Measuring Polarization in High-Dimensional Data

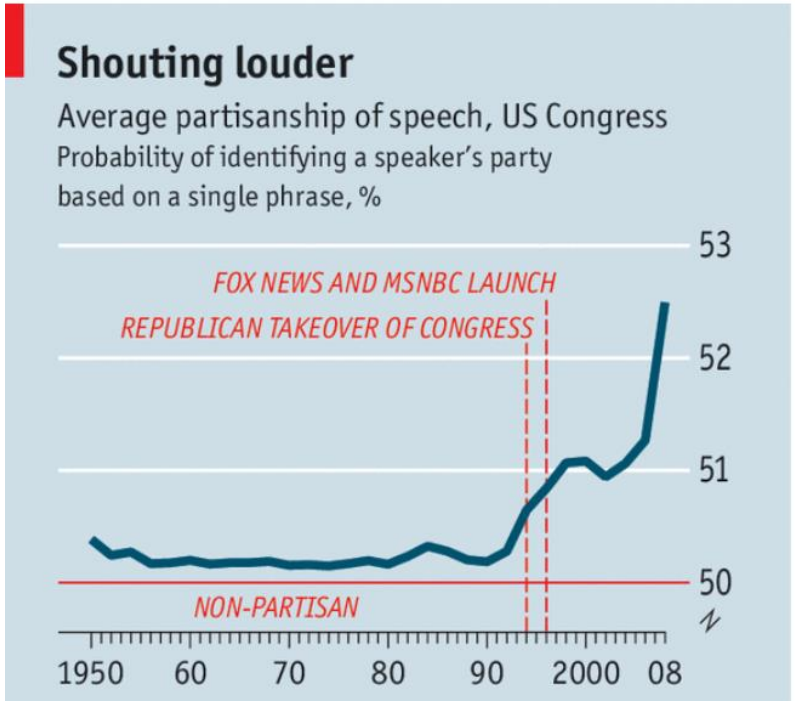
Method and Application to Congressional Speech

Matthew Gentzkow – Stanford and NBER
Jesse Shapiro – Brown and NBER
Matt Taddy — Microsoft Research and Chicago Booth

The New York Times

Republicans and Democrats in Congress Speak in Completely Different Languages

The Economist

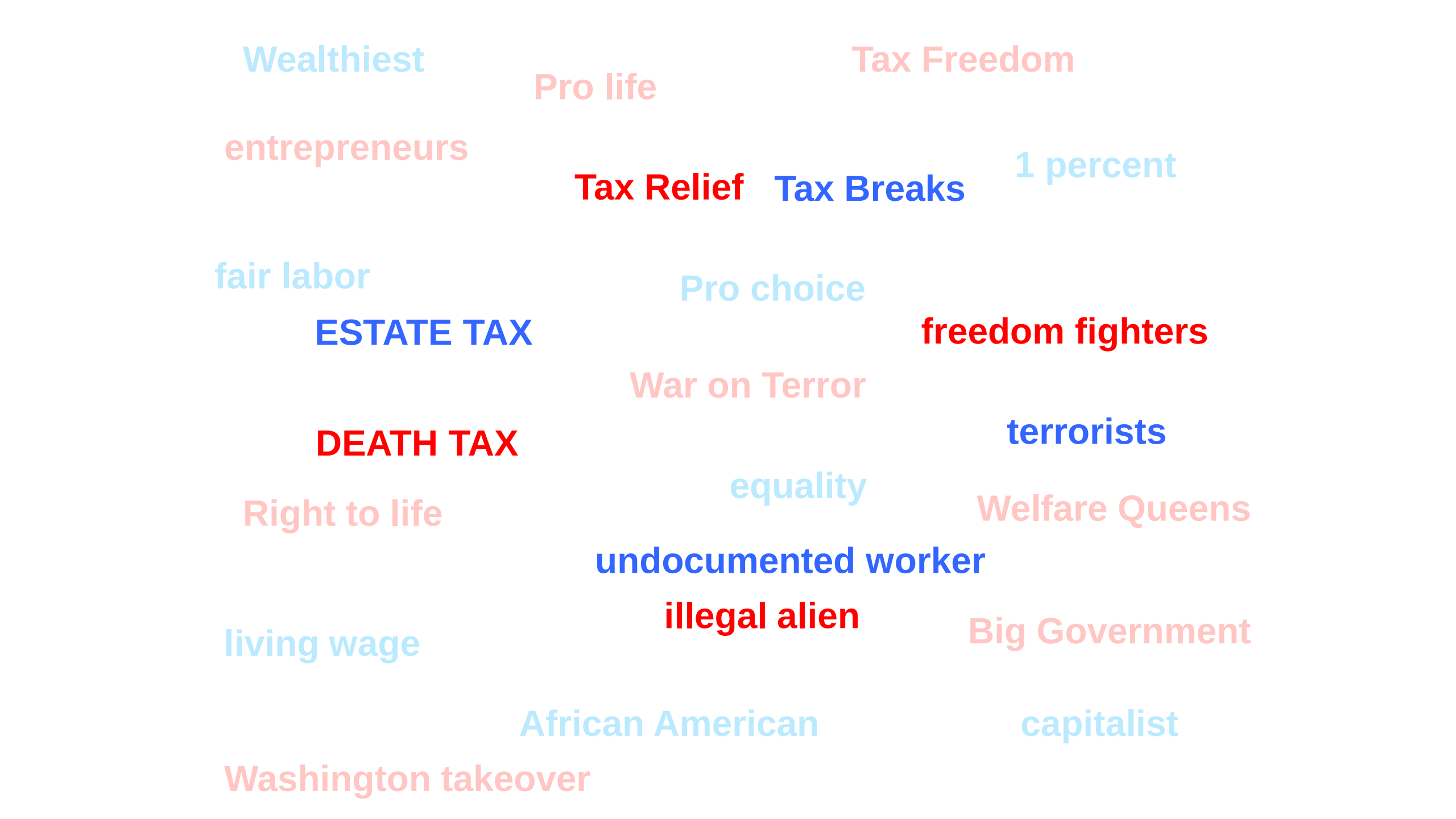


Why Democrats and Republicans Literally Speak Different Languages

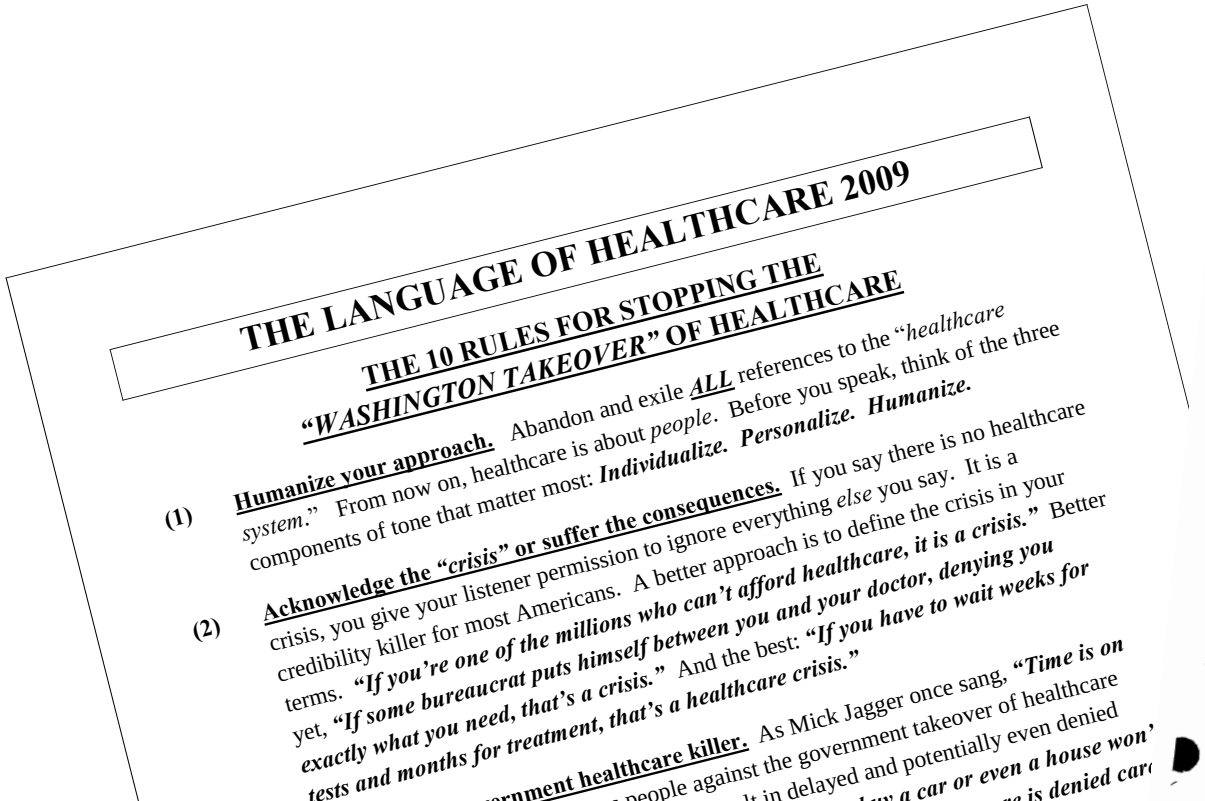
The Republican National Convention proved yet again that the GOP talks about America and U.S. policy with an entire unique vocabulary. It hasn't always been this way.

