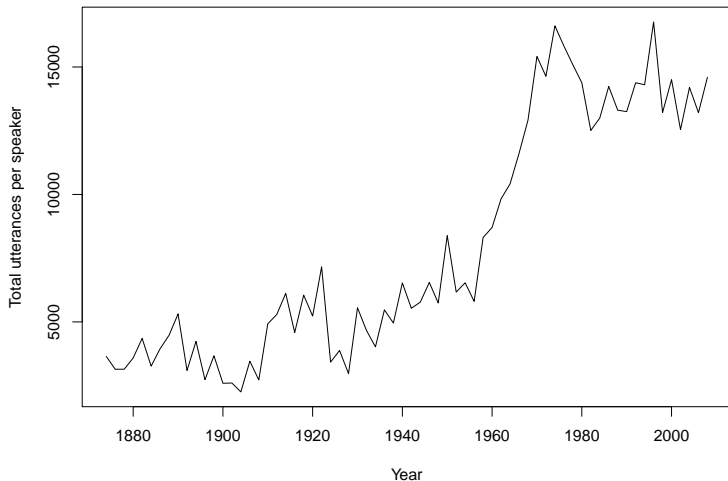
# Data

---

# Data

- US Congressional Record, 1872-2009
- Use automated script to identify speaker and tag with metadata
- Use some rules of thumb to remove procedural phrases
  - “I yield the remainder of my time...”
- Turn into counts of two-word phrases less stems and stopwords
  - “war on terrorism” and “war on terror” become “war terror”

# Trends in Verbosity



# Model

---

# Statistical Model

- Vector of phrase counts  $\mathbf{c}_{it}$  for members  $i$
- Party affiliation  $P(i) \in \{R, D\}$
- Speaker characteristics  $\mathbf{x}_{it}$
- Verbosity  $m_{it} = \sum_j c_{ijt}$
- Assume throughout that

$$\mathbf{c}_{it} \sim \text{MN} \left( m_{it}, \mathbf{q}_t^{P(i)}(\mathbf{x}_{it}) \right)$$

# Question

- *How different are choices of  $R$  and  $D$  at each  $t$ ?*
  - Translation: how different are  $\mathbf{q}_t^R()$  and  $\mathbf{q}_t^D()$ ?
- Approach: measure partisanship by diagnosticity
  - How much can I learn about your party from what you say?

# Posteriors

- Posterior belief of an observer with a neutral prior after hearing phrase  $j$

$$\rho_{jt}(\mathbf{x}) = \frac{q_{jt}^R(\mathbf{x})}{q_{jt}^R(\mathbf{x}) + q_{jt}^D(\mathbf{x})}$$



# Posteriors

- Posterior belief of an observer with a neutral prior after hearing phrase  $j$

$$\rho_{jt}(\mathbf{x}) = \frac{q_{jt}^R(\mathbf{x})}{q_{jt}^R(\mathbf{x}) + q_{jt}^D(\mathbf{x})}$$

- Posterior that the observer expects to assign to the speaker's true party

$$\pi_t(\mathbf{x}) = \frac{1}{2} \mathbf{q}_t^R(\mathbf{x})' \cdot \rho_t(\mathbf{x}) + \frac{1}{2} \mathbf{q}_t^D(\mathbf{x})' \cdot (1 - \rho_t(\mathbf{x}))$$

# Measure of Partisanship

$$\bar{\pi}_t = \frac{1}{N_t} \sum_i \pi_t(\mathbf{x}_{it})$$

- Between  $\frac{1}{2}$  (speech uninformative) and 1 (speech fully revealing)
- Close cousin of isolation (White 1986, Cutler et al 1999)

# Estimation

---

# Plug-In Estimator

- Empirical analogues

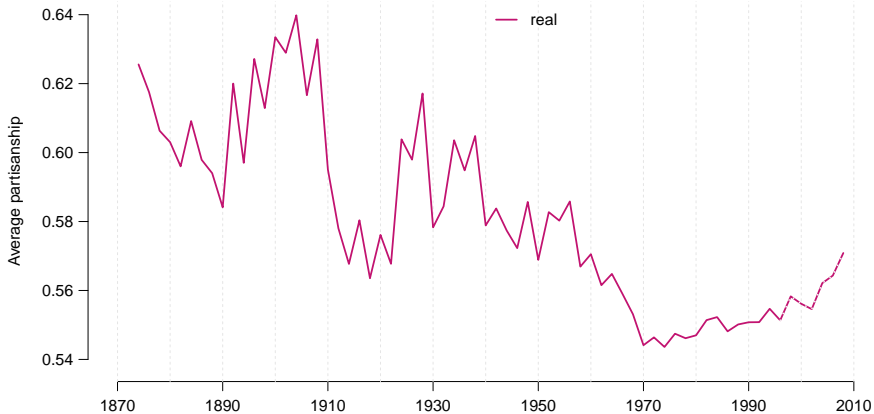
$$\hat{q}_{jt}^P = \frac{\sum_{i \in P} c_{ijt}}{\sum_{i \in P} m_{it}}$$

