SIEPR Discussion Paper No. 16-028

Measuring Polarization in High-Dimensional Data: Method and Application to Congressional Speech

By

Matthew Gentzkow, Jesse M. Shapiro, Matt Taddy

Stanford Institute for Economic Policy Research
Stanford University
Stanford, CA 94305
(650) 725-1874

Measuring Polarization in High-Dimensional Data: Method and Application to Congressional Speech

Matthew Gentzkow, *Stanford University and NBER*[*]
Jesse M. Shapiro, *Brown University and NBER*
Matt Taddy, *Microsoft Research and Chicago Booth*

July 2016

Abstract

We study trends in the partisanship of Congressional speech from 1873 to 2009. We define partisanship to be the ease with which an observer could infer a congressperson's party from a fixed amount of speech, and we estimate it using a structural choice model and methods from machine learning. The estimates reveal that partisanship is far greater today than at any point in the past. Partisanship was low and roughly constant from 1873 to the early 1990s, then increased dramatically in subsequent years. Evidence suggests innovation in political persuasion beginning with the *Contract with America*, possibly reinforced by changes in the media environment, as a likely cause. Naive estimates of partisanship are subject to a severe finite-sample bias and imply substantially different conclusions.

[*]E-mail: gentzkow@stanford.edu, jesse_shapiro_1@Brown.edu, taddy@microsoft.com. We acknowledge funding from the Initiative on Global Markets and the Stigler Center at Chicago Booth, the National Science Foundation, and Stanford Institute for Economic Policy Research (SIEPR). We thank numerous seminar audiences and our many dedicated research assistants for their contributions to this project.

1 Introduction

America's two political parties speak different languages. Democrats talk about "estate taxes," "undocumented workers," and "tax breaks for the wealthy," while Republicans refer to "death taxes," "illegal aliens," and "tax reform." The 2010 Affordable Care Act was "comprehensive health reform" to Democrats and a "Washington takeover of health care" to Republicans. Within hours of the 2016 killing of 49 people in a nightclub in Orlando, Democrats were calling the event a "mass shooting"—linking it to the broader problem of gun violence—while Republicans were calling it an act of "radical Islamic terrorism"—linking it to concerns about national security and immigration.[1] Partisan language diffuses into media coverage (Gentzkow and Shapiro 2010; Martin and Yurukoglu 2016) and other domains of public discourse (Greenstein and Zhu 2012; Jensen et al. 2012). Experiments and surveys show that partisan framing can have large effects on public opinion (Nelson et al. 1997; Graetz and Shapiro 2006; Chong and Druckman 2007), and language is one of the most basic determinants of group identity (Kinzler et al. 2007).

Is today's partisan language a new phenomenon? In one sense, the answer is clearly no: one can easily find examples of partisan terms in America's distant past.[2] Yet the magnitude of the differences, the deliberate strategic choices that seem to underlie them, and the expanding role of consultants, focus groups, and polls (Bai 2005; Luntz 2006; Issenberg 2012) suggest that what we see today might represent a consequential change (Lakoff 2003). If the language of politics is more partisan today than in the past, it could be contributing to deeper polarization and cross-party animus, both in Congress and in the broader public.

In this paper, we apply tools from structural estimation and machine learning to study the partisanship of language in the US Congress from 1873 to 2009. We specify a discrete-choice model of speech in which political actors choose phrases to influence an audience. We define the overall partisanship of speech in a given period to be the ease with which an observer could guess

[1] The use of individual phrases such as "estate taxes" and "undocumented workers" is based on our analysis of congressional speech data below. For discussion of the Affordable Care Act, see Luntz (2009) and Democratic National Committee (2016). For discussion of the Orlando shooting, see Andrews and Buchanan (2016).

[2] In the 1946 essay "Politics and the English Language," George Orwell discusses the widespread use of political euphemisms (Orwell 1946). Northerners referred to the American Civil War as the "War of the Rebellion" or the "Great Rebellion," while southerners called it the "War for Southern Independence" or, in later years, the "War of Northern Aggression" (McCardell 2004). The bulk of the land occupied by Israel during the Six-Day War in 1967 is commonly called the "West Bank," but some groups within Israel prefer the name "Judea and Samaria," which evokes historical Biblical connections (Newman 1985). The rebels fighting the Sandinista government in Nicaragua were commonly called "Contras," but were referred to as "freedom fighters" by Ronald Reagan and Republican politicians who supported them (Peace 2010).

a speaker's party based solely on the speaker's choice of words. We estimate the model using the text of speeches from the *United States Congressional Record*.

To compute an accurate estimate of partisanship, we must grapple with two methodological challenges. First, natural plug-in estimators of our model suffer from severe finite-sample bias. The bias arises because the number of phrases a speaker could choose is large relative to the total amount of speech we observe, meaning many phrases are said mostly by one party or the other purely by chance. Second, although our discrete-choice model takes a standard multinomial logit form, the large number of choices and parameters makes standard approaches to estimation computationally infeasible. We address these challenges by using an L_1 or lasso-type penalty on key model parameters to control bias, and a Poisson approximation to the multinomial logit likelihood to permit distributed computing.

We find that the partisanship of language has exploded in recent decades, reaching an unprecedented level. From 1873 to the early 1990s, partisanship was roughly constant and fairly small in magnitude: in the 43rd session of Congress (1873-75), the probability of correctly guessing a speaker's party based on a one-minute speech was 54 percent; by the 101st session (1989-1990) this figure had increased to 55 percent. Beginning with the congressional election of 1994, partisanship turned sharply upward, with the probability of guessing correctly based on a one-minute speech climbing to 83 percent by the 110th session (2007-09). Methods that do not correct for finite-sample bias—including both the maximum likelihood estimator of our model and estimates reported by Jensen et al. (2012)—imply instead that partisanship is no higher today than in the past, and that the recent upward trend is not unusual by historical standards.

We unpack the recent increase in partisanship along a number of dimensions. The most partisan phrases in each period—defined as those phrases whose removal from the vocabulary would most reduce partisanship—align well with the issues emphasized in party platforms and, in recent years, include well-known partisan phrases like those mentioned above. Manually classifying phrases into substantive topics shows that the increase in partisanship is due more to changes in the language used to discuss a given topic (e.g., "estate tax" vs. "death tax") than to changes in the topics parties emphasize (e.g., Republicans focusing more on taxes and Democrats focusing more on labor issues). The topics that show the sharpest increase in recent years include taxes, immigration, crime, health care, and the scope of government. Separating out phrases that first appear in the vocabulary after 1980, we find that such "neologisms" exhibit a particularly dramatic rise in partisanship, though they are not the sole driver of partisanship in the main series. Comparing

our measure to a standard measure of polarization based on roll-call votes, we find that the two are correlated in the cross section but exhibit very different dynamics in the time series.

While we cannot say definitively why partisanship of language increased when it did, the evidence points to innovation in political persuasion as a proximate cause. The 1994 inflection point in our series coincides precisely with the Republican takeover of Congress led by Newt Gingrich, under a platform called the *Contract with America*. This election is widely considered a watershed moment in political marketing, as consultants such as Frank Luntz applied novel focus group technologies to identify effective language and disseminate it broadly to candidates (Luntz 2004; Bai 2005; Lakoff 2004). Consistent with this, we show that phrases from the text of the *Contract with America* see a spike in usage in 1994, and then exhibit a particularly strong upward trend in partisanship. As a related factor, the years leading up to this inflection point had seen important changes in the media environment: the introduction of television cameras as a permanent presence in the chamber, the live broadcast of proceedings on the C-SPAN cable channels, and the rise of partisan cable and the twenty-four hour cable news cycle. Prior work suggests that these media changes strengthened the incentive to engineer language and impose party discipline on floor speeches, and made the new attention to language more effective than it would have been in earlier years (Frantzich and Sullivan 1996).

The methods we develop here can be applied to a broad class of problems in which the goal is to characterize the polarization or segregation of choices in high-dimensional data. Examples include measuring residential segregation across small geographies, polarization of web browsing or social media behavior, and between-group differences in consumption. Whenever the number of possible choices is large relative to the number of actual choices observed, naive estimates will tend to suffer from finite-sample bias similar to the one we have documented for speech, and our penalized estimator can provide an accurate and computationally feasible solution.

Our analysis relates most closely to recent work by Jensen et al. (2012). They use text from the *Congressional Record* to characterize the partisanship of language from the late nineteenth century to the present, applying a plug-in estimator of partisanship based on the observed correlation of phrases with party labels. They conclude that partisanship has been rising recently but was even higher in the past. We apply a new method that addresses finite-sample bias and leads to substantially different conclusions.

Methodologically, our work is most related to the literature on residential segregation, which is surveyed in Reardon and Firebaugh (2002). The finite-sample bias we highlight has been noted in

that context by Cortese et al. (1976) and Carrington and Troske (1997). Recent work has derived axiomatic foundations for segregation measures (Echenique and Fryer 2007; Frankel and Volij 2011), asking which measures of segregation satisfy certain intuitive properties.[3] Our approach is, instead, to specify a generative model of the data and to measure segregation using objects that have a well-defined meaning in the context of the model.[4] To our knowledge, ours is the first paper to estimate group differences based on preference parameters in a structural model.[5] It is also the first to use a penalization scheme to address the finite-sample bias arising in segregation measurement, which has previously been addressed by benchmarking against random allocation (Carrington and Troske 1997), applying asymptotic or bootstrap bias corrections (Allen et al. 2015), and estimating mixture models (Rathelot 2012). Within the literature on measuring document partisanship (e.g., Laver et al. 2003; Gentzkow and Shapiro 2010; Kim et al. 2015), our approach is closest to that of Taddy (2013), but unlike Taddy (2013), we allow for a rich set of covariates and we target faithful estimation of partisanship rather than classification performance.[6]

Substantively, our findings speak to a broader literature on trends in political polarization. A large body of work builds on the ideal point model of Poole and Rosenthal (1985) to analyze polarization in congressional roll-call votes, finding that inter-party differences fell from the late nineteenth to the mid-twentieth century, and have increased steadily since (McCarty et al. 2015). We show that the dynamics of polarization in language are very different, suggesting that language is a distinct dimension of party differentiation.[7]

The next sections introduce our data, model, and approach to estimation. We then present our main estimates, along with evidence that unpacks the recent increase in partisanship. In the following section, we discuss possible explanations for this change. We conclude by considering the wider implications of increasing partisanship and discussing other applications of our method.

[3] See also Mele (2013) and Ballester and Vorsatz (2014). Our measure is also related to measures of cohesiveness in preferences of social groups, as in Alcalde-Unzu and Vorsatz (2013).

[4] In this respect, our paper builds on Ellison and Glaeser (1997), who use a model-based approach to measure agglomeration spillovers in US manufacturing.

[5] Mele (2015) shows how to estimate preferences in a random-graph model of network formation and measures the degree of homophily in preferences. Bayer et al. (2002) use an equilibrium model of a housing market to study the effect of changes in preferences on patterns of residential segregation. Fossett (2011) uses an agent-based model to study the effect of agent preferences on the degree of segregation.

[6] More broadly, our paper relates to work in statistics on authorship determination (Mosteller and Wallace 1963), work in economics that uses text to measure the sentiment of a document (e.g., Antweiler and Frank 2004; Tetlock 2007), and work that classifies documents according to similarity of text (Blei and Lafferty 2007; Grimmer 2010).

[7] A related literature considers polarization among American voters, with most measures offering little support for the widespread view that voters are more polarized today than in the past (Fiorina et al. 2005; Fiorina and Abrams 2008; Glaeser and Ward 2006). An exception is measures of inter-party dislike or distrust, which do show a sharp increase in recent years (Iyengar et al. 2012).

2 Congressional speech data

Our primary data source is the complete non-extension text of the proceedings of the *United States Congressional Record* from the 43rd to 110th Congresses. The 43rd Congress was the first congressional session to be covered in its entirety by the *Congressional Record*.

The *Record* is printed by the U.S. Government Publishing Office (GPO). It was preceded by a publication called the *Congressional Globe*, which was printed privately through a contract with Congress from 1855 through 1872, initially on a weekly basis and then daily following an 1865 law (Amer 1993). The *Globe*, in turn, evolved from a series of predecessor publications, beginning in 1824 with the *Register of Debates,* which was in precis rather than verbatim format (Amer 1993).

We obtained the text for the 43rd to 104th Congresses (hereafter, "sessions") from Lexis-Nexis (LN), who performed optical character recognition on scanned print volumes. For the 104th to 110th sessions, we obtained the text from the website of the GPO (2011).[8] Throughout the paper, we estimate partisanship (and other objects) in the 104th session separately for each data source, and add the resulting difference to the entire GPO series so that the two series agree in the overlapping session. Our plots use dashed lines to indicate the portions of a series that have been adjusted in this way.

We use an automated script to parse the raw text into individual speeches. The script identifies when a new speech begins by looking for a declaration of the speaker's identity (e.g., "Mr. ALLEN of Illinois."). We obtain speaker attributes (e.g., party) by matching speakers to members of Congress in the Database of Congressional Historical Statistics (Swift et al. 2009) or the Voteview Roll Call Data (McCarty et al. 2009).[9] We include only Republicans and Democrats in our analysis.

We follow Gentzkow and Shapiro (2010) in pre-processing the text to increase the ratio of signal to noise. First, we remove stopwords and reduce words to their stems using the tools defined in Porter (2009). Second, we group words into two-word phrases. Third, we remove phrases that are likely to be procedural according to the definition in appendix A. Fourth, we remove phrases

[8] We obtained information on the dates associated with each congressional session from United States Senate (2013).

[9] The Voteview data include metadata from Martis (1989). We obtain state-level geographic data from United States Census Bureau (1990) and Ruggles et al. (2008). We use a speaker's name, chamber, state, district and gender to match the speaker to a corresponding member of Congress, manually resolving spelling mistakes and other sources of error when possible. We exclude speeches by speakers that we cannot successfully match to a unique member of Congress, as well as speeches made by speakers identified by office rather than name (e.g., the Speaker of the House). The online appendix presents statistics on the share of speeches successfully matched to a member of Congress in each session.

that include the name of a congressperson or state, or that consist of numbers, symbols, or a few other objects with low semantic meaning. Fifth, we restrict attention to phrases spoken at least 10 times in at least one session, at least 100 times across all sessions, and in at least 10 unique speaker-sessions.

The resulting vocabulary contains 529,980 unique two-word phrases spoken a total of 297 million times by 7,254 unique speakers. We analyze data at the level of the speaker-session, of which there are 33,373.[10] The online appendix reports additional summary statistics for our estimation sample.

We manually classify a subset of phrases into 22 non-mutually exclusive topics as follows. We begin with a set of partisan phrases which we group into 22 topics (e.g., taxes, defense, etc.).[11] For each topic, we create a set of keywords consisting of relevant words contained in one of the categorized phrases, plus a set of additional manually included words. Finally, we identify all phrases in the vocabulary that include one of the topic keywords, that are used more frequently than a given, topic-specific occurrence threshold, and that are not obvious false matches. The online appendix lists, for each topic, the keywords, the occurrence threshold, and a random sample of included and excluded phrases.

3 Model and measure of partisanship

3.1 Probability model

Let $\mathbf{c}_{it}$ be the J-vector of phrase counts for speaker i in session t, with $m_{it} = \sum_j c_{itj}$ denoting the total amount of speech by speaker i in session t. Let $P(i) \in \{R, D\}$ denote the party affiliation of speaker i, and let $D_t = \{i : P(i) = D, m_{it} > 0\}$ and $R_t = \{i : P(i) = R, m_{it} > 0\}$ denote the set of Democrats and Republicans, respectively, active in session t. Let $\mathbf{x}_{it}$ be a $K-$vector of (possibly time-varying) speaker characteristics.

We assume that:

$$\mathbf{c}_{it} \sim \text{MN}\left(m_{it}, \mathbf{q}_t^{P(i)}(\mathbf{x}_{it})\right), \tag{1}$$

where $\mathbf{q}_t^P(\mathbf{x}_{it}) \in (0,1)$ for all P, i, and t. The speech-generating process is fully characterized by the verbosity m_{it} and the probability $\mathbf{q}_t^P(\cdot)$ of speaking each phrase.

[10] In the rare cases in which a speaker switched chambers in a single session (usually from the House to the Senate), we treat the text from each chamber as a distinct speaker-session.

[11] These partisan phrases are drawn from an earlier version of this paper (Gentzkow et al. 2015).