The Washington Post

However divided you think our politics are, this chart shows that it's actually way worse


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

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."



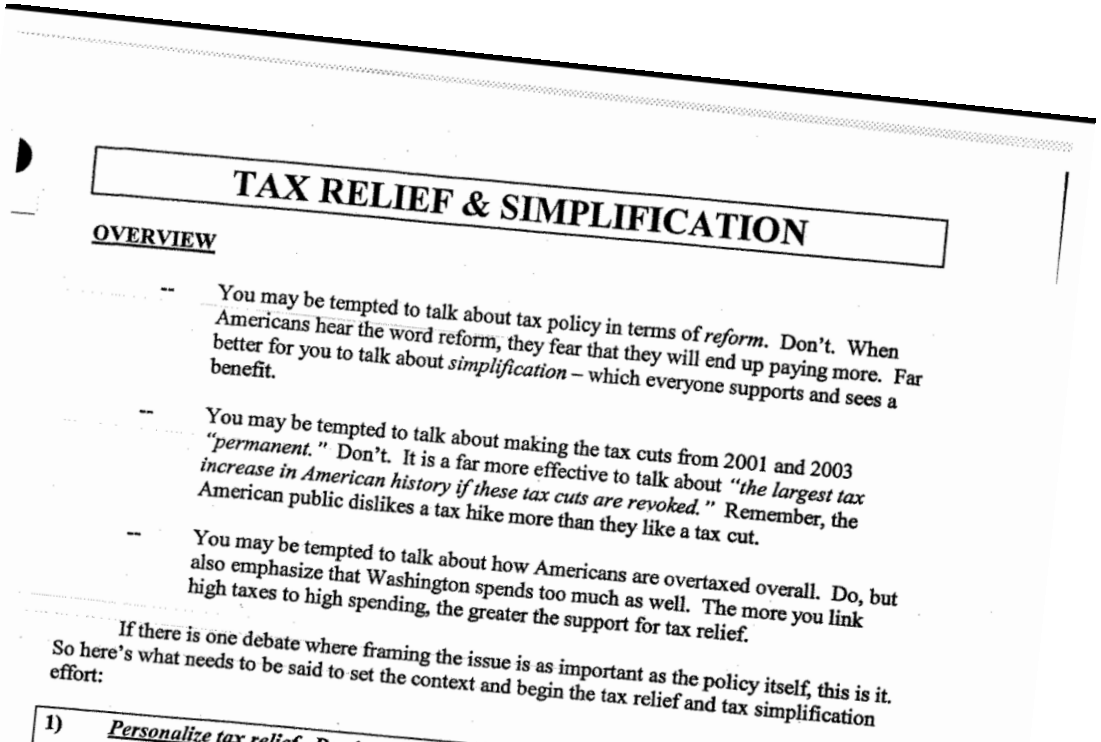
THE LANGUAGE OF HEALTHCARE 2009

THE 10 RULES FOR STOPPING THE "WASHINGTON TAKEOVER" OF HEALTHCARE

(1) **Humanize your approach.** 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.***

(2) **Acknowledge the "*crisis*" or suffer the consequences.** 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 bureaucrat puts 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."***

...rnment healthcare killer. As Mick Jagger once sang, ***"Time is on*** ...people against the government takeover of healthcare ...t in delayed and potentially even denied ...w a car or even a house won' ...e is denied car



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 a far more effective to talk about *"the largest tax increase in American history if these tax cuts are revoked."* 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 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) ***Personalize tax relief.***

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*)

Data

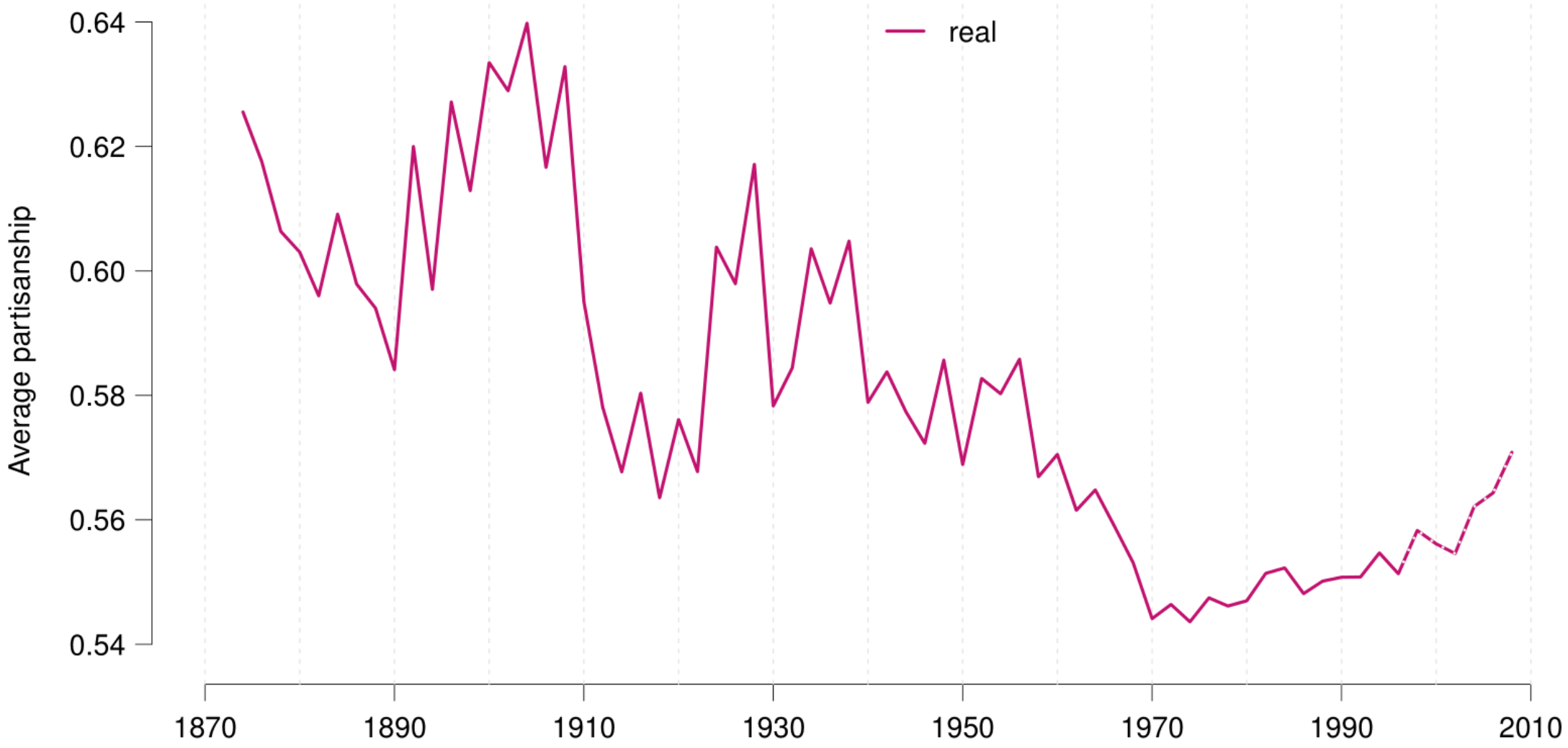
- US Congressional Record, 1873-2009
- Tokens (e.g. “war terror”) by speaker-session

Model

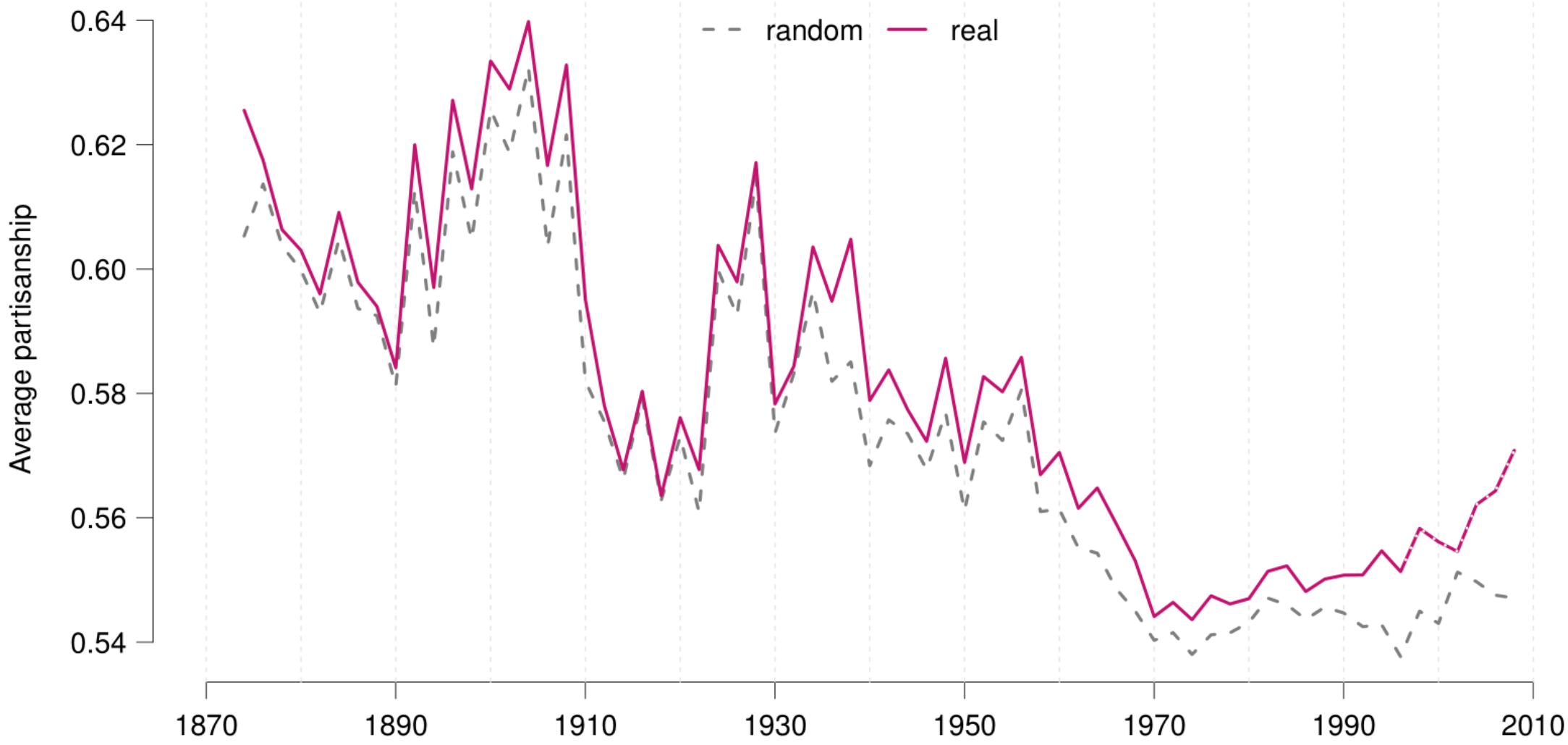
- A strategic model for language choice given party P
- Token counts are multinomial, with probabilities

