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

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

June 2015

Abstract

Standard measures of segregation or polarization are inappropriate for high-dimensional data such as Internet browsing histories, item-level purchase data, or text. We develop a model-based measure of polarization that can be applied to such data. We illustrate the measure with an application to the partisanship of speech in the US Congress from 1872 to the present. We find that speech has become more polarized across party lines over time, with a clear trend break around 1980.

[*]E-mail: gentzkow@ChicagoBooth.edu, jesse_shapiro_1@Brown.edu, taddy@ChicagoBooth.edu. We acknowledge funding from the Initiative on Global Markets and the Stigler Center at Chicago Booth, and the National Science Foundation. We thank our many dedicated research assistants for their contributions to this project.

1 Introduction

By many accounts, Democrats and Republicans in America live in different worlds. They differ not only in political attitudes and preferences, but also in beliefs about basic facts (Shapiro 2014). They live in different neighborhoods (Bishop 2008), consume different products (Chapman 2012), and get information from different media sources (Gentzkow and Shapiro 2011; Pew Research Center 2014). They use different language, with Democrats talking about "estate taxes," "undocumented workers," and "tax breaks for the rich," while Republicans refer to "death taxes," "illegal aliens," and "tax reform" (Gentzkow and Shapiro 2010; Martin and Yurukoglu 2014). These differences may contribute to cross-party animus, and ultimately to gridlock and dysfunction in the political system.

In this paper, we consider the problem of measuring such polarization in opinions or behavior, and characterizing its evolution over time. That is, given vectors of choices $\mathbf{c}_{it}$ (e.g., phrases spoken, websites visited, or neighborhoods chosen) for a sample of individuals $i \in R \cup D$ at time t, how different are the choices of individuals in R from individuals in D at each t? In some settings, this problem is simple, or at least well-understood. If $\mathbf{c}_{it}$ is a scalar—an indicator for supporting gay marriage, or the number of hours spent watching Fox News—the difference in sample means, or other moments or quantiles, provides a natural summary. If $\mathbf{c}_{it}$ has a small number of dimensions—indicators for which of 100 neighborhoods an individual chose, or the answers to 20 survey questions—we have the classic problem of measuring segregation. We could apply measures such as the isolation index (White 1986; Cutler et al. 1999), the dissimilarity index (Duncan and Duncan 1955; Cutler et al. 1999), the Atkinson index (Atkinson 1970; James and Taeuber 1985; Frankel and Volij 2011), or the mutual information index (Theil 1971; Frankel and Volij 2011), which aggregate differences across dimensions.

The situation changes, however, when the number of dimensions in $\mathbf{c}_{it}$ is large relative to the size of the sample. To characterize differences in the language used by the 535 Republicans and Democrats in the US Congress, for example, the vector $\mathbf{c}_{it}$ might be counts of uses of individual words (on the order of ten thousand elements), or counts of phrases (hundreds of thousands or millions of elements). The number of dimensions would be similarly large if we were studying visits to website URLs, purchases of products in the supermarket, or residence in small geographic areas such as census blocks. In such cases, most options will be chosen by at most a small number of individuals. As a result, standard measures of difference or segregation can be severely biased relative to their population analogues.

To see the intuition for this bias, suppose that we wish to measure the extent to which US zipcodes are segregated by political party. Suppose the population shares of Republicans and

Democrats are both 50 percent, and that people of both parties are in fact uniformly distributed across zipcodes. We have available a sample of individuals that is small relative to the number of zipcodes; for simplicity, suppose we sample exactly two individuals in each. Then, although the true level of segregation is zero, we would find that about *half* of the zipcodes in our sample are perfectly segregated, either 100 percent Republican or 100 percent Democrat. Standard segregation measures would record high levels of segregation as a result (Cortese et al. 1976; Carrington and Troske 1997). The root of the problem is that measures of segregation and polarization are essentially measures of variance. The variance of estimated party shares across elements of $\mathbf{c}_{it}$ will be biased upward by sampling error, just as the variance of a vector of estimated fixed effects can be an upwardly biased estimate of the variance of the true effects.

Figure 1 shows that this bias is extreme in the case of polarization in congressional speech. Actual segregation as measured by various standard measures is high, but only slightly higher than it would be if party labels were randomly assigned, and measured segregation grows mechanically as the size of the corpus shrinks relative to the size of the vocabulary. We show below that the same bias arises in estimates based on text analysis measures such as the slant index of Gentzkow and Shapiro (2010), or the measure of polarization in congressional speech employed by Jensen et al. (2012).

In this paper we develop a new approach to measuring polarization in high-dimensional data. Our measure corrects for small sample bias, and also provides meaningful estimates of which choices in the vector $\mathbf{c}_{it}$ display the most polarization, and how the polarization of particular choices has evolved over time. We apply the measure to data on congressional speech from 1872 to 2009. Whereas prior estimates suggest that the polarization of congressional speech is not unusually high today, and was in fact higher in the late nineteenth and early twentieth century (Jensen et al. 2012), our results suggest that polarization of speech was relatively low until the 1980s, when it began a rapid rise, and that current levels are unprecedented over the period of our data.

Conceptually our measurement approach is simple. We specify a multinomial logit model of speech in which the utility to a given speaker i from using a phrase j is determined by a variety of measured and unmeasured factors, one of which is the speaker's party affiliation. In the context of the model, it is natural to define a phrase's *partisanship* as the effect of party affiliation on the mean utility of using the phrase. We define a speaker's partisanship as the frequency-weighted mean partisanship of the phrases used by the speaker.

Our measure of speaker partisanship has both an economic and a statistical interpretation. Economically, it reflects the expected difference in utility between Republicans and Democrats for using the phrases chosen by the speaker. A speaker who uses phrases preferred by Repub-

licans will have a very high partisanship. A speaker who uses phrases preferred by Democrats will have a very low partisanship. A speaker who mainly uses phrases that are equally preferred by members of both parties will have a middling partisanship.

Statistically, under our model our measure of partisanship is a sufficient reduction of the data in the sense that, conditional on partisanship (and other model observables), phrase frequency is statistically independent of party (Taddy 2013). Under the model, partisanship captures all of the information in the data about a given speaker's tendency to "talk like" a Republican or Democrat.

To summarize trends in segregation, we define the partisanship of a given session as the difference in the mean partisanship of Republicans and the mean partisanship of Democrats. We find that partisanship has risen steadily, with a take-off around 1980. The behavior of our measure post-1980 is similar to that derived from roll-call-voting measures (e.g., Poole and Rosenthal 1985), but our measure does not exhibit an increase in partisanship in the late 19th and early 20th centuries as roll-call-voting measures do.

Using a model-based measure of segregation has several key advantages over the more descriptive approach commonly taken to segregation problems. First, having a statistical model of the sampling process allows us to distinguish naturally between sample estimates and population values. Second, having an economic model of speech allows us to account for both unobservable and observable shifters of speaking behavior that are not related to party.

Our model-based approach also has disadvantages. Our model is highly parametric and our conclusions need to be interpreted within the context of the model. Although we introduce flexibility where possible, with high-dimensional data there are some limits to what we can allow for. Using a model also increases the computational burden significantly relative to many common descriptive indices of segregation, which are typically closed-form functions of count vectors. We show how to use a Poisson approximation to the likelihood of our model to permit scalable estimation following Taddy (forthcoming). This scalability makes it possible to estimate our measure on even larger datasets than the ones we employ here.

This paper contributes to a large literature on the measurement of segregation, surveyed recently in Reardon and Firebaugh (2002). Recent approaches in economics have derived axiomatic foundations for segregation measures (Echenique and Fryer 2007; Frankel and Volij 2011), asking which measures of segregation satisfy certain intuitive properties.[1] 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. To our knowledge, ours is the first

[1] 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).

paper to propose a comparative measure of segregation that is based on preferences recovered from a structural model,[2] and the first to measure trends in segregation in high-dimensional data.

Our measure of partisanship relates to a growing literature, mostly in economics and political science, on measuring the partisanship of a document (e.g., Laver et al. 2003). Our measure of partisanship is conceptually similar to Gentzkow and Shapiro's (2010) slant measure, but unlike the slant measure our measure is derived from a nonlinear count model, which is more appropriate for textual data (Taddy 2013). We show below that naive measures based on the slant index yield different (and less robust) conclusions regarding trends in segregation over time, although these measures can be improved with an intuitive finite-sample adjustment. 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 to work that classifies documents according to similarity of text (Blei 2004; Grimmer 2010).

2 The congressional speech data

Our primary data source consists of 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*.[3]

We obtained the text for the 43rd to 104th Congresses from a set of XML files produced by Lexis-Nexis (LN) by performing Optical Character Recognition (OCR) on scanned print volumes. The XML tags identify the session and date at all points in the record, and permit us to link XML text back to its original location in the print volumes.

