NBER WORKING PAPER SERIES

MEASURING GROUP DIFFERENCES IN HIGH-DIMENSIONAL CHOICES: METHOD AND APPLICATION TO CONGRESSIONAL SPEECH

Matthew Gentzkow
Jesse M. Shapiro
Matt Taddy

Working Paper 22423
http://www.nber.org/papers/w22423

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
July 2016, Revised March 2019

Previously circulated as "Measuring Polarization in High-Dimensional Data: Method and Application to Congressional Speech." We acknowledge funding from the Initiative on Global Markets and the Stigler Center at Chicago Booth, the National Science Foundation, the Brown University Population Studies and Training Center, and the Stanford Institute for Economic Policy Research (SIEPR).We thank Egor Abramov, Brian Knight, John Marshall, Suresh Naidu, Vincent Pons, Justin Rao, and Gaurav Sood for their comments and suggestions. We thank Frances Lee for sharing her data on congressional communications staff. We also thank numerous seminar audiences and our many dedicated research assistants for their contributions to this project. This work was completed in part with resources provided by the University of Chicago Research Computing Center and the Stanford Research Computing Center. The data providers and funding agencies bear no responsibility for use of the data or for interpretations or inferences based upon such uses. The views expressed herein are those of the authors and do not necessarily reflect the views of the National Bureau of Economic Research.

At least one co-author has disclosed a financial relationship of potential relevance for this research. Further information is available online at http://www.nber.org/papers/w22423.ack

Measuring Group Differences in High-Dimensional Choices: Method and Application to Congressional Speech

Matthew Gentzkow, Jesse M. Shapiro, and Matt Taddy

NBER Working Paper No. 22423
July 2016, Revised March 2019
JEL No. D72

ABSTRACT

We study the problem of measuring group differences in choices when the dimensionality of the choice set is large. We show that standard approaches suffer from a severe finite-sample bias, and we propose an estimator that applies recent advances in machine learning to address this bias. We apply this method to measure trends in the partisanship of congressional speech from 1873 to 2016, defining partisanship to be the ease with which an observer could infer a congressperson's party from a single utterance. Our estimates imply that partisanship is far greater in recent years than in the past, and that it increased sharply in the early 1990s after remaining low and relatively constant over the preceding century.

Matthew Gentzkow
Department of Economics
Stanford University
579 Serra Mall
Stanford, CA 94305
and NBER
gentzkow@stanford.edu

Jesse M. Shapiro
Economics Department
Box B
Brown University
Providence, RI 02912
and NBER
jesse_shapiro_1@brown.edu

Matt Taddy
Amazon
mataddy@gmail.com

A online appendix is available at http://www.nber.org/data-appendix/w22423
A dataset is available at https://data.stanford.edu/congress_text

1 Introduction

In many settings, researchers seek to measure differences in the choices made by different groups, and the way such differences evolve over time. Examples include measuring the extent of racial segregation in residential choices (Reardon and Firebaugh 2002), of partisanship in digital media consumption (Gentzkow and Shapiro 2011; Flaxman et al. 2016), of geographic differences in treatment choices of physicians (Chandra et al. 2012), and of differences between demographic groups in survey responses (Bertrand and Kamenica 2018). We consider the problem of measuring such differences in settings where the dimensionality of the choice set is large—i.e., where the number of possible choices is large relative to the number of actual choices observed. We show that in such settings standard approaches suffer from a severe finite-sample bias, and we propose methods based on recent advances in machine learning that address this bias in a way that is computationally tractable with large-scale data.

Our approach is motivated by a specific application: measuring trends in party differences in political speech. It is widely apparent that America's two political parties speak different languages.[1] Partisan differences in language diffuse into media coverage (Gentzkow and Shapiro 2010; Martin and Yurukoglu 2017) and other domains of public discourse (Greenstein and Zhu 2012; Jensen et al. 2012), and partisan framing has been shown to have large effects on public opinion (Nelson et al. 1997; Graetz and Shapiro 2006; Chong and Druckman 2007).

Our main question of interest is to what extent the party differences in speech that we observe today are a new phenomenon. One can easily find examples of politically charged terms in America's distant past.[2] Yet the magnitude of the differences between parties, 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 the partisan differences in language that we see today might represent a consequential change (Lakoff 2003). If the two parties speak more differently today than in the past, these divisions could be contributing to deeper polarization in Congress and cross-party animus in the broader public.

[1] See, for example, Gentzkow and Shapiro (2010), Ball (2013), and *Economist* (2013). 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 (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).

We use data on the text of speeches in the US Congress from 1873 to 2016 to quantify the magnitude of partisan differences in speech, and to characterize the way these differences have evolved over time. We specify a multinomial model of speech with choice probabilities that vary by party. We measure partisan differences in speech in a given session of Congress by the ease with which an observer who knows the model could guess a speaker's party based solely on the speaker's choice of a single phrase. We call this measure *partisanship* for short.

To compute an accurate estimate of partisanship, we must grapple with two methodological challenges. The first is the finite-sample bias mentioned above. The bias arises because the number of phrases a speaker could choose is large relative to the total amount of speech we observe, so many phrases are said mostly by one party or the other purely by chance. Naive estimators interpret such differences as evidence of partisanship, leading to a bias we show can be many orders of magnitude larger than the true signal in the data. Second, although our model takes a convenient multinomial logit form, the large number of choices and parameters makes standard approaches to estimation computationally infeasible.

We use two estimation approaches to address these challenges. The first is a leave-out estimator that addresses the main source of finite-sample bias while allowing for simple inspection of the data. The second, our preferred estimator, uses 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. A permutation test and an out-of-sample validation both suggest that any bias that remains in these estimates is dramatically lower than in standard approaches, and small relative to the true variation in partisanship over time.

We find that the partisanship of language has exploded in recent decades, reaching an unprecedented level. From 1873 to the early 1990s, partisanship was nearly 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 57 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 73 percent by the 110th session (2007-09). Methods that do not correct for finite-sample bias, including the maximum likelihood estimator of our model, instead imply that partisanship is no higher today than in the past.

We unpack the recent increase in partisanship along a number of dimensions. The most partisan phrases in each period—defined as those phrases most diagnostic of the speaker's party—align well

with the issues emphasized in party platforms and, in recent years, include well-known partisan phrases such as "death tax" and "estate tax." 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).

While we cannot definitively say 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* (Gingrich and Armey 1994). This election is widely considered a watershed moment in political marketing, with consultants such as Frank Luntz applying novel techniques to identify effective language and disseminate it to candidates (Lakoff 2004; Luntz 2004; Bai 2005). We also discuss related changes such as the expansion of cable television coverage that may have provided further incentives for linguistic innovation.

This discussion highlights that partisanship of speech as we define it is a distinct phenomenon from other inter-party differences. In particular, the large body of work building on the ideal point model of Poole and Rosenthal (1985) finds that inter-party differences in roll-call voting fell from the late nineteenth to the mid-twentieth century, and have since steadily increased (McCarty et al. 2015). These dynamics are very different from those we observe in speech, consistent with our expectation that speech and roll-call votes respond to different incentives and constraints, and suggesting that the analysis of speech may reveal aspects of the political landscape that are not apparent from the analysis of roll-call votes.

We build on methods developed by Taddy (2013, 2015). Many aspects of the current paper, including our proposed leave-out estimator, our approaches to validation and inference, and the covariate specification of our model are novel with respect to that prior work. Most importantly, Taddy (2013, 2015) makes no attempt to define or quantify the divergence in language between groups either at a point in time or over time, nor does he discuss the finite sample biases that arise in doing so. Our paper also relates to other work on measuring document partisanship, including Laver et al. (2003), Groseclose and Milyo (2005), Gentzkow and Shapiro (2010), Kim et al. (2018), and Yan et al. (2018).[3]

Our paper contributes a recipe for using statistical predictability in a probability model of

[3]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).

speech as a metric of differences in partisan language between groups. Jensen et al. (2012) use text from the *Congressional Record* to characterize party differences in language from the late nineteenth century to the present. Their index, which is based on the observed correlation of phrases with party labels, implies that partisanship has been rising recently but was similarly high in the past. We apply a different method that addresses finite-sample bias and leads to substantially different conclusions. Lauderdale and Herzog (2016) specify a generative hierarchical model of floor debates and estimate the model on speech data from the Irish Dail and the US Senate. Studying the US Senate from 1995 to 2014, they find that party differences in speech have increased faster than party differences in roll-call voting. Peterson and Spirling (2018) study trends in the partisanship of speech in the UK House of Commons. In contrast to Lauderdale and Herzog's (2016) analysis (and ours), Peterson and Spirling (2018) do not specify a generative model of speech. Instead, Peterson and Spirling (2018) measure partisanship using the predictive accuracy of several machine-learning algorithms. They cite our article to justify using randomization tests to check for spurious trends in their measure. These tests (Peterson and Spirling 2018, Online Appendix C) show that their measure implies significant and time-varying partisanship even in fictitious data in which speech patterns are independent of party.

The recipe that we develop can be applied to a broad class of problems in which the goal is to characterize group differences in high-dimensional choices. A prominent example is the measurement of residential segregation (e.g., Reardon and Firebaugh 2002), where the groups might be defined by race or ethnicity and the choices might be neighborhoods or schools. The finite-sample bias that we highlight has been noted in that context by Cortese et al. (1976) and 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; D'Haultfœuille and Rathelot 2017).[4] Recent work has derived axiomatic foundations for segregation measures (Echenique and Fryer 2007; Frankel and Volij 2011), asking which measures of segregation satisfy certain properties.[5] Instead, our approach is to specify a generative model of the data and to measure group differences using objects that have a well-defined meaning in the context of the model.[6] In the body of the paper, we note some formal connections to the litera-

[4]Logan et al. (2018) develop methods for bias correction in the context of measuring residential segregation by income.

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

[6]In this respect, our paper builds on Ellison and Glaeser (1997), who use a model-based approach to measure agglomeration spillovers in US manufacturing. Davis et al. (forthcoming) use a structural demand model to estimate racial segregation in restaurant choices in a sample of New York City Yelp reviewers. Mele (2017) shows how to estimate

ture on residential segregation, and in an earlier draft we pursue a detailed application to trends in residential segregation by political affiliation (Gentzkow et al. 2017).