$$\text{prob}(word\ j \text{ by } speaker\ i \text{ at } time\ t) = q_{tj}(\boldsymbol{x}_{it}, P_i)$$

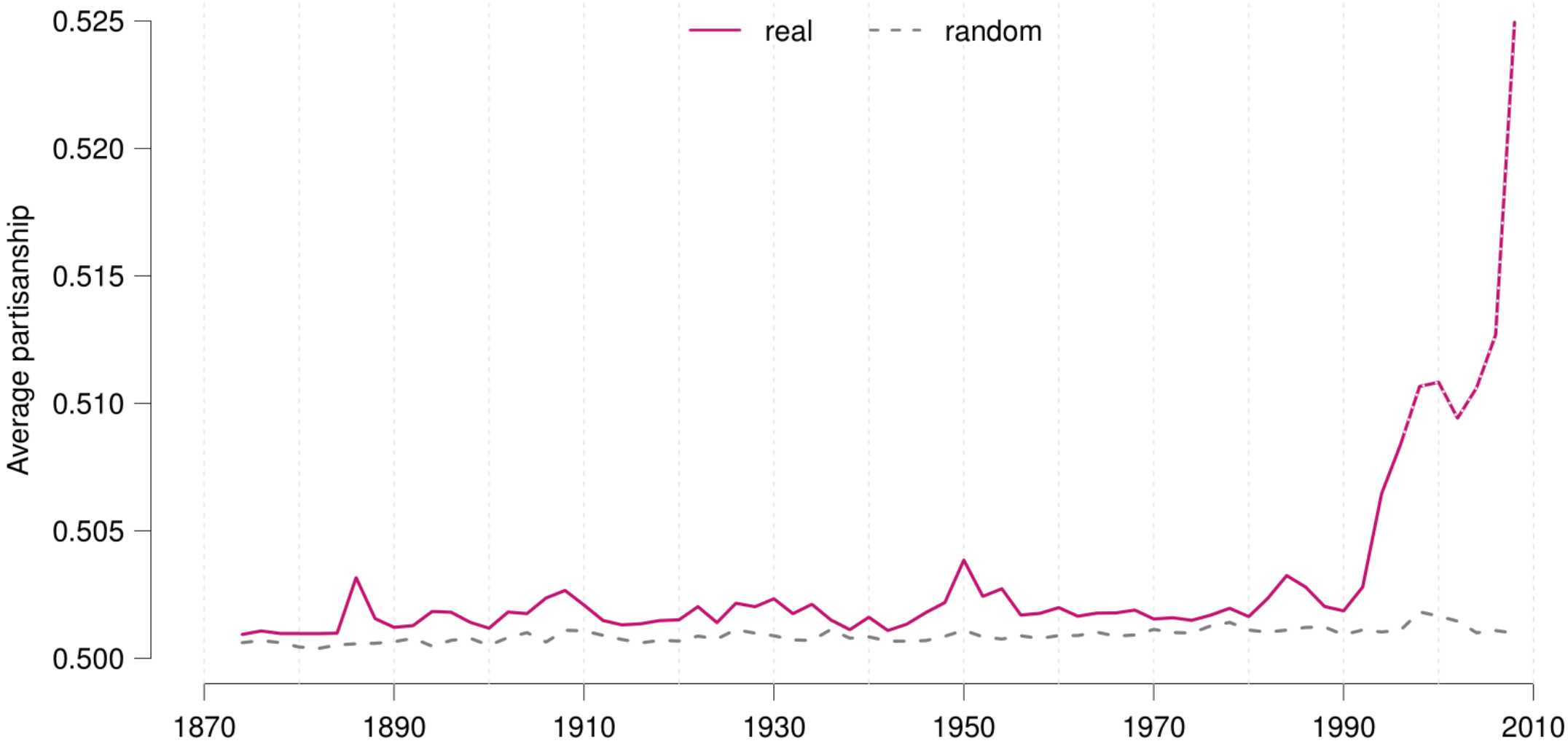
Naïve Estimation (MLE)



Naïve Estimation (MLE)



Machine Learning Estimator



Multinomial logit model: Each individual i in period t makes m_{it} choices over units j to maximize utility

$$\eta_{itj} + \varepsilon_{itj} = \alpha_{jt} + \mathbf{u}_{it}'\boldsymbol{\gamma}_{jt} + \boldsymbol{\varphi}_{jt}' r_{it} + \varepsilon_{ijt}$$

where:

- α_{jt} is unit-specific utility intercept
- $\mathbf{u}_{it}$ are covariates and $\boldsymbol{\gamma}_{jt}$ are associated loadings
- $r_{it} \in \{0, 1\}$ is an indicator for group membership and $\boldsymbol{\varphi}_{jt}$ are associated loadings.
- ε_{ijt} is T1EV random utility component

Alternatively, $\mathbf{c}_{it} \sim \mathrm{MN}(\mathbf{p}_{it}, m_{it})$ with $p_{itj} = e^{\eta_{itj}} / \sum_l e^{\eta_{itl}}$.

Distributed Multinomial Regression

We approximate the MN likelihood with *independent* Poissons:

$$c_{itj} \sim \text{Po}(\ m_{it} e^{\eta_{itj}}\ )$$

$\Rightarrow$ you can estimate each regression fully independently!

This works because MN dependence is *only induced by totals*.

DMR is equivalent to MN logit in a variety of simple examples, and is shown empirically to perform well in more complex settings.

Everything in distribution: estimation, penalization, selection ...

More precisely, start from the Poisson:

$$c_{ij} \overset{ind}{\sim} \mathrm{Pois}\left(\exp\left[\mu_i + \eta_{ij}\right]\right)$$

where μ_i is a 'verbosity' nuisance parameter.

This model leads to

$$\Pr\left(\mathbf{c}_i \mid m_i\right) = \frac{\prod_j \mathrm{Po}\left(c_{ij}; \exp\left[\mu_i + \eta_{ij}\right]\right)}{\mathrm{Po}\left(m_i; \sum_l \exp\left[\mu_i + \eta_{il}\right]\right)} = \mathrm{MN}(\mathbf{c}_i;\ \mathbf{q}_i, m_i)$$

Thus, given m_i, Poisson and MN imply the same model.
DMR fixes $\hat{\mu}_i = \log m_i$, so LHD factorizes to independent Poissons.

Big Data: focus computation on the bits that are hard to measure.

Penalization

We also place L_1 estimation penalties on key parameters.