For the 104th to 110th Congresses, we obtained the text from the website of the U.S. Government Publishing Office (GPO). Throughout the paper, we estimate partisanship in the 104th Congress 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.

We use an automated script to parse the text into individual speeches. The script identifies when a new speech begins by looking for a declaration of the speaker's identity. To illustrate,

[2] 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.

[3] Prior Congresses were covered by a predecessor publication called the *Congressional Globe*, which evolved over time from precis to verbatim format (Library of Congress 2015).

consider this typical selection from the proceedings of the House in the 77th Congress, on January 21, 1941:

> "Mr. ALLEN of Illinois. Are those reports available?
>
> Mr. ROBERTSON of Virginia. They are available. We mail them to every Member of Congress. Of course, the Members get so much literature they do not always read what they get. However, we always have additional copies that we can furnish whenever any Member wants them. We have requests from all the principal universities of this Nation. Our hearings are used as a textbook in the schools that teach game management and also used in their biology classes.
>
> Mr. ALLEN of Illinois. What was the expenditure of this committee last session?"

Here, the script recognizes three speeches: two by Leo Ellwood Allen (R, IL) and one by A. Willis Robertson (D, VA). The expression "Mr. SO-AND-SO of STATE" tells the script a new speech is beginning, and delivers the name of the speaker.

We match the speaker of each line of text to a congressperson from the universe of members of Congress defined in the Database of Congressional Historical Statistics (Swift et al. 2009) and the Voteview Roll Call Data (www.voteview.com). We require a perfect match of speaker name, chamber, state, and gender. When such a match does not exist or is not unique (e.g., there are two MR. ALLENs and the state is not declared) we exclude that speech from our analysis. We also exclude speeches made by speakers identified by office rather than name, such as the Speaker of the House or the President of the Senate. In the average Congress of the LN series, we match 86 percent of speeches to a congressperson; in all but one Congress the match rate exceeds 75 percent. In the average Congress of the GPO series, we match 70 percent of speeches to a congressperson. For each congressperson in each session, we obtain data on political party and we construct an indicator for whether the congressperson was in the majority party for her chamber in a given session.

We pre-process the text by removing stopwords and word stems using the stopword list and stemming algorithms defined in the Snowball package (http://snowball.tartarus.org/). The advantage of these pre-processing steps is that they help to aggregate text with similar meaning. To illustrate, "war on terror," "war on terrorism," and "war against terror" would all be resolved to "war terror." The corresponding disadvantage is that text with different meaning may be aggregated, as in "the war on terror," "war and terror," and "not war but terror."

We count the frequency of every bigram (two-word phrase) in every speech. For the purposes of our analysis a speech is represented by a vector of bigram counts; this is sometimes

called the "bag of words" approach as it treats words as an unordered jumble. More sophisticated approaches exist (Mikolov et al. 2013), but most text mining uses some version of the "bag of words" approach because it typically makes it possible to incorporate more data into the analysis (see discussion in Mikolov et al. 2013). For our purposes, an added advantage of modeling speech as a vector of counts is that doing so makes the speech partisanship problem more similar to canonical segregation measurement problems, and therefore makes our model more portable outside of our setting.

We consider the universe of all speeches made by Republicans or Democrats. We then filter out all procedural phrases identified based on Robert's Rules of Order, a parliamentary manual outlining the procedures of assemblies, and Riddick's Senate Procedure, a book containing contemporary precedents and practices of the US Senate. We parse all bigrams in the two manuals as procedural phrases. We identify additional procedural phrases as phrases that appear in many highly procedural speeches, or phrases that occur often and speeches that contain them are highly procedural on average. The exact procedure is described in appendix A.

We eliminate phrases that include a congressperson's name or a state's name. We exclude phrases that only contain numbers and symbols, phrases that identify a legislature number or a zipcode, and a few other categories of phrases with low semantic meaning. Finally, we trim the data by eliminating phrases that appear in every Congress fewer than 10 times or that are used by fewer than 75 speakers. We confirm that these are typically procedural phrases and phrases that do not carry meaningful content.[4]

Our final sample has 723,198 unique phrases spoken a total of 271 million times by 7,285 unique speakers. We will model phrase frequencies at the level of the speaker-session.[5] The 7,285 speakers and 68 sessions combine for 33,486 speaker-sessions.

In the online appendix, we show for each Congress the number of unique speeches, the number of unique speakers, the number of unique phrases, the total phrase counts for Democratic speakers, the total phrase counts for Republican speakers, and the match rate of speeches.

3 Preliminary estimates of partisanship

In this section we present some preliminary estimates of partisanship. These provide a feel for the data and a sense of the challenges involved in measuring trends in segregation convincingly.

[4]There are also phrases that are made of two independent words that commonly take turn after each other. For example, "scienc literatur" appears in the 100th Congress in the following sentence: "Greek Americans have contributed a considerable amount to our culture, with their architecture, art, science, and literature."

[5]In the rare case where a speaker switched chambers in a single session (usually from the house to senate), text from each chamber is treated as a distinct speaker-session.

Throughout we will use the following notation. Let $\mathbf{c}_{it}$ be the vector of phrase counts by speaker i in session t, with $m_{it} = \sum_j c_{itj}$ denoting the total amount of speech by speaker i in session t and $m_{jt} = \sum_i c_{itj}$ denoting the total number of uses of phrase j in session t. Let r_{it} be an indicator for whether speaker i in session t is Republican.

In figure 2, we present trends in partisanship of speech as implied by four measures of segregation: the isolation index (White 1986; Cutler et al. 1999), the dissimilarity index (Duncan and Duncan 1955; Cutler et al. 1999), the Atkinson index (Atkinson 1970; James and Taeuber 1985; Frankel and Volij 2011), and the mutual information index (Theil 1971; Frankel and Volij 2011). Each plot shows the given measure for both the actual data and for hypothetical data in which we randomly assign political party to each speaker. All four measures imply that partisanship of speech has been steadily declining. However, all four measures imply nearly identical dynamics when we randomly assign parties.

In figure 3, we present trends as implied by two measures that were developed specifically to measure the partisanship of speech. The first is based on the slant index of Gentzkow and Shapiro (2010) and shows the difference in the average slant of Republicans and the average slant of Democrats over time. The second is based on Jensen et al. (2012) and shows the correlation of the (count-weighted) average phrase with political party. As in figure 2, both measures imply a decline in partisanship over time, but in each case the dynamics are similar between the real data and the data in which party has been reassigned at random.

Some authors have suggested finite-sample corrections to segregation indices so that they imply zero segregation when groups are assigned at random. Figure 4 presents three such measures: the adjusted dissimilarity index of Carrington and Troske (1997), the leave-out isolation index of Gentzkow and Shapiro (2011), and a leave-out analogue of the slant index (Gentzkow and Shapiro 2010). By construction these measures do not inherit the dynamics of the "random" series. They imply different dynamics of partisanship from one another and from the measures in figures 2 and 3. The adjusted dissimilarity index exhibits a post-1980 rise in partisanship, but also shows large fluctuations in the early- to mid-1900s. The leave-out isolation index shows a rise and decline of partisanship in the early 1900s, followed by a rise post-1980. The leave-out slant index exhibits a post-1980 rise and no large swings or trends before that.

4 Model

We assume that, for those sessions t which have speaker i as a member, $\mathbf{c}_{it} \sim \mathrm{MN}(\mathbf{q}_{it}, m_{it})$ with

$$q_{itj} = e^{\eta_{itj}} / \sum_l e^{\eta_{itl}} \tag{1}$$
$$\eta_{itj} = \alpha_{jt} + \mathbf{u}_{it}' \boldsymbol{\gamma}_{jt} + \varphi_{jt} r_{it}.$$

This can be interpreted as a multinomial logit choice model in which η_{itj} is the mean utility of phrase j for speaker i in session t. Here α_{jt} are phrase-time-specific intercepts and $\mathbf{u}_{it}$ are attributes of speaker i in session t, excluding the Republican party membership indicator r_{it}. In our baseline specification, $\mathbf{u}_{it}$ consists of indicators for state, chamber, gender, Census region, and whether the party is in the majority. The coefficients $\boldsymbol{\gamma}_{jt}$ on control attributes $\mathbf{u}_{it}$ are static in time (i.e., $\gamma_{jtk} := \gamma_{jk}$) except for those on Census region, which are allowed to vary across sessions. We also explore specifications in which $\mathbf{u}_{it}$ includes unobserved speaker-level preference shocks.