$$\hat{\rho}_{jt} = \frac{\hat{q}_{jt}^R}{\hat{q}_{jt}^R + \hat{q}_{jt}^D}$$

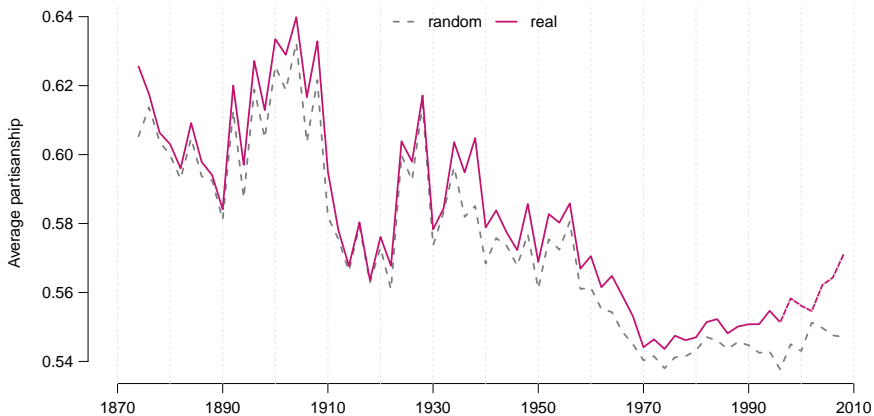
$$\hat{\pi}_t^{PLUGIN} = \frac{1}{2} (\hat{\mathbf{q}}_t^R)' \hat{\rho}_t + \frac{1}{2} (\hat{\mathbf{q}}_t^D)' (1 - \hat{\rho}_t)$$

- This is the MLE when  $\mathbf{x}_{it}$  is constant
- Consistent as quantity of speech grows large holding size of vocabulary fixed

# Maximum Likelihood Estimator



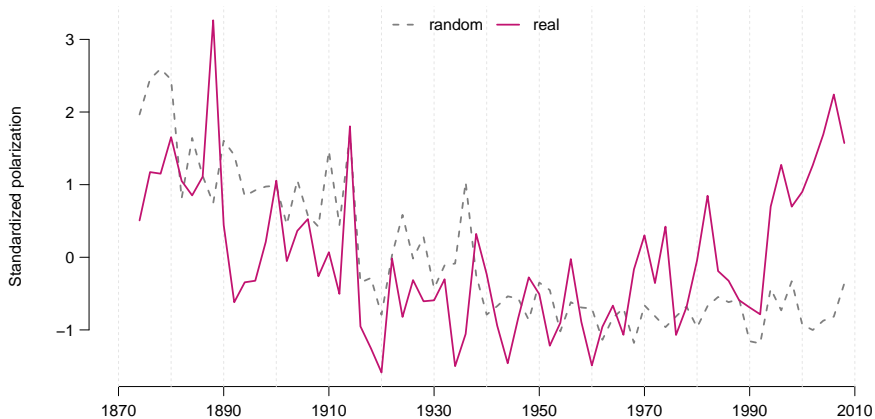
# Maximum Likelihood Estimator



$$\begin{aligned} \mathbb{E} \left[ (\hat{\mathbf{q}}_t^R)' \hat{\rho}_t - (\mathbf{q}_t^R)' \rho_t \right] &= (\mathbf{q}_t^R)' \mathbb{E} (\hat{\rho}_t - \rho_t) + \\ &\quad \text{Cov} \left[ (\hat{\mathbf{q}}_t^R - \mathbf{q}_t^R)', (\hat{\rho}_t - \rho_t) \right] \end{aligned}$$

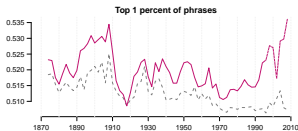
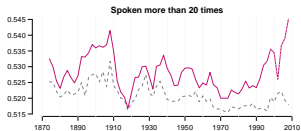
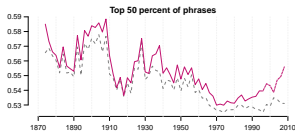
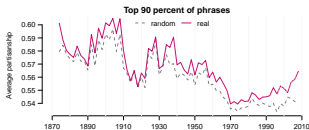
- $\hat{\mathbf{q}}_t^P$  is unbiased for  $\mathbf{q}_t^P$
- First term non-zero because  $\hat{\rho}_t$  is a non-linear function of  $\hat{\mathbf{q}}_t^P$
- Second term non-zero because  $\hat{\rho}_t$  is an increasing function of  $\hat{\mathbf{q}}_t^R$