2 Congressional Speech Data

Our primary data source is the text of the *United States Congressional Record* (hereafter, the *Record*) from the 43rd Congress to the 114th Congress. We obtain digital text from HeinOnline, who performed optical character recognition (OCR) on scanned print volumes. The *Record* is a "substantially verbatim" record of speech on the floor of Congress (Amer 1993). We exclude Extensions of Remarks, which are used to print unspoken additions by members of the House that are not germane to the day's proceedings.[7]

The modern *Record* is issued in a daily edition, printed at the end of each day that Congress is in session, and in a bound edition that collects the content for an entire Congress. These editions differ in formatting and in some minor elements of content (Amer 1993). Our data contains bound editions for the 43rd to 111th Congresses, and daily editions for the 97th to 114th Congresses. We use the bound edition in the sessions where it is available and the daily edition thereafter. The Online Appendix shows results from an alternative data build that uses the bound edition through the 96th Congress and the daily edition thereafter.

We use an automated script to parse the raw text into individual speeches. Beginnings of speeches are demarcated in the *Record* by speaker names, usually in all caps (e.g., "Mr. ALLEN of Illinois."). We determine the identity of each speaker using a combination of manual and automated procedures, and append data on the state, chamber, and gender of each member from historical sources.[8] We exclude any speaker who is not a Republican or a Democrat, speakers who

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.

[7] The *Record* seeks to capture speech as it was intended to have been said (Amer 1993). Speakers are allowed to insert new remarks, extend their remarks on a specific topic, and remove errors from their own remarks before the *Record* is printed. The rules for such insertions and edits, as well as the way they appear in print, differ between the House and Senate, and have changed to some degree over time (Amer 1993; Johnson 1997; Haas 2015). We are not aware of any significant changes that align with the changing partisanship we observe in our data. We present our results separately for the House and Senate in the Online Appendix.

[8] Our main source for information on congresspeople is the congress-legislators GitHub repository https://github.com/unitedstates/congress-legislators/tree/1473ea983d5538c25f5d315626445ab038d8141b accessed on November 15, 2016. We make manual corrections, and add additional information from ICPSR and McKibbin (1997), the Voteview Roll Call Data (Carroll et al. 2015a, b), and the King (1995) election returns. Some of these sources include metadata from Martis (1989).

are identified by office rather than name, non-voting delegates, and speakers whose identities we cannot determine.[9] The Online Appendix presents the results of a manual audit of the reliability of our parsing.

The input to our main analysis is a matrix $\mathbf{C}_t$ whose rows correspond to speakers and whose columns correspond to distinct two-word phrases or bigrams (hereafter, simply "phrases"). An element c_{ijt} thus gives the number of times speaker i has spoken phrase j in session (Congress) t. To create these counts, we first perform the following pre-processing steps: (i) delete hyphens and apostrophes; (ii) replace all other punctuation with spaces; (iii) remove non-spoken parenthetical insertions; (iv) drop a list of extremely common words;[10] and (v) reduce words to their stems according to the Porter2 stemming algorithm (Porter 2009). We then drop phrases that are likely to be procedural or have low semantic meaning according to criteria we define in the Online Appendix. Finally, we restrict attention to phrases spoken at least 10 times in at least one session, spoken in at least 10 unique speaker-sessions, and spoken at least 100 times across all sessions. The Online Appendix presents results from a sample in which we tighten each of these restrictions by 10 percent. The Online Appendix also presents results from an alternative construction of $\mathbf{C}_t$ containing counts of three-word phrases or trigrams.

The decision to represent text as a matrix of phrase counts is fairly common in text analysis, as is the decision to reduce the dimensionality of the data by removing word stems and non-word content (Gentzkow et al. forthcoming). We remove procedural phrases because they appear frequently and their use is likely not informative about the inter-party differences that we wish to measure (Gentzkow and Shapiro 2010). We remove infrequently used phrases to economize on computation (Gentzkow et al. forthcoming).

[9] In the rare case in which a speaker switches parties during a term, we assign the new party to all the speech in that term. We handle the similarly rare case in which a speaker switches chambers in a single session (usually from the House to the Senate) by treating the text from each chamber as a distinct speaker-session. If a speaker begins a session in the House as a non-voting delegate of a territory and receives voting privileges after the territory gains statehood, we treat the speaker as a voting delegate for the entirety of that speaker-session. If a non-voting delegate of the House later becomes a senator, we treat each position as a separate speaker-session. We obtain data on the acquisition of statehood from http://www.thirty-thousand.org/pages/QHA-02.htm (accessed on January 18, 2017) and data on the initial delegates for each state from https://web.archive.org/web/20060601025644/http://www.gpoaccess.gov/serialset/cdocuments/hd108-222/index.html. When we assign a majority party in each session, we count the handful of independents that caucus with the Republicans or Democrats as contributing to the party's majority in the Senate. Due to path dependence in our data build, such independents are omitted when computing the majority party in the House. The Online Appendix shows the results of a specification in which we exclude from the sample any speaker whose party changes between sessions.

[10] The set of these "stopwords" we drop is defined by a list obtained from http://snowball.tartarus.org/algorithms/english/stop.txt on November 11, 2010.

The resulting vocabulary contains 508,352 unique phrases spoken a total of 287 million times by 7,732 unique speakers. We analyze data at the level of the speaker-session, of which there are 36,161. The Online Appendix reports additional summary statistics for our estimation sample and vocabulary.

We identify 22 substantive topics based on our knowledge of the *Record*. We associate each topic with a non-mutually exclusive subset of the vocabulary. To do this, we begin by grouping a set of partisan phrases into the 22 topics (e.g., taxes, defense, etc.). For each topic, we form a set of keywords by (i) selecting relevant words from the associated partisan phrases and (ii) manually adding other topical words. Finally, we identify all phrases in the vocabulary that include one of the topic keywords, are used more frequently than a topic-specific occurrence threshold, and 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 Model of Speech

The observed outcome is a $J-$vector $\mathbf{c}_{it}$ of phrase counts for speaker i, which we assume comes from a multinomial distribution

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

with $m_{it} = \sum_j c_{ijt}$ denoting the total amount of speech by speaker i in session t, $P(i) \in \{R, D\}$ denoting the party affiliation of speaker i, $\mathbf{x}_{it}$ denoting a $K-$vector of (possibly time-varying) speaker characteristics, and $\mathbf{q}_t^P(\mathbf{x}_{it}) \in (0,1)^J$ denoting the vector of choice probabilities. We let $R_t = \{i : P(i) = R, m_{it} > 0\}$ and $D_t = \{i : P(i) = D, m_{it} > 0\}$ denote the set of Republicans and Democrats, respectively, active in session 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.

We suppose further that the choice probabilities are

$$q_{jt}^{P(i)}(\mathbf{x}_{it}) = e^{u_{ijt}} / \sum_l e^{u_{ilt}} \tag{2}$$

$$u_{ijt} = \alpha_{jt} + \mathbf{x}_{it}'\boldsymbol{\gamma}_{jt} + \varphi_{jt}\mathbf{1}_{i \in R_t}.$$

Here α_{jt} is a scalar parameter capturing the baseline popularity of phrase j in session t, $\boldsymbol{\gamma}_{jt}$ is a $K-$vector capturing the effect of characteristics $\mathbf{x}_{it}$ on the propensity to use phrase j in session t, and φ_{jt} is a scalar parameter capturing the effect of party affiliation on the propensity to use phrase j in session t. If $\mathbf{x}_{it} := \mathbf{x}_t$, any phrase probabilities $\left(\mathbf{q}_t^R(\cdot), \mathbf{q}_t^D(\cdot)\right)$ can be represented with appropriate choice of parameters in equation (2).

The model in (1) and (2) is restrictive, and it ignores many important aspects of speech. For example, it implies that the propensity to use a given phrase is not related to other phrases used by speaker i in session t, and need not be affected by the speaker's verbosity m_{it}. We adopt this model because it is tractable and has proved useful in extracting meaning from text in many related contexts (Groseclose and Milyo 2005; Taddy 2013, 2015).

The model also implies that speaker identities matter only through party affiliation $P(i)$ and the characteristics $\mathbf{x}_{it}$. Specification of $\mathbf{x}_{it}$ is therefore important for our analysis. We consider specifications of $\mathbf{x}_{it}$ with different sets of observable characteristics, as well as a specification with unobserved speaker characteristics (i.e., speaker random effects).

We assume throughout that if a phrase (or set of phrases) is excluded from the choice set, the relative frequencies of the remaining phrases are unchanged. We use this assumption in Sections 6 and 7 to compute average partisanship for interesting subsets of the full vocabulary. This assumption encodes the independence of irrelevant alternatives familiar from other applications of the multinomial logit model. It is a restrictive assumption, as some phrases are clearly better substitutes than others, but it provides a useful benchmark for analysis absent a method for estimating flexible substitution patterns in a large vocabulary.

3.2 Measure of Partisanship

For given characteristics $\mathbf{x}$, we define partisanship of speech to be 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 say that partisanship is low. When these vectors are far from each other, the parties speak differently and we 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 utter a single phrase.

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

$$\pi_t(\mathbf{x}) = \frac{1}{2}\mathbf{q}_t^R(\mathbf{x}) \cdot \boldsymbol{\rho}_t(\mathbf{x}) + \frac{1}{2}\mathbf{q}_t^D(\mathbf{x}) \cdot (1 - \boldsymbol{\rho}_t(\mathbf{x})), \tag{3}$$

where

$$\rho_{jt}(\mathbf{x}) = \frac{q_{jt}^R(\mathbf{x})}{q_{jt}^R(\mathbf{x}) + q_{jt}^D(\mathbf{x})}. \tag{4}$$

Average partisanship in session t is:

$$\overline{\pi}_t = \frac{1}{|R_t \cup D_t|} \sum_{i \in R_t \cup D_t} \pi_t(\mathbf{x}_{it}). \tag{5}$$

To understand these definitions, note 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. Average partisanship is defined with respect to a given vocabulary of J phrases.

There are many possible measures of the divergence between between $\mathbf{q}_t^R(\mathbf{x})$ and $\mathbf{q}_t^D(\mathbf{x})$. We show in the Online Appendix that the time series of partisanship looks qualitatively similar if we replace our partisanship measure with either the Euclidean distance between $\mathbf{q}_t^R(\mathbf{x})$ and $\mathbf{q}_t^D(\mathbf{x})$ or the implied mutual information between party and phrase choice, though the series for Euclidean distance is noisier.

Partisanship is closely related to the isolation index, a common index of residential segregation (White 1986; Cutler et al. 1999).[11] 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]

[11] To see this, imagine that choices are neighborhoods rather than phrases, and let $m_{it} = 1$ for all i and t, so that each individual chooses one and only one neighborhood. 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 share of Republicans and Democrats, all with characteristics $\mathbf{x}$, this is simply $2\pi_t(\mathbf{x}) - 1$.