3.2 Choice model

As a micro-foundation for equation (1), suppose that at the end of session t speaker i receives a payoff:

$$u_{it} = \begin{cases} \delta y_t + (1-\delta)\left(\boldsymbol{\alpha}_t' + \mathbf{x}_{it}'\boldsymbol{\gamma}_t\right)\mathbf{c}_{it}, & i \in R_t \\ -\delta y_t + (1-\delta)\left(\boldsymbol{\alpha}_t' + \mathbf{x}_{it}'\boldsymbol{\gamma}_t\right)\mathbf{c}_{it}, & i \in D_t \end{cases} \tag{2}$$

where

$$y_t = \boldsymbol{\varphi}_t' \sum_i \mathbf{c}_{it} \tag{3}$$

indexes public opinion. Here, $\boldsymbol{\varphi}_t$ is a $J-$vector mapping speech into public opinion, δ is a scalar denoting the relative importance of public opinion in speakers' utility, $\boldsymbol{\alpha}_t$ is a $J-$vector denoting the baseline popularity of each phrase at time t, and $\boldsymbol{\gamma}_t$ is a $K \times J$ matrix mapping speaker characteristics into the utility of using each phrase.

Each speaker chooses each phrase she speaks to maximize u_{it} up to a choice-specific i.i.d. type 1 extreme value shock, so that:

$$q_{jt}^{P(i)}(\mathbf{x}_{it}) = e^{u_{ijt}} / \sum_l e^{u_{ilt}} \tag{4}$$

$$u_{ijt} = \delta\left(2 \cdot \mathbf{1}_{i \in R_t} - 1\right)\varphi_{jt} + (1-\delta)\left(\alpha_{jt} + \mathbf{x}_{it}'\boldsymbol{\gamma}_{jt}\right).$$

Note that if $\mathbf{x}_{it}$ is a constant ($\mathbf{x}_{it} := \mathbf{x}_t$), any interior phrase probabilities $\mathbf{q}_t^P(\cdot)$ are consistent with equation (4). In this sense, the choice model in this subsection only restricts the model in equation (1) by pinning down how phrase probabilities depend on speaker characteristics. Note also that, according to equation (4), if a phrase (or set of phrases) is excluded from the choice set, the relative frequencies of the remaining phrases are unchanged. We exploit this fact in sections 5 and 6 to compute average partisanship for interesting subsets of the full vocabulary.

3.3 Measure of partisanship

For given characteristics $\mathbf{x}$, the problem of measuring partisanship can be restated as one of measuring the divergence between $\mathbf{q}_t^R(\mathbf{x})$ and $\mathbf{q}_t^D(\mathbf{x})$. When these vectors are close, Republicans and Democrats speak similarly and we would say that partisanship is low. When they are far from each other, languages diverge and we would say that partisanship is high.

We choose a particular measure of this divergence that has a clear interpretation in the context

of our model: the posterior probability that an observer with a neutral prior expects to assign to a speaker's true party after hearing the speaker speak a single phrase.

Definition. The *partisanship* of speech at $\mathbf{x}$ is:

$$\pi_t(\mathbf{x}) = \frac{1}{2}\mathbf{q}_t^R(\mathbf{x}) \cdot \boldsymbol{\rho}_t(\mathbf{x}) + \frac{1}{2}\mathbf{q}_t^D(\mathbf{x}) \cdot (1 - \boldsymbol{\rho}_t(\mathbf{x})), \tag{5}$$

where

$$\rho_{jt}(\mathbf{x}) = \frac{q_{jt}^R(\mathbf{x})}{q_{jt}^R(\mathbf{x}) + q_{jt}^D(\mathbf{x})}. \tag{6}$$

Average partisanship in session t is:

$$\overline{\pi}_t = \frac{1}{|R_t \cup D_t|} \sum_{i \in R_t \cup D_t} \pi_t(\mathbf{x}_{it}). \tag{7}$$

To understand these definitions, note that that $\rho_{jt}(\mathbf{x})$ is the posterior belief that an observer with a neutral prior assigns to a speaker being Republican if the speaker chooses phrase j in session t and has characteristics $\mathbf{x}$. Partisanship $\pi_t(\mathbf{x})$ averages $\rho_{jt}(\mathbf{x})$ over the possible parties and phrases: if the speaker is a Republican (which occurs with probability $\frac{1}{2}$), the probability of a given phrase j is $q_{jt}^R(\mathbf{x})$ and the probability assigned to the true party after hearing j is $\rho_{jt}(\mathbf{x})$; if the speaker is a Democrat, these probabilities are $q_{jt}^D(\mathbf{x})$ and $1 - \rho_{jt}(\mathbf{x})$, respectively. Average partisanship $\overline{\pi}_t$, which is our target for estimation, averages $\pi_t(\mathbf{x}_{it})$ over the characteristics $\mathbf{x}_{it}$ of speakers active in session t.

Partisanship is closely related to isolation, a common measure of residential segregation (White 1986; Cutler et al. 1999). To see this, redefine the model so that j indexes neighborhoods and $m_{it} = 1$ for all i and t. Isolation is the difference in the share Republican of the average Republican's neighborhood and the average Democrat's neighborhood. In an infinite population with an equal number of Republicans and Democrats, all with characteristics $\mathbf{x}$, this is:

$$\begin{aligned} Isolation_t(\mathbf{x}) &= \mathbf{q}_t^R(\mathbf{x}) \cdot \boldsymbol{\rho}_t(\mathbf{x}) - \mathbf{q}_t^D(\mathbf{x}) \cdot \boldsymbol{\rho}_t(\mathbf{x}) \\ &= 2\pi_t(\mathbf{x}) - 1. \end{aligned} \tag{8}$$

Thus, isolation is an affine transformation of partisanship.

Frankel and Volij (2011) characterize a large set of segregation indices based on a set of ordinal axioms. Ignoring covariates $\mathbf{x}$, our measure satisfies six of these axioms: Non-triviality, Continuity,

Scale Invariance, Symmetry, Composition Invariance, and the School Division Property. It fails to satisfy one axiom: Independence.[12]

4 Estimation

4.1 Plug-in estimators

Ignoring covariates $\mathbf{x}$, a straightforward way to estimate partisanship is to plug in empirical analogues for the terms that appear in equation (5). This approach yields the maximum likelihood estimator (MLE) of our model.

More precisely, let $\hat{\mathbf{q}}_{it} = \mathbf{c}_{it}/m_{it}$ be the empirical phrase frequencies for speaker i. Let $\hat{\mathbf{q}}_t^P = \sum_{i \in P_t} \mathbf{c}_{it} / \sum_{i \in P_t} m_{it}$ be the empirical phrase frequencies for party P, and let $\hat{\rho}_{jt} = \hat{q}_{jt}^R / \left(\hat{q}_{jt}^R + \hat{q}_{jt}^D\right)$. Then the MLE of $\overline{\pi}_t$ when $\mathbf{x}_{it} := \mathbf{x}_t$ is:[13]

$$\hat{\pi}_t^{MLE} = \frac{1}{2}\left(\hat{\mathbf{q}}_t^R\right) \cdot \hat{\boldsymbol{\rho}}_t + \frac{1}{2}\left(\hat{\mathbf{q}}_t^D\right) \cdot \left(1 - \hat{\boldsymbol{\rho}}_t\right). \tag{9}$$

Standard results imply that this estimator is consistent and efficient in the limit as the amount of speech grows, fixing the number of phrases.

The MLE can, however, be severely biased in finite samples. As $\hat{\pi}_t^{MLE}$ is a convex function of $\hat{\mathbf{q}}_t^R$ and $\hat{\mathbf{q}}_t^D$, Jensen's inequality implies that it has a positive bias. To build intuition for the form of the bias, use the fact that $\mathrm{E}\left(\hat{\mathbf{q}}_t^R, \hat{\mathbf{q}}_t^D\right) = \left(\mathbf{q}_t^R, \mathbf{q}_t^D\right)$ to decompose the bias of a generic term $\left(\hat{\mathbf{q}}_t^R\right) \cdot \hat{\boldsymbol{\rho}}_t$ as:

$$\mathrm{E}\left(\left(\hat{\mathbf{q}}_t^R\right) \cdot \hat{\boldsymbol{\rho}}_t - \left(\mathbf{q}_t^R\right) \cdot \boldsymbol{\rho}_t\right) = \left(\mathbf{q}_t^R\right) \cdot \mathrm{E}\left(\hat{\boldsymbol{\rho}}_t - \boldsymbol{\rho}_t\right) + \mathrm{Cov}\left(\left(\hat{\mathbf{q}}_t^R - \mathbf{q}_t^R\right), \left(\hat{\boldsymbol{\rho}}_t - \boldsymbol{\rho}_t\right)\right). \tag{10}$$

The first term is nonzero because $\hat{\boldsymbol{\rho}}_t$ is a nonlinear transformation of $\left(\hat{\mathbf{q}}_t^R, \hat{\mathbf{q}}_t^D\right)$.[14] The second term is nonzero because the sampling error in $\hat{\boldsymbol{\rho}}_t$ is mechanically related to the sampling error in $\left(\hat{\mathbf{q}}_t^R, \hat{\mathbf{q}}_t^D\right)$.

A similar bias arises for plug-in estimators of polarization measures other than partisanship,

[12] In our context, Independence would require that the ranking in terms of partisanship of two years t and s remain unchanged if we add a new set of phrases J^* to the vocabulary whose probabilities are the same in both years ($q_{jt}^P = q_{js}^P \forall P, j \in J^*$). Frankel and Volij (2011) list one other axiom, Group Division Property, which is only applicable for indices where the number of groups (i.e., parties in our case) is allowed to vary.

[13] A requirement for this interpretation is that we exclude from the choice set any phrases that are not spoken in a given session.

[14] Suppose that there are two speakers, one Democrat and one Republican, each with $m_{it} = 1$. There are two phrases. The Republican says phrase two with certainty and the Democrat says phrase two with probability 0.01. Then $\mathrm{E}(\hat{\rho}_{2t}) = 0.01(\frac{1}{2}) + 0.99(1) = 0.995 > \rho_{2t} = 1/1.01 \approx 0.990$.

because sampling variability means that $\hat{\mathbf{q}}_t^R$ and $\hat{\mathbf{q}}_t^D$ will tend to differ by more than $\mathbf{q}_t^R$ and $\mathbf{q}_t^D$. This is especially transparent if we use a norm such as Euclidian distance as a metric: Jensen's inequality implies that for any norm $\|\cdot\|$, $\mathrm{E}\left\|\hat{\mathbf{q}}_t^R - \hat{\mathbf{q}}_t^D\right\| \geq \left\|\mathbf{q}_t^R - \mathbf{q}_t^D\right\|$. Similar issues arise for the measure of Jensen et al. (2012), which is given by $\frac{1}{m_t}\sum_j m_{jt}\left|corr\left(c_{ijt}, \mathbf{1}_{i\in R_t}\right)\right|$. If speech is independent of party ($\mathbf{q}_t^R = \mathbf{q}_t^D$), then the population value of $corr\left(c_{ijt}, \mathbf{1}_{i\in R_t}\right)$, conditional on total verbosity m_i, is zero. But in any finite sample the correlation will be nonzero with positive probability, so the measure may imply some amount of polarization even when speech is unrelated to party.

One appealing approach to addressing finite-sample bias in $\hat{\pi}_t^{MLE}$ is to use different samples to estimate $\hat{\mathbf{q}}_t^P$ and $\hat{\boldsymbol{\rho}}_t$, making the errors in the former orthogonal to the errors in the latter and so eliminating the second bias term in equation (10). This leads naturally to a leave-out estimator:

$$\hat{\pi}_t^{LO} = \frac{1}{2}\frac{1}{|R_t|}\sum_{i\in R_t}\hat{\mathbf{q}}_{i,t}\cdot\hat{\boldsymbol{\rho}}_{-i,t} + \frac{1}{2}\frac{1}{|D_t|}\sum_{i\in D_t}\hat{\mathbf{q}}_{i,t}\cdot\left(1-\hat{\boldsymbol{\rho}}_{-i,t}\right), \tag{11}$$

where $\hat{\boldsymbol{\rho}}_{-i,t}$ is the analogue of $\hat{\boldsymbol{\rho}}_t$ computed from the empirical frequencies $\hat{\mathbf{q}}_{-i,t}^P$ of all speakers other than i.[15] This estimator is consistent in the limit as the amount of speech grows, fixing the number of phrases. It is still biased for $\overline{\pi}_t$ (because $\hat{\boldsymbol{\rho}}_{-i,t}$ is biased for $\boldsymbol{\rho}_t$), but we show below that the bias appears small in practice.

4.2 Penalized estimator

Our preferred estimation method draws on the structure of the choice model in section 3.2, allowing us to include speaker characteristics $\mathbf{x}_{it}$ and to control bias through penalization. Rewrite u_{ijt} as:[16]

$$u_{ijt} = \tilde{\alpha}_{jt} + \mathbf{x}_{it}'\tilde{\boldsymbol{\gamma}}_{jt} + \tilde{\varphi}_{jt}\mathbf{1}_{i\in R_t}. \tag{12}$$

The $\tilde{\alpha}_{jt}$ are phrase-time-specific intercepts and the $\tilde{\varphi}_{jt}$ are phrase-time-specific party loadings. In our baseline specification, $\mathbf{x}_{it}$ consists of indicators for state, chamber, gender, Census region, and whether the party is in the majority. The coefficients $\tilde{\boldsymbol{\gamma}}_{jt}$ on these attributes are static in time (i.e., $\tilde{\gamma}_{jtk} := \tilde{\gamma}_{jk}$) except for those on Census region, which are allowed to vary across sessions. We also explore specifications in which $\mathbf{x}_{it}$ includes unobserved speaker-level preference shocks.

[15] Implicitly, in each session t we exclude any phrase that is spoken only by a single speaker.

[16] This parametrization is observationally equivalent to equation (4), with $\tilde{\varphi}_{jt} = 2\delta\varphi_{jt}$, $\tilde{\boldsymbol{\gamma}}_{jt} = \boldsymbol{\gamma}_{jt}(1-\delta)$, $\tilde{\alpha}_{jt} = (1-\delta)\alpha_{jt} - \delta\varphi_{jt}$.

We estimate the parameters $\{\tilde{\boldsymbol{\alpha}}_t, \tilde{\boldsymbol{\gamma}}_t, \tilde{\boldsymbol{\varphi}}_t\}_{t=1}^{T}$ of equation (12) by minimization of the following penalized objective function:

$$\sum_j \left\{ \sum_t \sum_i \left[m_{it} \exp(\tilde{\alpha}_{jt} + \mathbf{x}_{it}' \tilde{\boldsymbol{\gamma}}_{jt} + \tilde{\varphi}_{jt} \mathbf{1}_{i \in R_t}) - c_{ijt} (\tilde{\alpha}_{jt} + \mathbf{x}_{it}' \tilde{\boldsymbol{\gamma}}_{jt} + \tilde{\varphi}_{jt} \mathbf{1}_{i \in R_t}) \right] + \lambda_j \left| \tilde{\varphi}_{jt} \right| \right\}, \quad (13)$$

where the summation of the term in square brackets is an approximation to the negative log-likelihood of our model. We form an estimate $\hat{\pi}_t^*$ of $\overline{\pi}_t$ by substituting estimated parameters into the probability objects in equation (7).

The minimand in equation (13) encodes two key decisions. First, we approximate the likelihood of our multinomial logit model with the likelihood of a Poisson model (Palmgren 1981; Baker 1994; Taddy 2015), where $c_{ijt} \sim \text{Pois}\left(\exp\left[\mu_{it} + u_{ijt}\right]\right)$, and we use the plug-in estimate $\hat{\mu}_{it} = \log m_{it}$ of μ_{it}. Because the Poisson and the multinomial logit share the same conditional likelihood $\Pr(\mathbf{c}_{it} \mid m_{it})$, their MLEs coincide when $\hat{\mu}_{it}$ is the MLE. Although our plug-in is not the MLE, Taddy (2015) shows that our approach often performs well in related settings. In the online appendix, we show that our estimator performs well on data simulated from the multinomial logit model.

We adopt the Poisson approximation because, fixing $\hat{\mu}_{it}$, the likelihood of the Poisson is separable across phrases. This feature allows us to use distributed computing to estimate the model parameters (Taddy 2015). Without the Poisson approximation, computation of our estimator would be infeasible due to the cost of repeatedly calculating the denominator of the logit choice probabilities.