Let the *partisanship* of speaker i in session t be

$$z_{it} = \boldsymbol{\varphi}_t' \mathbf{c}_{it} / m_{it}. \tag{2}$$

Partisanship is the expected utility gain to a Republican relative to a Democrat from speaking exactly like speaker i in session t. Partisanship is also a sufficient reduction (Taddy 2013) in the sense that, under the model in (1),

$$r_{it} \perp\!\!\!\perp \mathbf{c}_{it} \mid z_{it}, \mathbf{u}_{it}, m_{it}. \tag{3}$$

That is, conditional on covariates, a speaker's party r_{it} contains no statistical information about her speech beyond that which is contained in partisanship z_{it}.

5 Penalized maximum likelihood estimation via Poisson approximation

We estimate the model via penalized maximum likelihood, using a Poisson approximation to the multinomial model to allow distributed computing.

5.1 Distributed estimation via Poisson approximation of the likelihood

Given the number of distinct phrases and attributes that we consider, estimation of the multinomial logistic regression defined by (1) is computationally intractable due to the cost of repeatedly computing the denominator $\sum_l e^{\eta_{itl}}$. We therefore approximate the multinomial logit model with a Poisson model that is amenable to distributed computing.

Suppose that $c_{itj} \sim \text{Pois}\left(\exp\left[\mu_{it} + \eta_{itj}\right]\right)$, where μ_{it} is a nuisance parameter that determines speaker i's overall verbosity in session t (regardless of phrase) and η_{itj} is the speaker mean utility from (1). It follows that $m_{it} \sim \text{Pois}\left(\sum_j \exp\left[\mu_{it} + \eta_{itj}\right]\right)$ and therefore that

$$\Pr\left(\mathbf{c}_{it} \mid m_{it}\right) = \frac{\prod_{\text{j}} \text{Po}\left(\text{c}_{\text{itj}}; \exp\left[\mu_{\text{it}} + \eta_{\text{itj}}\right]\right)}{\text{Po}\left(\text{m}_{\text{it}}; \sum_{\text{l}} \exp\left[\mu_{\text{it}} + \eta_{\text{itl}}\right]\right)} = \text{MN}(\mathbf{c}_{it};\ \mathbf{q}_{it}, m_{it}), \tag{4}$$

where $\text{Po}\left(\right)$ is the Poisson likelihood and we recall that $\mathbf{q}_{it}$ is the vector of phrase probabilities defined by (1). The last equality of (4), which follows from some algebraic manipulation, states that, conditional on m_{it}, the Poisson and multinomial models imply the same likelihood for phrase counts $\mathbf{c}_{it}$.

Next observe that:

$$\Pr\left(\mathbf{c}_j \mid m_{it}, \mu_{it}\right) = \prod_{i,t} \text{Po}\left(c_{itj};\ \exp[\mu_{it} + \eta_{itj}]\right). \tag{5}$$

That is, given the value of the nuisance parameter, the likelihood of the counts $\mathbf{c}_j$ for phrase j does not depend on the counts for other phrases.

We estimate our model by penalized maximization of (5), replacing the true μ_{it} with the plug-in estimate $\hat{\mu}_{it} = \log m_{it}$. The disadvantage of this approach is that it requires an estimate of the nuisance parameter μ_{it}. The advantage is that it allows us to treat the likelihood for each phrase separately, which means that we can distribute estimation across many machines.

If $\hat{\mu}_{it}$ is the maximum likelihood estimate, then our approach is not an approximation, it is exact. Taddy (forthcoming) shows some special cases in which this equivalence holds. More generally, we should expect our approximation to be best when $\hat{\mu}_{it}$ is close to the maximum likelihood estimate. Taddy (forthcoming) provides some empirical examples in which the approximation works well. Appendix figure 1 shows a parametric bootstrap in which our model performs well on data simulated from the multinomial model. In the online appendix we show that in small-scale experiments (where it is practical to estimate the multinomial logit) the Poisson approximation recovers good estimates of φ_{jt}.

5.2 Penalization

Because of the large number of phrases, our model is susceptible to over-fit: without any further structure, a large fraction of the relationships between phrase use and party will be spurious. This is the problem that leads to the finite-sample biases of the estimates in section 3.

We address this issue by adopting penalties for key parameters. To write the penalized likelihood, begin by decomposing $\boldsymbol{\varphi}_t$ as

$$\varphi_{jt} = \bar{\varphi}_j + \sum_k \tilde{\varphi}_{jk} \mathbf{1}_{t>k}. \tag{6}$$

We will estimate by minimizing a penalized negative log likelihood:[6]

$$\sum_j \left\{ \sum_t \sum_i \left[m_{it} \exp(\alpha_{jt} + \mathbf{u}_{it}' \boldsymbol{\gamma}_{jt} + \varphi_{jt} r_{it}) - c_{itj}(\alpha_{jt} + \mathbf{u}_{it}' \boldsymbol{\gamma}_{jt} + \varphi_{jt} r_{it}) \right] + \lambda_j \left[|\bar{\varphi}_j| + \sum_t |\tilde{\varphi}_{jt}| \right] \right\}. \tag{7}$$

Ignoring dynamics across sessions, the objective in (7) amounts to adding an L_1 or lasso-type penalty to the party phrase loadings $\bar{\varphi}_j$ (Tibshirani 1996). This imposes sparsity on the loadings; some will be exactly zero. By penalizing the time dynamics implied by $\tilde{\varphi}_{jk}$, we also restrict the partisan "meaning" of a given phrase to evolve slowly over time, as in Blei and Lafferty (2006). As we impose no penalty on the phrase-specific intercepts α_{jt} or on the static covariate coefficients $\boldsymbol{\gamma}_j$, we allow phrases to grow more or less popular over time with no restriction, and we flexibly fit the impact of geography, chamber of Congress, etc.

We determine the penalties λ by regularization path estimation as follows. For each phrase j we find λ_j^1 large enough so that $\bar{\varphi}_j$ and the $\tilde{\varphi}_{jk}$ are all estimated at zero, then incrementally decrease $\lambda_j^2, ..., \lambda_j^G$, and update parameter estimates to minimize the penalized deviance at each new weight specification. Given the path of estimates for each phrase regression, the single optimal specification can be chosen to minimize the Bayesian information criterion (BIC), $\sum_{i,t} \log \mathrm{Po}\left(c_{itj};\ \exp[\hat{\mu}_i + \eta_{itj}]\right) + df \log n$.[7] A useful computational property of this approach

[6] For reasons detailed in Haberman (1973) and summarized in Silva and Tenreyro (2010), the existence of posterior maximizing $\hat{\boldsymbol{\gamma}}_{jt}$ without penalization on these parameters (i.e., under an improper prior) is not straightforward to establish in each Poisson regression. A sufficient condition for existence is that the controls design (i.e., the part of the regression involving $\mathbf{u}_{it}$) is full rank on the subset of observations where $c_{itj} > 0$; 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 $c_{itj} > 0$. To avoid remaining nonexistence-related issues, we then add a very small (10^{-6}) L_1 penalty on the $\boldsymbol{\gamma}_{jt}$ to ensure posterior propriety and numerical convergence.

[7] The degrees of freedom here, df, are available following Zou et al. (2007) as the number of parameters estimated with nonzero values (excluding the $\hat{\mu}_{it}$, as outlined in Taddy forthcoming). We actually apply versions of the criterion which are *corrected* for high dimensions, multiplying each κ by $n/(n - df - 1)$ to account for estimation

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.

6 Application

Examples of the regularized estimates are shown in figure 5. The algorithm proceeded from right to left in these plots, moving from simple to complex representations.

Figure 6 shows the evolution of party loadings over time for select groups of phrases. The figure shows, for example, that phrases like "estate tax" and "tax break" did not become partisan until late in the twentieth century.

Figure 7 shows the difference in the average value of speaker partisanship z_{it} between parties over time. This difference can be taken as a measure of the segregation or polarization of speech by party. For comparison, we show the measure for hypothetical data in which we randomly reassign speakers to parties and re-estimate the model. Unlike many of the measures in section 3, our measure has very different dynamics when estimated on real data and on data in which party has been randomly reassigned. Our measure implies a rapid increase in the partisanship of speech that began around 1980, with a slight secular increase before.

To give an idea of magnitude, the partisanship in 1980 of 0.007 corresponds to a two-speaker, two-phrase example in which the Republican speaker uses one phrase 53 percent of the time and the Democratic speaker uses the other phrase 53 percent of the time. By comparison, the partisanship in 2008 of 0.031 corresponds to a preferred phrase use of 56 percent. For another comparison, the change in partisanship from 1980 to 2008 is equivalent to 4.75 standard deviations of partisanship in 1980.

Figure 8 shows nonparametric bootstraps on the results from figure 7.