[12] In our context, Independence would require that the ranking in terms of partisanship of two years t and s remains unchanged if we add a new set of phrases J^* to the vocabulary whose probabilities are the same in both years

Average partisanship $\overline{\pi}_t$ summarizes how well an observer can predict a hypothetical speaker's party given a single realization and knowledge of the true model. This is distinct from the question of how well an econometrician can predict a given speaker's party in a given sample of text.

4 Estimation, Inference, and Validation

4.1 Plug-in Estimators

Maximum likelihood estimation is straightforward in our context. Ignoring covariates $\mathbf{x}$, the maximum likelihood estimator (MLE) can be computed by plugging in empirical analogues for the terms that appear in equation (3).

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 $\hat{\rho}_{jt} = \hat{q}_{jt}^R / \left(\hat{q}_{jt}^R + \hat{q}_{jt}^D \right)$, excluding from the choice set any phrases that are not spoken in session t. Then the MLE of $\overline{\pi}_t$ when $\mathbf{x}_{it} := \mathbf{x}_t$ is:

$$\hat{\pi}_t^{MLE} = \frac{1}{2} \left(\hat{\mathbf{q}}_t^R \right) \cdot \hat{\boldsymbol{\rho}}_t + \frac{1}{2} \left(\hat{\mathbf{q}}_t^D \right) \cdot (1 - \hat{\boldsymbol{\rho}}_t) \,. \tag{6}$$

An important theme of our paper is that this and related estimators can be severely biased in finite samples even if $\mathbf{x}_{it} := \mathbf{x}_t$. Intuitively, partisanship will be high when the dispersion of the posteriors ρ_{jt} is large—i.e., when some phrases are spoken far more by Republicans and others are spoken far more by Democrats. The MLE estimates the ρ_{jt} using their sample analogues $\hat{\rho}_{jt}$. However, sampling error will tend to increase the dispersion of the $\hat{\rho}_{jt}$ relative to the dispersion of the true ρ_{jt}. When the number of phrases is large relative to the volume of speech observed, many phrases will be spoken only a handful of times, and so may be spoken mainly by Republicans ($\hat{\rho}_{jt} \approx 1$) or mainly by Democrats ($\hat{\rho}_{jt} \approx 0$) by chance even if the true choice probabilities do not differ by party.

To see the source of the bias more formally, note that $\hat{\pi}_t^{MLE}$ is a convex function of $\hat{\mathbf{q}}_t^R$ and $\hat{\mathbf{q}}_t^D$, and so Jensen's inequality implies that it has a positive bias. We can also 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:

$$\mathrm{E}\left(\left(\hat{\mathbf{q}}_t^R \right) \cdot \hat{\boldsymbol{\rho}}_t - \left(\mathbf{q}_t^R \right) \cdot \boldsymbol{\rho}_t \right) = \left(\mathbf{q}_t^R \right) \cdot \mathrm{E}\left(\hat{\boldsymbol{\rho}}_t - \boldsymbol{\rho}_t \right) + \mathrm{Cov}\left(\left(\hat{\mathbf{q}}_t^R - \mathbf{q}_t^R \right), \left(\hat{\boldsymbol{\rho}}_t - \boldsymbol{\rho}_t \right) \right). \tag{7}$$

($q_{jt}^P = q_{js}^P \forall P, j \in J^*$). Frankel and Volij (2011) list one other axiom, the Group Division Property, which is only applicable for indices where the number of groups (i.e., parties in our case) is allowed to vary.

The second term will typically be far from zero because the sampling error in $\hat{\boldsymbol{\rho}}_t$ is mechanically related to the sampling error in $\left(\hat{\mathbf{q}}_t^R, \hat{\mathbf{q}}_t^D\right)$. Any positive residual in $\hat{\mathbf{q}}_t^R$ will increase both terms inside the covariance; any negative residual will do the reverse. The first term is also nonzero because $\hat{\boldsymbol{\rho}}_t$ is a nonlinear transformation of $\left(\hat{\mathbf{q}}_t^R, \hat{\mathbf{q}}_t^D\right)$,[13] though this component of the bias tends to be small in practice.

The bias we highlight is not specific to the MLE, but will tend to arise for any measure of group differences that uses observed choices as a direct approximation of true choice probabilities. This is especially transparent if we measure the difference between $\mathbf{q}_t^R$ and $\mathbf{q}_t^D$ using a norm such as Euclidean distance: Jensen's inequality implies that for any norm $\|\cdot\|$, $\mathrm{E}\left\|\hat{\mathbf{q}}_t^R - \hat{\mathbf{q}}_t^D\right\| > \left\|\mathbf{q}_t^R - \mathbf{q}_t^D\right\|$. 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 ($\mathbf{q}_t^R = \mathbf{q}_t^D$) and verbosity is fixed, then the population value of $corr\left(c_{ijt}, \mathbf{1}_{i\in R_t}\right)$ is zero. But in any finite sample the correlation will be nonzero with positive probability, so the measure may imply party differences even when speech is unrelated to party.

4.2 Leave-Out Estimator

The first approach we propose to addressing this bias is a leave-out estimator that uses different samples to estimate $\hat{\mathbf{q}}_t^P$ and $\hat{\boldsymbol{\rho}}_t$. This makes the errors in the former independent of the errors in the latter by construction, and so eliminates the second bias term in equation (7).

The leave-out estimator is given by:

$$\hat{\pi}_t^{LO} = \frac{1}{2}\frac{1}{|R_t|}\sum_{i\in R_t}\hat{\mathbf{q}}_{i,t}\cdot\hat{\boldsymbol{\rho}}_{-i,t} + \frac{1}{2}\frac{1}{|D_t|}\sum_{i\in D_t}\hat{\mathbf{q}}_{i,t}\cdot\left(1-\hat{\boldsymbol{\rho}}_{-i,t}\right), \tag{8}$$

where $\hat{\boldsymbol{\rho}}_{-i,t}$ is the analogue of $\hat{\boldsymbol{\rho}}_t$ computed from the speech of all speakers other than i.[14] This estimator is biased for $\overline{\pi}_t$, even if $\mathbf{x}_{it} := \mathbf{x}_t$, because of the first term in equation (7), but we expect (and find) that this bias is small in practice.

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

[14] For each i, j, and t, define $\hat{q}^P_{-i,j,t} = \frac{\sum_{l\in\{P_t\setminus i\}} c_{ljt}}{\sum_{l\in\{P_t\setminus i\}} m_{lt}}$ for $P \in \{R, D\}$ and

$$\hat{\rho}_{-i,j,t} = \frac{\hat{q}^R_{-i,j,t}}{\hat{q}^R_{-i,j,t} + \hat{q}^D_{-i,j,t}}.$$

Implicitly, in each session t we exclude from the calculation in (8) any phrase that is spoken only by a single speaker.

The leave-out estimator is simple to compute and provides a direct look at the patterns in the data. It also has important limitations. In particular, it does not allow us to incorporate covariates. In addition, it does not recover the underlying parameters of the model and so does not directly provide estimates of objects such as the most partisan phrases, which we rely on heavily in our application.

4.3 Penalized Estimator

The second approach we propose uses a penalized estimator to fully estimate the model and incorporate covariates. We estimate the parameters $\{\boldsymbol{\alpha}_t, \boldsymbol{\gamma}_t, \boldsymbol{\varphi}_t\}_{t=1}^{T}$ of equation (2) by minimization of the following penalized objective function:

$$\sum_j \Bigg\{ \sum_t \sum_i \Big[m_{it} \exp(\alpha_{jt} + \mathbf{x}_{it}' \boldsymbol{\gamma}_{jt} + \varphi_{jt} \mathbf{1}_{i \in R_t}) - c_{ijt} (\alpha_{jt} + \mathbf{x}_{it}' \boldsymbol{\gamma}_{jt} + \varphi_{jt} \mathbf{1}_{i \in R_t}) + \psi \left(|\alpha_{jt}| + \left\| \boldsymbol{\gamma}_{jt} \right\|_1 \right) + \lambda_j |\varphi_{jt}| \Big] \Bigg\}. \quad (9)$$

We form an estimate $\hat{\pi}_t^*$ of $\overline{\pi}_t$ by substituting estimated parameters into the probability objects in equation (5).

Because partisanship is defined as a function of the characteristics $\mathbf{x}$, the choice of characteristics to include in the model affects our target for estimation. We wish to include those characteristics that are likely to be related both to party and to speech but whose relationship with speech would not generally be thought of as a manifestation of party differences. A leading example of such a confound is geographic region: speakers from different parts of the country will tend to come from different parties and to use different phrases, but regional differences in language would not generally be thought of as a manifestation of party differences.

In our baseline specification, $\mathbf{x}_{it}$ consists of indicators for state, chamber, gender, Census region, and whether the party is in the majority for the entirety of the session. The coefficients $\boldsymbol{\gamma}_{jt}$ on these attributes are static in time (i.e., $\gamma_{jtk} := \gamma_{jk}$) except for those on Census region, which are allowed to vary freely across sessions to allow more flexibly for regional variation in speech. The online appendix shows results from a specification in which $\mathbf{x}_{it}$ includes unobserved speaker-level preference shocks (i.e. speaker random effects), from a specification in which $\mathbf{x}_{it}$ includes no covariates, and from a specification in which $\mathbf{x}_{it}$ includes several additional covariates.

The minimand in (9) 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 the parameter 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 $\lambda_j \left|\varphi_{jt}\right|$, which imposes sparsity on the party loadings and shrinks them toward zero (Tibshirani 1996). Sparsity and shrinkage limit the effect of sampling error on the dispersion of the estimated posteriors ρ_{jt}, which is the source of the bias in $\hat{\pi}_t^{MLE}$. We determine the penalties $\boldsymbol{\lambda}$ by regularization path estimation, first finding λ_j^1 large enough so that φ_{jt} is estimated to be 0, and then incrementally decreasing $\lambda_j^2, ..., \lambda_j^G$ and updating parameter estimates accordingly. 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. We then choose the value of λ_j that minimizes a Bayesian Information Criterion.[15] The Online Appendix reports a qualitatively similar time series of partisanship when we use 5- or 10-fold cross-validation to select the λ_j that minimizes average out-of-sample deviance.

We also impose a minimal penalty of $\psi = 10^{-5}$ on the phrase-specific intercepts α_{jt} and the covariate coefficients $\boldsymbol{\gamma}_{jt}$. We do this to handle the fact that some combinations of data and covariate design do not have an MLE in the Poisson model (Haberman 1973; Santos Silva and Tenreyro 2010). A small penalty allows us to achieve numerical convergence while still treating the covariates in a flexible way.[16]

[15]The Bayesian Information Criterion we use is $-2\sum_{i,t} \log \text{Pois}\left(c_{ijt}; \exp[\hat{\mu}_i + u_{ijt}]\right) + df \log n$, where $n = \sum_t \left(|D_t| + |R_t|\right)$ 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).