# Jensen et al. (2012)





# Restrict to Commonly Occurring Phrases?



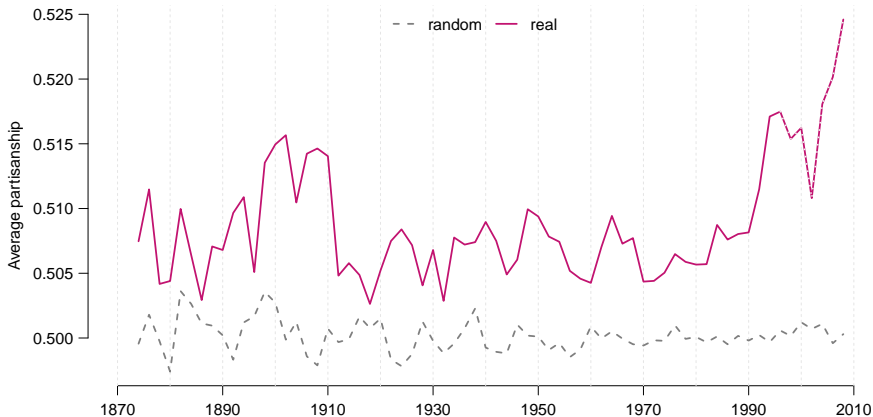
# Leave-Out Estimator

- Define  $\hat{\rho}_{-i,t}$  which leaves out  $i$
- Define

$$\hat{\pi}_t^{LOE} = \frac{1}{2} \frac{1}{|R_t|} \sum_{i \in R_t} \hat{\mathbf{q}}'_{i,t} \cdot \hat{\rho}_{-i,t} + \frac{1}{2} \frac{1}{|D_t|} \sum_{i \in D_t} \hat{\mathbf{q}}'_{i,t} \cdot (1 - \hat{\rho}_{-i,t})$$

- Enforces independence of  $\hat{q}$  and  $\hat{\rho}$
- Still biased because of non-linear  $\hat{\rho}$

# Leave-Out Estimator

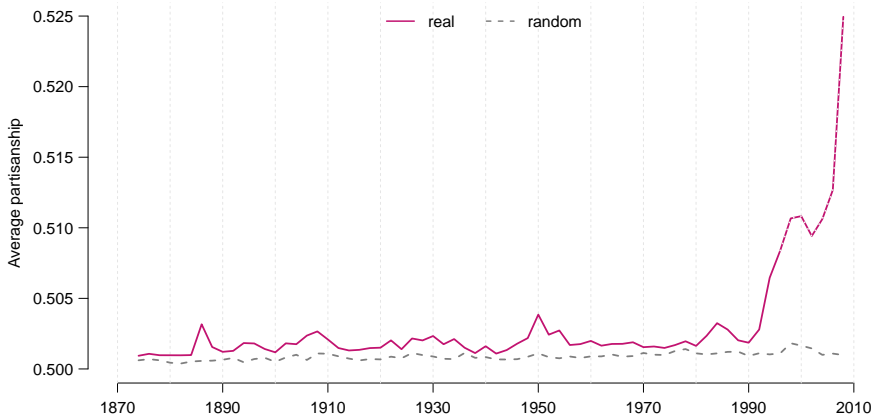


- Controlling bias
  - Add lasso type penalty to likelihood
  - Shrink  $\hat{\alpha}$  toward 1