Partisanship loadings are decomposed

$$\varphi_{jt} = \bar{\varphi}_j + \sum_{k=1}^{T} \tilde{\varphi}_{jk} \mathbb{1}_{t>k}$$

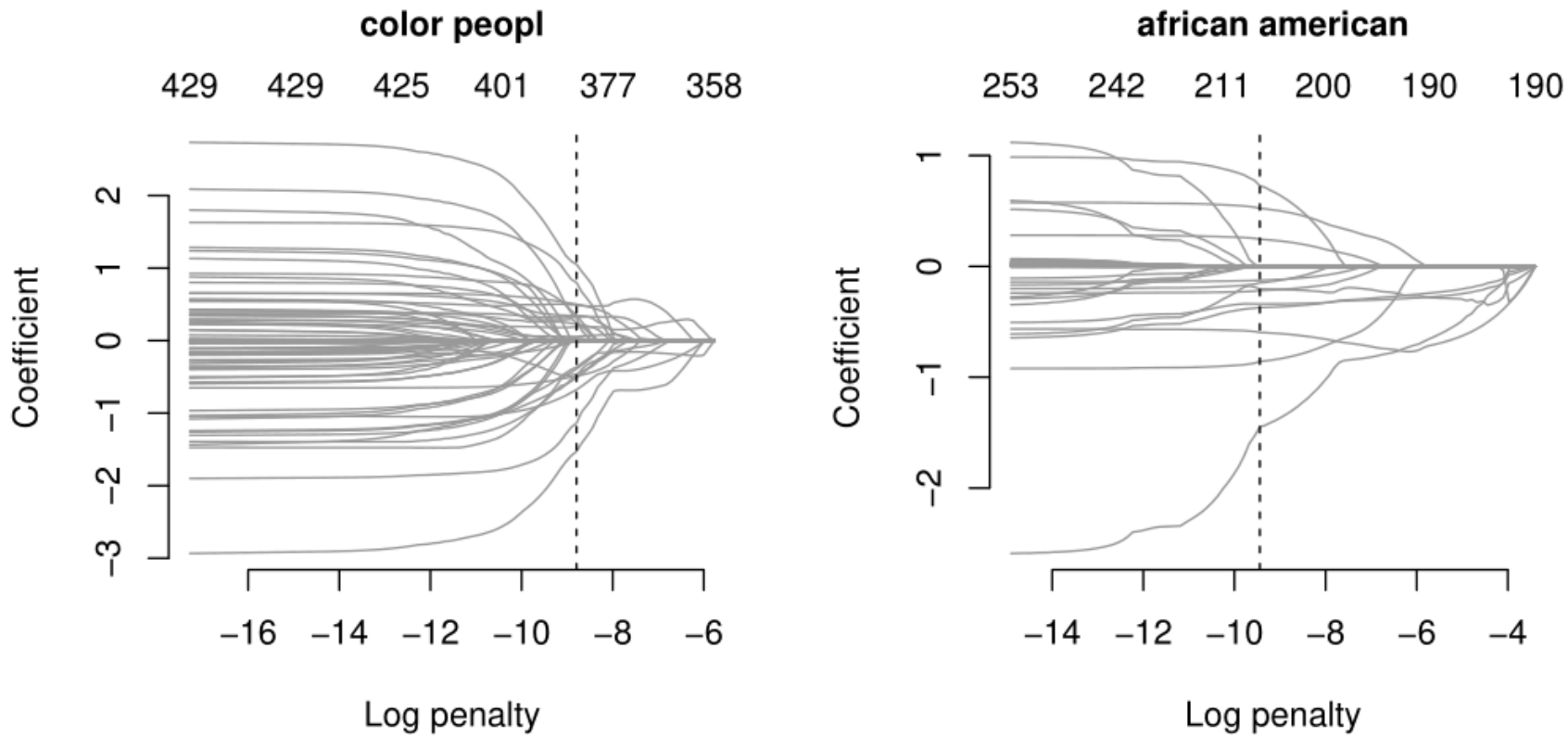
And the coefficients in this spline are penalized

$$c\left(\varphi_{tj}\right) = \lambda_j \left(|\bar{\varphi}_j| + \sum_k |\tilde{\varphi}_{jk}| \right)$$

We select λ_j using a BIC within each Poisson regression.

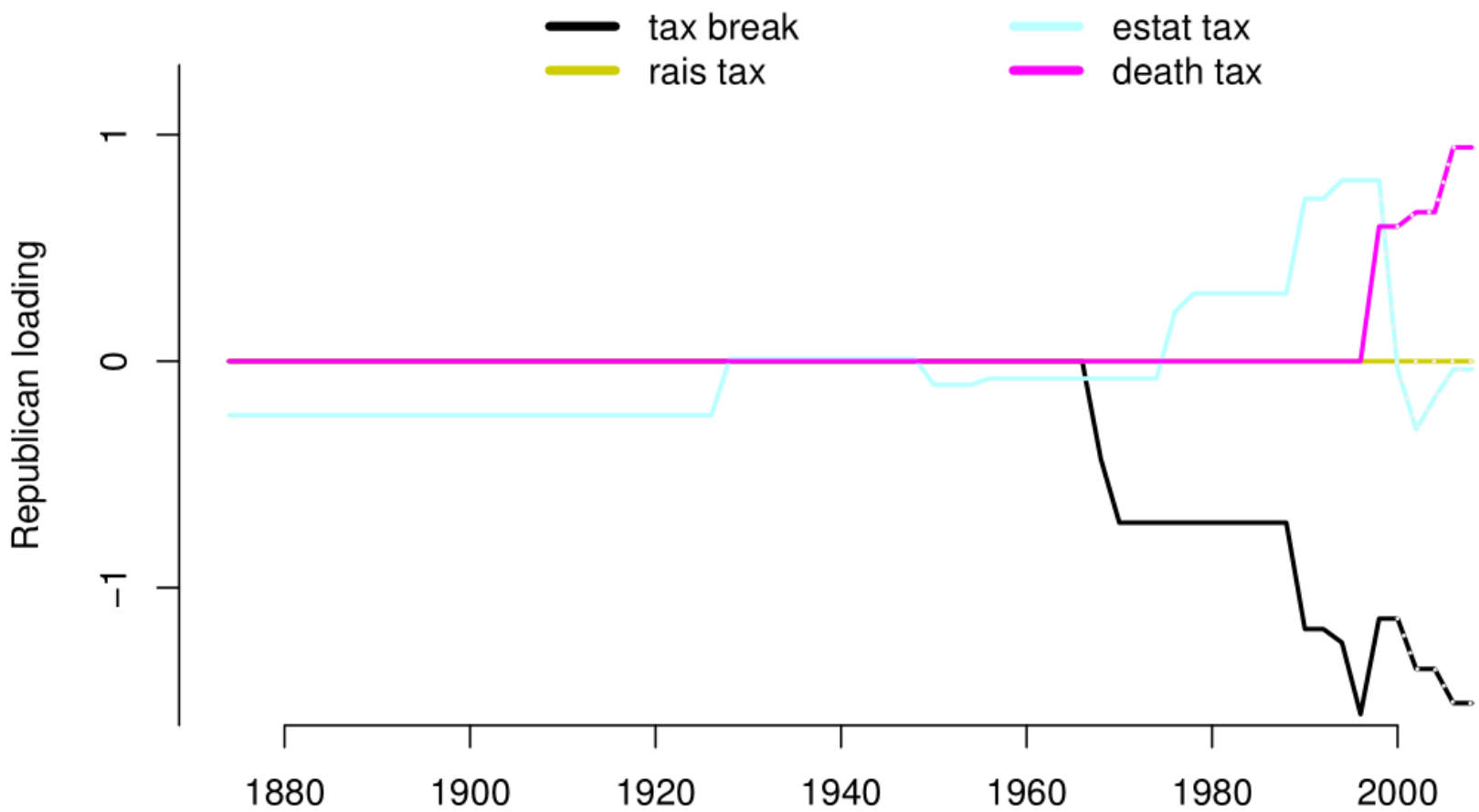
When $\mathbf{u}_{it}$ also gets really HD, we'll penalize elements of γ_{jt}.

Poisson regression regularization paths



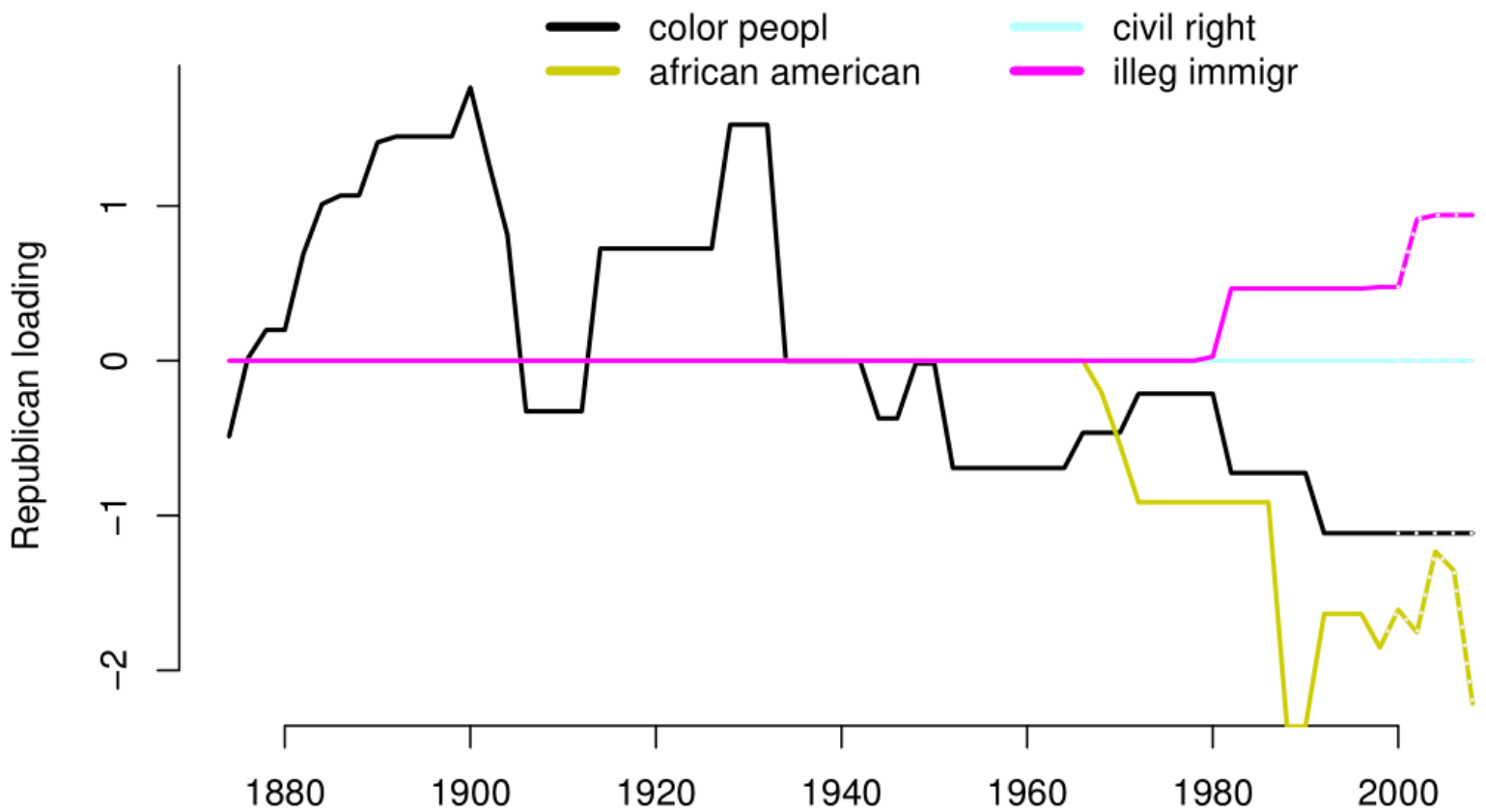
We've run a separate Poisson regression for each phrase.
The code runs a MapReduce routine using `dmr` for R.
BIC selection occurs within reducer, and is marked here.