The second key decision is the use of an L_1 penalty $\lambda_j \left| \tilde{\varphi}_{jt} \right|$, which imposes sparsity on the party loadings and shrinks them toward 0 (Tibshirani 1996). We determine the penalties $\boldsymbol{\lambda}$ by regularization path estimation, first finding λ_j^1 large enough so that $\tilde{\varphi}_{jt}$ is estimated to be 0, and then incrementally decreasing $\lambda_{jt}^2, ..., \lambda_{jt}^G$ and updating parameter estimates accordingly.[17] We then choose the value of λ_j that minimizes a Bayesian Information Criterion.[18] We do not penalize the

[17] An attractive computational property of this approach is that the coefficient estimates change smoothly along the path of penalties, so each segment's solution acts as a hot-start for the next segment and the optimizations are fast to solve.

[18] The Bayesian Information Criterion we use is $\sum_{i,t} \log \text{Po}\left(c_{ijt};\ \exp[\hat{\mu}_i + u_{ijt}]\right) + (n/(n - df - 1))\, df \log n$, where $n = \sum_t (|D_t| + |R_t|)$ is the number of speaker-sessions and df is a degrees-of-freedom term that (following Zou et al. 2007) is given by the number of parameters estimated with nonzero values (excluding the $\hat{\mu}_{it}$, as outlined in Taddy 2015). The adjustment $n/(n - df - 1)$ corrects for the potential for overfit given the high dimensionality of the problem (Flynn et al. 2013, Taddy forthcoming).

phrase-specific intercepts $\tilde{\alpha}_{jt}$ or the covariate coefficients $\tilde{\boldsymbol{\gamma}}_{jt}$.[19]

If the penalty shrinks sufficiently quickly with the sample size, then under standard regularity conditions our estimator is root-n consistent like the MLE (Fan and Li 2001). However, as we will see, our estimator is far less biased than the MLE. Moreover, for fixed penalty $\boldsymbol{\lambda}$, our estimator can be interpreted as the unique posterior mode of a Bayesian model with a diffuse Laplace prior on the coefficients $\tilde{\boldsymbol{\gamma}}$, an informative Laplace prior on the party loadings $\tilde{\boldsymbol{\varphi}}$, and an uninformative prior on the intercepts $\tilde{\boldsymbol{\alpha}}$. In this sense, our parameter estimates are optimal in finite samples against a "0-1" loss function (Murphy 2012).

5 Results

5.1 Trends in partisanship

We now turn to our main results on trends in the partisanship of speech over time. As a diagnostic for finite-sample bias, we present, for each measure, a placebo series where we reassign parties to speakers at random and then re-estimate the measure on the resulting data. In this "random" series, $\mathbf{q}_t^R = \mathbf{q}_t^D$ by construction, so the true value of π_t is equal to $\frac{1}{2}$ in all years. We thus expect the random series for an unbiased estimator of π_t to have value $\frac{1}{2}$ in each session t, and we can measure the bias of an estimator by its deviation from $\frac{1}{2}$.

Figure 1 presents results for two plug-in estimators: the maximum likelihood estimator $\hat{\pi}_t^{MLE}$ of our model, and the measure reported by Jensen et al. (2012) computed using their publicly available data. In panel A, we see that the random series for $\hat{\pi}^{MLE}$ is far from $\frac{1}{2}$, indicating that the bias in the MLE is severe in practice. Variation over time in the magnitude of the bias dominates the series, leading the random series and the real series to be highly correlated. Taking the MLE at face value, we would conclude that language was much more partisan in the past and that the upward trend in recent years is small by historical standards.

Because bias is a finite-sample property, it is natural to expect that the severity of the bias in $\hat{\pi}_t^{MLE}$ in a given session t depends on the amount of speech, i.e., on the verbosity $\mathbf{m}_t$ of speakers

[19] For reasons detailed in Haberman (1973) and summarized in Silva and Tenreyro (2010), it is not straightforward to establish the existence of a single maximizing argument $\hat{\boldsymbol{\gamma}}_{jt}$ in each Poisson regression. A sufficient condition for existence is that the controls design (i.e., the part of the regression involving $\mathbf{x}_{it}$) is full rank on the subset of observations where $c_{ijt} > 0$; however, this is overly conservative and will remove variables which do have a measurable (even large) effect on the likelihood. Instead, we build a controls design that is full rank on the entire dataset and has no columns that are all zero when $c_{ijt} > 0$. To avoid remaining nonexistence-related issues, we then add a very small (10^{-6}) L_1 penalty on the $\tilde{\boldsymbol{\gamma}}_{jt}$ to ensure numerical convergence.

in that session. The online appendix shows that this is indeed the case: a first-order approximation to the bias in $\hat{\pi}_t^{MLE}$ as a function of verbosity follows a similar path to the random series in panel A of figure 1, and the dynamics of $\hat{\pi}_t^{MLE}$ are similar to those in the real series when we allow verbosity to follow its empirical distribution but fix phrase frequencies $\left(\mathbf{q}_t^R, \mathbf{q}_t^D\right)$ at those observed in a particular session t'. The online appendix also shows that excluding infrequently used phrases is not a satisfactory solution to the bias in $\hat{\pi}^{MLE}$: while the severity of the bias falls as we exclude less frequently spoken phrases, there remains a severe and time-varying bias even when we exclude 99 percent of phrases from our calculations.

In panel B of figure 1, we see that the Jensen et al. (2012) measure behaves similarly to the MLE. The plot for the real series replicates the published version. The random series is again far from $\frac{1}{2}$, and the real and random series both trend downward in the first part of the sample period. Jensen et al. (2012) conclude that polarization has been increasing recently, but that it was as high or higher in the late-nineteenth century. The results in panel B suggest that the second part of this conclusion could be an artifact of finite-sample bias.[20]

Figure 2 shows the leave-out estimator $\hat{\pi}_t^{LO}$. The random series suggests that the leave-out correction largely purges the estimator of bias: the series is close to $\frac{1}{2}$ throughout the period. The real series suggests that partisanship is roughly constant for much of the sample period, then rises rapidly beginning in the 1990s.

Figure 3 presents our main result: the time series of partisanship from our preferred penalized estimator described in section 4.2. Panel A shows the full series, and panel B zooms in on the most recent years, indicating some events of interest. These estimates have two important advantages relative to $\hat{\pi}_t^{LO}$: they control for observables $\mathbf{x}_{it}$, and they use penalization to control bias and reduce variance. The results show that this approach essentially eliminates bias, and dramatically reduces the amount of noise in the series.[21] Looking at the data through this sharper lens reveals that partisanship was low and roughly constant until the early 1990s, then exploded, reaching unprecedented heights in the recent years of our sample. The appendix figure shows two variants of our baseline model, one in which we exclude covariates and one in which we allow speaker-specific random effects. These series look broadly similar to the baseline but show a smaller

[20] In the online appendix, we show that the dynamics of $\hat{\pi}_t^{MLE}$ in Jensen et al.'s (2012) data are similar to those in our own data, which is reassuring as Jensen et al. (2012) obtain the *Congressional Record* from a different source, use different processing algorithms, and use a vocabulary of three-word phrases rather than two-word phrases.

[21] In the online appendix, we show results from a specification in which we do not penalize the party loadings (i.e., set $\boldsymbol{\lambda} \approx 0$). The random and real series look similar to those for $\hat{\pi}_t^{MLE}$, with large apparent bias and high variance over time. This confirms that the penalties are the crucial elements for controlling bias and reducing variance.

increase in partisanship in recent years.

The recent increase in partisanship implied by our baseline estimates is large. Recall that average partisanship is the posterior that a neutral observer expects to assign to a speaker's true party after hearing a single phrase. Figure 4 extends this concept to show the expected posterior for speeches of various lengths. An average one-minute speech in our data contains around 33 phrases (after pre-processing). In 1874, an observer hearing such a speech would be expected to have a posterior of around .54 on the speaker's true party. In 1990, this value remained almost equivalent at around .55. In 2008, the value was .83.

In figure 5, we compare our speech-based measure of partisanship to the standard measure of ideological polarization based on roll-call votes (McCarty et al. 2015). The latter is based on an ideal-point model that places both speakers and legislation in a latent space; polarization is the distance between the average Republican and the average Democrat along the first dimension. Panel A shows that the dynamics of these two series are very different: though both indicate a large increase in recent years, the roll-call series is about as high in the late nineteenth and early twentieth century as it is today, and its current upward trend begins around 1950 rather than 1990. We conclude from this that speech and roll-call votes should not be seen as two different manifestations of a single underlying ideological dimension. Rather, speech appears to respond to a distinct set of incentives and constraints. Panel B shows that a measure of the Republican-ness of an individual's speech from our model and the individual NOMINATE scores from the roll-call voting data are nevertheless strongly correlated in the cross-section. Across all sessions, the correlation between speech and roll-call based partisanship measures is .331 ($p = .000$). After controlling for party, the correlation is .053 and remains highly statistically significant ($p = .000$).[22] Thus, members who vote more conservatively also use more conservative language on average, even though the time-series dynamics of voting and speech diverge.

5.2 Partisan phrases

Our model provides a natural way to define the partisanship of an individual phrase: the extent to which removing the phrase from the vocabulary would reduce the ability of an observer to predict a speaker's party. We implement this concept by calculating the change in estimated average

[22] These correlations are .371 ($p = .000$) and .077 ($p = .000$), respectively, when we use data only on speakers who speak an average of at least 1000 phrases across the sessions in which they speak.

partisanship if we remove a phrase from the choice set.[23]

Table 1 lists the ten most partisan phrases in every tenth session of Congress according to this definition. (The online appendix shows the list for all sessions.) These lists illustrate the underlying variation driving our measure, and give a sense of how partisan speech has changed over time. We now argue that the top phrases align closely with the policy positions of the parties, confirming that our measure is indeed picking up partisanship rather than some other dimension correlated with it or spurious noise. Our discussion draws mainly on the original congressional text and on the national party platforms (from Peters and Woolley 2016). We cite proceedings in Congress using the format "CR Date," with a hyperlink to ProQuest Congressional, a gated service to which many universities subscribe.

In the 50th session (1887-88), many tariff-related phrases ("increase duti," "high protect," "tariff tax") are highly Democratic. Party cleavage on this issue is consistent with the 1888 Democratic Party platform, which endorses tariff reduction in its first sentence, and with the Republican platform of the same year, which says the party is "uncompromisingly in favor of the American system of protection." The highly Republican phrase "fisheri treati" relates to a then-controversial international agreement regarding fishing in Canadian waters.[24] Controversies over the appropriation of land grants are reflected in the partisanship of terms like "public domain" and "navig compani."[25]

The Republican Party platform of 1908 devotes a section to the need for "generous provision" for veterans, and indeed Republicans make much of phrases related to veteran compensation ("infranti war," "indian war," "mount volunt," "spain pay," "war pay") in the 60th session (1907-08).[26]

[23] Let

$$\pi_t^R(\mathbf{x}) = \mathbf{q}_t^R(\mathbf{x}) \cdot \boldsymbol{\rho}_t(\mathbf{x})$$
$$\pi_t^D(\mathbf{x}) = \mathbf{q}_t^D(\mathbf{x}) \cdot (1 - \boldsymbol{\rho}_t(\mathbf{x})).$$

Then the partisanship ζ_{jt} of phrase j in session t is defined as the average across the speakers in session t of

$$\frac{1}{2}\left(\frac{q_{jt}^R(\mathbf{x}_{it})}{1-q_{jt}^R(\mathbf{x}_{it})}\right)\left(\rho_{jt}(\mathbf{x}_{it}) - \pi_t^R(\mathbf{x}_{it})\right) + \frac{1}{2}\left(\frac{q_{jt}^D(\mathbf{x}_{it})}{1-q_{jt}^D(\mathbf{x}_{it})}\right)\left((1-\rho_{jt}(\mathbf{x}_{it})) - \pi_t^D(\mathbf{x}_{it})\right).$$

Note that this definition holds the posterior $\rho_{j't}(\mathbf{x})$ for phrases $j' \neq j$ constant when removing phrase j. Allowing the posterior to adjust is computationally more expensive but produces essentially identical results: in three sessions for which we computed both measures, the 25 most partisan phrases agreed between the two measures in all but 1% of cases.

[24] The Republican platform says, "we arraign the present Democratic Administration for its weak and unpatriotic treatment of the fisheries question" and fulminates at length about the need to assert US fishing rights.

[25] In CR December 5, 1888 members of Congress discuss the Supreme Court's decision in Wolcott vs. Des Moines [Navigation] Company.

[26] CR May 12, 1908 includes an extensive amendment specifying pension benefits to accrue to individuals, for exam-

The Democratic Party platform of that same year has as its key theme to "free the Government from the grip of those who have made it a business asset of the favor-seeking corporations." Several Democratic phrases of the 60th session relate to this theme, for example "bureau corp," a reference to the Bureau of Corporations, the predecessor to the Federal Trade Commission, and "powder trust," a reference to appropriations to private manufacturers of gunpowder.[27] The Democratic platform also emphasizes trade and shipping issues, declaring support for the Panama Canal ("republ panama," "level canal," "lock canal"), and "demanding" the repeal of tariffs on several commodities including "print paper," and supporting the "upbuilding of the American merchant marine without...bounties from the public treasury" ("ship subsidi").[28]

Veterans' benefits, now including those related to the Great War, continued to play an important role in US policy in the 70th session (1927-1928), with both parties devoting whole sections of their platforms to the subject. Republicans supported a measure that became "The Emergency Officers' Retirement Act of 1928" (Public Law 506), seeking to equalize retirement benefits for certain categories of veterans. Language related to this act ("pay period," "rate advantage," "construct service," "war regular," "fourth pay") marks out Republicans in Congress (CR January 4, 1929). Language related to commercial and military aviation ("air corp," "air mail") also pervades Republican speech, and indeed the Republican platform discusses development of commercial aviation at some length. Both parties devote attention in their platforms to the newly developing radio market and associated regulation; Democratic partisan phrases like "radio commiss" and "wave length" reflect discussions about regulating this market.

The 80th session (1947-1948) convened in the wake of the Second Wold War and saw precursors of many modern partisan fissures. Many Republican-leaning phrases relate to the war and war financing ("coast guard," "stop communism," "lend leas," "zone germani," "british loan," "unit kingdom"). The 1948 Democratic Party Platform advocates amending the Fair Labor Standards Act to raise the minimum wage from 40 to 75 cents an hour ("labor standard," "standard act," "cent hour"). By contrast, the Republican platform of the same year does not even mention the Fair Labor Standards Act or the minimum wage. Likewise, the Democratic platform supports the school

ple: "The name of Annie A. Robbins, late nurse, Medical Department, United States Army, war with Spain, and pay her a pension at the rate of $12 per month." CR February 11, 1908 includes discussion of legislation to provide "bounty land" to those who fought in certain "Indian Wars."

[27] In CR January 21, 1909, William Cox (D-IN) complains that "the entire United States is now being held up by a great hydra-headed monster, known in ordinary parlance as a 'powder trust'."

[28] On this latter point, in CR May 23, 1908, William Sulzer (D-NY) says "Now, Mr. Speaker, the Republicans in Congress have been advocating ever since I have been here the restoration of the American merchant marine by ship subsidies, by gratuities, that rob all the people in order to foster a special industry."

lunch program ("school lunch"), of which the Republican platform makes no mention. Both party platforms discuss support for agriculture, reflected in phrases like "depart agricultur" (Republican) and "soil conserv" (Democratic).

The school lunch program ("lunch program, "wholesom meat," "school lunch") remains a signature theme of Democratic speech in the 90th session (1967-68), as do elements of the expanding modern welfare state such as the Food Stamp Program ("food stamp," "stamp program"). The Democrats stand out for vocally supporting the president ("support president," "commend presid"). The Republicans are marked by support for the Human Investment Act ("human invest," "invest act"), which provided tax credits to employers for training and hiring certain kinds of workers. Although both parties' 1968 platforms emphasize transportation issues, transportation-related phrases are only associated with Republicans ("aid highway," "highway program"). In this period, Congress discussed the UN Conventions on Forced Labor and the Political Rights of Women (e.g., CR February 8, 1968), both of which connect to frequent Democratic terms ("forc labor," "polit right").

Partisan language in the 100th session (1987-88) centers around familiar Cold War themes. Both parties focus on the Iran-Contra scandal and the related conflict in Nicragua. Democrats refer to the insurgents as "Contras" ("contra aid," "contra war," "support contra") and their opponents the Sandinistas as simply the "nicaraguan govern[ment]." Republicans instead call the insurgents "freedom fighter[s]" (e.g., CR March 17, 1988) and emphasize that the Sandinistas are communists ("communist govern"). The Democrats criticize the Strategic Defense Initiative ("star war"), and the Republicans emphasize the need for national defense ("incom ballist," e.g., CR May 12, 1987). Republicans' most partisan phrases have more of a domestic focus than those of Democrats, covering labor relations ("double breast"),[29] tax policy ("heifer tax"), and abortion policy ("abort industri," "abort demand").[30] Republicans are also identified in this session with some procedural language ("demand second," "withdraw reserve," "reserv object").[31]

Language in the 110th session (2007-2008) follows clear partisan divides. Republicans focus on taxes ("tax increas," "higher tax"), immigration ("illeg immigr," "illeg alien"), and energy policy

[29]The phrase "double breast" refers to "double-breasted" corporate forms in which a company has two components, one a union firm and one an open shop. The 100th session considered legislation to eliminate double-breasting in the construction industry (Lamy 1988).