The behavior of our measure depends critically on penalization. Figure 9 shows the same measure under the no-penalty specification (the far left of the paths depicted in figure 5). Without any penalty, our measure partially replicates the secular decline in partisanship seen in many measures in section 3, and inherits the dynamics of the random data.

Figure 10 explores the robustness of our results under different specifications. We consider four variants: (i) assuming that vocabulary has consistent meaning over time, in the sense that $\varphi_{jt} := \varphi_j$, (ii) dropping covariates $\mathbf{u}_{it}$ from the model by imposing that $\boldsymbol{\gamma}_{jt} := 0$, (iii) keeping the amount of "data information" constant at one million phrase utterances per session, by letting $\tilde{c}_{itj} = c_{itj}10^6/m_t$, where $m_t = \sum_{i,j} c_{itj}$, and (iv) allowing Laplace-distributed speaker random effects with shape parameter calibrated to match the degree of overdispersion in the

overfit. See Flynn et al. (2013) and Taddy (2015).

baseline model.[8] All of these specifications show dynamics similar to those of our baseline model in figure 7.

7 Conclusions

Measurement of polarization or segregation is a core topic in quantitative social science. Traditional measures behave poorly in high-dimensional applications. We present a model-based alternative that has good finite-sample properties and intuitive economic and statistical interpretations. We illustrate the method with an application to the partisanship of Congressional speech.

References

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

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

Atkinson, Anthony B. 1970. On the measurement of inequality. *Journal of Economic Theory* 2(3): 244–263.

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

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. University of Toronto mimeo. Accessed at <http://homes.chass.utoronto.ca/∼mcmillan/bmr2.pdf> on June 10, 2015.

Bishop, Bill. 2008. *The Big Sort: Why the Clustering of Like-Minded America is Tearing Us Apart*. New York: Houghton Mifflin.

Blei, David Meir. 2004. Probabilistic models of text and images. Princeton mimeo.

Blei, David Meir and John D. Lafferty. 2006. Dynamic topic models. *Proceedings of the 23rd International Conference on Machine learning*: 113–120.

[8] Formally, we allow that the $\mathbf{u}_{it}$ includes a speaker random effect υ_{itj} that is independent of covariates and is distributed as $Laplace\,(0,\tau)$ so that its standard deviation is $\sqrt{2}/\tau$. We set $\hat{\tau} = \sqrt{2}/\mathrm{sd}(\hat{e}_{itj})$ where $\hat{e}_{itj} = \log(c_{itj}/\exp[\hat{\mu}_i + \eta_{itj}])$ are the observed Poisson residuals from our baseline model when $c_{itj} > 0$. We then estimate the random effects υ_{itj} 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 τ/n (see, e.g., Taddy 2015).

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.

Chapman, Mary M. 2012. Party affiliations in car-buying choices: A thorny patch of consumer analysis. *New York Times*, March 30, 2012. Accessed at <http://wheels.blogs.nytimes.com/2012/03/30/party-affiliations-in-car-buying-choices-a-thorny-patch-of-consumer-analysis/> on June 10, 2015.

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.

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

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.

Duncan, Otis Dudley and Beverly Duncan. 1955. A methodological analysis of segregation indexes. *American Sociological Review* 20(2): 210–217.

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

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.

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.

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. *The Annals of Statistics* 1(4): 617–632.

James, David R. and Karl E. Taeuber. 1985. Measures of segregation. *Sociological Methodology* 15: 1–32.

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* 2: 1–81.

Laver, Michael, Kenneth Benoit, and John Garry. 2003. Extracting policy positions from political texts using words as data. *The 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.

Martin, Gregory J. and Ali Yurukoglu. 2014. Bias in cable news: Real effects and polarization. NBER Working Paper No. 20798.

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

Mele, Angelo. 2015. A structural model of segregation in social networks. Johns Hopkins University mimeo.

Mikolov, Tomas, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. arXiv:1301.3781v3.

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

Pew Research Center. 2014. Political polarization & media habits. Accessed at <http://www.journalism.org/2014/10/21/political-polarization-media-habits/> on June 5, 2015.

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

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 September 23, 2010.

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

Shapiro, Jesse. 2014. Special interests and the media: Theory and an application to climate change. Chicago Booth mimeo.

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> on June 29, 2009.

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

Taddy, Matt. 2015. One-step estimator paths for concave regularization. arXiv:1308.5623v6.

Taddy, Matt. Forthcoming. Distributed multinomial regression. *The Annals of Applied Statistics.*

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

Theil, Henri. 1971. *Principles of Econometrics*. New York: Wiley and Sons.

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

Voteview Roll Call Data. Accessed at <http://voteview.com/> on October 18, 2010.

White, Michael J. 1986. Segregation and diversity measures in population distribution. *Population Index* 52(2): 198–221.

Zou, Hui, Trevor Hastie, and Robert Tibshirani. 2007. On the "degrees of freedom" of the lasso. *The 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.[9] We also obtain an electronic copy of the appendix from "Riddick's Senate Procedure" for the 101st Congress (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.[10] 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% 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 that a phrase occurs in.[11] We have two separate rules to identify occurrence 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 Congresses and one of: 1) it appears in more than 5,200 highly Robert speeches, and at least 1.75% of speeches it appears in are highly Robert; or 2) it appears in more than 100 highly Robert speeches, and at least 7.5% of speeches it appears in are highly Robert; or 3) it appears in more than 50 highly Robert speeches, and more than 30% of speeches it appears in are highly Robert.

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

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

[10] The PDF version is downloaded from http://www.gpoaccess.gov/riddick/1441-1608.pdf and converted into text using OCR with metadata cleaned out.

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

- It appears in at least 5 highly Riddick speeches in more than 10 Congresses and one of: 1) it appears in at least 3,000 Riddick speeches, and at least 1.75% of speeches it appears in are highly Riddick; or 2) it appears in at least 100 Riddick speeches, and at least 7% of speeches it appears in are highly Riddick; or 3) it appears in at least 50 highly Riddick speeches, and at least 20% of speeches it appears in are highly Riddick.

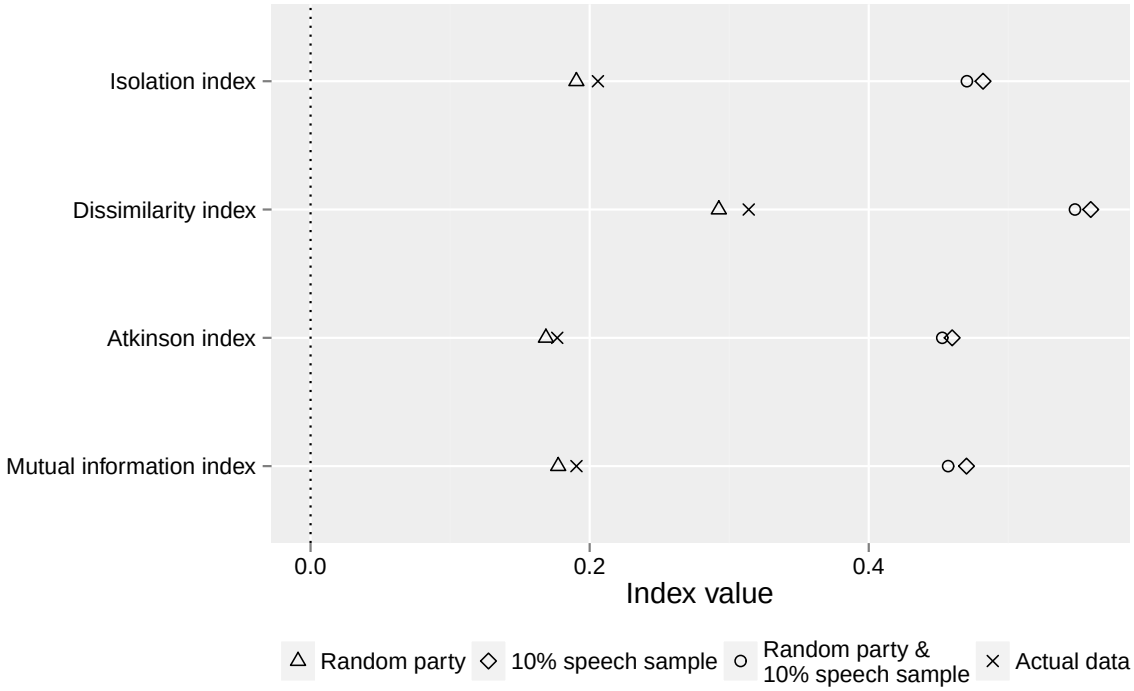
We compute, for every phrase, the average percentage of Robert's procedural phrases/Riddick's procedural phrases in all speeches that the phrase appears in. Of the phrases that are not identified by our first rule, a phrase qualifies as procedural using our second rule if one of the following sets of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