Dynamic Phrase-Party Loadings: Tax



The resulting fit has φ_{tj} changing as a step function in t.

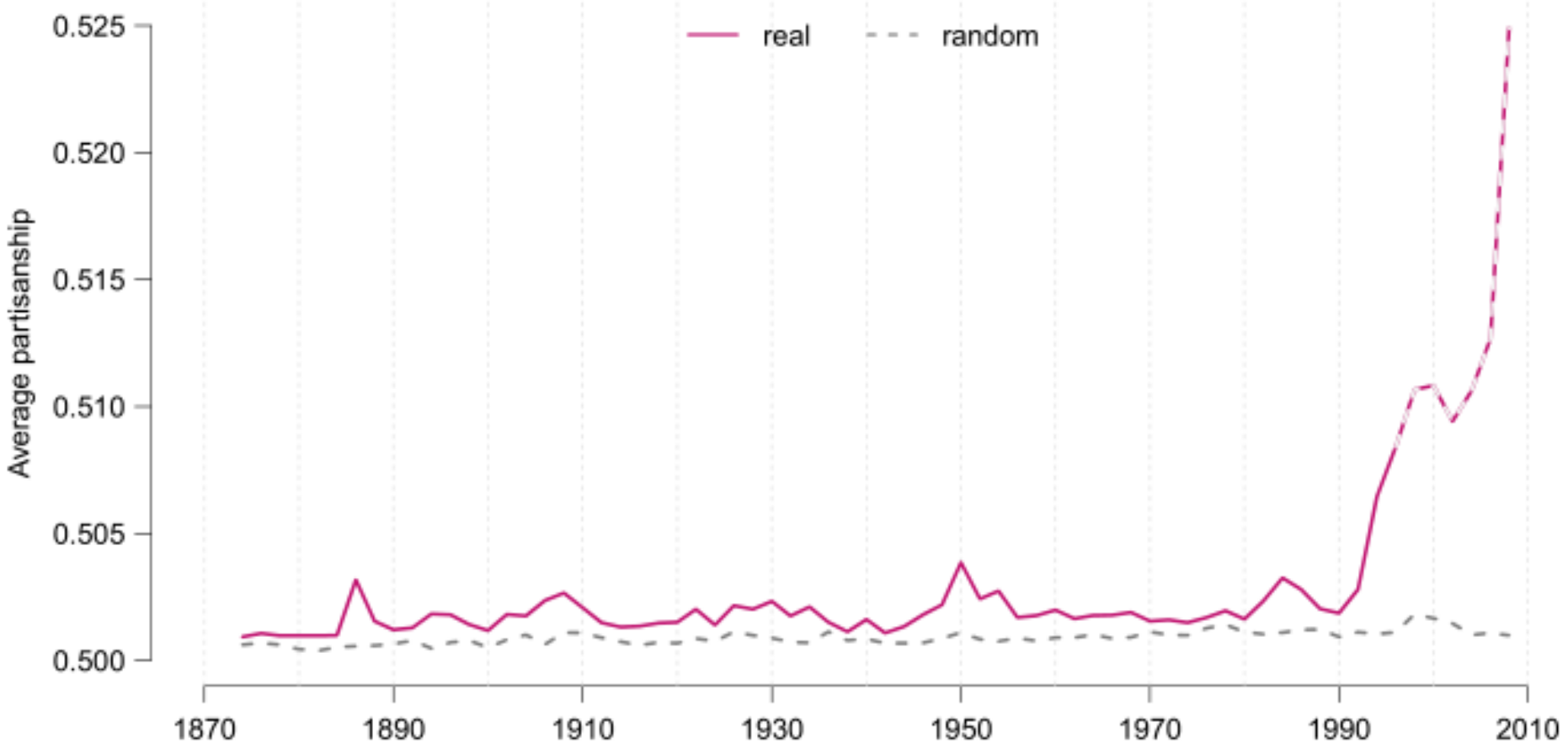
Dynamic Phrase-Party Loadings: Race



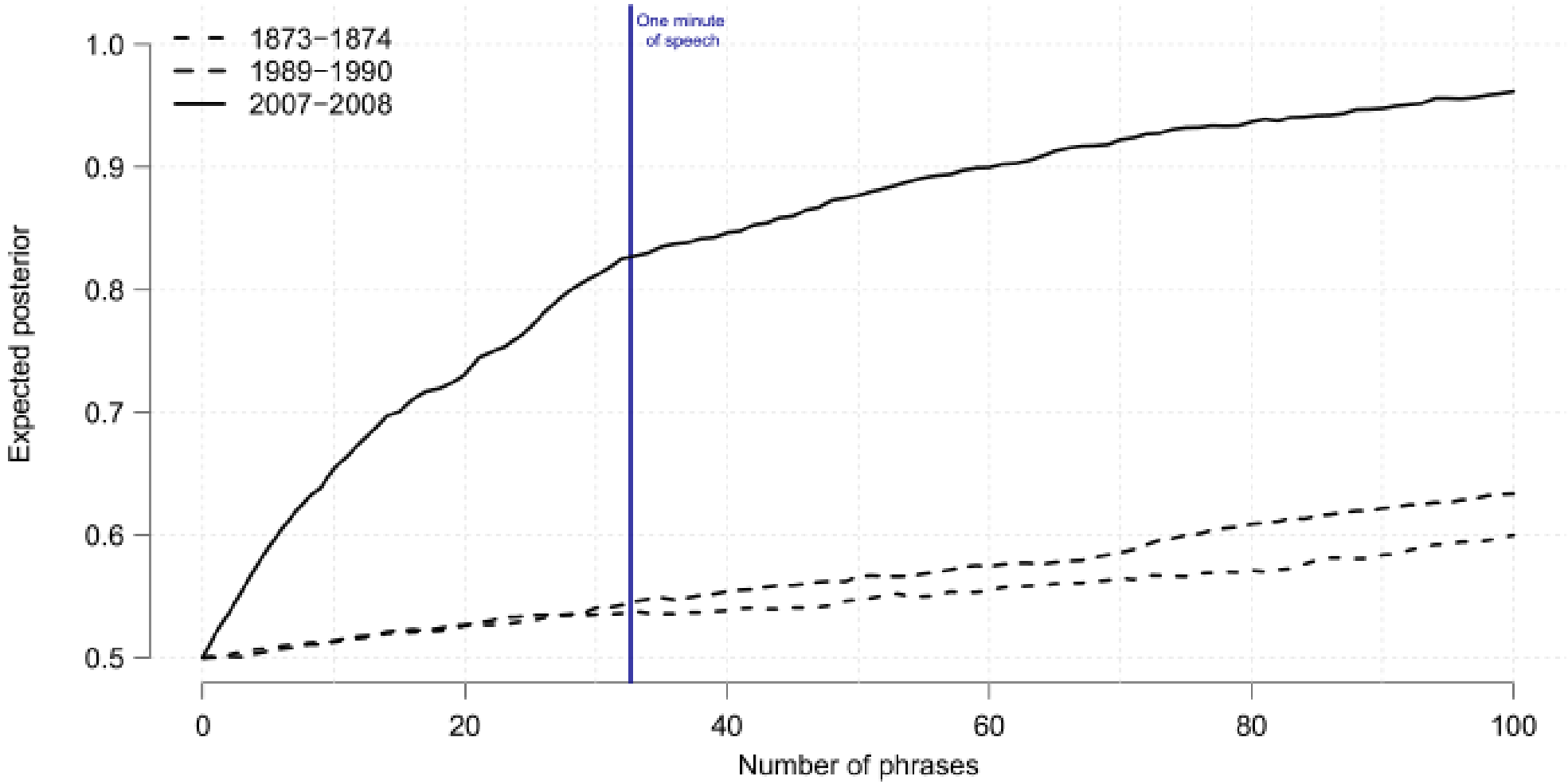
For this example, partisanship is robust to fixing $\varphi_{tj} = \varphi_j$.