[30]See CR July 25, 1988 and CR October 21, 1987.

[31]For "I demand a second," see CR October 3, 1988. For "I withdraw my reservation of objection," see CR October 21, 1988.

("age oil," "american energi," "increas american").[32] Democrats focus on the aftermath of the wars in Iraq and Afghanistan ("afghanistan veteran," "respons redeploy," "contract hallibutron," "redeploy iraq").[33]

5.3 Decomposing the trends

In this section, we explore several implications of the estimated model, focusing on the drivers of the large increase in partisanship in recent decades.

Figure 6 shows that the recent increase in partisanship is driven by phrases that account for about 10 percent of speech. The figure plots quantiles of the estimated average value of $\rho_{jt}(\mathbf{x}_{it})$ in each session, weighted by frequency of phrase use. These quantiles thus show the distribution of the informativeness of phrases about the party of a speaker. The plot shows a fanning out of the distribution in the quantiles at or above 0.95 and at or below 0.05. The sharp change in 1994 is especially pronounced in the extreme tail (quantiles 0.999 or 0.001) of the distribution.

We next consider how much of the increase in partisanship is driven by the introduction of newly coined phrases. We define these "neologisms" to be phrases that are first spoken in our data after 1980.[34] Figure 7 plots our baseline average partisanship alongside two adjusted series: one in which the choice set includes only these neologisms, and one in which it excludes them. The first series reveals that neologisms are extremely partisan, showing a much larger increase than the baseline measure. At the same time, the second series shows that a large increase in partisanship remains even when we exclude neologisms from the choice set. This counterfactual is imperfect as our model does not account for the fact that the introduction of a given phrase (e.g., "death tax") may change the usage of another phrase (e.g., "estate tax"), meaning the series excluding neologisms is not completely purged of their effect.

Our final set of results presents a decomposition of partisanship using our 22 manually defined topics. Our baseline measure of partisanship captures changes both in the topics speakers choose to discuss and in the phrases they use to discuss them. Whether a speech about taxes includes the phrases "tax relief" or "tax breaks" will help an observer to guess the speaker's party; so, too, will

[32] One member of Congress, Joe Wilson (R-SC), ended many speeches with "In conclusion, God bless our troops, and we will never forget September 11," hence the unusual pattern of the phrase "conclus god." "Ms. Ginni" refers to Ginni Brown-Waite (R-FL), a reference that our algorithm to exclude names of congresspeople failed to catch.

[33] The phrase "engag unfortun" is used repeatedly by Evan Bayh in the context of remembering members of the Armed Forces, as in "When I think about this just cause in which we are engaged, and the unfortunate pain that comes with the loss of our heroes..." (e.g., CR September 4, 2007).

[34] The online appendix shows that results are very similar when we instead define a neologism to be a phrase such that at least 99% of the occurrences of the phrase are after 1980.

whether the speaker chooses to talk about taxes or about the environment. To separate these, we define between-topic partisanship to be the expected posterior a neutral observer expects to assign to a speaker's true party when the observer knows only the topic a speaker chooses, not the particular phrases chosen within the topic. Partisanship within a specific topic is the expected posterior when the choice set consists only of phrases in that topic. The overall within-topic partisanship in a given session is the average of partisanship across all topics, weighting each topic by its frequency of occurrence.

Figure 8 shows that the rise in partisanship is driven mainly by divergence in how the parties talk about a given substantive topic, rather than by divergence in which topics they talk about. According to our estimates, choice of topic encodes much less information about a speaker's party than does choice of phrases within a topic.

Figure 9 shows estimated partisanship for phrases within each of the 22 topics. Partisanship has increased within many topics in recent years, with the largest increases in the tax and immigration topics. Other topics with large increases include budget and fiscal issues, crime, health, and government. Not all topics have increased recently however. Alcohol, for example, was fairly partisan in the Prohibition Era but is not especially partisan today. Figure 9 also shows that the partisanship of a topic is not strongly related in general to the frequency with which the topic is discussed. For example, the world wars are associated with a surge in frequency of discussion of defense, but not with an increase in the partisanship of that topic.

To illustrate the underlying variation at the phrase level, the top panel of figure 10 shows the evolution of the informativeness of various tax-related phrases. The plot shows that these phrases become consistently more informative about a speaker's party over time. Some phrases, such as "tax break" and "tax spend," were partisan in the 1980s but became more so in the 1990s. Others, such as "death tax," only emerge as partisan in the 1990s. The bottom panel aggregates the Republican and Democratic phrases, showing a sharp fanning out after 1990.

In summary, in the mid-1990s members of Congress began to use speech in a way that was tightly tied to their political party. This trend was initially concentrated in a relatively small number of phrases, then expanded to affect a broader set. It involved both the opening of fissures on existing language, and the invention of new and highly partisan terms. And it occurred within rather than between topics, particularly within those focused on key areas of domestic policy.

6 Discussion

What caused the dramatic increase in the partisanship of speech? We cannot provide a definitive answer, but the timing of the change shown in panel B of figure 3 suggests two possible hypotheses: innovation in political persuasion coinciding with the 1994 Republican takeover of the House of Representatives, and changes in the media environment including the introduction of live broadcasts of congressional proceedings on the C-SPAN cable network.

The inflection point in the partisanship series occurs in the 104th session (1995-1996), the first following the 1994 election. This election was a watershed event in the history of the US Congress. It brought a Republican majority to the House for the first time in more than forty years, and was the largest net partisan gain since 1948. It "set off a political earthquake that [would] send aftershocks rumbling through national politics for years to come" (Jacobson 1996). The Republicans were led by future Speaker of the House Newt Gingrich, who succeeded in uniting the party around a platform called the *Contract with America.* It specified the actions Republicans would take upon assuming control, focusing the contest around a set of national domestic issues including taxes, crime, and government efficiency.

Innovation in language and persuasion was, by many accounts, at the center of this victory. Assisted by the consultant Frank Luntz—who was hired by Gingrich to help craft the *Contract with America,* and became famous in significant part because of his role in the 1994 campaign—the Republicans used focus groups and polling to identify rhetoric that resonated with voters (Bai 2005).[35] Important technological advances included the use of instant feedback "dials" that allowed focus group participants to respond to the content they were hearing in real time.[36] Asked in an interview whether "language can change a paradigm," Luntz replied:

> 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).

Among the specific persuasive innovations credited to Luntz are the use of "death tax" in place of "estate tax" and of "climate change" in place of "global warming" (Luntz 2004). A 2006

[35] By his own description, Luntz specializes in "testing language and finding words that will help his clients... turn public opinion on an issue or a candidate" (Luntz 2004).

[36] Luntz said, "[The dial technology is] like an X-ray that gets inside [the subject's] head... it picks out every single word, every single phrase [that the subject hears], and you know what works and what doesn't" (Luntz 2004).

memorandum written by Luntz and distributed to Republican congressional candidates provides detailed advice on language to use on topics including taxes, budgets, social security, and trade (Luntz 2006).

We can use our data to look directly at the importance of the *Contract with America* in shaping congressional speech. We extract all phrases that appear in the text of the *Contract* and treat them as a single "topic," computing both their frequency and their partisanship in each session. Figure 11 reports the results. As expected, the frequency of these phrases spikes in the 104th session (1995-1996). Their partisanship rises sharply in that year and continues to increase even as their frequency declines.[37]

In the years after 1994, Democrats sought to replicate what they perceived to have been a highly successful Republican strategy. George Lakoff, a linguist who advised many Democratic candidates, writes: "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). Strategic imitation and continued diffusion of the Republicans' linguistic innovation is consistent with the continuing increase in partisanship we see up to the end of our data in 2009.

The new attention to crafting language coincided with attempts to impose greater party discipline in speech. In the 101st session (1989-1991), the Democrats established the "Democratic Message Board" which would "defin[e] a cohesive national Democratic perspective" (quoted from party documents in Harris 2013). The "Republican Theme Team" formed in the 102nd session (1991-1993) sought likewise to "develop ideas and phrases to be used by all Republicans" (Michel 1993 and quoted in Harris 2013).

Changes in the media environment may also have contributed to the increase in partisanship.[38] Prior to the late 1970s, television cameras were only allowed on the floor of Congress for special hearings or other events. With the introduction of the C-SPAN cable network to the House in 1979, and the C-SPAN2 cable network to the Senate in 1986, every speech was recorded and broadcast live. While live viewership of these networks has always been limited, they created a video record of speeches that could be used for subsequent press coverage and in candidates' own advertising. This plausibly increased the return to carefully crafted language, both by widening the

[37] According to the metric defined in table 1, the most Republican phrases in the 104th session (1995-1996) that appear in the *Contract* are "tax relief," "save account," "term limit," "item veto," "care insur," "term care," "fiscal respons," "tax increas," "vote term," and "legal reform." We accessed the text of the *Contract* at http://wps.prenhall.com/wps/media/objects/434/445252/DocumentsLibrary/docs/contract.htm on May 18, 2016.

[38] Our discussion of C-SPAN is based on Frantzich and Sullivan (1996).

reach of successful sound bites, and by dialing up the cost of careless mistakes. The subsequent introduction of the Fox News cable network and the increasing partisanship of cable news more generally (Martin and Yurukoglu 2016) may have further increased this return.

The timing shown in figure 3 is inconsistent with the C-SPAN networks being the proximate cause of increased partisanship. But it seems likely that they provided an important complement to linguistic innovation in the 1990s. Gingrich particularly encouraged the use of "special order" speeches outside of the usual legislative debate protocol, which allowed congresspeople to speak directly for the benefit of the television cameras. The importance of television in this period is underscored by Frantzich and Sullivan (1996): "When asked whether he would be the Republican leader without C-SPAN, Ginigrich... [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."

7 Conclusion

A consistent theme of much prior literature is that political polarization today—both in Congress and among voters—is not that different from what existed in the past (McCarty et al. 2015; Fiorina and Abrams 2008; Glaeser and Ward 2006). We find that language is a striking exception: Democrats and Republicans now speak different languages to a far greater degree than ever before. The fact that partisan language diffuses widely through media and public discourse (Gentzkow and Shapiro 2010; Martin and Yurukoglu 2016; Greenstein and Zhu 2012; Jensen et al. 2012) implies that this could be true not only for congresspeople but for the American electorate more broadly.

Does growing partisanship of language matter? Although measuring the effects of language is beyond the scope of this paper, existing evidence suggests that these effects could be profound. Laboratory experiments show that varying the way political issues are "framed" can have large effects on public opinion across a wide range of domains including free speech (Nelson et al. 1997), immigration (Druckman et al. 2013), climate change (Whitmarsh 2009), and taxation (Birney et al. 2006; Graetz and Shapiro 2006). Politicians routinely hire consultants to help them craft messages for election campaigns (Johnson 2015) and policy debates (Lathrop 2003), an investment that only makes sense if language matters. Field studies reveal effects of language on outcomes including marriage (Caminal and Di Paolo 2015), political preferences (Clots-Figueras and Masella 2013), and savings and risk choices (Chen 2013).

Language is also one of the most fundamental cues of group identity, with differences in lan-

guage or accent producing own-group preferences even in infants and young children (Kinzler et al. 2007). Imposing a common language was a key factor in the creation of a common French identity (Weber 1976), and Catalan-language education has been effective in strengthening a distinct Catalan identity within Spain (Clots-Figueras and Masella 2013). That the two political camps in the US increasingly speak different languages may contribute to the striking increase in inter-party hostility evident in recent years (Iyengar et al. 2012).

Beyond our substantive findings, we introduce a method that can be applied to the many settings in which researchers wish to characterize differences in behavior between groups and the space of possible choices is high-dimensional. One such domain is residential segregation. Abrams and Fiorina (2012) argue that to evaluate recent claims that Americans are increasingly segregating according to their political views ideally requires sub-county data. Even large surveys will have few respondents in any given zipcode or Census tract, so the finite-sample issues that our methodology confronts are likely to arise in such data. Another potential application is to media consumption online. Gentzkow and Shapiro (2011) show that Internet news media are only slightly more segregated by political ideology than traditional media. They provide only limited evidence on how segregation online has evolved over time, and their data is at the domain level. Studying trends over time and accounting for sub-domain behavior would almost certainly require confronting the finite-sample issues that we highlight here.

More broadly, measuring trends in group differences, especially racial or gender differences, is a core topic in quantitative social science. Finite-sample issues are pervasive in such measurement problems, which range from studies of racial differences in first names (Fryer and Levitt 2004) to studies of racial and gender segregation in the workplace (Carrington and Troske 1997; Bayard et al. 2003; Hellerstein and Neumark 2008). This paper introduces a new method that is grounded in a choice model and designed to control finite-sample bias through penalization.

References

Abrams, Samuel J. and Morris P. Fiorina. 2012. "The big sort" that wasn't: A skeptical reexamination. *PS: Political Science & Politics* 45(2): 203–222.

Alcalde-Unzu, Jorge and Marc Vorsatz. 2013. Measuring the cohesiveness of preferences: An axiomatic analysis. *Social Choice and Welfare* 41(4): 965–988.

Allen, Rebecca, Simon Burgess, Russel Davidson, and Frank Windmeijer. 2015. More reliable inference for the dissimilarity index of segregation. *Econometrics Journal* 18(1): 40–66.

Amer, Mildred L. 1993. The congressional record; content, history, and issues. Washington DC: Congressional Research Service. CRS Report No. 93-60 GOV.

Andrews, Wilson and Larry Buchanan. 2016. Mass shooting or terrorist attack? Depends on your party. *New York Times*, June 13, 2016. Accessed at <http://www.nytimes.com/interactive/2016/06/13/us/politics/politicians-respond-to-orlando-nightclub-attack.html?_r=1> on June 24, 2016.

Antweiler, Werner and Murray Z. Frank. 2004. Is all that talk just noise? The information content of internet stock message boards. *Journal of Finance* 59(3): 1259–1294.

Bai, Matt. 2005. The framing wars. *New York Times*, July 17, 2005. Accessed at <http://www.nytimes.com/2005/07/17/magazine/the-framing-wars.html?_r=0> on June 16, 2016.

Baker, Stuart G. 1994. The multinomial-Poisson transformation. *The Statistician*: 495-504.

Ballester, Coralio and Marc Vorsatz. 2014. Random walk-based segregation measures. *Review of Economics and Statistics* 96(3): 383–401.

Bayard, Kimberly, Judith Hellerstein, David Neumark, and Kenneth Troske. 2003. New evidence on sex segregation and sex differences in wages and matched employee-employer data. *Journal of Labor Economics* 21(4): 887–922.

Bayer, Patrick, Robert McMillan, and Kim Rueben. 2002. An equilibrium model of sorting in an urban housing market: A study of the causes and consequences of residential segregation. NBER Working Paper No. 10865.

Birney, Mayling, Michael J. Graetz, and Ian Shapiro. 2006. Public opinion and the push to repeal the estate tax. *National Tax Journal* 59(3): 439-461.

Blei, David M., and John D. Lafferty. 2007. A correlated topic model of science. *The Annals of Applied Statistics*: 17-35.

Caminal, Ramon and Antonio Di Paolo. 2015. Your language or mine? Barcelona GSE Working Paper No. 852.

Carrington, William J. and Kenneth R. Troske. 1997. On measuring segregation in samples with small units. *Journal of Business & Economic Statistics* 15(4): 402–409.

Chen, M. Keith. 2013. The effect of language on economic behavior: Evidence from savings rates, health behaviors, and retirement assets. *The American Economic Review* 103(2): 690-731.

Chong, Dennis and James N. Druckman. 2007. Framing public opinion in competitive democracies. *American Political Science Review* 101(4): 637–655.

Clots-Figueras, Irma, and Paolo Masella. 2013. Education, language and identity. *The Economic Journal* 123(570): F332-F357.

Congressional Record 43-104. Available from LexisNexis® Congressional. Accessed in April, 2009.

Cortese, Charles F., R. Frank Falk, and Jack K. Cohen. 1976. Further considerations on the methodological analysis of segregation indices. *American Sociological Review* 41(4): 630–637.