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

We choose the cut-off points such that phrases that just make the cut-offs are subjectively procedural,[12] whereas phrases that do not make the cut-offs are subjectively not procedural.[13]

[12]Examples: phrases with bill numbers, committee names, etc.
[13]Examples: veteran associ, war time, victim hurrican

Figure 1: Segregation measures for the 109th congress

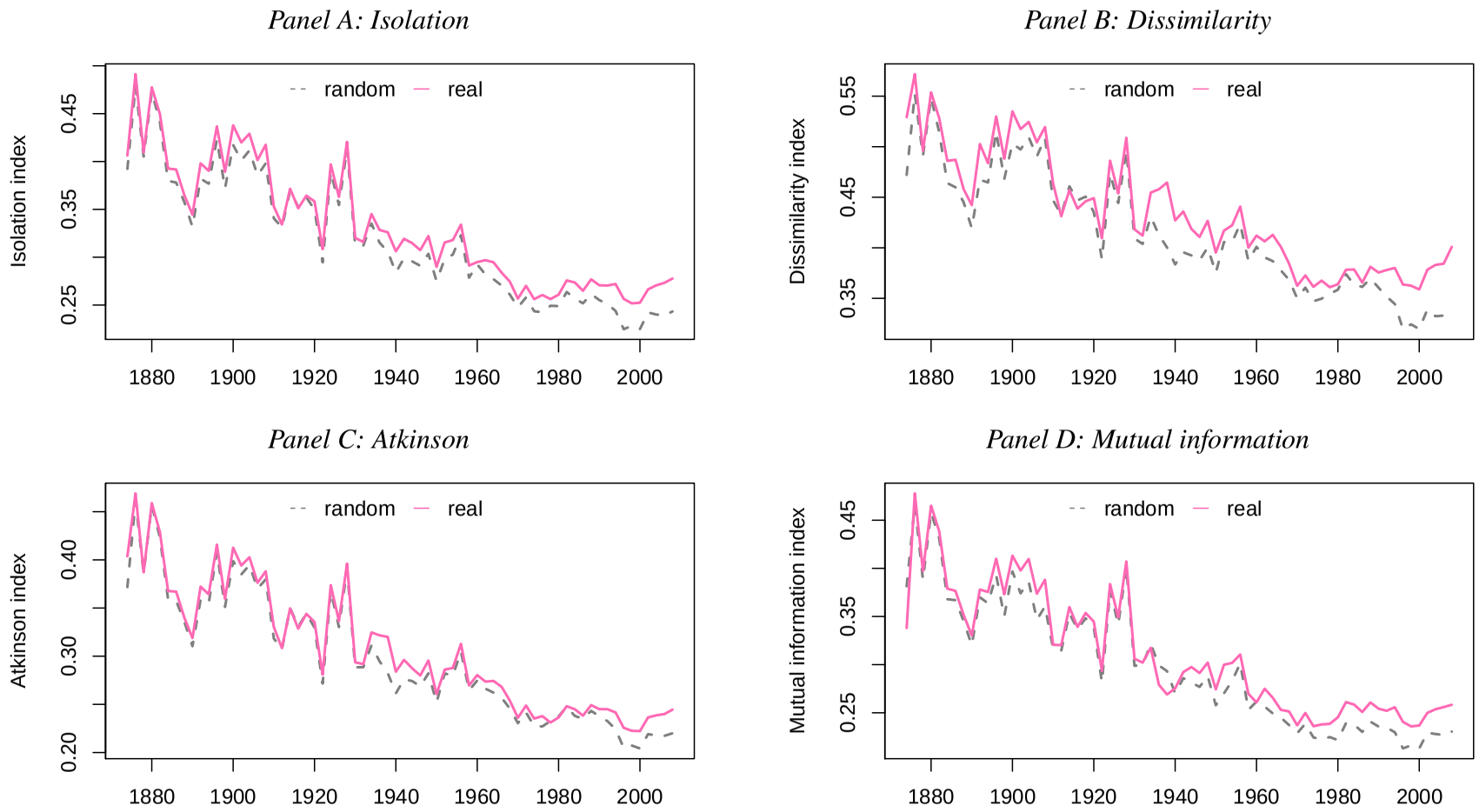


Notes: Plot shows segregation measures (isolation index, dissimilarity index, Atkinson index, and mutual information index) for the 109th congress, with party affiliations (Republican, Democrat) as "groups" and phrases as "neighborhoods." "Random party" reports mean indices across 1,000 simulations where speakers' party labels were randomly assigned Republican with probability equal to the fraction of Republicans in the 109th congress. "10% speech sample" reports mean indices across 1,000 simulations where random 10% subsets of utterances were used as samples. "Random party & 10% speech sample" reports mean indices using combined samples from the 1,000 simulations of "random party" and "10% speech sample."

Let $Rep_{jt} = c_{jt,r=Republican}$ be the number of utterances of phrase j by Republicans, $Rep_t = \sum_j Rep_{jt}$ be the total number of Republican utterances, and define Dem_{jt} and Dem_t similarly. Let $q_{jt} = \frac{Rep_{jt}}{Rep_{jt}+Dem_{jt}}$ be the Republican share of utterances of phrase j and $q_t = \frac{Rep_t}{Rep_t+Dem_t}$ be the Republican share of total utterances. Finally, define the *entropy* of a Bernoulli process with probability of success q as $e(q) = -q\log_2(q) - (1-q)\log_2(1-q)$. The indices are then computed as follows:

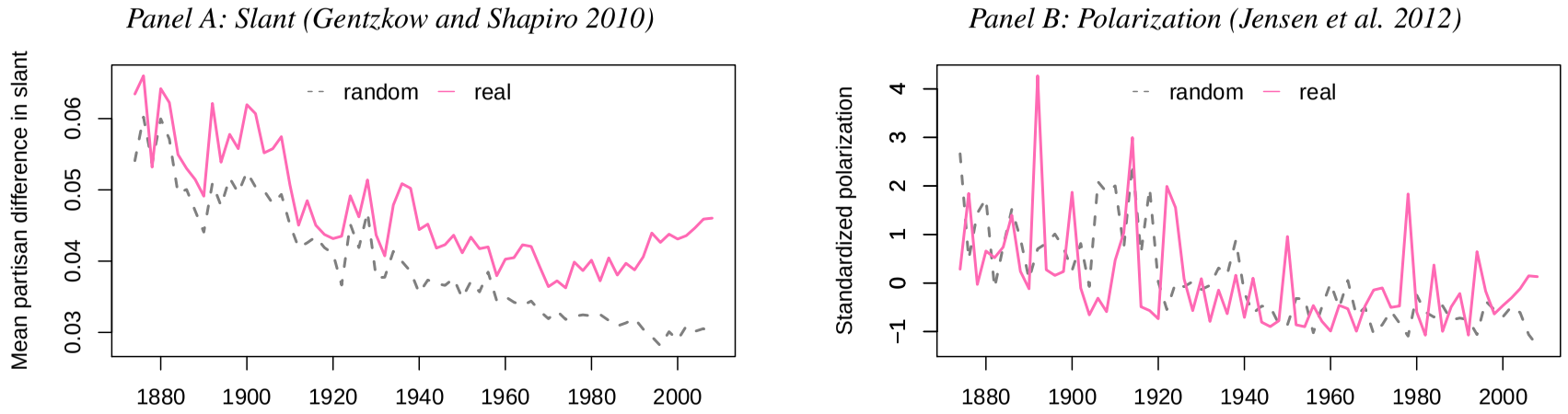
$$
\begin{aligned}
s_t^{Atkinson} &= 1 - \sum_j \left(\frac{Rep_{jt}}{Rep_t}\right)^{\frac{1}{2}} \left(\frac{Dem_{jt}}{Dem_t}\right)^{\frac{1}{2}} \\
s_t^{dissimilarity} &= \frac{1}{2}\sum_j \left|\frac{Rep_{jt}}{Rep_t} - \frac{Dem_{jt}}{Dem_t}\right| \\
s_t^{isolation} &= \sum_j \left(\frac{Rep_{jt}}{Rep_t} q_{jt}\right) - \sum_j \left(\frac{Dem_{jt}}{Dem_t} q_{jt}\right) \\
s_t^{mutual\ information} &= e(q_t) - \sum_j \frac{Rep_{jt}+Dem_{jt}}{Rep_t+Dem_t} e(q_{jt}).
\end{aligned}
$$

Figure 2: Partisanship of speech over time implied by traditional measures of segregation



Notes: "Real" series is from actual data; "random" series is from hypothetical data in which each speaker has her party label randomly assigned Republican with probability equal to the average fraction of Republicans in the sessions she appeared in. See the notes to figure 1 for definitions of the measures.