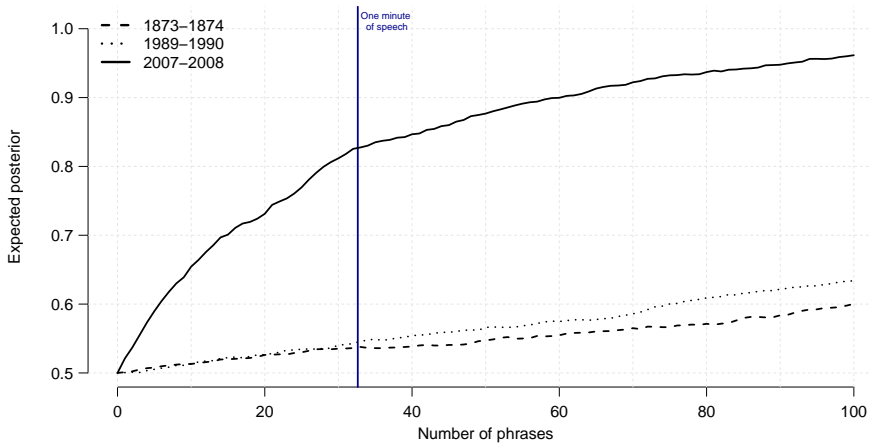
## **Main Results**

---

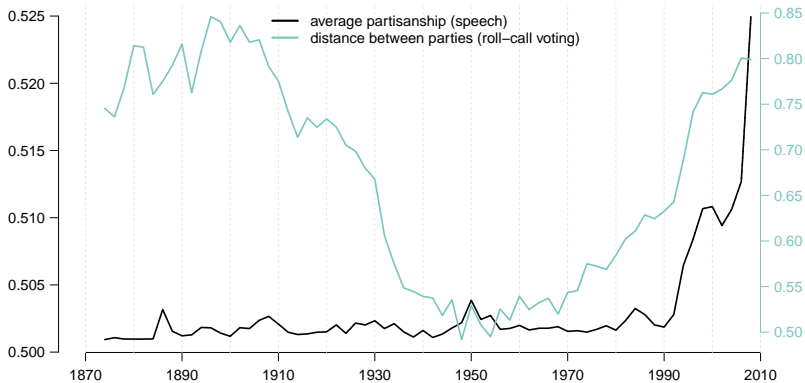
# Baseline Specification



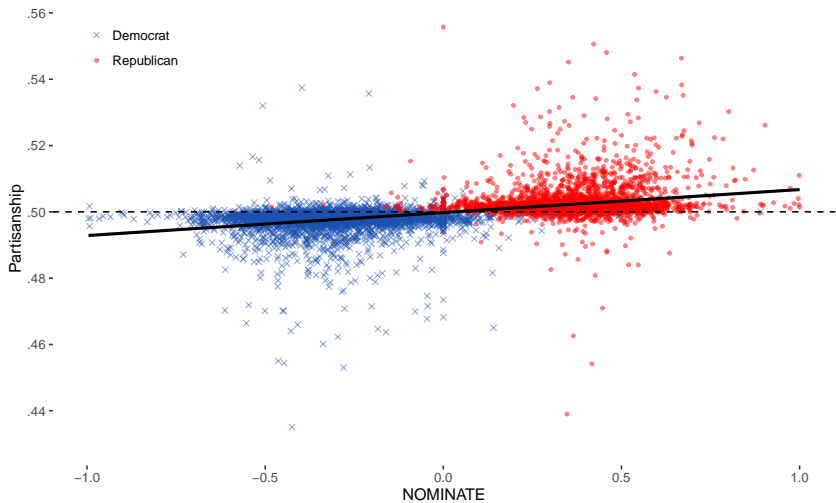
# Magnitude



# Comparison: Roll Call Votes



# Comparison: Roll Call Votes





# Unpacking Partisanship

---

# Most Partisan Phrases

- Define the *partisanship of phrase  $j$*  in session  $t$  to be the effect on  $\bar{\pi}_t$  of removing phrase  $j$  from the vocabulary (redistributing probability mass to other phrases proportionally)
  - Let  $\tilde{q}_{kt}^P$  equal  $q_{kt}^P / (1 - q_{jt}^P)$  if  $k \neq j$  and 0 otherwise
  - Recompute  $\bar{\pi}_t$  replacing  $\mathbf{q}_t^P$  with  $\tilde{\mathbf{q}}_t^P$  and holding  $\rho_t$  constant

## 60th Congress (1907-08)

<b>Most Republican</b>	<b>Most Democratic</b>
infantri war	section corner
indian war	ship subsidi
mount volunt	republ panama
feet thenc	level canal
postal save	powder trust
spain pay	print paper
war pay	lock canal
first regiment	bureau corpor
soil survey	senatori term
nation forest	remove wreck

## 60th Congress (1907-08)

<b>Most Republican</b>	<b>Most Democratic</b>
<b>infantri war</b>	section corner
<b>indian war</b>	ship subsidi
<b>mount volunt</b>	republ panama
feet thenc	level canal
postal save	powder trust
<b>spain pay</b>	print paper
<b>war pay</b>	lock canal
<b>first regiment</b>	bureau corpor
soil survey	senatori term
nation forest	remove wreck

- 1908 Rep platform: Calls for “generous provision” for veterans of Spanish-American and Indian wars

## 60th Congress (1907-08)

Most Republican	Most Democratic
infantri war	section corner
indian war	<b>ship subsidi</b>
mount volunt	republ panama
feet thenc	level canal
postal save	<b>powder trust</b>
spain pay	<b>print paper</b>
war pay	lock canal
first regiment	<b>bureau corpor</b>
soil survey	senatori term
nation forest	remove wreck

- 1908 Dem platform: "Free the Government from the grip of those who have made it a business asset of the favor-seeking corporations."
- William Cox (D-IN): "the entire United States is now being held up by a great hydra-headed monster, known in ordinary parlance as a '*powder trust*'."