Cutler, David M., Edward L. Glaeser, and Jacob L. Vigdor. 1999. The rise and decline of the American ghetto. *Journal of Political Economy* 107(3): 455–506.

Democratic National Committee. 2016. Health Care. Accessed at <https://www.democrats.org/issues/health-care> on June 24, 2016.

Druckman, James N., Erik Peterson, and Rune Slothuus. 2013. How elite partisan polarization affects public opinion formation. *American Political Science Review* 107(1): 57-79.

Echenique, Frederico and Roland G. Fryer Jr. 2007. A measure of segregation based on social interactions. *Quarterly Journal of Economics* 122(2): 441–485.

Ellison, Glenn and Edward L. Glaeser. 1997. Geographic concentration in US manufacturing industries: A dartboard approach. *Journal of Political Economy* 105(5): 889–927.

Fan, Jianqing and Runze Li. 2001. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association* 96(456): 1348–1360.

Fiorina, Morris P. and Samuel J. Abrams. 2008. Political polarization in the American public. *Annual Review of Political Science* 11: 563-588.

Fiorina, Morris P., Samuel J. Abrams, and Jeremy C. Pope. 2005. *Culture war? The myth of a polarized America.* New York: Pearson Longman.

Flynn, Cheryl J., Clifford M. Hurvich, and Jeffrey S. Simonoff. 2013. Efficiency for regularization parameter selection in penalized likelihood estimation of misspecified models. *Journal of the American Statistical Association* 108(503): 1031–1043.

Fossett, Mark. 2011. Generative models of segregation: Investigating model-generated patterns of residential segregation by ethnicity and socioeconomic status. *Journal of Mathematical Sociology* 35(1–3): 114–145.

Frankel, David M. and Oscar Volij. 2011. Measuring school segregation. *Journal of Economic Theory* 146(1): 1–38.

Frantzich, Stephen and John Sullivan. 1996. *The C-SPAN revolution*. Norman OK: University of Oklahoma Press.

Fryer, Roland G., Jr. and Steven D. Levitt. 2004. The causes and consequences of distinctively black names. *Quarterly Journal of Economics* 119(3): 767–805.

Gentzkow, Matthew and Jesse M. Shapiro. 2010. What drives media slant? Evidence from U.S. daily newspapers. *Econometrica* 78(1): 35–71.

Gentzkow, Matthew and Jesse M. Shapiro. 2011. Ideological segregation online and offline. *Quarterly Journal of Economics* 126(4): 1799–1839.

Gentzkow, Matthew, Jesse M. Shapiro, and Matt Taddy. 2015. Measuring polarization in high-dimensional data: Method and application to Congressional speech. Stanford University mimeo. Accessed at <http://web.stanford.edu/∼gentzkow/research/politext.pdf> on July 6, 2016.

Glaeser, Edward L. and Bryce A. Ward. 2006. Myths and realities of American political geography. *The Journal of Economic Perspectives* 20(2): 119-144.

Graetz, Michael J. and Ian Shapiro. 2006. *Death by a thousand cuts: The fight over taxing inherited wealth*. Princeton, NJ: Princeton University Press.

Greenstein, Shane and Feng Zhu. 2012. Is Wikipedia Biased? *American Economic Review: Papers and Proceedings.* 102(3): 343-348.

Grimmer, Justin. 2010. A bayesian hierarchical topic model for political texts: Measuring expressed agendas in Senate press releases. *Political Analysis* 18(1): 1–35.

Haberman, Shelby J. 1973. Log-linear models for frequency data: Sufficient statistics and likelihood equations. *Annals of Statistics* 1(4): 617–632.

Harris, Douglas B. 2013. Let's play hardball: Congressional partisanship in the television era. In *Politics to the extreme: American political institutions in the twenty-first century*, ed. Scott A. Frisch and Sean Q. Kelly, 93-115. New York: Palgrave MacMillan.

Hellerstein, Judith K. and David Neumark. 2008. Workplace segregation in the United States: Race, ethnicity, and skill. *Review of Economics and Statistics* 90(3): 459–477.

Issenberg, Sasha. 2012. The death of the hunch. *Slate*, May 22, 2012. Accessed at <http://www.slate.com/articles/news_and_politics/victory_lab/2012/05/obama_campaign_ads_how_the_analyst_institute_is_helping_him_hone_his_message_.html> on June 16, 2016.

Iyengar, Shanto, Gaurav Sood, and Yphtach Lelkes. 2012. Affect, not ideology a social identity perspective of polarization. *Public Opinion Quarterly* 76(3): 405-431.

Jacobson, Gary C. 1996. The 1994 House elections in perspective. *Political Science Quaterly* 111(2): 203–223.

Jensen, Jacob, Suresh Naidu, Ethan Kaplan, and Laurence Wilse-Samson. 2012. Political polarization and the dynamics of political language: Evidence from 130 years of partisan speech. *Brookings Papers on Economic Activity*: 1–81.

Johnson, Dennis W. 2015. *Political consultants and American elections: Hired to fight, Hired to*

win. Routledge.

Kim, In Song, John Londregan, and Marc Ratkovic. 2015. Voting, speechmaking, and the dimensions of conflict in the US Senate. Princeton mimeo. Accessed at <https://www.princeton.edu/∼ratkovic/public/sfa.pdf> on June 16, 2016.

Kinzler, Katherine D., Emmanuel Dupoux, and Elizabeth S. Spelke. 2007. The native language of social cognition. *Proceedings of the National Academy of Sciences* 104(30): 12577-12580.

Lakoff, George. 2003. Framing the issues: UC Berkeley professor George Lakoff tells how conservatives use language to dominate politics. *UC Berkeley News*, Octoboer 27, 2003. Accessed at <http://www.berkeley.edu/news/media/releases/2003/10/27_lakoff.shtml> on June 16, 2016.

—. 2004. *Don't think of an elephant! Know your values and frame the debate the essential guide for progressives.* White River Junction, VT: Chelsea Green.

—. 2014. *The all new don't think of an elephant!: Know your values and frame the debate.* White River Junction, VT: Chelsea Green.

Lamy, Lynne C. 1988. Recent developments in construction industry bargaining: Doublebreasting and prehire agreements. *Missouri Law Review* 53(3): 465-495.

Lathrop, Douglas A. 2003. *The campaign continues: How political consultants and campaign tactics affect public policy.* ABC-CLIO.

Laver, Michael, Kenneth Benoit, and John Garry. 2003. Extracting policy positions from political texts using words as data. *American Political Science Review* 97(2): 311–331.

Library of Congress. 2015. Congressional Globe. Accessed at <http://memory.loc.gov/ammem/amlaw/lwcg.html> on June 11, 2015.

Luntz, Frank I. 2004. Interview Frank Luntz. *Frontline*, November 9, 2004. Accessed at <http://www.pbs.org/wgbh/pages/frontline/shows/persuaders/interviews/luntz.html> on June 16, 2016.

—. 2006. The new American lexicon. *Luntz Research Companies*. Accessed at *<https://drive.google.com/file/d/0B_2I29KBujFwNWY2MzZmZjctMjdmOS00ZGRhLWEyY2MtMGE1MDMyYzVjYWM2/view>* on June 16, 2016.

—. 2009. The language of healthcare. Accessed at <http://thinkprogress.org/wp-content/uploads/2009/05/frank-luntz-the-language-of-healthcare-20091.pdf> on June 24, 2016.

Martin, Gregory J. and Ali Yurukoglu. 2016. Bias in cable news: Persuasion and polarization. Stanford University mimeo. Accessed at *<http://web.stanford.edu/∼ayurukog/cable_news.pdf>* on July 7, 2016.

Martis, Kenneth C. 1989. *The Historical Atlas of Political Parties in the United States Congress, 1789-1989.* New York: Macmillan Publishing Company.

McCardell, John M., Jr. 2004. Reflections on the Civil War. *Sewanee Review* 122(2): 295-303.

McCarty, Nolan, Keith Poole, and Jeff Lewis. 2009. NOMINATE and related data. *Voteview*. Accessed at <http://www.voteview.com/> in April, 2009.

McCarty, Nolan, Keith Poole, and Howard Rosenthal. 2015. The polarization of congressional parties. *Voteview*, March 21, 2015. Accessed at <http://voteview.com/political_polarization_2014.html> on June 16, 2016.

Mele, Angelo. 2013. Poisson indices of segregation. *Regional Science and Urban Economics* 43(1): 65–85.

—. 2015. A structural model of segregation in social networks. Johns Hopkins University mimeo.

Michel, Robert Henry. 2013. The theme team. Accessed at <http://www.robertmichel.name/RHM_blueprint/blueprint_Theme%20Team.pdf> on June 16, 2016.

Mosteller, Frederick and David L. Wallace. 1963. Inference in an authorship problem. *Journal of the American Statistical Association* 58(302): 275–309.

Murphy, Kevin P. 2012. *Machine Learning: A Probabilistic Perspective.* Cambridge, MA: MIT Press.

Nelson, Thomas E., Rosalee A. Clawson, and Zoe M. Oxley. 1997. Media framing of a civil liberties conflict and its effect on tolerence. *American Political Science Review* 91(3): 567–583.

Newman, David. 1985. The evolution of a political landscape: Geographical and territorial implications of Jewish colonization in the West Bank. *Middle Eastern Studies* 21(2): 192-205.

Orwell, George. 1946. Politics and the English language. *Horizon* 13(76): 252–265.

Palmgren, Juni. 1981. The Fisher information matrix for log linear models arguing conditionally on observed explanatory variable. *Biometrika* 68 (2): 563-566.

Peace, Roger. 2010. Winning hearts and minds: The debate over U.S. intervention in Nicaragua in the 1980s. *Peace & Change* 35(1): 1-38.

Peters, Gerhard and John T. Woolley. 2016. Political party platforms of parties receiving electoral votes. *American Presidency Project*. Accessed at <http://www.presidency.ucsb.edu/platforms.php> on June 16, 2016.

Politis, Dimitris N., Joseph P. Romano, and Michael Wolf. 1999. *Subsampling*. New York: Springer series in statistics.

Poole, Keith T. and Howard Rosenthal. 1985. A spatial model for legislative roll call analysis. *American Journal of Political Science* 29(2): 357–384.

Porter, Martin. 2009. Snowball. Accessed at <http://snowball.tartarus.org/> on November 11, 2010.

ProQuest Congressional. 2016. Legislative and executive publications. Accessed at

<http://congressional.proquest.com/congressional/search/basic/basicsearch> on June 16, 2016.

Rathelot, Roland. 2012. Measuring segregation when units are small: A parametric approach. *Journal of Business & Economic Statistics* 30(4): 546–533.

Reardon, Sean F. and Glenn Firebaugh. 2002. Measures of multigroup segregation. *Sociological Methodology* 32(1): 33–67.

Riddick, Floyd. 1992. Riddick's Senate procedure: Precedents and practices. Accessed at <http://www.gpoaccess.gov/riddick/1441-1608.pdf> on August 11, 2010.

Robert, Henry M. 1876. *Robert's rules of order.* Chicago, IL: S. C. Griggs & Company.

Ruggles, Steven, Katie Genadek, Ronald Goeken, Josiah Grover, and Matthew Sobek. 2008. *Integrated Public Use Microdata Series: Version 4* [Machine-readable database]. Minneapolis: University of Minnesota.

Silva, J.M.C. Santos and Silvana Tenreyro. 2010. On the existence of the maximum likelihood estimates in Poisson regression. *Economics Letters* 107(2): 310–312.

Swift, Elaine K., Robert G. Brookshire, David T. Canon, Evelyn C. Fink, John R. Hibbing, Brian D. Humes, Michael J. Malbin, and Kenneth C. Martis. 2009. Database of Congressional Historical Statistics, 1789–1989. *ICPSR Study No. 3371.* Accessed at <http://www.icpsr.umich.edu/cocoon/ICPSR/STUDY/03371.xml> in April, 2009.

Taddy, Matt. 2013. Multinomial inverse regression for text analysis. *Journal of the American Statistical Association* 108(503): 755–770.

—. 2015. Distributed multinomial regression. *The Annals of Applied Statistics* 9(3): 1394-1414.

—. Forthcoming. One-step estimator paths for concave regularization. *Journal of Computational and Graphical Statistics.*

Tetlock, Paul C. 2007. Giving content to investor sentiment: The role of media in the stock market. *Journal of Finance* 62(3): 1139–1168.

Tibshirani, Robert. 1996. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B (Methodological)* 58(1): 267–288.

United States Census Bureau. 1990. 1990 TIGER/GICS (Geography Identification Code Scheme) Census File. Washington, D.C. CD Rom.

United States Government Publishing Office. 2011. Congressional Record. Accessed at http://gpoaccess.gov/ in December 2011.

United States Senate. 2013. Dates of the sessions of the Congress. Accessed at <http://www.senate.gov/reference/Sessions/sessionDates.htm> on September 13, 2011.

Weber, Eugen. 1976. *Peasants into Frenchmen: The modernization of rural France 1870-1914*. Stanford, CA: Stanford University Press.

White, Michael J. 1986. Segregation and diversity measures in population distribution. *Population*

Index 52(2): 198–221.

Whitmarsh, Lorraine. 2009. What's in a name? Commonalities and differences in public understanding of "climate change" and "global warming". *Public understanding of science* 18(4): 401-420.

Zou, Hui, Trevor Hastie, and Robert Tibshirani. 2007. On the "degrees of freedom" of the lasso. *Annals of Statistics* 35(5): 2173–2192.

Appendices

A Filtering procedural phrases

We start by obtaining an electronic copy of "Robert's Rules of Order" (1876), a widely accepted manual that explains procedures of assemblies.[39] We also obtain an electronic copy of the appendix from "Riddick's Senate Procedure" for the 101st session (1989–1991), a parliamentary authority explaining the rules, ethics, and customs governing meetings and other operations of the United States Senate, arranged in a glossary style.[40] All the bigrams that we parse from the two documents are then considered procedural phrases. If a speech contains many procedural phrases, it is likely to be a procedural speech. We use this fact to filter out more procedural phrases using some occurrence rules. We define any speech in which 30 percent phrases are procedural according to Riddick's or Robert's manual as a highly procedural speech with respect to that manual. A procedural speech is one that is highly procedural with respect to at least one manual. We then count the number of times a phrase appears in a highly procedural speech, the number of times a phrase is used in total, and the percentage of procedural speeches in which a phrase occurs.[41] We have two separate rules to identify procedural phrases.

A phrase qualifies as procedural using our first rule if one of the following sets of conditions applies:

- It appears in at least 5 procedural speeches in more than 5 sessions and one of: 1) it appears in more than 5,200 highly Robert speeches, and at least 1.75 percent of speeches it appears in are highly Robert; or 2) it appears in more than 100 highly Robert speeches, and at least 7.5 percent of speeches it appears in are highly Robert; or 3) it appears in more than 50 highly Robert speeches, and more than 30 percent of speeches it appears in are highly Robert.

- It appears in at least 5 highly Robert speeches in more than 10 sessions and one of: 1) it appears in more than 2,000 highly Robert speeches, and at least 1 percent of speeches it appears in are highly Robert; or 2) it appears in more than 100 highly Robert speeches, and at least 5 percent of speeches it appears in are highly Robert; or 3) it appears in more than 50 highly Robert speeches, and at least 20 percent of speeches it appears in are highly Robert.

- It appears in at least 5 highly Riddick speeches in more than 10 sessions and one of: 1) it appears in at least 3,000 Riddick speeches, and at least 1.75 percent of speeches it appears

[39] The text version is downloaded from Project Gutenberg http://www.gutenberg.org/etext/9097. The file was obtained in early August 2009 and is the original 1876 version of the document. There have since been ten additional editions.

[40] The PDF version is downloaded from http://www.gpoaccess.gov/riddick/1441-1608.pdf on August 11, 2010 and converted into text using Optical Character Recognition with metadata cleaned out.

[41] For computational purposes, we drop all phrases that appear at most once in each session.

in are highly Riddick; or 2) it appears in at least 100 Riddick speeches, and at least 7 percent of speeches it appears in are highly Riddick; or 3) it appears in at least 50 highly Riddick speeches, and at least 20 percent of speeches it appears in are highly Riddick.