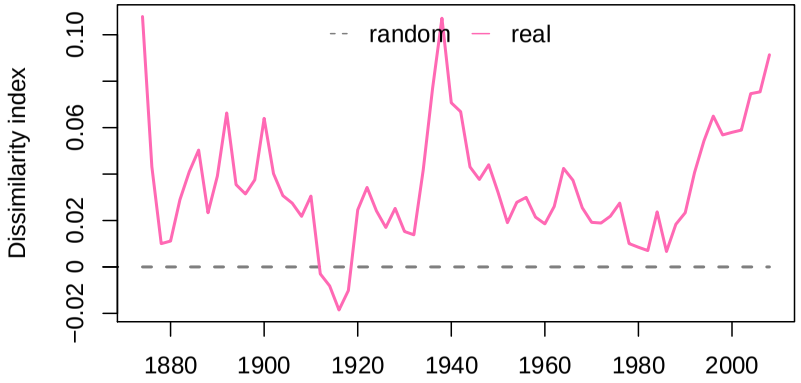
Figure 3: Partisanship of speech over time implied by text-based measures of polarization



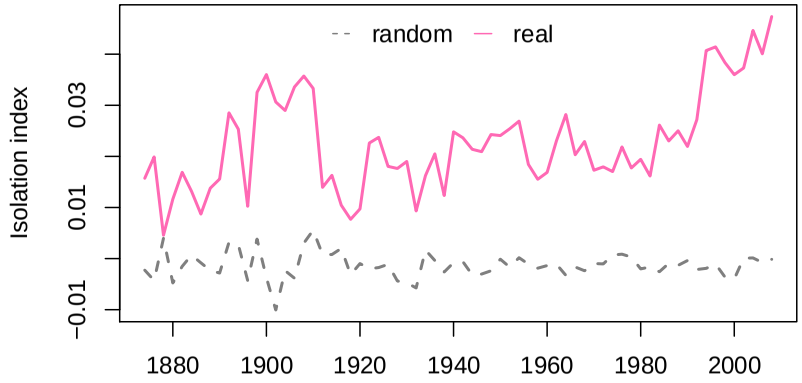
Notes: "Real" series is from actual data; "random" series is from hypothetical data in which each speaker has her party label randomly assigned Republican with probability equal to the average fraction of Republicans in the sessions she appeared in. Let $\rho_{jt} = \text{corr}(c_{itj}, r_{it})$. (Panel A) A speaker's slant in session t is $\sum_j c_{itj}\rho_{jt} / m_{it}$; the plot shows the difference between the average slant of Republicans and the average slant of Democrats in each session. (Panel B) Polarization in session t is $\sum_j m_{jt}\,|\rho_{jt}| \,/ \sum_j m_{jt}$; the plot shows polarization that has been standardized across the time series. Following Jensen et al. (2012), we restrict the sample of phrases to the 10,000 phrases with the greatest values of Pearson's χ^2 statistic in each session of congress.

Figure 4: Partisanship of speech over time with finite-sample corrections

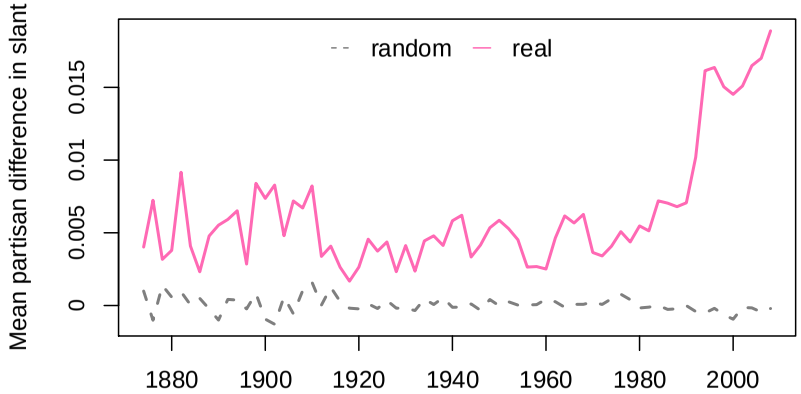
Panel A: Adjusted dissimilarity (Carrington and Troske 1997)



Panel B: Leave-out isolation (Gentzkow and Shapiro 2011)

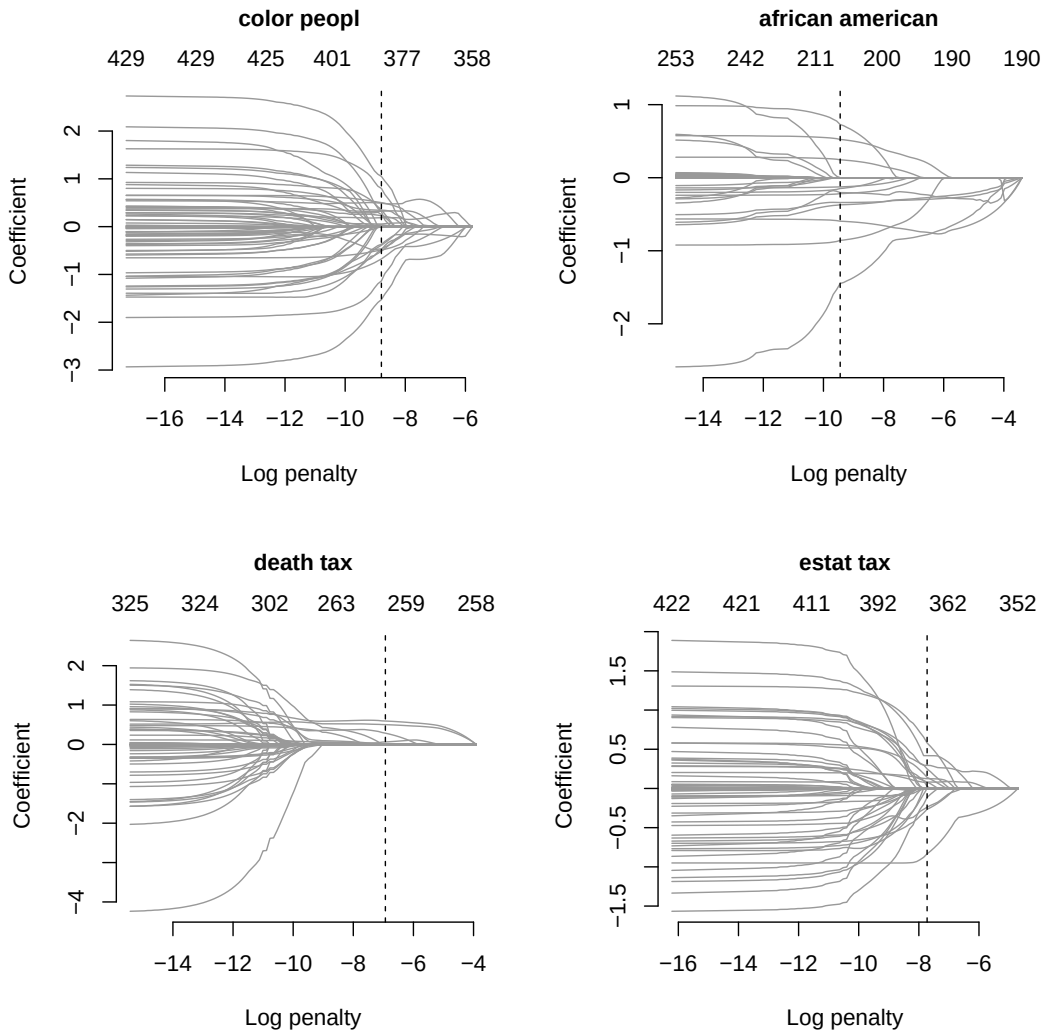


Panel C: Leave-out slant (Gentzkow and Shapiro 2010)