Main Results

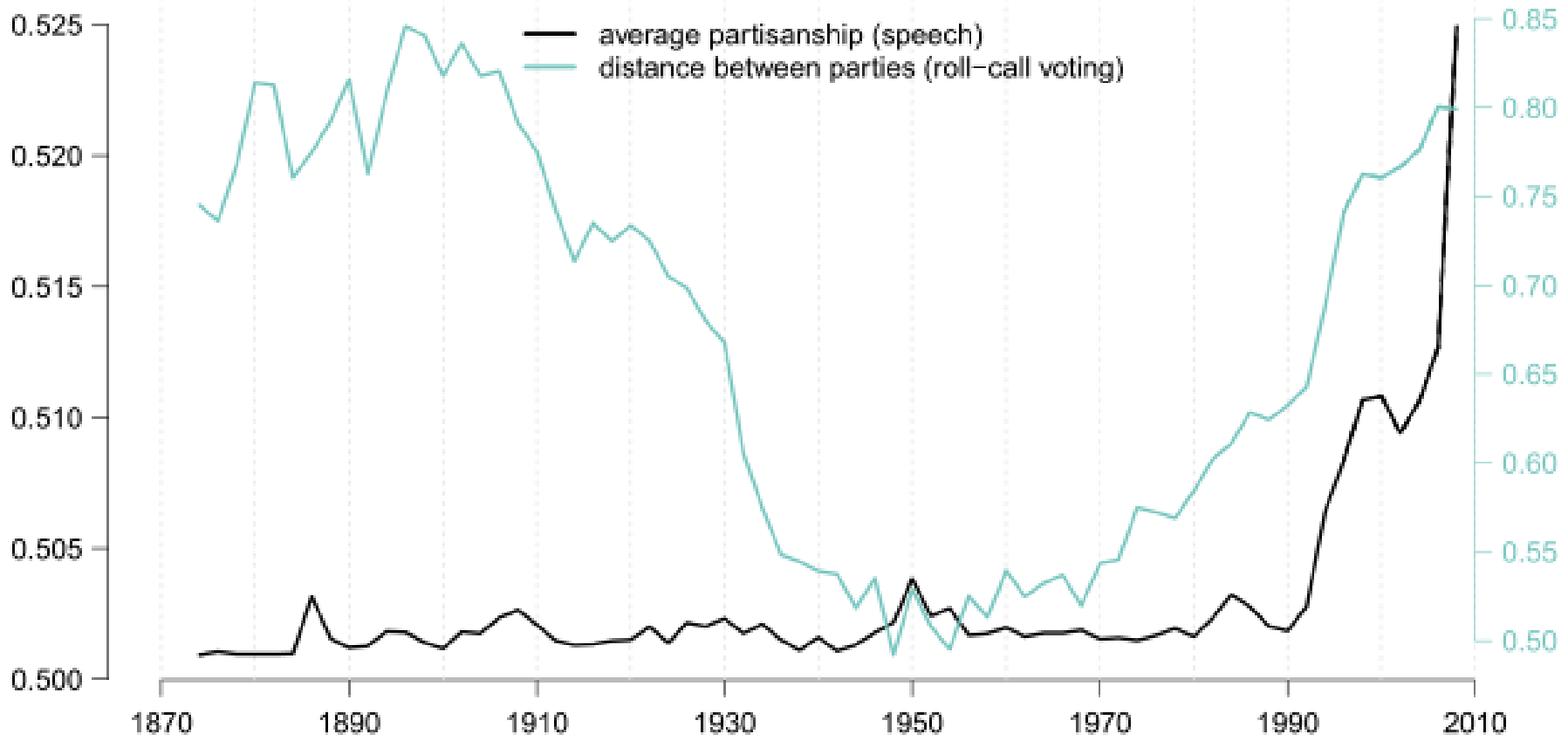
Our preferred estimator



Magnitude


1873–1874
1989–1990
2007–2008
One minute of speech
Expected posterior
1.0
0.9
0.8
0.7
0.6
0.5
0
20
40
60
80
100
Number of phrases

Comparison: rollcall votes



Most Partisan Phrases

60th Congress (1907-08)

Most Republican	Most Democratic
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

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

Most Partisan Phrases

60th Congress (1907-08)

Most Republican	Most Democratic
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

- 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*’.”

Most Partisan Phrases

100th Congress (1987-88)

Most Republican	Most Democratic
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

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

Most Partisan Phrases

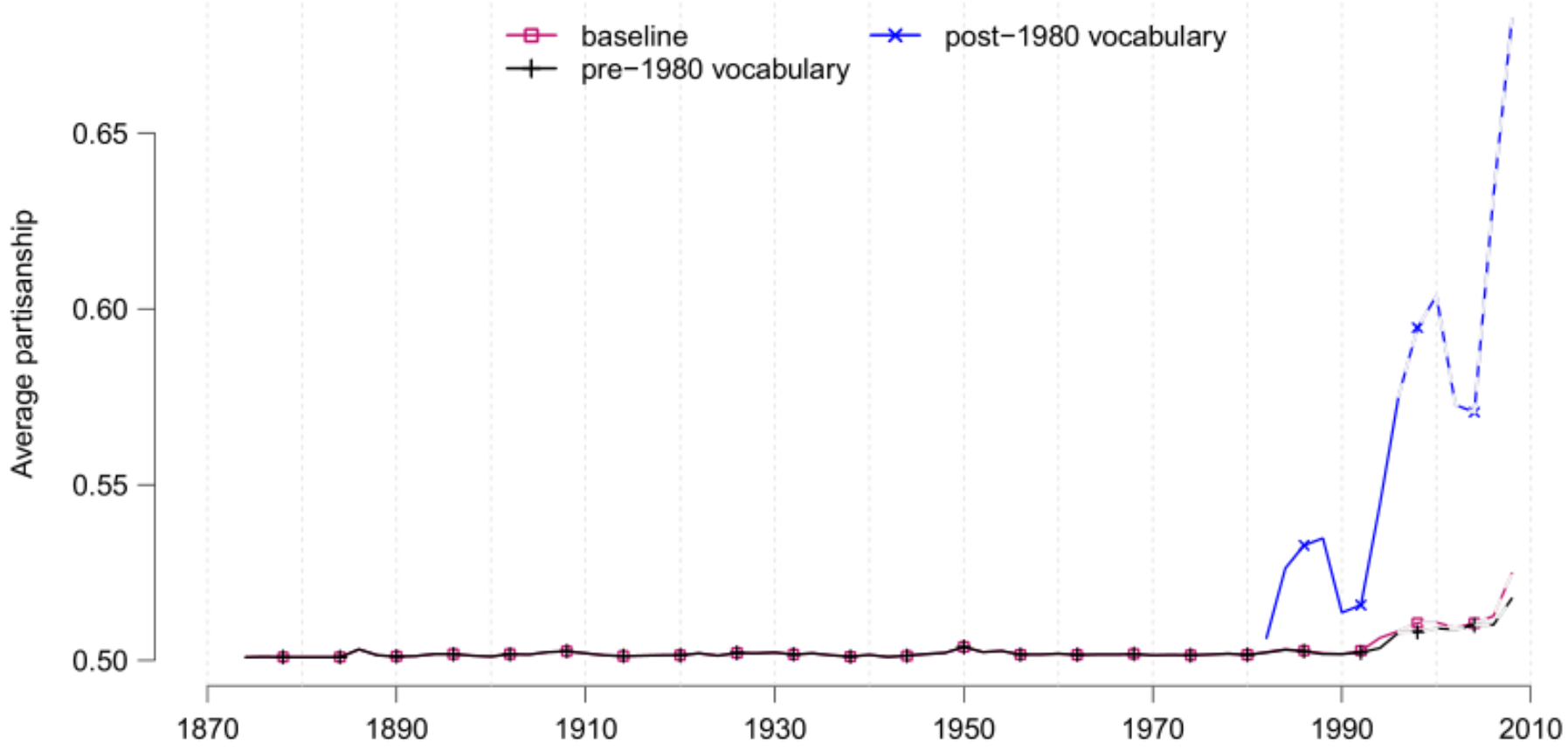
104th Congress (1995-96)

Most Republican	Most Democratic
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

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

Unpacking Partisanship

Neologisms


baseline
pre−1980 vocabulary
post−1980 vocabulary
Average partisanship
0.65
0.60
0.55
0.50
1870
1890
1910
1930
1950
1970
1990
2010