We compute, for every phrase, the average percentage of Robert's procedural phrases/Riddick's procedural phrases in all speeches in which the phrase appears. Of the phrases that are not identified by our first rule, a phrase qualifies as procedural using our second rule if one of the following sets of conditions applies:

- 1) It is mentioned over 500 times; and 2) it appears in more than 5 sessions; and 3) speeches that it occurs in average over 5 percent Robert procedural phrases.
- 1) It is mentioned over 20,000 times; and 2) it appears in more than 10 sessions; and 3) speeches that it occurs in average over 7.5 percent Riddick procedural phrases.
- 1) It is mentioned over 500 times; and 2) it appears in more than 10 sessions; and 3) speeches that it occurs in average over 9.6 percent Riddick procedural phrases.

We choose the cut-off points such that phrases that just make the cut-offs are subjectively procedural,[42] whereas phrases that do not make the cut-offs are subjectively not procedural.[43]

[42] Examples: phrases with bill numbers, committee names.
[43] Examples: veteran associ, war time, victim hurrican.

Table 1: Most partisan phrases by session

Session 50 (1887-1888)

Republican	m^R_{jt}	m^D_{jt}	Democrat	m^R_{jt}	m^D_{jt}
st loui	229	225	cutleri compani	29	36
dispens follow	5	3	public domain	248	281
street along	12	4	increas duti	120	207
intellig charact	7	3	high protect	32	149
sixth street	65	9	tariff tax	63	157
open poll	17	3	state vs	26	67
eleventh street	17	1	high tariff	43	147
navig compani	103	14	articl xviii	4	49
fisheri treati	147	52	bulwer treati	5	41
station servic	5	0	public defens	44	68

Session 60 (1907-1908)

Republican	m^R_{jt}	m^D_{jt}	Democrat	m^R_{jt}	m^D_{jt}
infantri war	71	9	section corner	1	5
indian war	106	19	ship subsidi	12	89
mount volunt	34	0	republ panama	27	34
feet thenc	29	0	level canal	36	51
postal save	345	38	powder trust	11	31
spain pay	50	6	print paper	65	89
war pay	57	5	lock canal	7	74
first regiment	191	21	bureau corpor	19	55
soil survey	134	95	senatori term	3	12
nation forest	315	56	remov wreck	2	6

Session 70 (1927-1928)

Republican	m^R_{jt}	m^D_{jt}	Democrat	m^R_{jt}	m^D_{jt}
pay period	49	4	wish announc	68	108
rate advantag	9	1	radio commiss	61	191
nation guard	289	119	wave length	30	168
construct servic	21	2	necessarili detain	5	73
organ reserv	85	22	coast guard	66	173
governor veto	16	0	citi los	53	94
war regular	13	4	report valu	1	13
air corp	226	65	feder trade	194	400
air mail	137	27	feder govern	801	1269
fourth pay	14	4	desir announc	137	173

Session 80 (1947-1948)

Republican	m^R_{jt}	m^D_{jt}	Democrat	m^R_{jt}	m^D_{jt}
steam plant	248	216	admir denfeld	3	13
coast guard	318	113	public busi	57	371
stop communism	239	49	labor standard	374	395
depart agricultur	1014	381	intern labor	16	125
lend leas	298	109	tax refund	25	124
zone germani	124	29	concili servic	91	176
british loan	99	33	standard act	289	303
approv compact	102	13	soil conserv	324	421
unit kingdom	164	45	school lunch	140	222
union shop	271	106	cent hour	31	112

Session 90 (1967-1968)

Republican	m^R_{jt}	m^D_{jt}	Democrat	m^R_{jt}	m^D_{jt}
delinqu employ	50	0	food stamp	353	1093
red china	375	322	support presid	114	533
secretari wirtz	76	47	stamp program	180	661
human invest	221	12	forc labor	15	357
invest act	206	42	polit right	25	345
aid highway	232	159	commend presid	17	276
demonstr citi	182	69	lunch program	118	374
highway program	260	193	wholesom meat	23	198
oblig author	232	179	school lunch	188	522
nation home	137	28	farmer home	77	302

Session 100 (1987-1988)

Republican	m^R_{jt}	m^D_{jt}	Democrat	m^R_{jt}	m^D_{jt}
freedom fighter	1118	185	star war	53	446
doubl breast	170	84	contra aid	354	1021
abort industri	9	2	nuclear weapon	776	1664
demand second	451	5	contra war	9	148
heifer tax	62	37	support contra	152	357
reserv object	694	77	nuclear wast	554	1259
incom ballist	17	0	agent orang	86	514
communist govern	376	34	central american	518	999
withdraw reserv	636	78	nicaraguan govern	123	324
abort demand	58	6	haitian peopl	10	124

Session 110 (2007-2008)

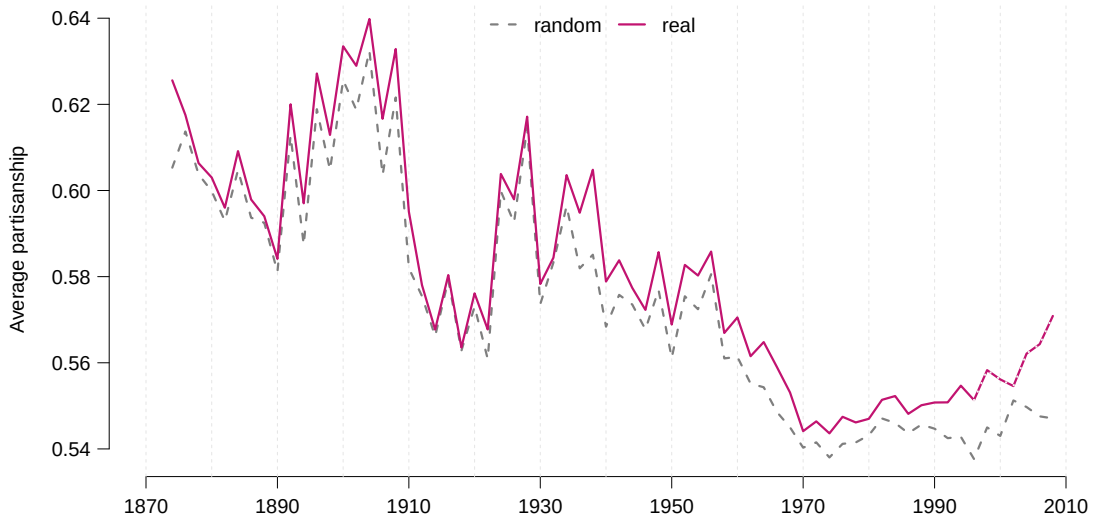
Republican	m^R_{jt}	m^D_{jt}	Democrat	m^R_{jt}	m^D_{jt}
conclus god	169	0	baccalaur cours	7	70
tax increas	3636	677	afghanistan veteran	6	85
age oil	83	0	respons redeploy	3	169
american energi	986	105	contract halliburton	1	24
illeg immigr	1146	344	engag unfortun	1	23
higher tax	415	73	bear sorrow	8	28
island brooklyn	13	0	cbc budget	1	106
illeg alien	619	97	nomin lifetim	2	40
increas american	660	63	neutral reserv	31	476
ms ginni	64	31	redeploy iraq	15	170

Notes: Calculations are based on our preferred specification in panel A of figure 3. The table shows the Republican and Democrat phrases with the greatest estimated partisanship ζ_{jt}, where Republican phrases are those for which $\tilde{\varphi}_{jt} > 0$ and $m^R_{jt} > m^D_{jt}$, and Democrat phrases are those for which $\tilde{\varphi}_{jt} < 0$ and $m^D_{jt} > m^R_{jt}$. Let $\pi^R_t(\mathbf{x}) = \mathbf{q}^R_t(\mathbf{x}) \cdot \boldsymbol{\rho}_t(\mathbf{x})$ and $\pi^D_t(\mathbf{x}) = \mathbf{q}^D_t(\mathbf{x}) \cdot (1 - \boldsymbol{\rho}_t(\mathbf{x}))$. Then the partisanship ζ_{jt} of phrase j in session t is defined as the average across the speakers in session t of:

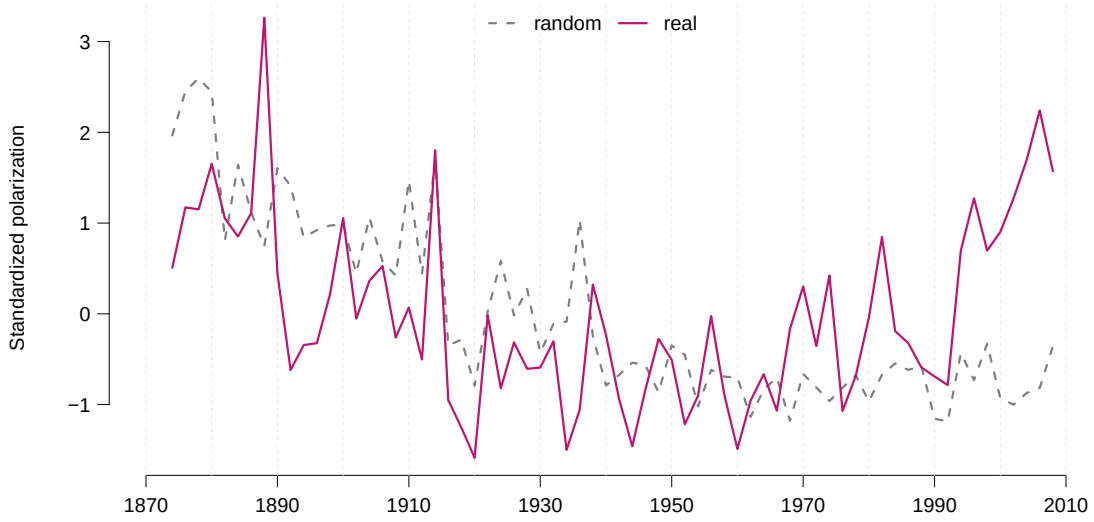
$$\frac{1}{2}\left(\frac{q^R_{jt}(\mathbf{x}_{it})}{1 - q^R_{jt}(\mathbf{x}_{it})}\right)\left(\rho_{jt}(\mathbf{x}_{it}) - \pi^R_t(\mathbf{x}_{it})\right) + \frac{1}{2}\left(\frac{q^D_{jt}(\mathbf{x}_{it})}{1 - q^D_{jt}(\mathbf{x}_{it})}\right)\left((1 - \rho_{jt}(\mathbf{x}_{it})) - \pi^D_t(\mathbf{x}_{it})\right).$$

Figure 1: Average partisanship and polarization of speech, plug-in estimates

Panel A: Partisanship from maximum likelihood estimator ($\hat{\pi}_t^{MLE}$)

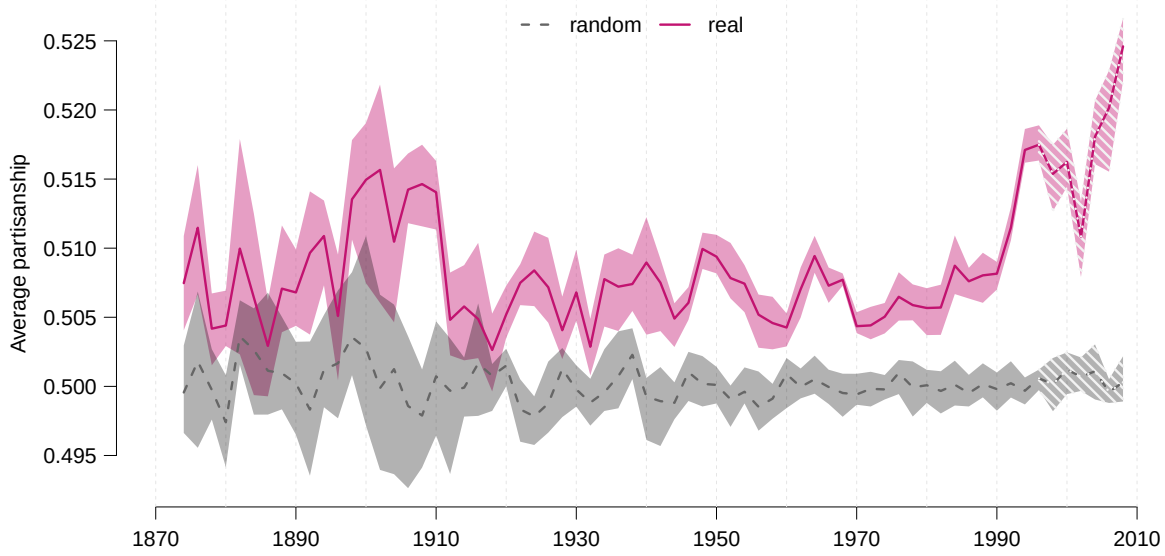


Panel B: Polarization from Jensen et al. (2012)



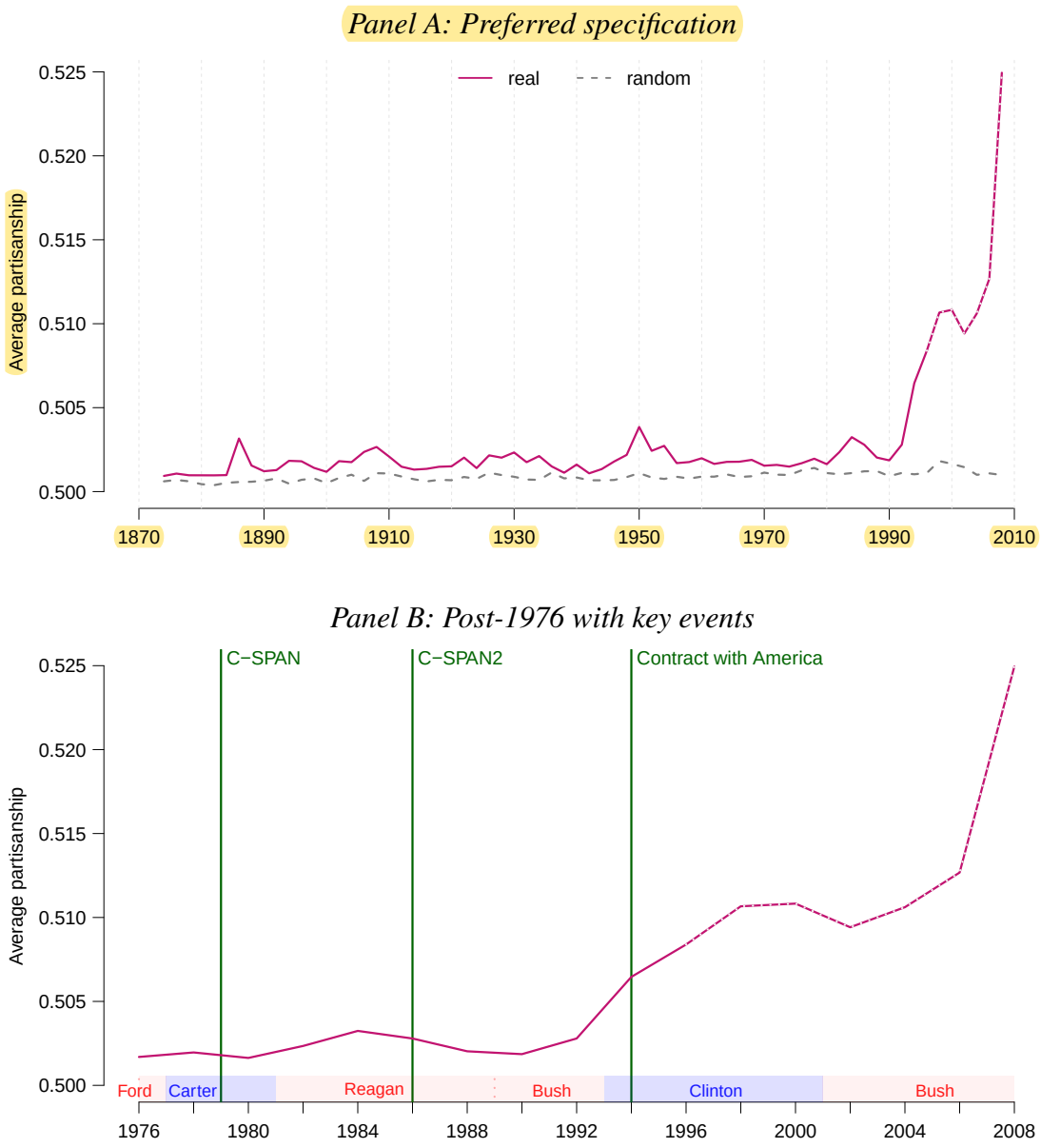
Notes: Panel A plots the average partisanship series from the maximum likelihood estimator $\hat{\pi}_t^{MLE}$ defined in section 4.1. "Real" series is from actual data; "random" series is from hypothetical data in which each speaker's party is randomly assigned with the probability that the speaker is Republican equal to the average share of speakers who are Republican in the sessions in which the speaker is active. Panel B plots the standardized measure of polarization from Jensen et al. (2012). Polarization in session t is defined as $\sum_j \left(m_{jt} \left| \rho_{jt} \right| \Big/ \sum_j m_{jt} \right)$ where $\rho_{jt} = corr(c_{ijt}, \mathbf{1}_{i \in R_t})$; the series is standardized by subtracting its mean and dividing by its standard deviation. "Real" series reproduces the polarization series in figure 3B of Jensen et al. (2012) using the replication data for that paper; "random" series uses the same data but randomly assigns each speaker's party with the probability that the speaker is Republican equal to the average share of speakers who are Republican in the sessions in which the speaker is active.