## 80th Congress (1947-48)

<b>Most Republican</b>	<b>Most Democratic</b>
steam plant	admir denfeld
coast guard	public busi
stop communism	labor standard
depart agricultur	intern labor
lend leas	tax refund
zone germani	concili service
british loan	standard act
approv compact	soil conserv
unit kingdom	school lunch
union shop	cent hour

## 80th Congress (1947-48)

Most Republican	Most Democratic
steam plant	admiral denfeld
<b>coast guard</b>	public busi
<b>stop communism</b>	labor standard
depart agricultur	intern labor
<b>lend leas</b>	tax refund
<b>zone germani</b>	concili service
<b>british loan</b>	standard act
approv compact	soil conserv
<b>unit kingdom</b>	school lunch
union shop	cent hour

- Aftermath of WWII

## 80th Congress (1947-48)

Most Republican	Most Democratic
steam plant	admir denfeld
coast guard	public busi
stop communism	<b>labor standard</b>
depart agricultur	<b>intern labor</b>
lend leas	tax refund
zone germani	concili service
british loan	<b>standard act</b>
approv compact	soil conserv
unit kingdom	<b>school lunch</b>
union shop	<b>cent hour</b>

- 1948 Dem platform: Advocates amending Fair Labor Standards Act to raise the federal minimum wage to 75 cents per hour; also advocates school lunch program



# 100th Congress (1987-88)

<b>Most Republican</b>	<b>Most Democratic</b>
freedom fighter	star war
doubl breast	contra aid
abort industri	nuclear weapon
demand second	contra war
heifer tax	support contra
reserv object	nuclear wast
incom ballist	agent orang
communist govern	central american
withdraw reserv	nicaraguan govern
abort demand	hatian peopl

# 100th Congress (1987-88)

Most Republican	Most Democratic
<b>freedom fighter</b>	star war
doubl breast	<b>contra aid</b>
abort industri	nuclear weapon
demand second	<b>contra war</b>
heifer tax	<b>support contra</b>
reserv object	nuclear wast
incom ballist	agent orang
<b>communist govern</b>	<b>central american</b>
withdraw reserv	<b>nicaraguan govern</b>
abort demand	hatian peopl

- Debate over support for Contra rebels fighting Sandinista government in Nicaragua; Iran-Contra affair

# 100th Congress (1987-88)

<b>Most Republican</b>	<b>Most Democratic</b>
freedom fighter	<b>star war</b>
doubl breast	contra aid
abort industri	<b>nuclear weapon</b>
demand second	contra war
heifer tax	support contra
reserv object	<b>nuclear wast</b>
<b>incom ballist</b>	agent orang
communist govern	central american
withdraw reserv	nicaraguan govern
abort demand	hatian peopl

- Debate over Reagan's "Star Wars" missile defense initiative & nuclear weapons policy

# 104th Congress (1995-96)

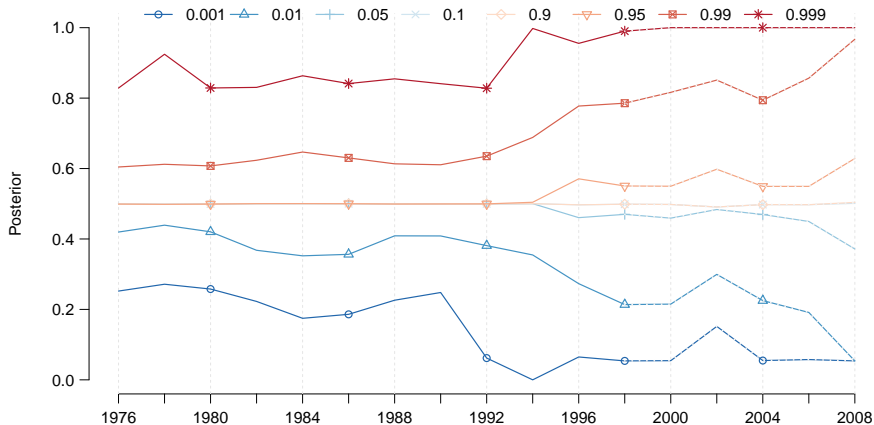
<b>Most Republican</b>	<b>Most Democratic</b>
medic save	tax break
partialbirth abort	nurs home
big govern	comp time
feder debt	break wealthi
tax increas	break wealthiest
tax relief	communiti polic
term limit	million children
nation debt	assault weapon
tax freedom	deficit reduct
item veto	head start