[16]The Online Appendix shows how our results vary with alternative values of ψ. Larger values of ψ decrease computational time for a given problem. Note that in practice we implement our regularization path computationally as

4.4 Inference

For all of our main results, we perform inference via subsampling. We draw without replacement 100 random subsets of size equal to one-tenth the number of speakers (up to integer restrictions) and re-estimate on each subset. We then report confidence intervals based on the distribution of the estimator across these subsets, under the assumption of $\sqrt{n}$ convergence. We center these confidence intervals around the estimated series and report uncentered bias-corrected confidence intervals for our main estimator in the Online Appendix.

Politis et al. (1999, Theorem 2.2.1) show that this procedure yields valid confidence intervals under the assumption that the distribution of the estimator converges weakly to some non-degenerate distribution at a $\sqrt{n}$ rate. In the Appendix, we extend a result of Knight and Fu (2000) to show that this property holds, with fixed vocabulary and a suitable rate condition on the penalty, for the penalized maximum likelihood estimator of our multinomial logit model. This is the estimator that we approximate with the Poisson distribution in equation 9. Though we do not pursue formal results for the case where the vocabulary grows with the sample size, we note that such asymptotics might better approximate the finite-sample behavior of our estimators.

In the Online Appendix, we report the results of several exercises designed to probe the accuracy of our confidence intervals. First, we consider three alternative subsampling strategies: (i) doubling the number of speakers in each subsample, (ii) using 10 non-overlapping subsamples rather than 100 overlapping subsamples, and (iii) using 5 non-overlapping subsamples. Second, we compute confidence intervals based on a parametric bootstrap, repeatedly simulating data from our estimated model and re-estimating the model on the simulated data. Third, we compute confidence intervals using a sample-splitting procedure that uses one half of the model to perform variable selection and then estimates the selected model with minimal penalty across repeated bootstrap replicates on the second half of the sample. All of these procedures yield qualitatively similar conclusions. Note that we do not report results for a standard nonparametric bootstrap; the standard nonparametric bootstrap is known to be invalid for lasso regression (Chatterjee and Lahiri 2011).

$\psi\tilde{\lambda}_j^2, ..., \psi\tilde{\lambda}_j^G$ where $\tilde{\lambda}_j^G = \iota\tilde{\lambda}_j^1$, $\iota = 10^{-5}$, and $G = 100$. To ensure that the choice of $\tilde{\lambda}_j$ is not constrained by the regularization path, we recommend that users choose values of ψ and ι small enough that forcing $\tilde{\lambda}_j = \tilde{\lambda}_j^G$ for all j either leads to $\hat{\pi}_t^* \approx \hat{\pi}_t^{MLE}$ or to an estimator $\hat{\pi}_t^*$ that substantially differs from the one chosen by BIC.

4.5 Validation

As usual with non-linear models, none of the estimators proposed here are exactly unbiased in finite samples. Our goal is to reduce bias to the point that it is dominated by the signal in the data. We gauge our success in three main ways.

First, we consider a permutation test in which we randomly reassign parties to speakers and then re-estimate each 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. Thus the random series for an unbiased estimator of π_t has expected value $\frac{1}{2}$ in each session t, and the deviation from $\frac{1}{2}$ provides a valid measure of bias under the permutation.

Second, in the Online Appendix we present results from exercises in which we apply our estimators to two types of simulated data. The first exercise is a Monte Carlo in which we simulate data from our estimated model. The second exercise is a falsification test in which we simulate data from a model in which $\mathbf{q}_t^R$ and $\mathbf{q}_t^D$ (and hence partisanship) are constant over time but verbosity m_{it} is allowed to follow its empirical distribution.

Third, we perform an out-of-sample validation in which our hypothetical observer learns the partisanship of phrases from one sample of speech and attempts to predict the party of speakers in another. In particular, we divide the sample of speakers into five mutually exclusive partitions. For each partition k and each estimator, we estimate the $\boldsymbol{\rho}_t(\mathbf{x})$ terms in equation (3) using the given estimator on the sample excluding the k^{th} partition, and the $\mathbf{q}_t^R(\mathbf{x})$ and $\mathbf{q}_t^D(\mathbf{x})$ terms using their empirical frequencies within the k^{th} partition. We then average the estimates across partitions and compare to our in-sample estimates.

5 Main Results

Figure 1 presents the time series of the maximum likelihood estimator $\hat{\pi}_t^{MLE}$ of our model, and of the index reported by Jensen et al. (2012) computed from their publicly available data.[17] Panel A shows 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

[17] Downloaded from http://www.brookings.edu/~/media/Projects/BPEA/Fall-2012/Jensen-Data.zip?la=en on March 25, 2016. 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* independently, use different processing algorithms, and use a vocabulary of three-word phrases rather than two-word phrases.

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 verbosities m_{it} of speakers 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 while the severity of the bias falls as we exclude less frequently spoken phrases, very severe sample restrictions are needed to control bias, and a significant time-varying bias remains even when we exclude 99 percent of phrases from our calculations.[18]

Panel B of Figure 1 shows that the Jensen et al. (2012) polarization measure behaves similarly to the MLE. The plot for the real series replicates the published version. The random series is far from 0, 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 earlier years. The results in Panel B suggest that the second part of this conclusion could be an artifact of the finite-sample mechanics of their index.

Figure 2 presents our main estimates. Panel A 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.

Panel B presents our preferred penalized estimator, including controls for covariates $\mathbf{x}_{it}$. Estimates for the random series indicate minimal bias. The Online Appendix shows that the use of regularization is the key to the performance of this estimator: imposing only a minimal penalty (i.e., setting $\boldsymbol{\lambda} \approx 0$) leads, as expected, to behavior similar to that of the MLE. The Online Appendix also shows that, in contrast to the MLE, the dynamics of our proposed estimators cannot be explained by changes in verbosity over time.

Looking at the data through the sharper lens of the leave-out and penalized estimators reveals that partisanship was low and relatively constant until the early 1990s, then exploded, reaching unprecedented heights in recent years. This is a dramatically different picture than one would infer from the MLE or the Jensen et al. (2012) series. The sharp increase in partisanship is much larger

[18] Across the sessions in our data, the 99th percentile phrase is spoken between 40 and 192 times per session.

than the width of the subsampling confidence intervals.

The increase is also large in magnitude. 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 3 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 expect to have a posterior of around 0.54 on the speaker's true party, only slightly above the prior of 0.5. By 1990, this value increased slightly to 0.57. Between 1990 and 2008, however, it leaped up to 0.73.

Figure 4 presents the out-of-sample validation exercise described in Section 4.5 for the MLE, leave-out, and penalized estimators. We find that the MLE greatly overstates partisanship relative to its out-of-sample counterpart. Based on the in-sample estimate one would expect an observer to be able to infer a speaker's party with considerable accuracy, but when tested out of sample the predictive power turns out to be vastly overstated. In contrast, both the leave-out and penalized estimators achieve values quite close to their out-of-sample counterparts, as desired.

In Figure 2, the penalized estimates in Panel B imply lower partisanship than the leave-out estimates in Panel A. Sampling experiments in the Online Appendix show that the bias in the leave-out estimator is slightly positive, likely due to excluding controls for covariates, and that the bias in the penalized estimator is negative, possibly due to conservative overpenalization.

The Online Appendix presents a range of alternative series based on variants of our baseline model, estimator, and sample. Removing covariates leads to greater estimated partisanship while adding more controls or speaker random effects leads to lower estimated partisanship, though all of these variants imply a large rise in partisanship following the 1990s. Dropping the South from the sample does not meaningfully change the estimates, nor does excluding data from early decades. Using only the early decades or holding constant the number of congresspeople in each session somewhat increases our estimates of partisanship and bias, leaving the difference between the real and random series in line with our preferred estimates.

6 Unpacking Partisanship

6.1 Partisan Phrases

Our model provides a natural way to define the partisanship of an individual phrase. For an observer with a neutral prior, the expected posterior that a speaker with characteristics $\mathbf{x}_{it}$ is Repub-

lican is $\frac{1}{2} = \frac{1}{2}\left(\mathbf{q}_t^R(\mathbf{x}_{it}) + \mathbf{q}_t^D(\mathbf{x}_{it})\right) \cdot \boldsymbol{\rho}_t(\mathbf{x}_{it})$. If, unbeknownst to the observer, phrase j is removed from the vocabulary, the change in the expected posterior is

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

We define the partisanship ζ_{jt} of phrase j in session t to be the average of this value across all active speakers i in session t. This measure has both direction and magnitude: positive numbers are Republican phrases, negative numbers are Democratic phrases, and the absolute value gives the magnitude of partisanship.

Table 1 lists the ten most partisan phrases in every tenth session plus the most recent session. 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. In the Online Appendix, we argue in detail that the top phrases in each of these sessions align closely with the policy positions and narrative strategies of the parties, confirming that our measure is indeed picking up partisanship rather than some other dimension that happens to be correlated with it. In this section, we highlight a few illustrative examples.

The 50th session of Congress (1887-88) occurred in a period where the cleavages of the Civil War and Reconstruction Era were still fresh. Republican phrases like "union soldier" and "confeder soldier" relate to the ongoing debate over provision for veterans, echoing the 1888 Republican platform's commitment to show "[the] gratitude of the Nation to the defenders of the Union." The Republican phrase "color men" reflects the ongoing importance of racial issues. Many Democratic phrases from this Congress ("increase duti," "ad valorem," "high protect," "tariff tax," "high tariff") reflect a debate over reductions in trade barriers. The 1888 Democratic platform endorses tariff reduction in its first sentence, whereas the Republican platform says Republicans are "uncompromisingly in favor of the American system of protection."

The 80th session (1947-1948) convened in the wake of the Second World War. Many Republican-leaning phrases relate to the war and national defense ("arm forc," "air forc," "coast guard," "stop communism," "foreign countri"), whereas "unit nation" is the only foreign-policy-related phrase in the top ten Democratic phrases in the 80th session. 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," "depart labor," "collect bargain," "concili servic").[19] By

[19]The Federal Mediation and Conciliation Service was created in 1947 and was "given the mission of preventing

contrast, the Republican platform of the same year does not mention the Fair Labor Standards Act or the minimum wage.