Figure 2: Average partisanship of speech, leave-out estimate ($\hat{\pi}_t^{LO}$)



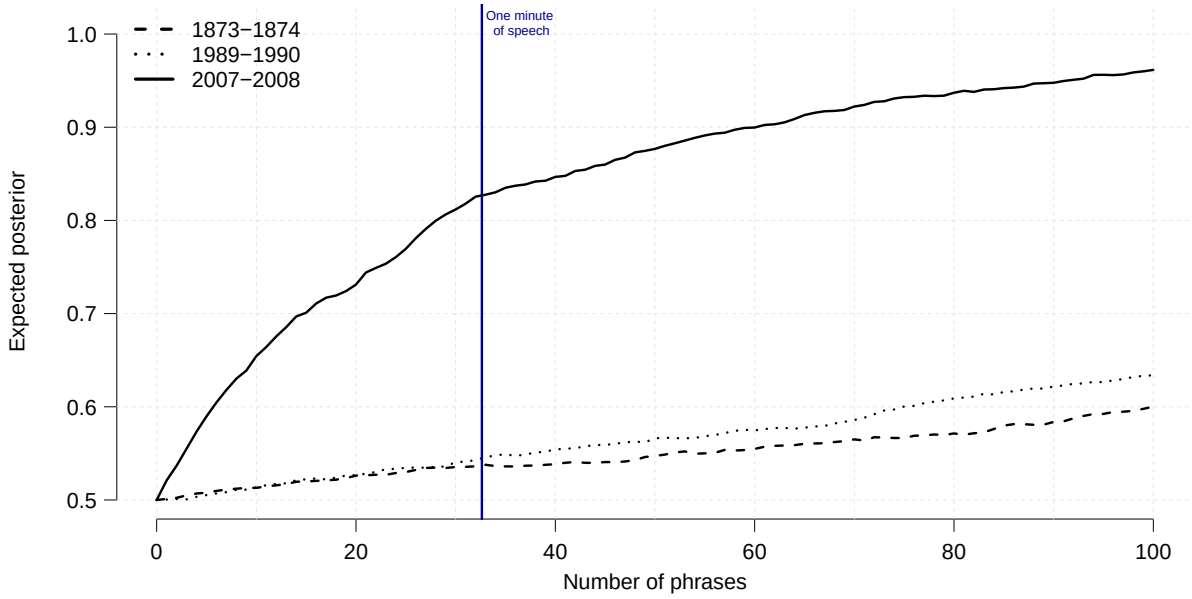
Notes: The figure plots the average partisanship series from the leave-out estimator $\hat{\pi}_t^{LO}$ defined in section 4.1. "Real" series is from actual data; "random" series is from hypothetical data in which each speaker's party is randomly assigned with the probability that the speaker is Republican equal to the average share of speakers who are Republican in the sessions in which the speaker is active. The shaded region around each series represents a pointwise confidence interval obtained by subsampling (Politis et al. 1999). Specifically, we randomly partition the set of speakers into 10 equal-sized subsamples (up to integer restrictions) and, for each subsample k, we compute the leave-out estimate $\hat{\pi}_t^k$. Define $Q_t^k = \sqrt{\tau_k}\left(\hat{\pi}_t^k - \frac{1}{10}\sum_{k=1}^{10}\hat{\pi}_t^k\right)$ where τ_k is the number of speakers in the k^{th} subsample. Our confidence interval is $(\hat{\pi}_t^{LO} - (Q_t^k)_{(9)}/\sqrt{\tau}, \hat{\pi}_t^{LO} - (Q_t^k)_{(2)}/\sqrt{\tau})$ where $\tau = |R_i \cup D_i|$ is the number of speakers in the full sample and $(Q_t^k)_{(b)}$ is the bth order statistic of Q_t^k.

Figure 3: Average partisanship of speech, penalized estimates



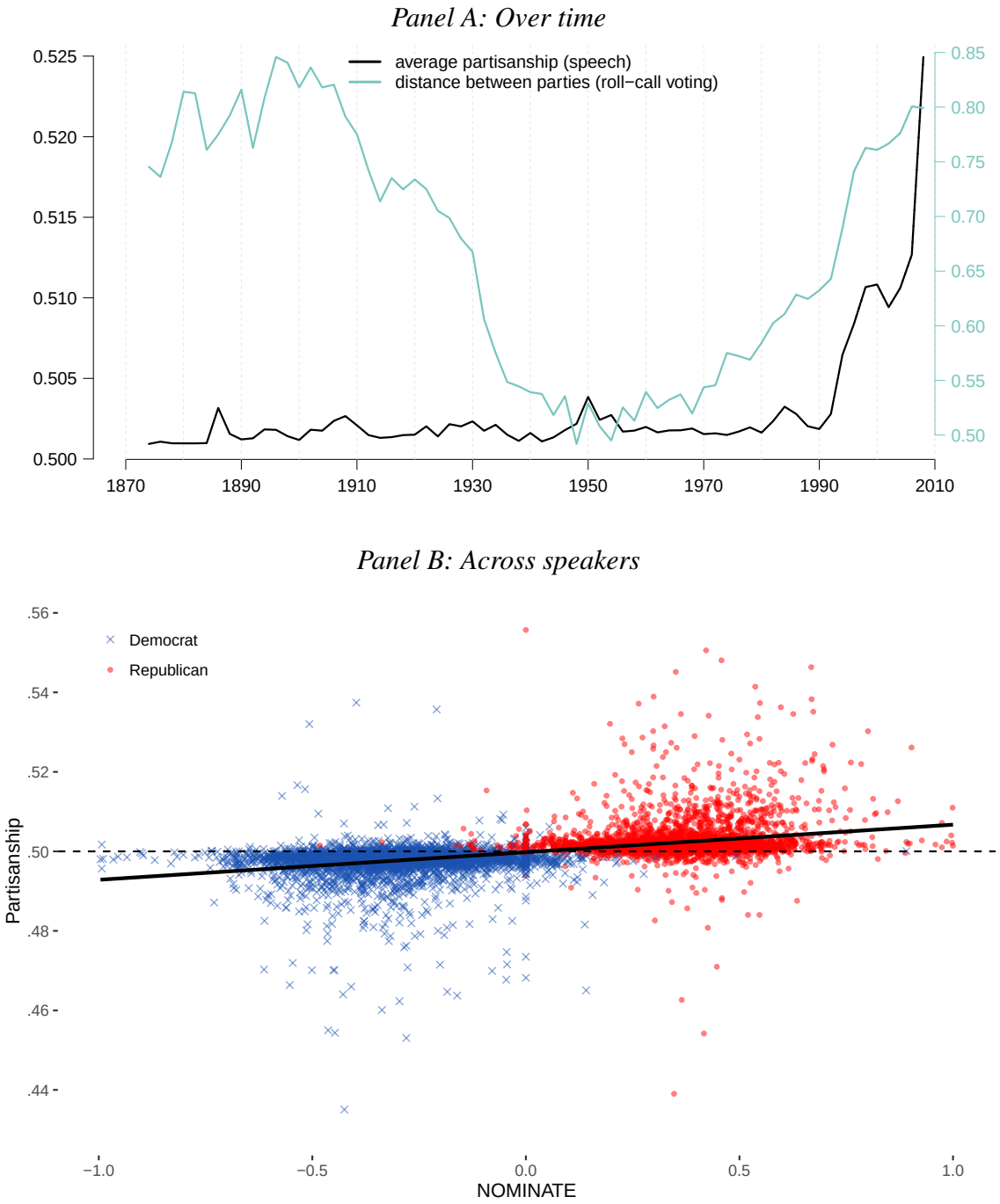
Notes: Panel A shows the results from our preferred penalized estimator defined in section 4.2. "Real" series is from actual data; "random" series is from hypothetical data in which each speaker's party is randomly assigned with the probability that the speaker is Republican equal to the average share of speakers who are Republican in the sessions in which the speaker is active. Panel B zooms in on average partisanship from the "real" series in panel A and includes party-coded shading for presidential terms and line markers for select events.

Figure 4: Informativeness of speech by speech length and session



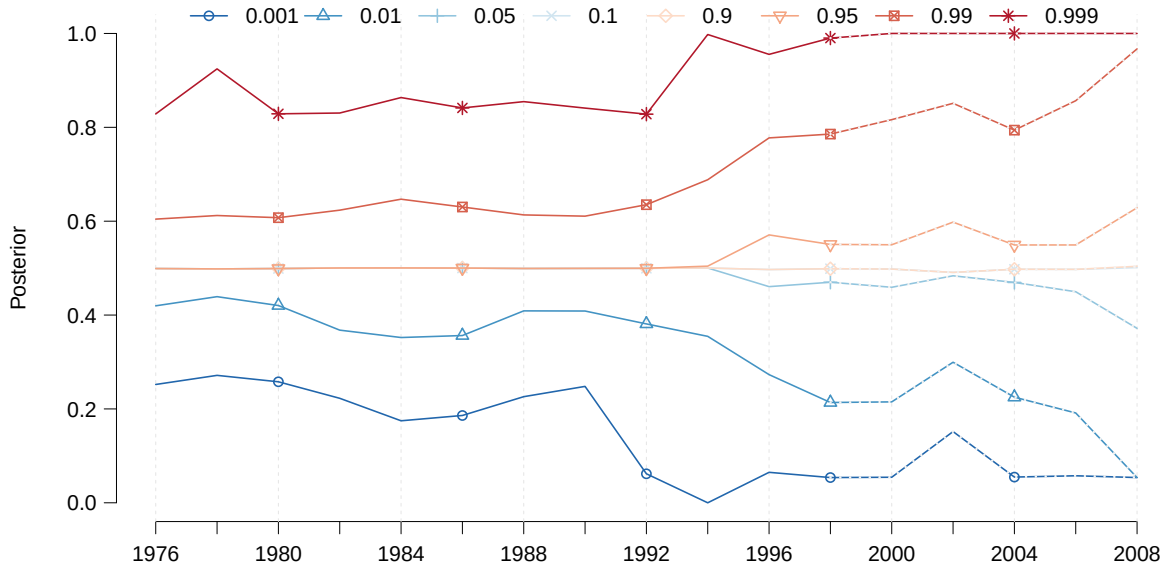
Notes: For each speaker i and session t we calculate, given characteristics $\mathbf{x}_{it}$, the expected posterior that an observer with a neutral prior would place on a speaker's true party after hearing a given number of phrases drawn according to the estimates in our preferred specification in panel A of figure 3. We perform this calculation by Monte Carlo simulation and plot the average across speakers for each given session and length of speech. We calculate the average number of phrases in one minute of speech as one-fifth of the average number of phrases in a sample of five-minute speeches. Our sample of five-minute speeches collects the ten speeches following any "special order" or "morning house debate" title in the GPO Congressional Record files and excludes speeches with fewer than 500 or more than 1000 words.

Figure 5: Partisanship vs. roll-call voting



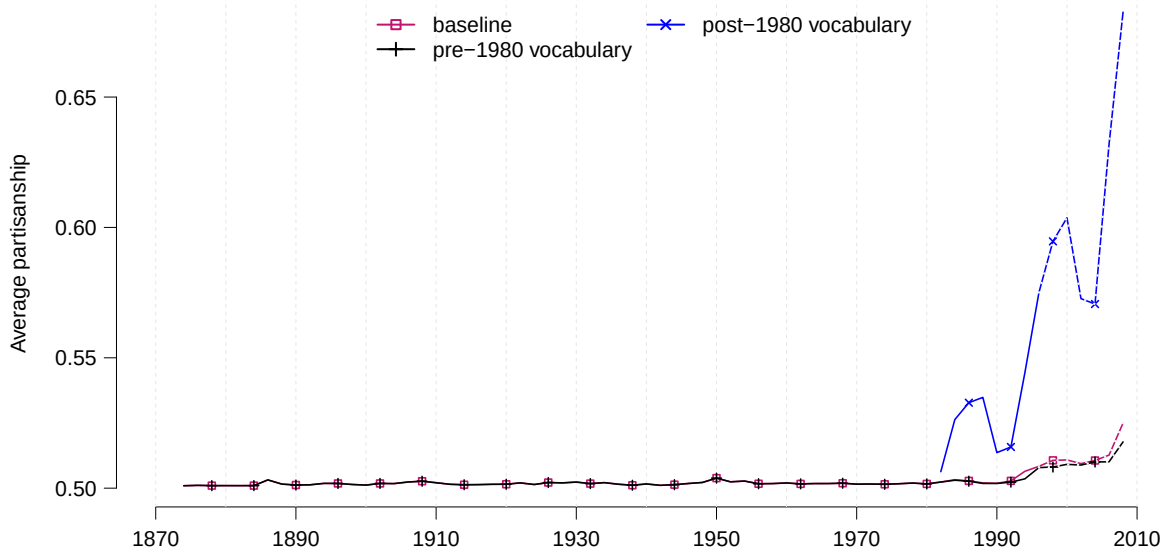
Notes: Panel A shows our preferred estimate of average partisanship from panel A of figure 3 and the difference between Republicans and Democrats in the first dimension of the CommonSpace DW-NOMINATE score from McCarty et al. (2009). Panel B plots a speaker's posterior probability $\hat{\rho}_i$ of being Republican based on speech against the first dimension of the CommonSpace DW-NOMINATE score (McCarty et al. 2009). To compute $\hat{\rho}_i$, we first define $\hat{\rho}_{it} = \hat{\mathbf{q}}_{it} \cdot \hat{\boldsymbol{\rho}}_t^*(\mathbf{x}_{it})$, where we recall that $\hat{\mathbf{q}}_{it} = \mathbf{c}_{it}/m_{it}$ are the empirical phrase frequencies for speaker i in session t and where we define $\hat{\boldsymbol{\rho}}_t^*(\mathbf{x}_{it})$ as the estimated value of $\boldsymbol{\rho}_t(\mathbf{x}_{it})$ from our baseline penalized estimates. We then let $\hat{\rho}_i = \frac{1}{|T_i|}\sum_{t \in T_i} \hat{\rho}_{it}$ where T_i is the set of all sessions in which speaker i appears. Four outliers with $\hat{\rho}_i > .6$ and positive DW-NOMINATE are excluded from the plot. The solid black line denotes the linear best fit among the points plotted.

Figure 6: Trends in the distribution of phrase informativeness



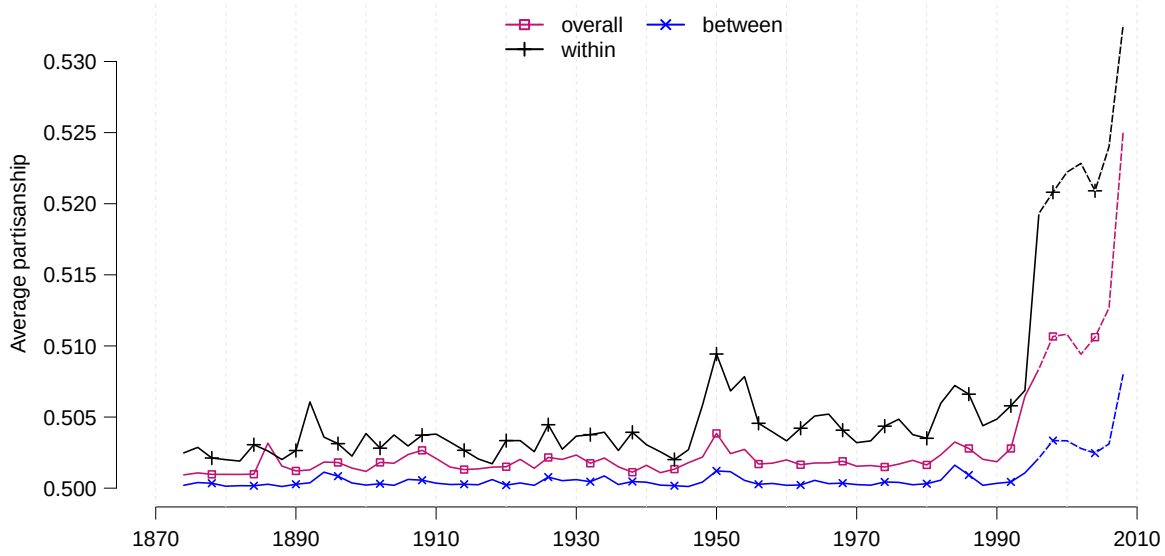
Notes: Calculations are based on our preferred estimates from panel A of figure 3. The solid lines denote the Qth quantile of the average estimated value of $\rho_{jt}(\mathbf{x}_{it})$ weighted by the midpoint of the average estimated values of $q^R_{jt}(\mathbf{x}_{it})$ and $q^D_{jt}(\mathbf{x}_{it})$. Averages are taken with respect to the set of speakers active in session t. We truncate values to lie on $[0,1]$ in cases where the LN/GPO data series reconciliation leads to values outside this range. We perform the same truncation to enforce that each quantile is bounded by the higher and lower quantiles.