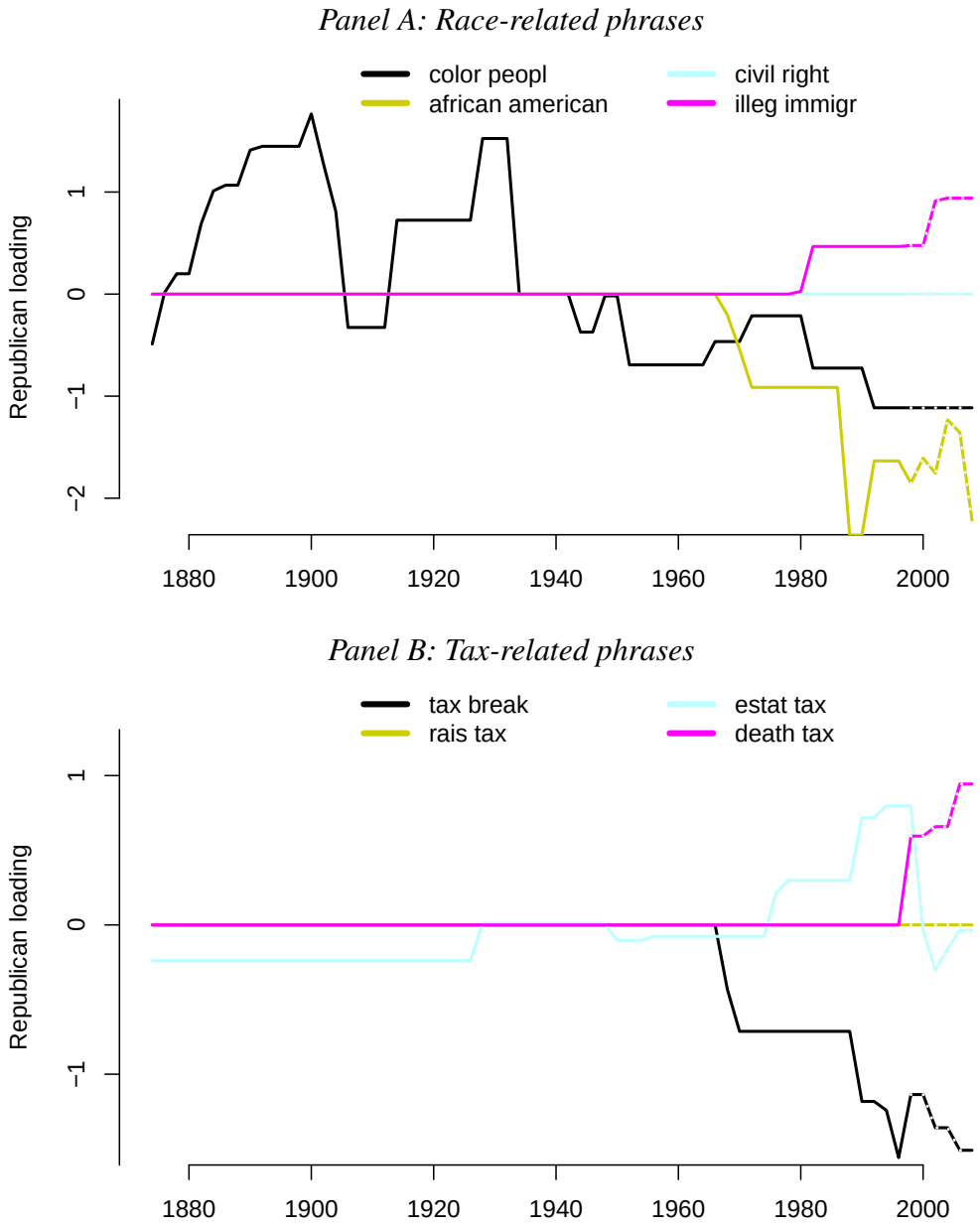
Notes: "Real" series is from actual data; "random" series is from hypothetical data in which each speaker has her party label randomly assigned Republican with probability equal to the average fraction of Republicans in the sessions she appeared in. (Panel A) Let $s_{t,rand}^{dissimilarity}$ be the dissimilarity index for session t implied by the "random" series. The plot shows the adjusted dissimilarity index, defined as $\left(s_t^{dissimilarity} - s_{t,rand}^{dissimilarity}\right) / \left(1 - s_{t,rand}^{dissimilarity}\right)$ if $s_t^{dissimilarity} \geq s_{t,rand}^{dissimilarity}$ and $\left(s_t^{dissimilarity} - s_{t,rand}^{dissimilarity}\right) / s_{t,rand}^{dissimilarity}$ if $s_t^{dissimilarity} < s_{t,rand}^{dissimilarity}$. (Panel B) For speaker i in session t, let the *Republican exposure* be $r_{\sim i,t} \equiv \left[\sum_j c_{itj}\left(\frac{\sum_{k\neq i} c_{ktj} r_k}{\sum_{k\neq i} c_{ktj}}\right)\right] / m_{it}$, i.e., the frequency-weighted Republican share among those who use the same phrases as speaker i. The plot shows the difference between the average Republican exposure of Republicans and the average Republican exposure of Democrats in each session. (Panel C) For speaker i in session t, define $\rho_{itj} \equiv \text{corr}(c_{ktj}, r_{kt})$, where $k \neq i$. Her leave-out slant is then $\sum_j c_{itj}\rho_{itj} / m_{it}$; the plot shows the difference between the average leave-out slant of Republicans and the average leave-out slant of Democrats in each session.

Figure 5: Examples of regularization paths



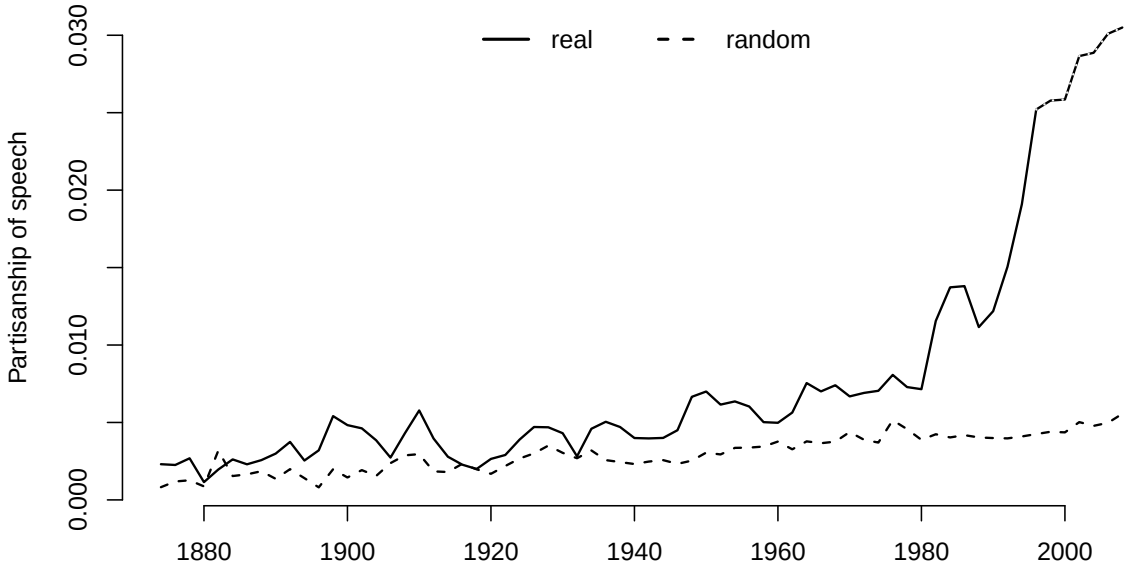
Notes: Plots show examples of regularization paths. Grey lines plot party loadings $\hat{\tilde{\varphi}}_{jt}$ for phrases j and sessions t and the dashed vertical lines indicate BIC selections. The row of integers at the top of each plot shows the number of parameters estimated to be nonzero at each penalty.

Figure 6: Party loadings over time for race- and tax-related phrases



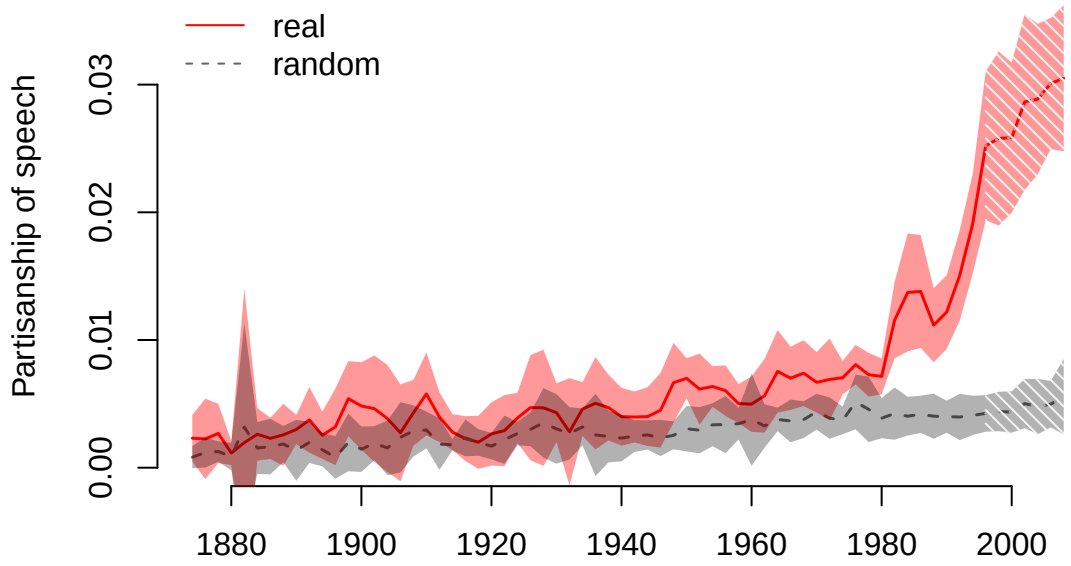
Notes: Plots show the party loadings $\hat{\varphi}_{jt}$ in time for selected phrases related to 'race' and 'tax.'