Language in the 110th session (2007-2008) follows familiar partisan divides. Republicans focus on taxes ("tax increas," "rais tax," "tax rate") and immigration ("illeg immigr"), while Democrats focus on the aftermath of the war in Iraq ("war iraq", "troop iraq") and social domestic policy ("african american," "children health," "middl class"). With regards to energy policy, Republicans focus on the potential of American energy ("natural gas," "american energi," "outer continent," "continent shelf"), while Democrats focus on the role of oil companies ("oil compani").

The phrases from the 114th session (2015-2016) relate to current partisan cleavages and echo themes in the 2016 presidential election. Republicans focus on terrorism, discussing "al qaeda" and using the phrase "radic islam," which echoes Donald Trump's use of the phrase "radical Islamic terrorism" during the campaign (Holley 2017). Democrats focus on climate change ("climat chang"), civil rights issues ("african american," "vote right"), and gun control ("gun violenc"). When discussing public health, Republicans focus on mental health ("mental health") in correspondence to the Republican-sponsored "Helping Familes in Mental Health Crisis Act of 2016," while Democrats focus on public health more broadly ("public health"), health insurance ("afford care"), and women's health ("plan parenthood").

6.2 Partisanship within and between 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. Knowing 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 knowing whether the speech is about taxes or about the environment. To separate these, we present a decomposition of partisanship into within- and between-topic components using our 22 manually defined topics.

We define between-topic partisanship to be the posterior that 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 vocabulary 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

or minimizing the impact of labor-management disputes on the free flow of commerce by providing mediation, conciliation and voluntary arbitration" (see https://www.fmcs.gov/aboutus/our-history/ accessed on April 15, 2017).

by its frequency of occurrence.

Figure 5 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 phrase within a topic.

Figure 6 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 immigration, crime, and religion topics. Other topics with large increases include taxes, environmental policy, and minorities. Not all topics have become increasingly partisan in recent years. For example, alcohol was fairly partisan in the Prohibition Era but is not especially partisan today. Figure 6 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 the 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, Figure 7 shows the evolution of the partisanship of the four most Republican and Democratic phrases in the "tax," "immigration," and "labor" topics. The plots show that the most partisan phrases become more informative about a speaker's party over time. Some phrases, such as "american taxpay," have been associated with one party since the 1950s. Others, like "tax relief" and "minimum wage," switch between parties before becoming strongly informative about one party during the 1990s and 2000s. A third group, including "immigr reform" and "job creator," is partisan only for a short period when it is relevant to congressional debate. The Online Appendix presents similar plots for the other 19 topics.

7 Discussion

What are we to make of the dramatic increase in the partisanship of speech? The pattern we observe suggests our language-based measure captures something quite different from ideological polarization as usually defined. In Figure 8, we compare our speech-based measure of partisanship to the standard measure of ideological polarization based on roll-call votes (Carroll et al. 2015a). 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 nine-

teenth and early twentieth century as it is today, and its current upward trend begins around 1950 rather than 1990. This finding reinforces our expectation that speech and roll-call votes respond to different incentives and constraints. Roll-call votes may be shaped by strategic considerations related to the passage of legislation, and may therefore not reflect legislators' sincere policy preferences. Speech may reflect party differences in values, goals, or persuasive tactics that are distinct from positions on specific pieces of legislation. And, related to our discussion below, speech may reflect innovations in rhetoric that have no counterpart in roll-call votes.

Panel B of Figure 8 shows that a measure of the Republican-ness of an individual's speech from our model and the individual Common Space DW-NOMINATE scores from the roll-call voting data are positively correlated in the cross section. Across all sessions, the correlation between speech and roll-call based partisanship measures is 0.537 ($p = 0.000$). After controlling for party, the correlation is 0.129 and remains highly statistically significant ($p = 0.000$).[20] Thus, members who vote more conservatively also use more conservative language on average, even though the time-series dynamics of voting and speech are very different. As another way to validate this relationship, we show in the Online Appendix that average partisanship exhibits a discontinuity in vote margin analogous to the discontinuity in vote margin of the non-Common-Space DW-NOMINATE scores (Lee et al. 2004; Carroll et al. 2015b). The Online Appendix also shows that the divergence in speech between parties in recent years is not matched by an equally large divergence in speech between the more moderate and more extreme wings within each party.

What caused the dramatic increase in the partisanship of speech beginning in the 1990s? We cannot provide a definitive answer, but the timing of the change shown in Panel A of Figure 9 suggests two natural 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 around the 104th session (1995-1996), the first following the 1994 midterm 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

[20] These correlations are 0.685 ($p = 0.000$) and 0.212 ($p = 0.000$), respectively, when we use data only on speakers who speak an average of at least 1000 phrases across the sessions in which they speak.

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 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).[21] Important technological advances used by Luntz included instant feedback "dials" that allowed focus group participants to respond to the content they were hearing in real time.[22] 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).

A 2006 memorandum written by Luntz and distributed to Republican congressional candidates provides detailed advice on the 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 language in the *Contract with America*. 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. Panel B of Figure 9 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.[23]

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 Democrats, writes:

[21] 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). A memo called "Language: A Key Mechanism of Control" circulated in 1994 to Republican candidates under a cover letter from Gingrich stating that the memo contained "tested language from a recent series of focus groups" (GOPAC 1994).

[22] 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).

[23] According to the metric defined in Table 1, the most Republican phrases in the 104th session (1995-1996) that appear in the *Contract* are "american peopl," "tax increas," "term limit," "lineitem veto," "tax relief," "save account," "creat job," "tax credit," "wast fraud," and "fiscal respons." 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.

"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).

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). Many scholars of the US Congress find that, over the last few decades, the two parties have increasingly aimed to have a disciplined and centralized strategy for public communication (Sinclair 2006; Malecha and Reagan 2012; Lee 2016a). A quantitative signal of this trend, displayed in Panel A of Figure 9, is the increasing fraction of Congressional leadership staff dedicated to communications roles, a fact that Lee (2016a) attributes in part to majority control of the chambers becoming more contested.

Consistent with a trend towards greater party discipline in language, the Online Appendix shows that the recent increase in partisanship is concentrated in a small minority of highly partisan phrases. The figure plots quantiles of the estimated average value of the partisanship of all individual phrases in each session. The plot shows a marked increase in the partisanship of the highest quantiles, while even the quantiles at 0.9 and 0.99 remain relatively flat.

In a similar vein, the Online Appendix shows that a vocabulary consisting of neologisms—which we define to be phrases first spoken in our data after 1980 (the 96th session)—exhibits very high and sharply rising partisanship. The figure also shows that a large increase in partisanship remains even when we exclude neologisms from the choice set.

Changes in the media environment may also have contributed to the increase in partisanship.[24] Prior to the late 1970s, television cameras were only allowed on the floor of Congress for special hearings and 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' advertising. This plausibly increased the return to carefully crafted language, both by widening the reach of success-

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

ful sound bites, and by dialing up the cost of careless mistakes.[25] The subsequent introduction of the Fox News cable network and the increasing partisanship of cable news more generally (Martin and Yurukoglu 2017) may have further increased this return.

The timing shown in Figure 9 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, Gingrich . . . [replied] 'No' . . . C-SPAN provided a group of media-savvy House conservatives in the mid-1980s with a method of . . . winning a prime-time audience."

The hypothesis that technological change strengthened the incentive for party discipline in language offers a possible explanation for the very different dynamics of inter-party differences in speech and in roll-call voting exhibited in Figure 8.

8 Conclusion

A consistent theme of much prior literature is that political partisanship today—both in Congress and among voters—is not that different from what existed in the past (Glaeser and Ward 2006; Fiorina and Abrams 2008; McCarty et al. 2015). 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; Greenstein and Zhu 2012; Jensen et al. 2012; Martin and Yurukoglu 2017) 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

[25] Mixon et al. (2001) and Mixon et al. (2003) provide evidence that the introduction of C-SPAN changed the nature of legislative debate.

makes sense if language matters. Field studies reveal effects of language on outcomes including marriage (Caminal and Di Paolo 2019), 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 language 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 propose 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. To illustrate the range of such settings in the political domain, the Online Appendix uses survey data to characterize the bias in plug-in estimates of the partisanship of respondents' choice of residential location, websites, and television programs, for various sample sizes.

References

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

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

Amer, M. L. (1993): The Congressional Record; content, history, and issues. Washington DC: Congressional Research Service. CRS Report No. 93-60 GOV.

Andrews, W., and L. 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> on June 24, 2016.

Antweiler, W., and M. 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, M. (2005): The framing wars. *New York Times*, July 17, 2005. Accessed at <http://www.nytimes.com/2005/07/17/magazine/the-framing-wars.html> on June 16, 2016.

Baker, S. G. (1994): The multinomial-Poisson transformation. *The Statistician*: 495-504.

Ball, M. (2013): Don't call it 'gun control'. *The Atlantic*, January 16, 2013. Accessed at <https://www.theatlantic.com/politics/archive/2013/01/dont-call-it-gun-control/267259/> on July 23, 2018.

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

Bayer, P., R. McMillan, and K. 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.

Bertrand, M., and E. Kamenica (2018): Coming apart? Cultural distances in the United States over time. University of Chicago mimeo.

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

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

Caminal, R., and A. Di Paolo (2019): Your language or mine? The noncommunicative benefits of language skills. *Economic Inquiry* 57(1): 726-750.

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

Carroll, R., J. Lewis, J. Lo, N. McCarty, K. Poole, and H. Rosenthal (2015a): "Common Space"

DW-NOMINATE scores with bootstrapped standard errors. Accessed at <http://www.voteview.com/dwnomin_joint_house_and_senate.htm> on November 18, 2016.

— (2015b): DW-NOMINATE scores with bootstrapped standard errors. Accessed at <http://www.voteview.com/dwnomin.htm> on March 30, 2017.

Chandra, A., D. Cutler, and Z. Song (2012): "Who ordered that? The economics of treatment choices in medical care," in *Handbook of Health Economics*, vol. 2, Mark V. Pauly, Thomas G. McGuire, and Pedro P. Barros, eds. (Amsterdam: Elsevier).

Chatterjee, A., and S. N. Lahiri (2011): Bootstrapping lasso estimators. *Journal of the American Statistical Association* 106(494): 608-625.

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

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

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

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

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

D'Haultfœuille, X., and R. Rathelot (2017): Measuring segregation on small units: A partial identification analysis. *Quantitative Economics* 8: 39-73.

Davis, D., J. I. Dingel, J. Monras, and E. Morales (Forthcoming): How segregated is urban consumption? *Journal of Political Economy*.

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

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

Economist (2013): The war of the words. *The Economist*, July 13, 2013. Accessed at <https://www.economist.com/united-states/2013/07/13/the-war-of-the-words> on July 23, 2018.