# 104th Congress (1995-96)

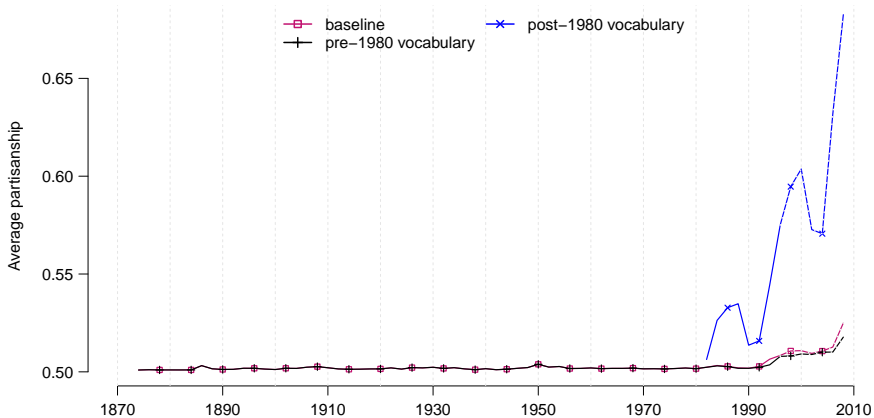
Most Republican	Most Democratic
medic save	<b>tax break</b>
partialbirth abort	nurs home
<b>big govern</b>	comp time
<b>feder debt</b>	<b>break wealthi</b>
<b>tax increas</b>	<b>break wealthiest</b>
<b>tax relief</b>	communiti polic
term limit	million children
<b>nation debt</b>	assault weapon
<b>tax freedom</b>	deficit reduct
item veto	head start

- Debate over taxes and fiscal policy; Republicans using language from Luntz memos and Contract with America

# Distribution of Phrase-Level Partisanship



# Neologisms



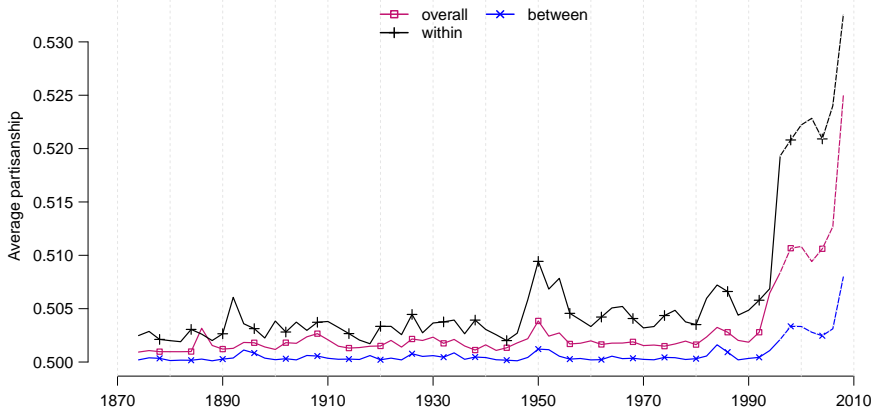
# Topic Decomposition

- Are trends in partisanship driven by
  - Divergence in *which* topics Dems/Reps emphasize?
  - Divergence in *how* the parties talk about a given topic?