Figure 7: Partisanship of speech from baseline model specification



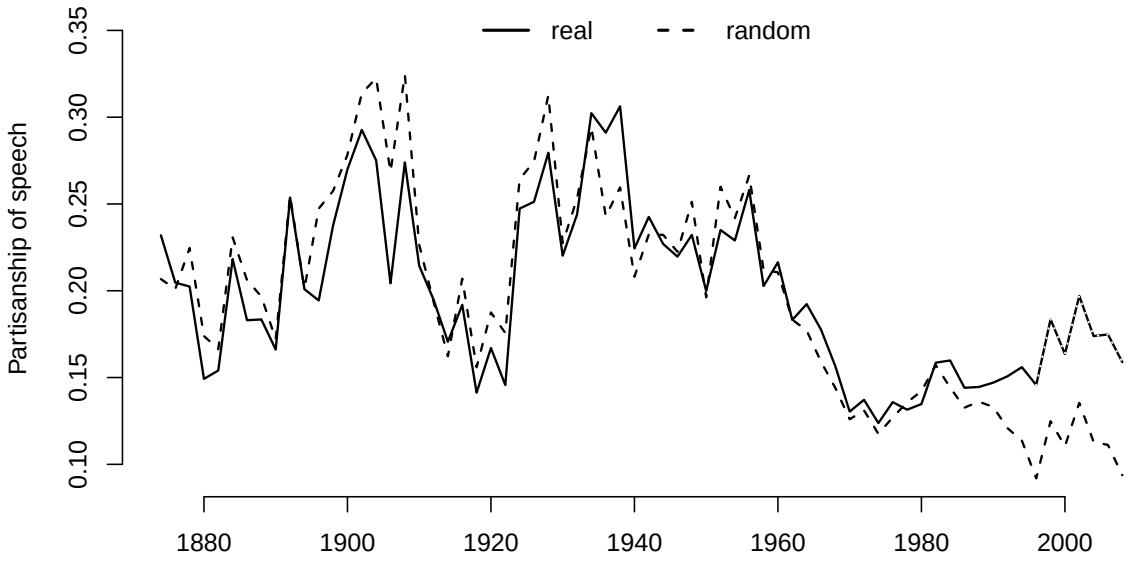
Notes: Plot shows the mean difference in partisanship z_{it} between Republicans and Democrats in each period. The solid line indicates partisanship measured based on actual party of speakers, and the dashed line indicates partisanship measured using hypothetical data in which each speaker has her party label randomly assigned Republican with probability equal to the average fraction of Republicans in the sessions she appeared in.

Figure 8: Nonparametric bootstrap of speaker partisanship



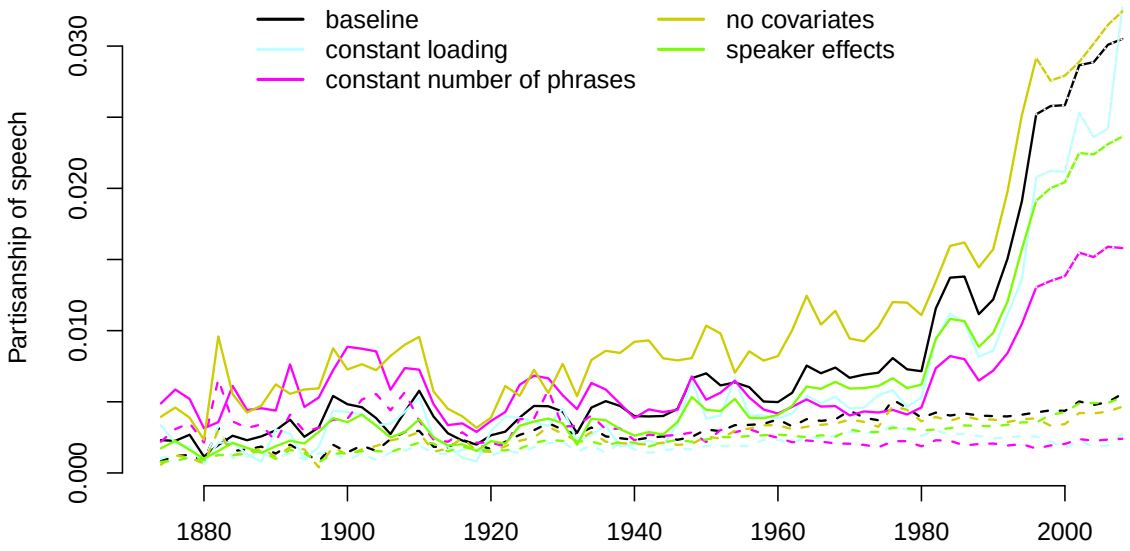
Notes: Plot shows a pointwise confidence interval for the plot in figure 7. The radius of the confidence interval is equal to two standard errors. Standard errors are estimated from a nonparametric bootstrap with 10 replicates. In each replicate we resample speakers with replacement and re-estimate the model.

Figure 9: Partisanship of speech from model with minimal penalty



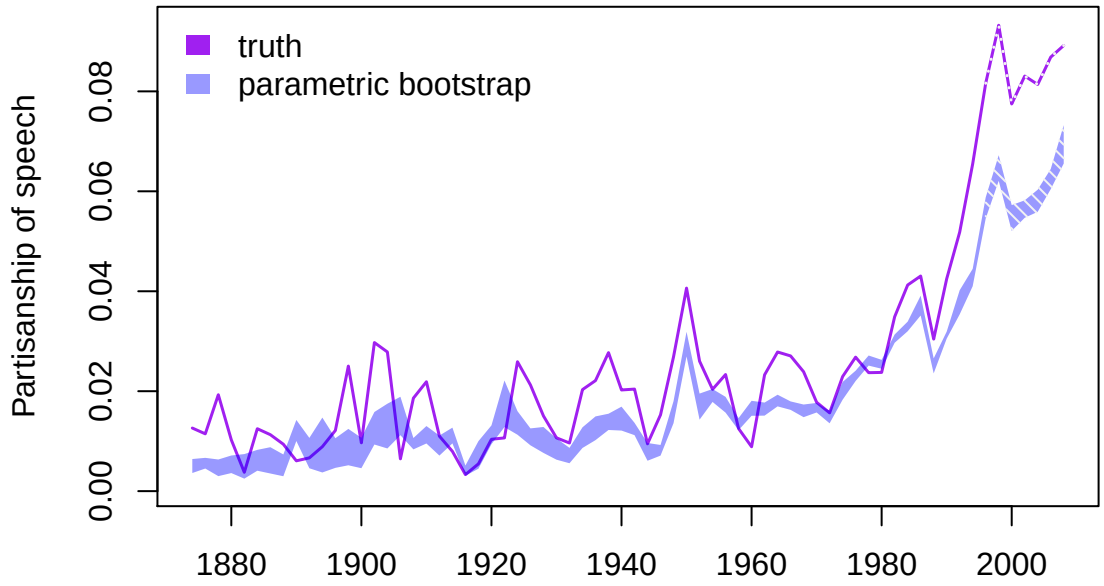
Notes: Plot is a version of figure 7 where partisanship of speech in each session is estimated with minimal penalty on party loadings φ_{jt}. The solid line indicates the partisanship measured based on actual party of speakers, and the dashed line indicates the partisanship measured using hypothetical data in which each speaker has her party label randomly assigned Republican with probability equal to the average fraction of Republicans in the sessions she appeared in.

Figure 10: Partisanship of speech from model variants



Notes: Solid line for each color indicates the partisanship measured based on actual party of speakers, and dashed line for each color indicates the partisanship measured using hypothetical data in which each speaker has her party label randomly assigned Republican with probability equal to the average fraction of Republicans in the sessions she appeared in. Baseline corresponds to the model shown in figure 7; other specifications are defined in section 6.

Appendix Figure 1: Parametric bootstrap of speaker partisanship



Notes: Plot shows parametric boostrap results for time path of partisanship. We restrict attention to the data from the most frequently spoken 1,000 phrases for computational reasons. We begin by estimating our baseline model on the restricted data. To produce each bootstrap replicate, we generate data for all speaker-sessions using our estimated model. We then re-estimate the model on the generated data. The plot shows the 10th–90th percentile range of 10 replicates.