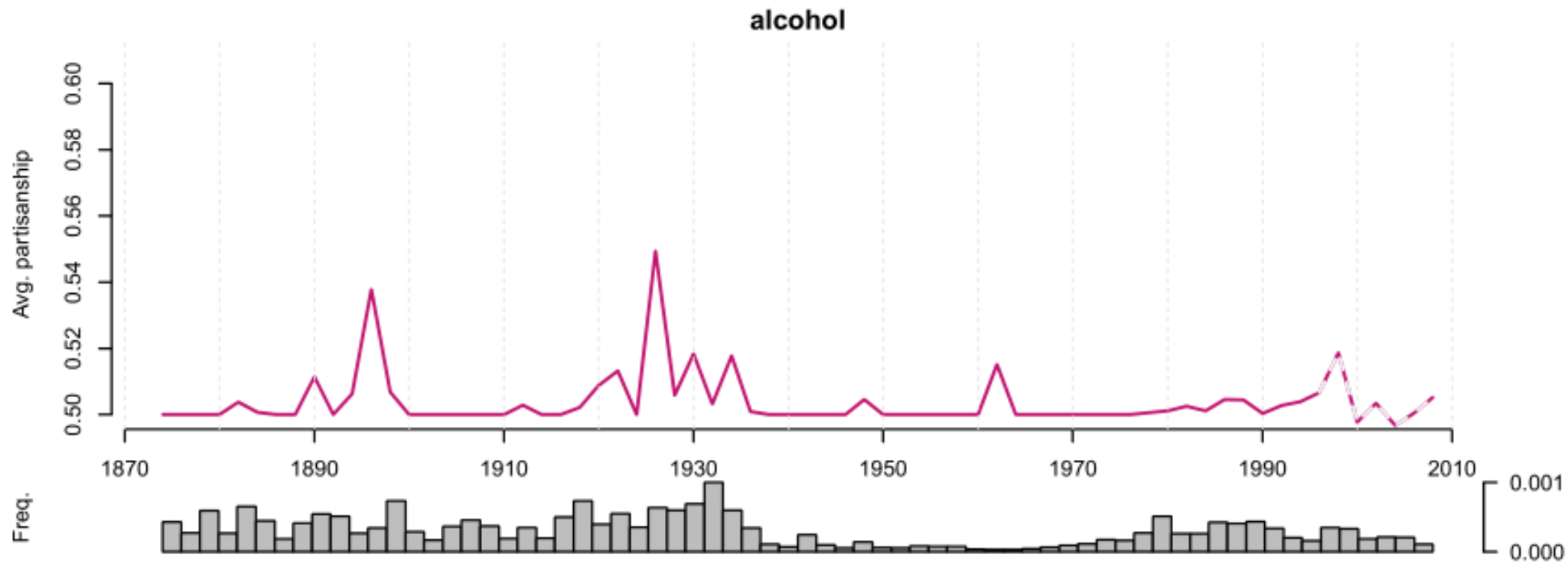
Topic decomposition

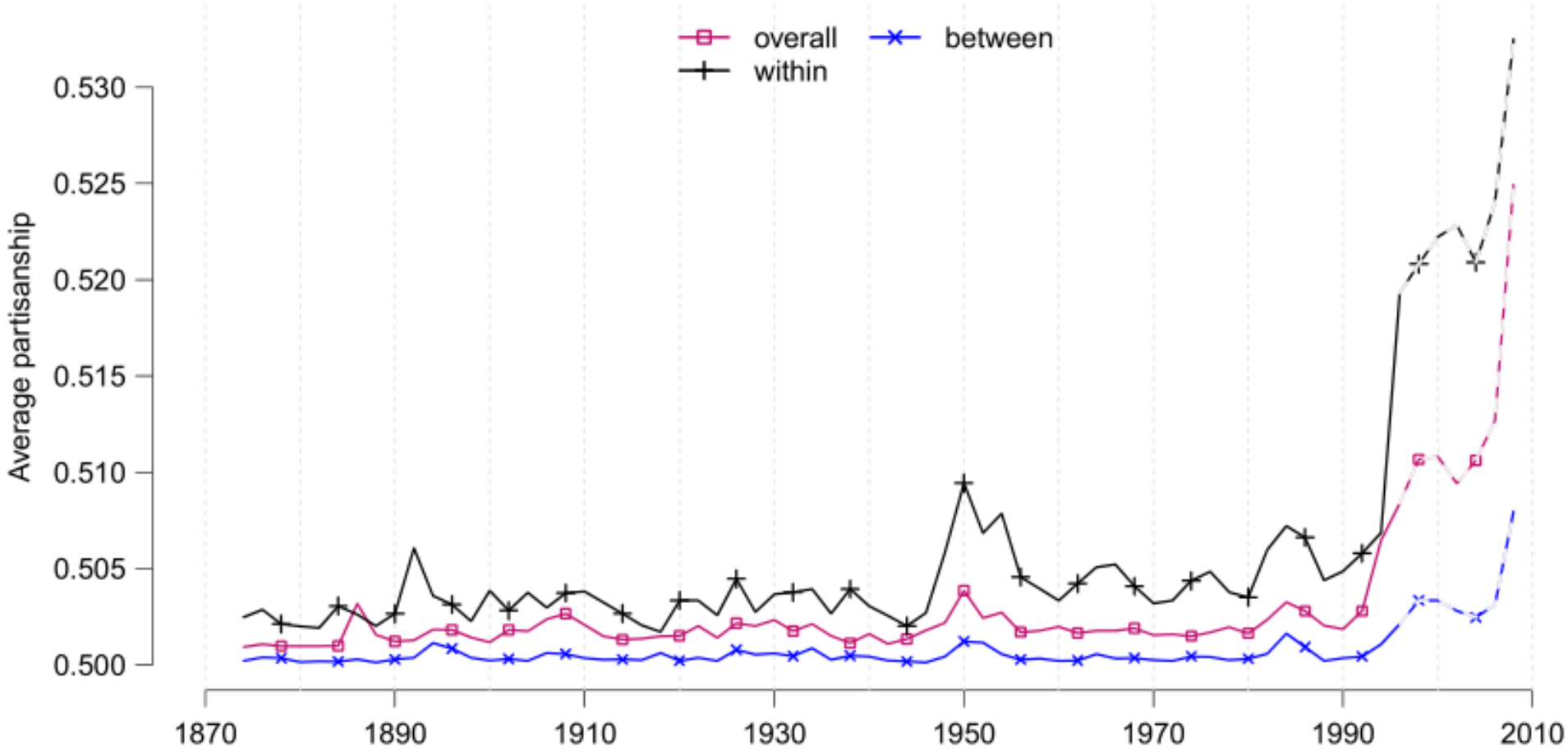
Are trends in partisanship driven by

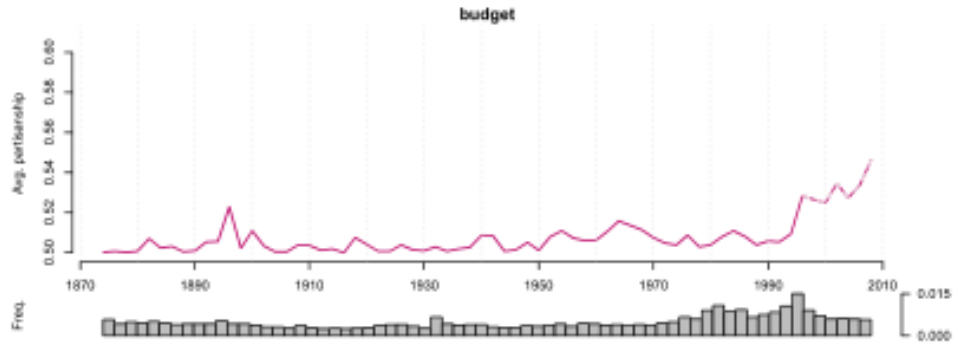
- *how parties talk about the same topic, or*
- *which topics they choose to emphasize?*

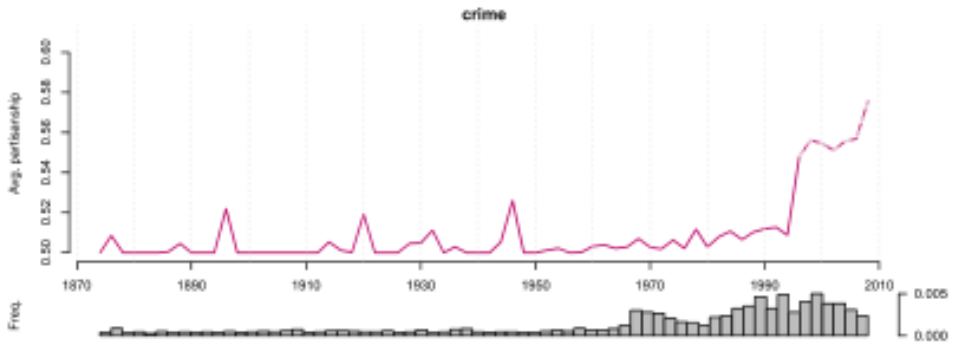
We look at some hand-coded topics…

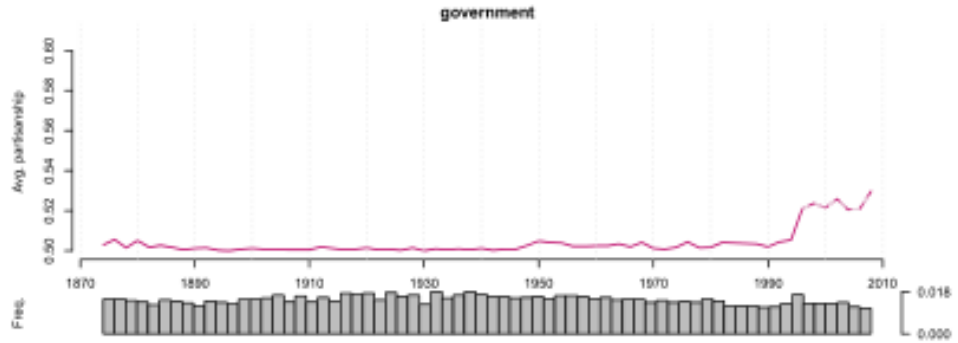
alcohol	environment	mail
budget	federalism	minorities
business	foreign	money
crime	government	religion
defense	health	tax
economy	immigration	trade
education	justice	
elections	labor	


alcohol
Avg. partisanship
0.50
0.52
0.54
0.56
0.58
0.60
1870
1890
1910
1930
1950
1970
1990
2010
Freq.
0.001
0.000