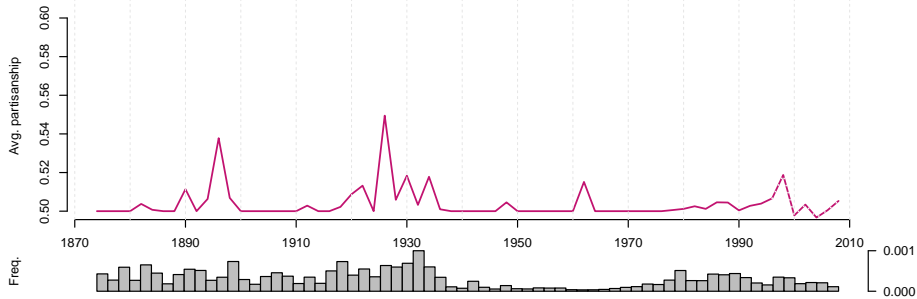


# Topics

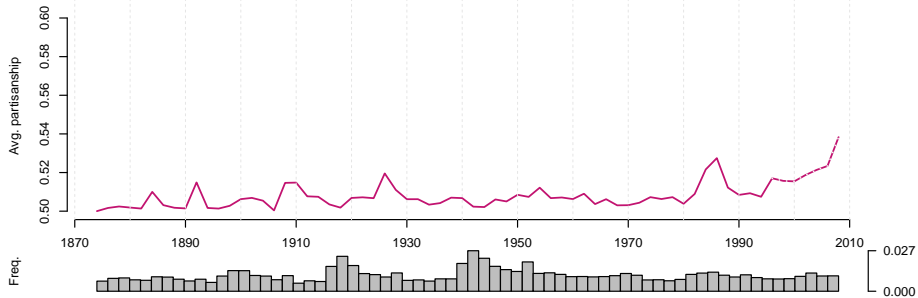
alcohol	environment	mail
budget	federalism	minorities
business	foreign	money
crime	government	religion
defense	health	tax
economy	immigration	trade
education	justice	
elections	labor	



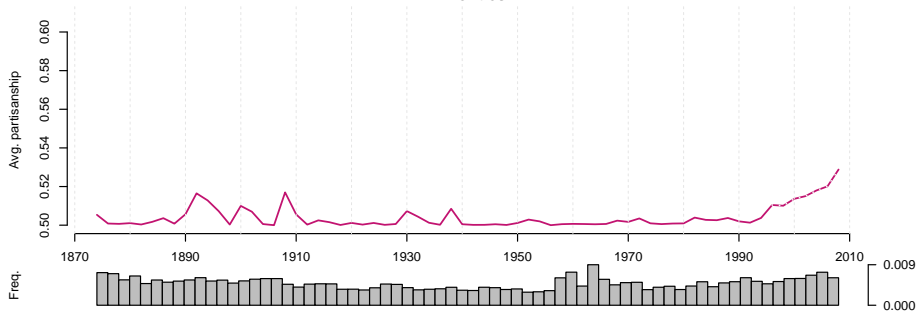
# alcohol

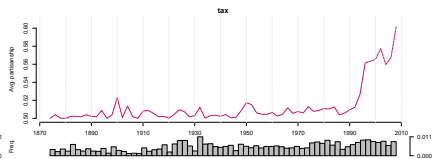
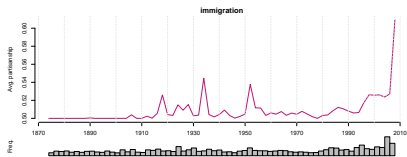
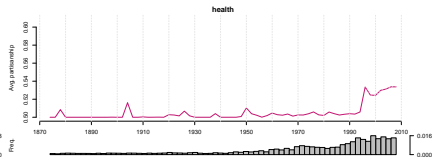
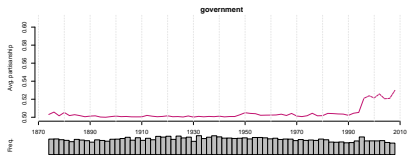
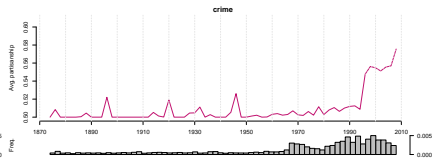
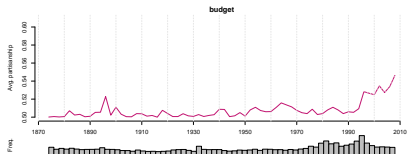