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

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

Flaxman, S., S. Goel, and J. Rao (2016): Filter bubbles, echo chambers, and online news consumption. *Public Opinion Quarterly*, 80: 298-320.

Fossett, M. (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, D. M., and O. Volij (2011): Measuring school segregation. *Journal of Economic Theory* 146(1): 1–38.

Frantzich, S., and J. Sullivan (1996): *The C-SPAN revolution*. Norman OK: University of Oklahoma Press.

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

— (2011): Ideological segregation online and offline. *Quarterly Journal of Economics* 126(4): 1799–1839.

Gentzkow, M., J. M. Shapiro, and M. Taddy (2017): Measuring polarization in high-dimensional data: Method and application to congressional speech. NBER Working Paper No. 22423.

Gentzkow, M., B. T. Kelly, and M. Taddy (Forthcoming): Text as data. *Journal of Economic Literature.*

Groseclose, T., and J. Milyo (2005): A measure of media bias. *Quarterly Journal of Economics* 120(4): 1191-1237.

Gingrich, N., and D. Armey (1994): *Contract with America.*

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

GOPAC (1994): Language: A key mechanism of control. Memo. Accessed at <http://fair.org/extra/language-a-key-mechanism-of-control/> on March 30, 2017.

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

Greenstein, S., and F. Zhu (2012): Is Wikipedia biased? *American Economic Review: Papers and Proceedings.* 102(3): 343-348.

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

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

Haas, K. L. (2015): Rules of the House of Representatives — one hundred fourteenth Congress. Accessed at <http://clerk.house.gov/legislative/house-rules.pdf> on March 1, 2017.

Harris, D. 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.

Holley, P. (2017): 'Radical Islamic terrorism': Three words that separate Trump from most of Washington. *Washington Post*, March 1, 2017. Accessed at

<https://www.washingtonpost.com/news/the-fix/wp/2017/02/28/radical-islamic-terrorism-three-words-that-separate-trump-from-most-of-washington/?utm term=.055e7d927bcf> on May 15, 2017.

Issenberg, S. (2012): The death of the hunch. *Slate*, May 22, 2012. Accessed at consistently <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.

Inter-university Consortium for Political and Social Research (ICPSR) and C. McKibbin (1997): Roster of United States congressional officeholders and biographical characteristics of members of the United States Congress, 1789-1996: merged data. ICPSR07803-v10. *ICPSR Study No. 7803*. Accessed at <http://www.icpsr.umich.edu/icpsrweb/ICPSR/studies/7803> on March 1, 2017.

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

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

Jensen, J., S. Naidu, E. Kaplan, and L. 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, C. W. (1997): House rules and manual—one hundred fifth Congress. Washington DC: U.S. Government Printing Office. H. Doc. no. 104-272.

Johnson, D. W. (2015): *Political consultants and American elections: Hired to fight, hired to win.* Routledge.

Kim, I. S., J, Londregan, and M. Ratkovic (2018): Estimating ideal points from votes and text. *Political Analysis* 26(2): 210-229.

King, G. (1995): Elections to the United States House of Representatives, 1898-1992. ICPSR version. Inter-university Consortium for Political and Social Research. Accessed at <http://www.icpsr.umich.edu/icpsrweb/ICPSR/studies/6311> on April 7, 2017.

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

Lakoff, G. (2003): Framing the issues: UC Berkeley professor George Lakoff tells how conservatives use language to dominate politics. *UC Berkeley News*, October 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.

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

Lauderdale, B. E., and A. Herzog (2016): Measuring political positions from legislative speech. *Political Analysis* 24(3): 374-394.

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

Lee, D. S., E. Moretti, and M. J. Butler (2004): Do voters affect or elect policies? Evidence from the U.S. House. *Quarterly Journal of Economics*. 119(3): 807–859.

Lee, F. (2016a): Legislative parties in an era of alternating majorities. Chapter. In *Governing in a Polarized Age: Elections, Parties, and Political Representation in America,* edited by Alan S. Gerber and Eric Schickler, 115-42. Cambridge: Cambridge University Press.

— (2016b): *Insecure Majorities: Congress and the Perpetual Campaign.* Chicago, IL: University of Chicago Press.

Logan, J. R., A. Foster, J. Ke, and F. Li (2018): The uptick in income segregation: Real trend or random sampling variation? *American Journal of Sociology* 124(1): 185-222.

Luntz, F. 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.

Malecha, G. L., and D. J. Reagan (2012): *The Public Congress: Congressional Deliberation in a New Media Age*. New York, NY: Routledge.

Martin, G. J., and A. Yurukoglu (2017): Bias in cable news: Persuasion and polarization. *American Economic Review* 107(9): 2565-2599.

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

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

McCarty, N., K. Poole, and H. 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, A. (2013): Poisson indices of segregation. *Regional Science and Urban Economics* 43(1): 65–85.

— (2017): A structural model of segregation in social networks. *Econometrica* 85(3): 825-850.

Michel, R. H. (1993): The theme team. Accessed at <http://www.robertmichel.name/RHM_blueprint/blueprint_Theme%20Team.pdf> on June 16,

2016.

Mixon, F. G., Jr., D. L. Hobson, and K. P. Upadhyaya (2001): Gavel-to-gavel congressional television coverage as political advertising: The impact of C-SPAN on legislative sessions. *Economic Inquiry* 39(3): 351-364.

Mixon, F. G., Jr., M. T. Gibson, and K. P. Upadhyaya (2003): Has legislative television changed legislator behavior? C-SPAN2 and the frequency of Senate filibustering. *Public Choice* 115(1): 139-162.

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

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

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

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

Peterson, A., and A. Spirling (2018): Classification accuracy as a substantive quantity of interest: Measuring polarization in Westminster systems. *Political Analysis* 26: 120-128.

Politis, D. N., J. P. Romano, and M. Wolf (1999): *Subsampling*. New York: Springer series in statistics.

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

Porter, M. (2009): The English (Porter2) stemming algorithm. Accessed at <http://snowball.tartarus.org/algorithms/english/stemmer.html> on March 31, 2017.

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

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

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

Sinclair, B. (2006): *Party Wars: Polarization and the Politics of National Policy Making*. Norman, OK: University of Oklahoma Press.

Taddy, M. (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.

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

Tibshirani, R. (1996): Regression shrinkage and selection via the lasso. *Journal of the Royal*

Statistical Society, Series B (Methodological) 58(1): 267–288.

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

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

Whitmarsh, L. (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.

Yan, H., S. Das, A. Lavoie, S. Li, and B. Sinclair (2018): The congressional classification challenge: Domain specificity and partisan intensity. Washington University of St. Louis Working Paper. Accessed at <https://www.cse.wustl.edu/∼sanmay/papers/partisanship-from-text.pdf> on January 11, 2019.

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

Table 1: Most Partisan Phrases by Session

Session 50 (1887-1888)

Republican	#R	#D	Democratic	#R	#D
sixth street	22	0	cutleri compani	0	72
union soldier	33	13	labor cost	11	37
color men	27	10	increas duti	11	34
railroad compani	85	70	cent ad	35	54
great britain	121	107	public domain	20	39
confeder soldier	18	4	ad valorem	61	78
other citizen	13	0	feder court	11	25
much get	12	1	high protect	6	18
paper claim	9	0	tariff tax	11	23
sugar trust	16	7	high tariff	6	16

Session 60 (1907-1908)

Republican	#R	#D	Democratic	#R	#D
postal save	39	3	canal zone	18	66
census offic	31	2	also petit	0	47
reserv balanc	36	12	standard oil	4	25
war depart	62	39	indirect contempt	0	19
secretari navi	62	39	bureau corpor	5	24
secretari agricultur	58	36	panama canal	23	41
pay pension	20	2	nation govern	12	30
boat compani	24	8	coal mine	9	27
twelfth census	14	0	revis tariff	8	26
forestri servic	20	7	feet lake	0	17

Session 70 (1927-1928)

Republican	#R	#D	Democratic	#R	#D
war depart	97	63	pension also	0	163
take care	105	72	american peopl	51	91
foreign countri	54	28	radio commiss	8	44
muscl shoal	97	71	spoken drama	0	30
steam plant	25	3	civil war	27	54
nation guard	39	18	trade commiss	19	46
air corp	32	12	feder trade	19	45
creek dam	25	6	wave length	6	25
cove creek	30	13	imperi valley	12	28
american ship	29	12	flowag right	5	20

Session 80 (1947-1948)

Republican	#R	#D	Democratic	#R	#D
depart agricultur	67	31	unit nation	119	183
foreign countri	49	22	calumet region	0	30
steam plant	34	7	concili servic	3	31
coast guard	34	9	labor standard	16	41
state depart	117	93	depart labor	24	46
air forc	88	69	collect bargain	15	35
stop communism	22	3	standard act	11	31
nation debt	43	25	polish peopl	4	20
pay roll	34	17	budget estim	22	38
arm forc	63	47	employ servic	25	41

Session 90 (1967-1968)

Republican	#R	#D	Democratic	#R	#D
job corp	35	20	human right	7	44
trust fund	26	14	unit nation	49	75
antelop island	11	0	men women	20	34
treasuri depart	23	12	world war	57	71
federalaid highway	13	2	feder reserv	26	39
tax credit	21	11	million american	15	27
state depart	45	35	arm forc	25	37
oblig author	14	4	high school	19	30
highway program	14	4	gun control	10	22
invest act	11	1	air pollut	18	29

Session 100 (1987-1988)

Republican	#R	#D	Democratic	#R	#D
judg bork	226	14	persian gulf	30	47
freedom fighter	36	8	contra aid	12	28
state depart	59	35	star war	1	14
human right	101	78	central american	17	30
minimum wage	37	19	aid contra	17	30
reserv object	23	8	nuclear wast	14	27
demand second	13	1	american peopl	97	109
tax increas	20	10	interest rate	24	35
pay rais	21	11	presid budget	11	21
plant close	37	28	feder reserv	12	22

Session 110 (2007-2008)

Republican	#R	#D	Democratic	#R	#D
tax increas	87	20	dog coalit	0	90
natur gas	77	20	war iraq	18	78
reserv balanc	147	105	african american	6	62
rais tax	44	10	american peopl	230	278
american energi	34	3	oil compani	20	65
illeg immigr	34	7	civil war	17	45
side aisl	132	106	troop iraq	11	39
continent shelf	33	8	children health	17	42
outer continent	32	8	nobid contract	0	24
tax rate	26	4	middl class	15	39