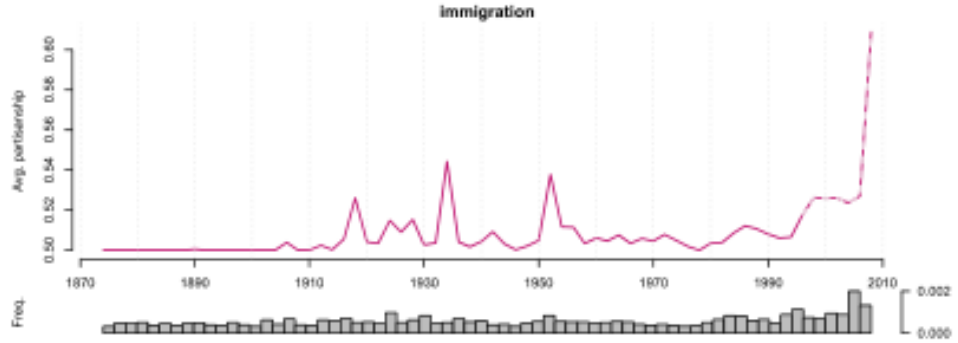

overall
between
within
Average partisanship
0.530
0.525
0.520
0.515
0.510
0.505
0.500
1870
1890
1910
1930
1950
1970
1990
2010


budget
Avg. partisanship
0.50
0.52
0.54
0.56
0.58
0.60
1870
1890
1910
1930
1950
1970
1990
2010
Freq.
0.015
0.000


crime
Avg. partisanship
0.50
0.52
0.54
0.56
0.58
0.60
1870
1890
1910
1930
1950
1970
1990
2010
Freq.
0.005
0.000


government
Avg. partisanship
0.50
0.52
0.54
0.56
0.58
0.60
1870
1890
1910
1930
1950
1970
1990
2010
Freq.
0.018
0.000

health
Avg. partisanship
0.50
0.52
0.54
0.56
0.58
0.60
1870
1890
1910
1930
1950
1970
1990
2010
Freq.
0.016
0.000


immigration
Avg. partisanship
0.50
0.52
0.54
0.56
0.58
0.60
1870
1890
1910
1930
1950
1970
1990
2010
Freq.
0.002
0.000

tax
Avg. partisanship
0.50
0.52
0.54
0.56
0.58
0.60
1870
1890
1910
1930
1950
1970
1990
2010
Freq.
0.011
0.000

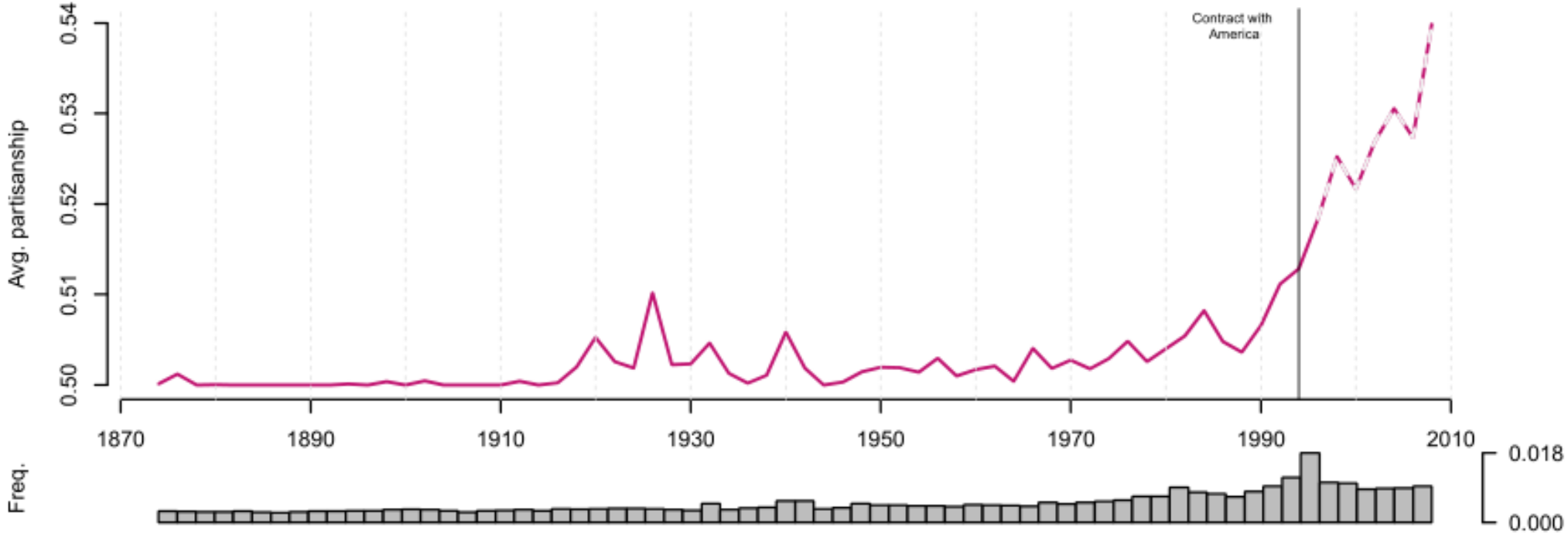
Explanations

Political Innovation

- **Contract with America** (1994)
  - Republicans take control of Congress for first time since 1952
  - Nationalized congressional election
  - Novel polling / focus group methods

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)*

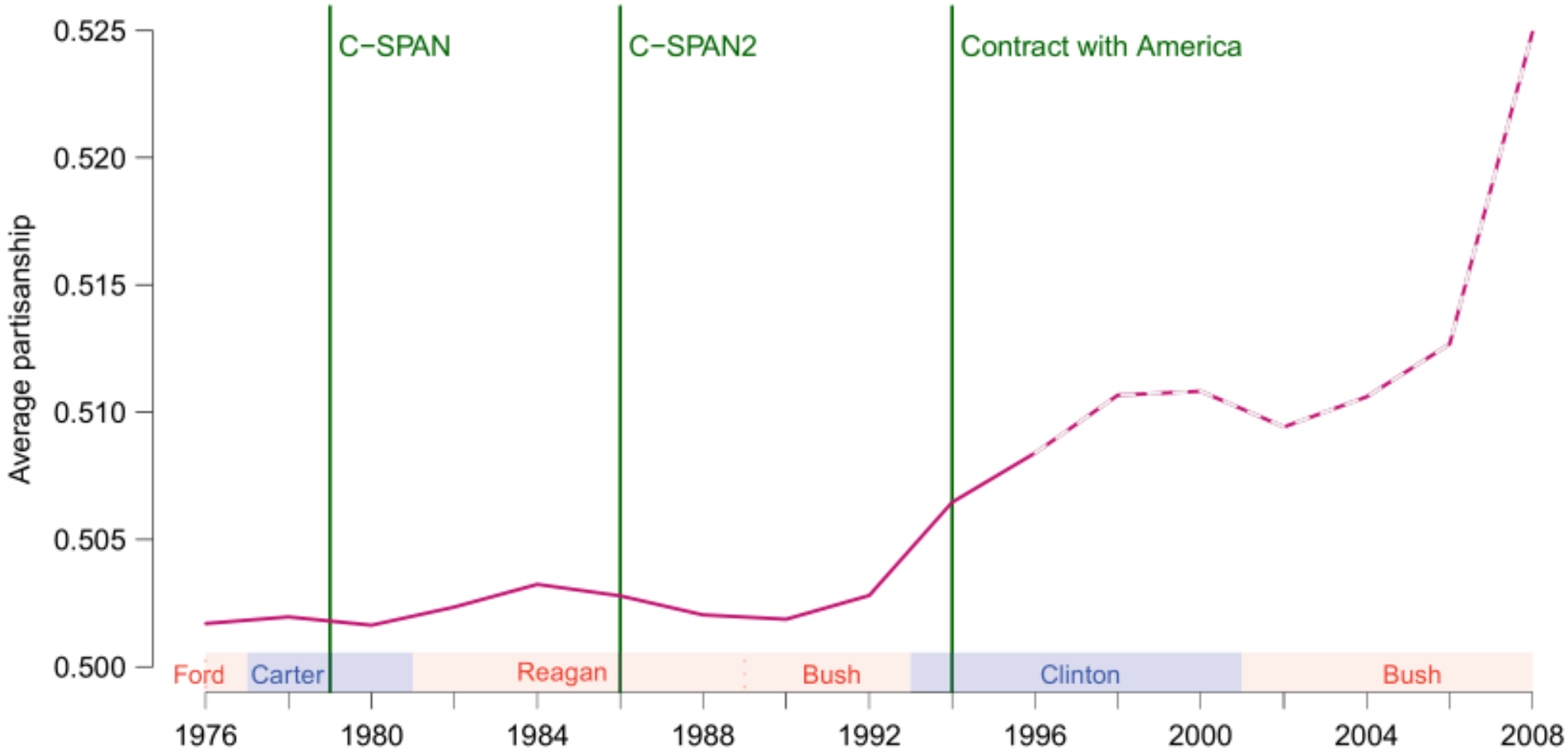
Phrases from the CWA



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)


C−SPAN
C−SPAN2
Contract with America
Average partisanship
0.525
0.520
0.515
0.510
0.505
0.500
Ford
Carter
Reagan
Bush
Clinton
Bush
1976
1980
1984
1988
1992
1996
2000
2004
2008

Conclusion: language matters

- 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)

Thanks!!