# defense

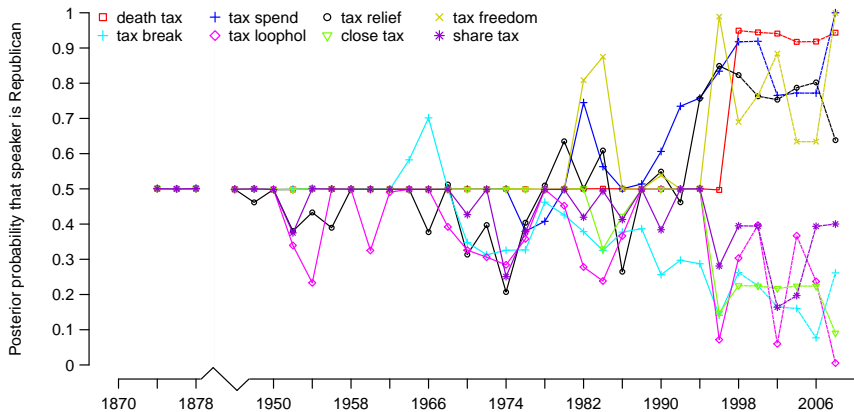


# minorities





# Individual Tax Phrases



# Explanations

---



# Political Innovation

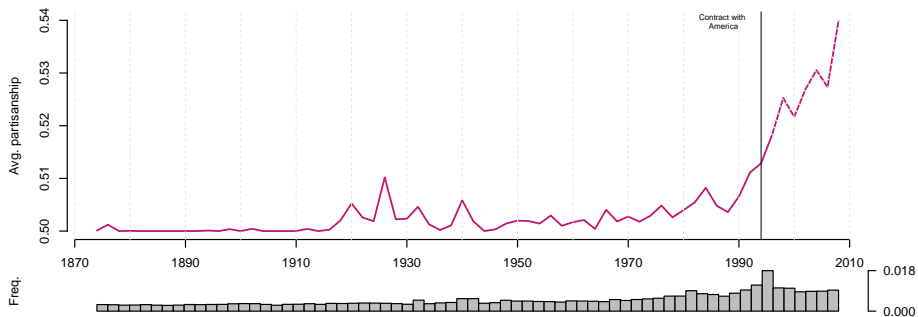
- **Contract with America (1994)**

- Republicans take control of Congress for first time since 1952
- Frank Luntz: novel polling techniques, memos to Republican candidates
- In the aftermath, Democrats launch an effort to improve their own choice of language

***You believe language can change a paradigm?*** “I don’t believe it – I know it. I’ve seen it with my own eyes...I watched in 1994 when the group of Republicans got together and said: ‘We’re going to do this completely differently than it’s ever been done before.’...Every politician and every political party issues a platform, but only these people signed a contract.” - **Luntz (2004)**

“Republican framing superiority had played a major role in their takeover of Congress in 1994. I and others had hoped that... a widespread understanding of how framing worked would allow Democrats to reverse the trend.” - **Lakoff (2014)**

# Phrases from CWA

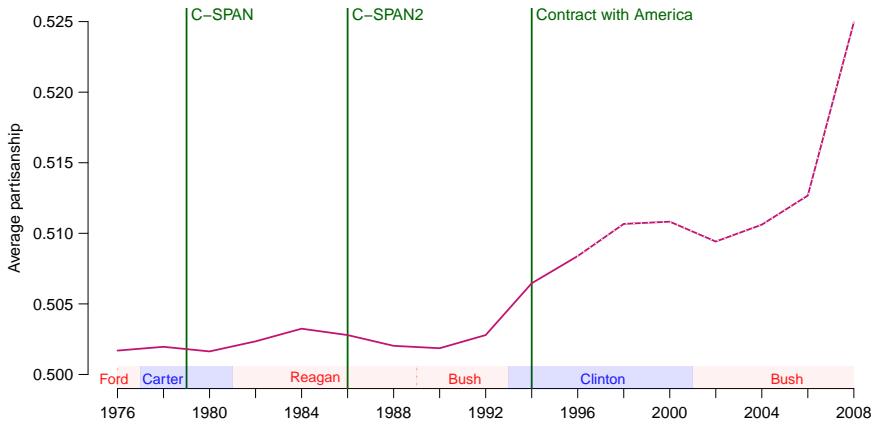


## Broader Context

- Party discipline in speech
  - Democratic Message Board (1989-1991)
  - Republican Theme Team (1991-1993): “develop ideas and phrases to be used by all Republicans”
- Changing media environment
  - 1979: C-SPAN (House of Representatives)
  - 1983: C-SPAN2 (Senate)

*“When asked whether he would be the Republican leader without C-SPAN, Gingrich... [replied] ‘No’... C-SPAN provided a group of media-savvy House conservatives in the mid-1980s with a method of... winning a prime-time audience.” (Frantzich & Sullivan 1996)*

# Summary



## Conclusion

---

# Does Language Matter?

- Partisan language in Congress diffuses to broader public
  - Gentzkow & Shapiro 2010; Martin & Yurukoglu 2016; Greenstein & Zhu 2012
- Issue framing affects public opinion
  - Lathrop 2003; Graetz and Shapiro 2006; Druckman et al. 2013
- Language affects group identity
  - Kinzler et al 2007, Clots-Figueas and Masella 2013
- *"Human beings do not live in the objective world alone, nor alone in the world of social activity as ordinarily understood, but are very much at the mercy of the particular language which has become the medium of expression."* (Sapir 1954)
- *"When we successfully reframe public discourse, we change the way the public sees the world. We change what counts as common sense.... Thinking differently requires speaking differently."* (Lakoff 2014)