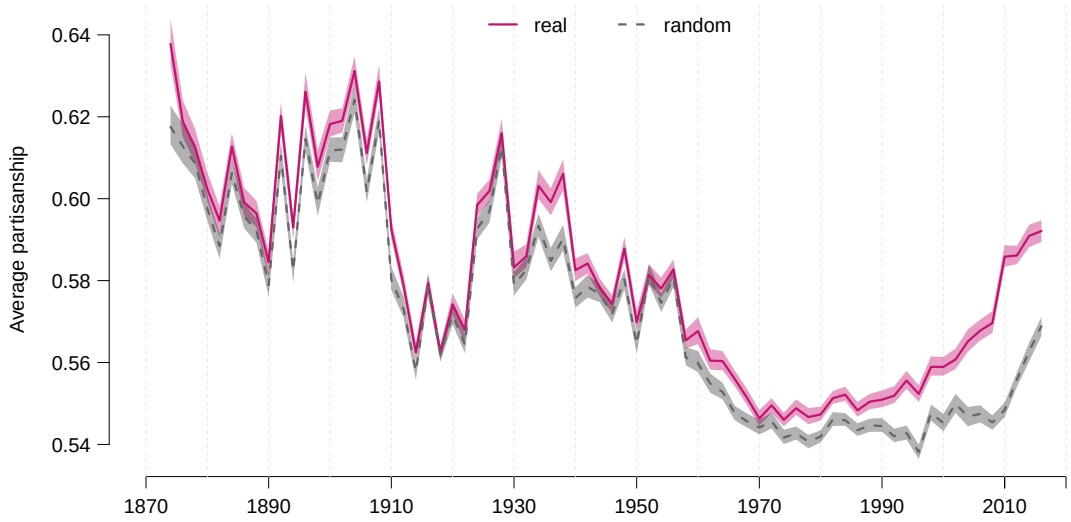
Session 114 (2015-2016)

Republican	#R	#D	Democratic	#R	#D
american peopl	327	205	homeland secur	96	205
al qaeda	50	7	climat chang	23	94
men women	123	83	gun violenc	3	74
side aisl	133	93	african american	11	71
human traffick	60	26	vote right	2	62
colleagu support	123	89	public health	24	83
religi freedom	34	4	depart homeland	48	93
taxpay dollar	47	19	plan parenthood	66	104
mental health	59	32	afford care	40	77
radic islam	22	0	puerto rico	42	79

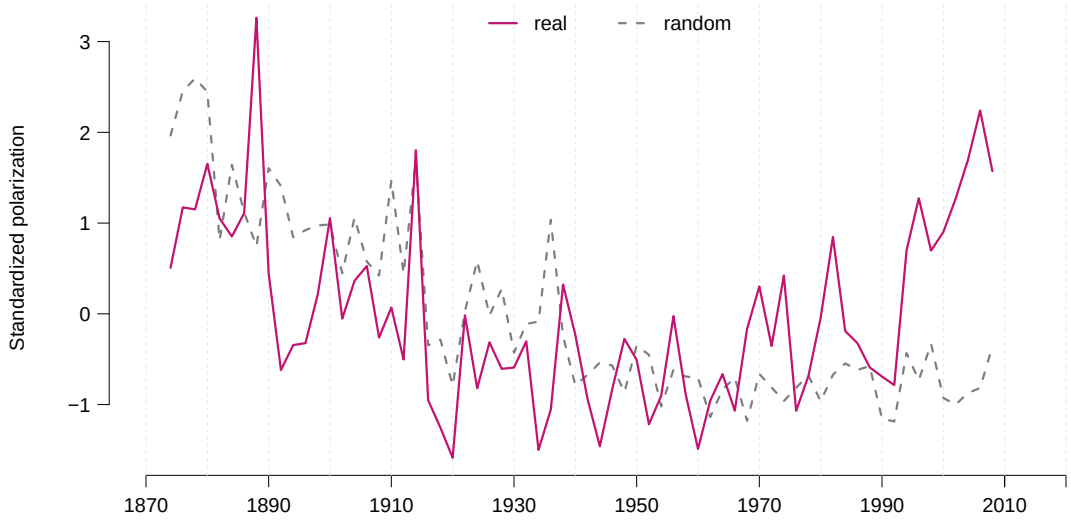
Notes: Calculations are based on our preferred specification in Panel B of Figure 2. The table shows the Republican and Democratic phrases with the greatest magnitude of estimated partisanship ζ_{jt}, as defined in Section 6.1, alongside the predicted number of occurrences of each phrase per 100,000 phrases spoken by Republicans or Democrats. Phrases with positive values of ζ_{jt} are listed as Republican and those with negative values are listed as Democratic.

Figure 1: Average Partisanship and Polarization of Speech, Plug-in Estimates

Panel A: Partisanship from Maximum Likelihood Estimator ($\hat{\pi}_t^{MLE}$)

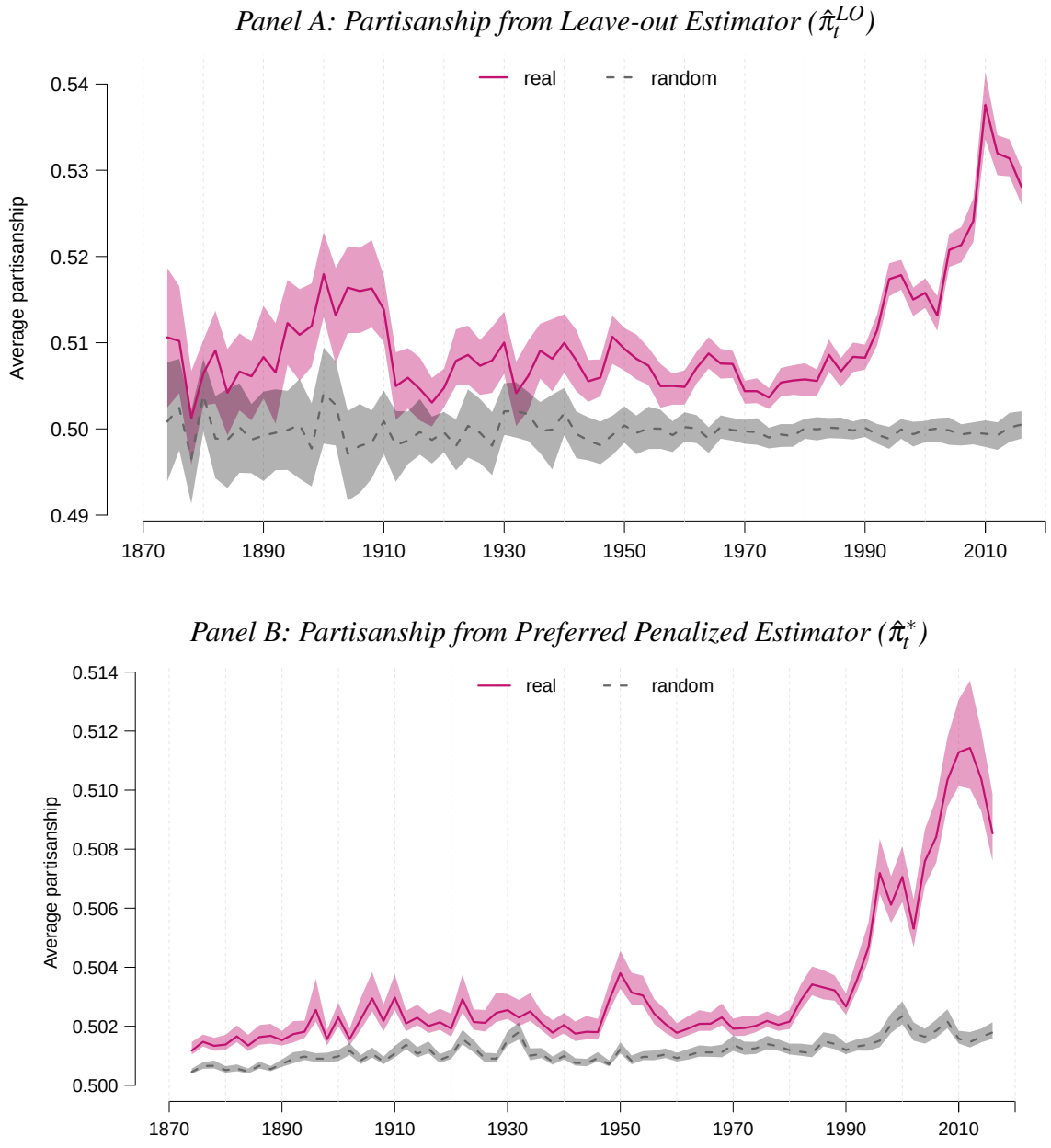


Panel B: Polarization from Jensen et al. (2012)