Figure 7: Evidence on the role of neologisms



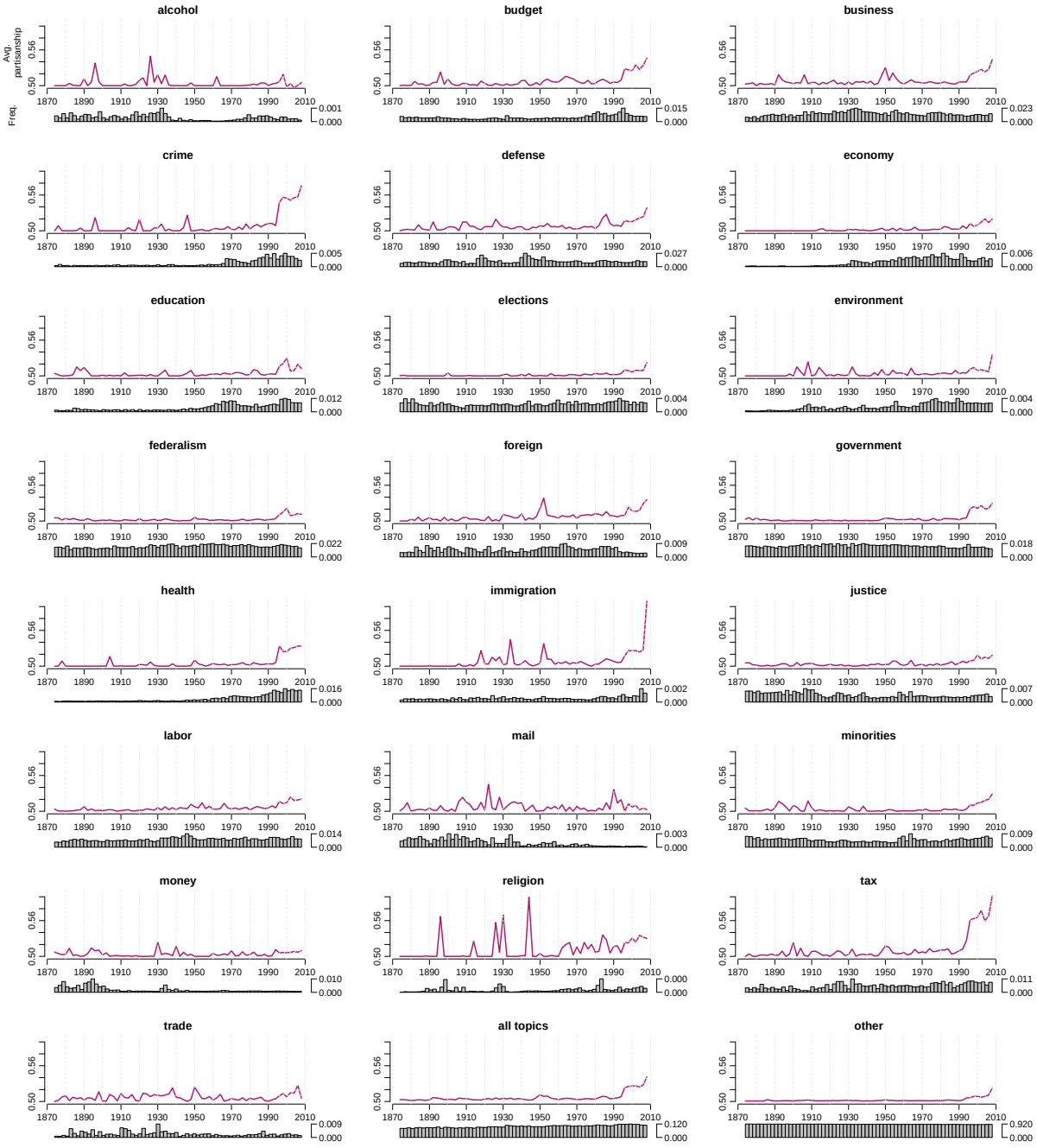
Notes: The "baseline" series is our preferred estimate of average partisanship from panel A of figure 3. The other two series are based on the same parameter estimates. The "pre-1980 vocabulary" series recomputes average partisanship exclusively on phrases spoken at least once during or prior to 1980, while the "post-1980 vocabulary" does so for phrases only spoken after 1980.

Figure 8: Partisanship within and between topics



Notes: "Overall" average partisanship is our preferred estimate from panel A of figure 3. The other two series are based on the same parameter estimates. Between-topic average partisanship is defined as the expected posterior that an observer with a neutral prior would assign to a speaker's true party after learning which of our manually-defined topics a speaker's chosen phrase belongs to. Average partisanship within a topic is defined as average partisanship if a speaker is required to use phrases in that topic. Within-topic average partisanship is then the mean of average partisanship across topics, weighting each topic by its total frequency of occurrence across all sessions. Phrases that are not part of one of our manually defined topics are excluded from these calculations.

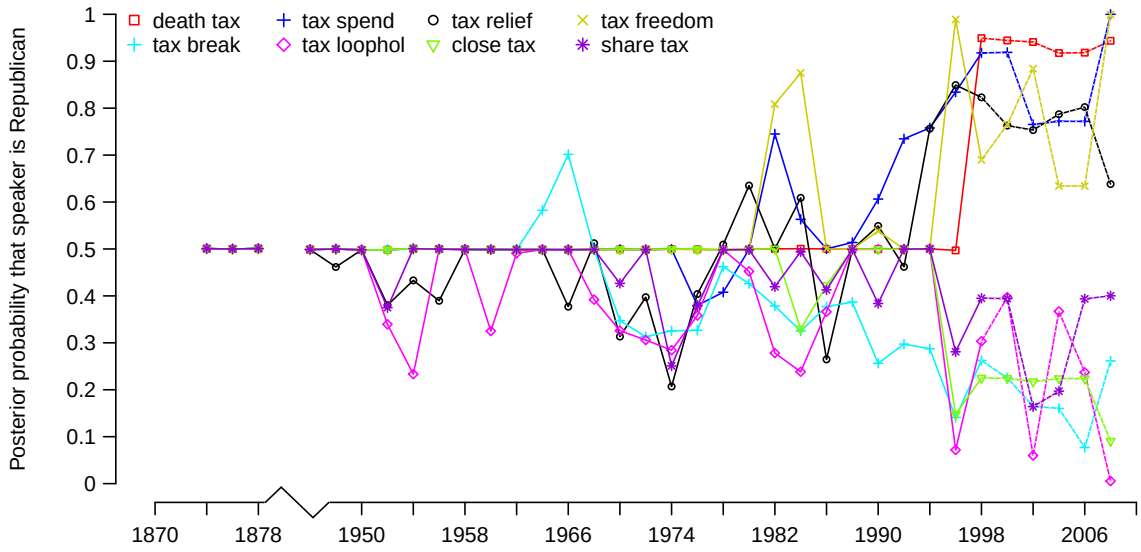
Figure 9: Partisanship by topic



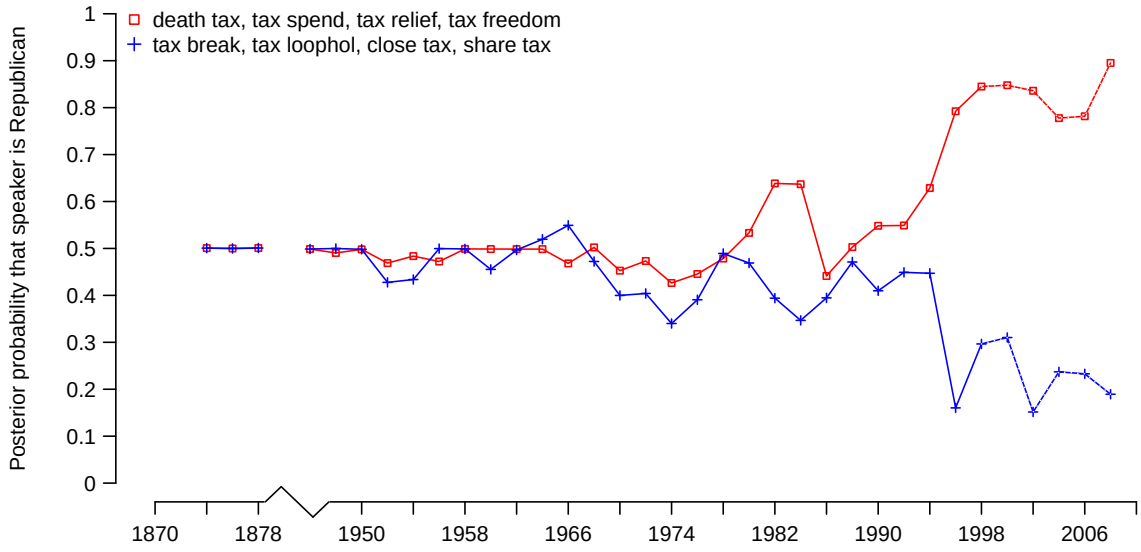
Notes: Calculations are based on our preferred estimates from panel A of figure 3. In each panel, the top (line) plot shows estimated average partisanship for a given topic, and the bottom (bar) plot shows the share of all phrase utterances that are accounted for by members of that topic in a given session. Average partisanship within a topic is defined as average partisanship if a speaker is required to use phrases in that topic. "All topics" includes all phrases classified into any of our substantive topics; "other" includes all phrases not classified into any of our substantive topics.

Figure 10: Partisanship over time for tax-related phrases

Panel A: By phrase

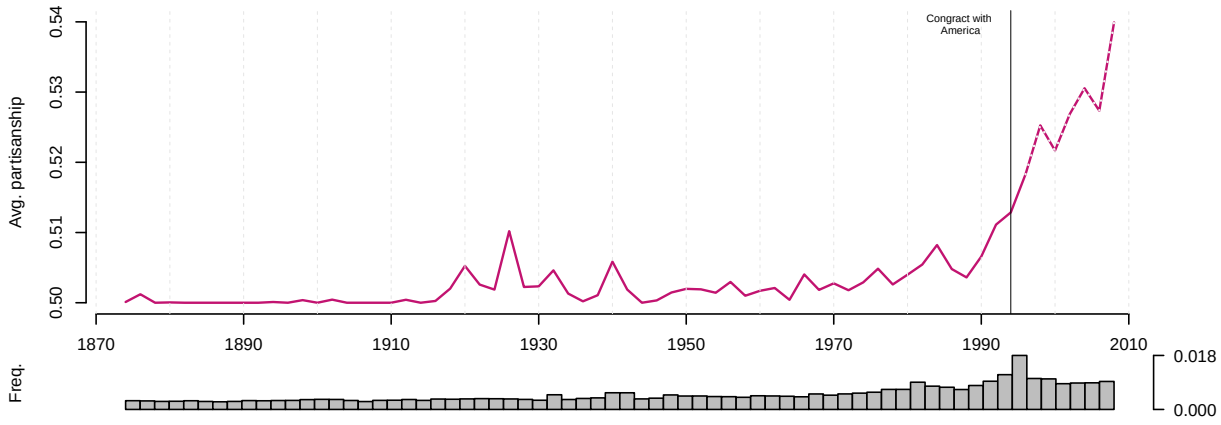


Panel B: Aggregated across phrases



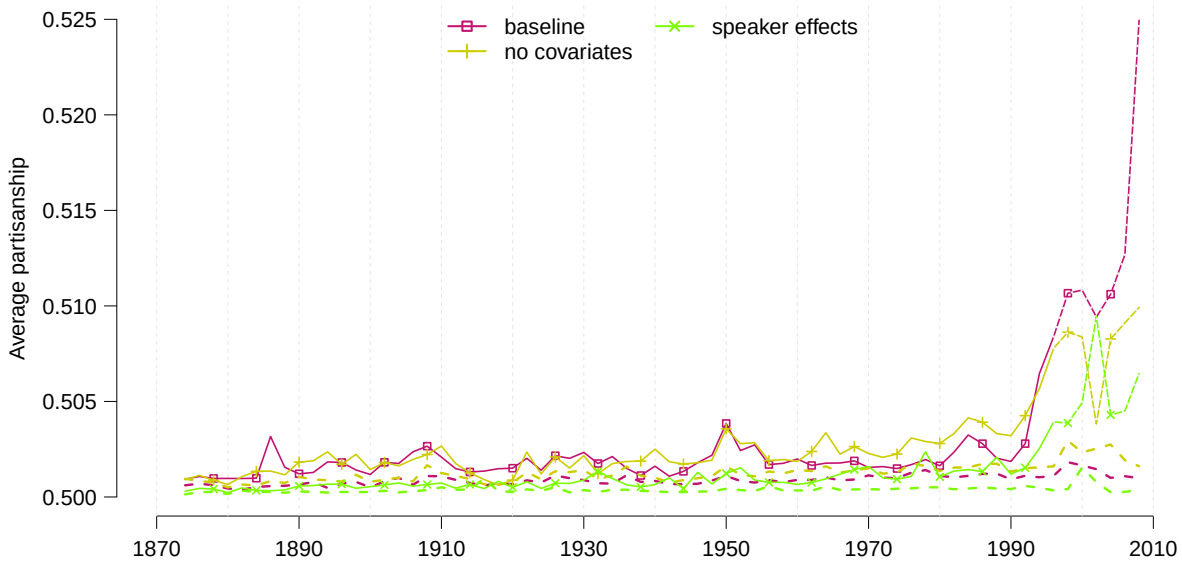
Notes: Calculations are based on our preferred estimates from panel A of figure 3. Panel A shows the mean estimated value of $\rho_{jt}(\boldsymbol{x})$ for selected phrases j related to 'tax.' When the LN and GPO data series reconciliation leads the estimated $\rho_{jt}(\boldsymbol{x})$ to be greater than 1 or less than 0, we truncate the estimated $\rho_{jt}(\boldsymbol{x})$ to 1 or 0, respectively. Panel B aggregates the values in panel A by taking the mean across groups of phrases, where the phrases are manually grouped based on the sign of the estimated $\rho_{jt}(\boldsymbol{x})$.

Figure 11: Partisanship and the *Contract with America*



Notes: Calculations are based on our preferred estimates from panel A of figure 3. The top (line) plot shows estimated average partisanship if a speaker is required to use phrases contained in the *Contract with America* (1994). The bottom (bar) plot shows the share of all phrase utterances that are accounted for by phrases in the *Contract* in a given session.

Appendix Figure: Partisanship of speech from model variants



Notes: For each specification, the solid line shows the average estimated partisanship on the real data and the corresponding dashed line shows the average estimated partisanship on hypothetical data in which each speaker's party is randomly assigned with the probability that the speaker is Republican equal to the average share of speakers who are Republican in the sessions in which the speaker is active. The "baseline" specification corresponds to the specification in panel A of figure 3. The "no covariates" specification imposes that $\tilde{\boldsymbol{\gamma}}_{jt} := 0$. The "speaker effects" specification includes in $\mathbf{x}_{it}$ a speaker random effect υ_{ijt} that is independent of covariates and is distributed as $Laplace(0,\kappa)$ so that its standard deviation is $\sqrt{2}/\kappa$. We set $\hat{\kappa} = \sqrt{2}/\text{sd}(\hat{e}_{ijt})$ where $\hat{e}_{ijt} = \log(c_{ijt}/\exp[\hat{\mu}_i + u_{ijt}])$ are the observed Poisson residuals from our baseline model when $c_{ijt} > 0$. We then estimate the random effects υ_{ijt} and the remaining parameters of the model by exploiting the fact that posterior maximization under the Laplace assumption is equivalent to L_1-penalized deviance minimization with cost κ/n, where n is the number of speaker-sessions (see, e.g., Taddy forthcoming).