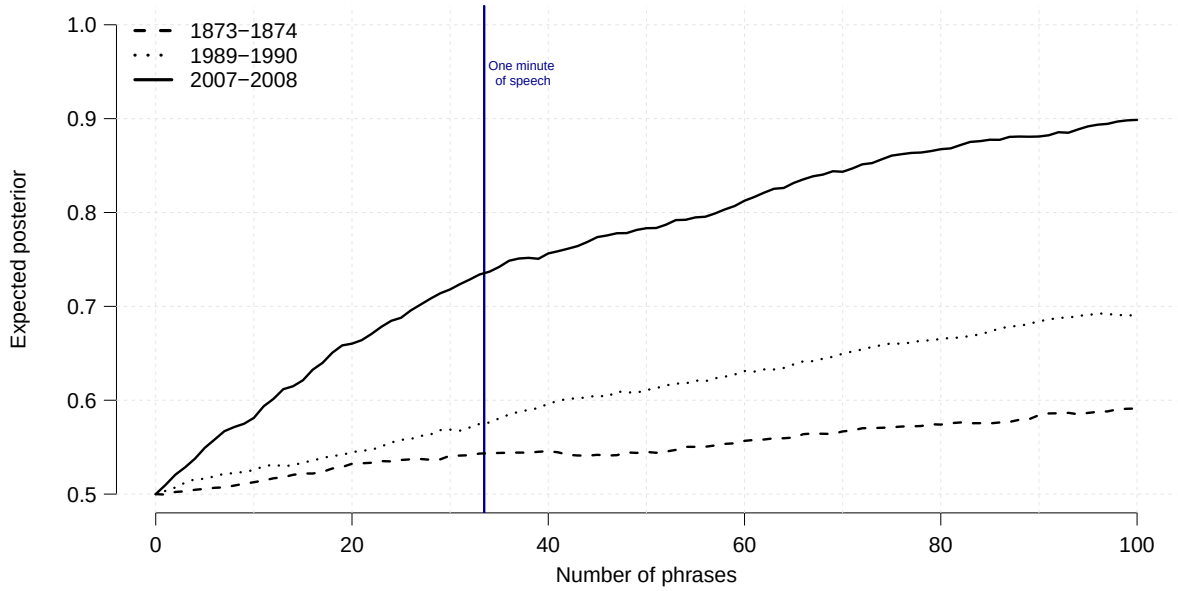
Notes: Panel A plots the average partisanship series from the maximum likelihood estimator $\hat{\pi}_t^{MLE}$ defined in Section 4.1. "Real" series is from actual data; "random" series is from hypothetical data in which each speaker's party is randomly assigned with the probability that the speaker is Republican equal to the average share of speakers who are Republican in the sessions in which the speaker is active. The shaded region around each series represents a pointwise confidence interval obtained via subsampling (Politis et al. 1999). Specifically, we randomly draw speakers without replacement to create 100 subsamples each containing (up to integer restrictions) one-tenth of all speakers and, for each subsample k, we compute the MLE estimate $\hat{\pi}_t^k$. Let τ_k be the number of speakers in the kth subsample and let τ be the number of speakers in the full sample. Then the confidence interval on the MLE is $\frac{1}{2} + (\exp[\log(\hat{\pi}_t - \frac{1}{2}) - (Q_t^k)_{(90)}/\sqrt{\tau}], \exp[\log(\hat{\pi}_t - \frac{1}{2}) - (Q_t^k)_{(11)}/\sqrt{\tau}])$, where $(Q_t^k)_{(b)}$ is the bth order statistic of $Q_t^k = \sqrt{\tau_k}\left(\log(\hat{\pi}_t^k - \frac{1}{2}) - \log([\frac{1}{100}\sum_{l=1}^{100}\hat{\pi}_t^l] - \frac{1}{2})\right)$. Panel B plots the standardized measure of polarization from Jensen et al. (2012). Polarization in session t is defined as $\sum_j \left(m_{jt} \left| \rho_{jt} \right| \Big/ \sum_l m_{lt} \right)$ where $\rho_{jt} = corr\left(c_{ijt}, \mathbf{1}_{i \in R_t}\right)$; 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 and Penalized Estimates



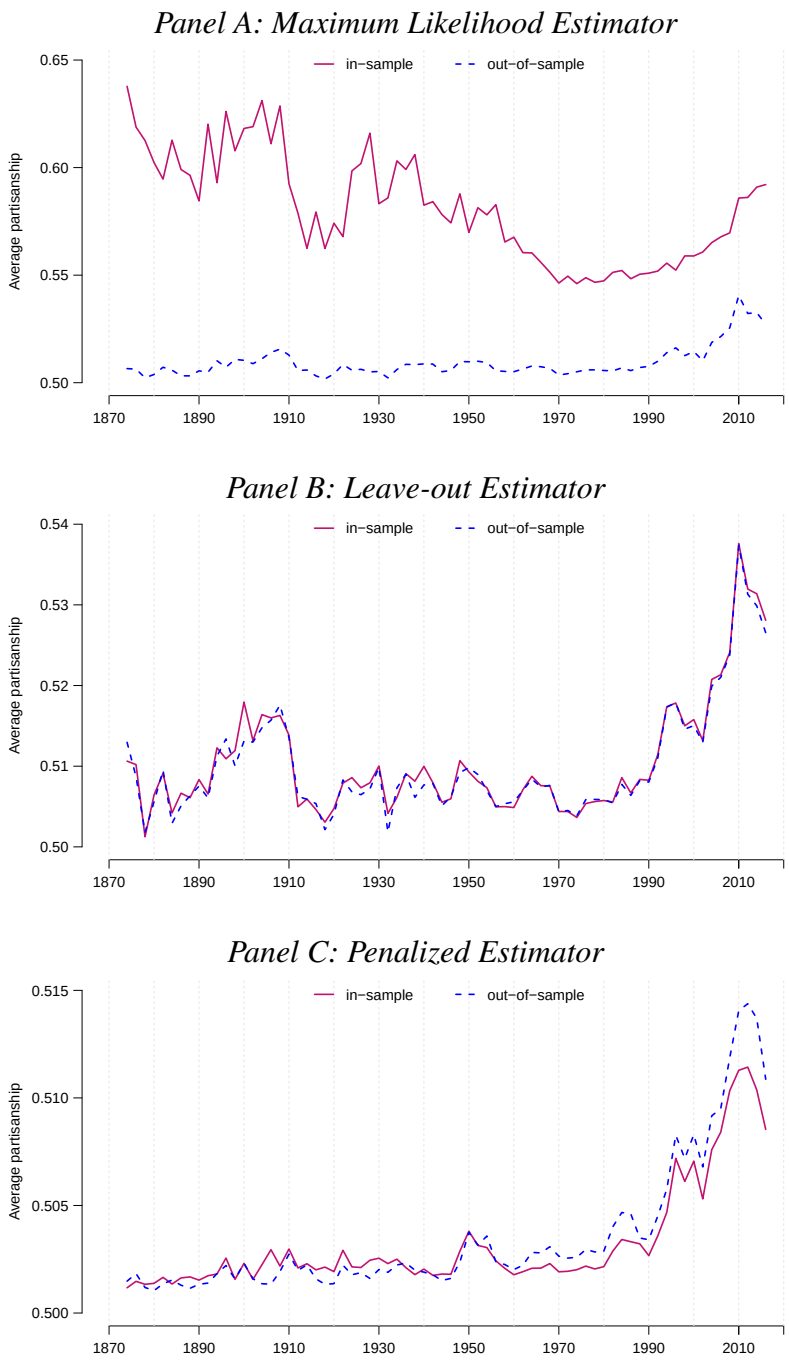
Notes: Panel A plots the average partisanship series from the leave-out estimator $\hat{\pi}_t^{LO}$ defined in Section 4.2. Panel B plots the average partisanship series from our preferred penalized estimator $\hat{\pi}_t^{*}$ defined in Section 4.3. In each plot, the "real" series is from actual data and the "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 via subsampling (Politis et al. 1999). Specifically, we randomly draw speakers without replacement to create 100 subsamples each containing (up to integer restrictions) one-tenth of all speakers and, for each subsample k, we compute the leave-out estimate $\hat{\pi}_t^k$ and the penalized estimate $\hat{\pi}_t^{*k}$. Let τ_k be the number of speakers in the kth subsample and let τ be the number of speakers in the full sample. Then the confidence interval on the leave-out estimator is $(\hat{\pi}_t^{LO} - (Q_t^k)_{(90)}/\sqrt{\tau}, \hat{\pi}_t^{LO} - (Q_t^k)_{(11)}/\sqrt{\tau})$, where $(Q_t^k)_{(b)}$ is the bth order statistic of $Q_t^k = \sqrt{\tau_k}\left(\hat{\pi}_t^k - \frac{1}{100}\sum_{l=1}^{100}\hat{\pi}_t^l\right)$. The confidence interval on the penalized estimator is $\frac{1}{2} + (\exp[\log(\hat{\pi}_t^* - \frac{1}{2}) - (Q_t^{*k})_{(90)}/\sqrt{\tau}], \exp[\log(\hat{\pi}_t^* - \frac{1}{2}) - (Q_t^{*k})_{(11)}/\sqrt{\tau}])$ where $(Q_t^{*k})_{(b)}$ is the bth order statistic of $Q_t^{*k} = \sqrt{\tau_k}\left(\log(\hat{\pi}_t^{*k} - \frac{1}{2}) - \log([\frac{1}{100}\sum_{l=1}^{100}\hat{\pi}_t^{*l}] - \frac{1}{2})\right)$.

Figure 3: Informativeness of Speech by Speech Length and Session



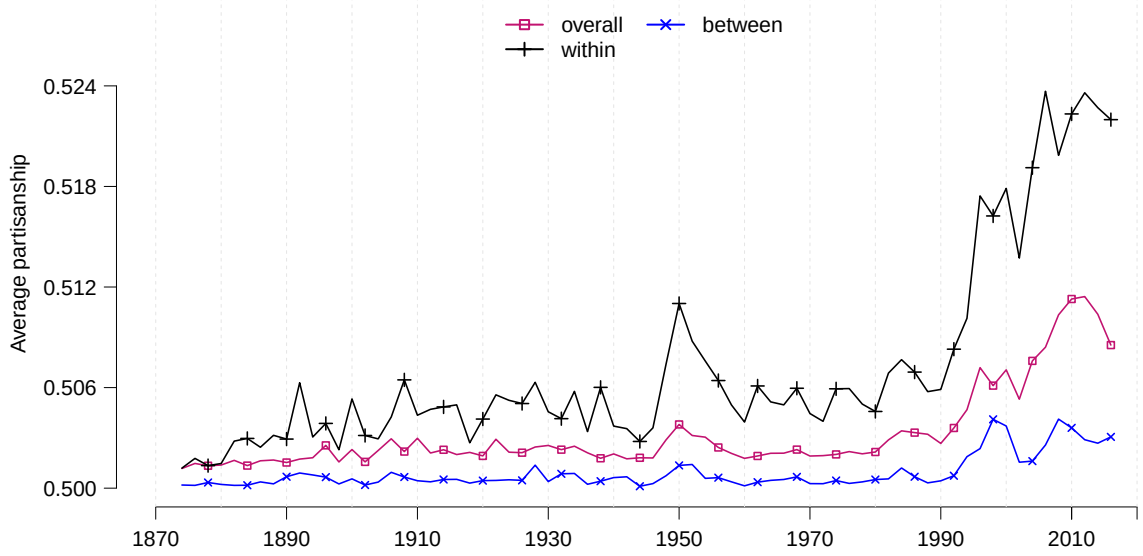
Notes: For each speaker i and session t we calculate, given characteristics $\mathbf{x}_{it}$, the expected posterior that an observer with a neutral prior would place on a speaker's true party after hearing a given number of phrases drawn according to our preferred specification in Panel B of Figure 2. We perform this calculation by Monte Carlo simulation and plot the average across speakers for each given session and length of speech. The vertical line shows the average number of phrases in one minute of speech. We calculate this by sampling 95 morning-hour debate speeches across the 2nd session of the 111th Congress and the 1st session of the 114th Congress. We use https://www.c-span.org/ to calculate the time-length of each speech and to obtain the text of the *Congressional Record* associated with each speech, from which we obtain the count of phrases in our main vocabulary following the procedure outlined in Section 2. The vertical line shows the average ratio, across speeches, of the phrase count to the number of minutes of speech.

Figure 4: Out-of-sample Validation



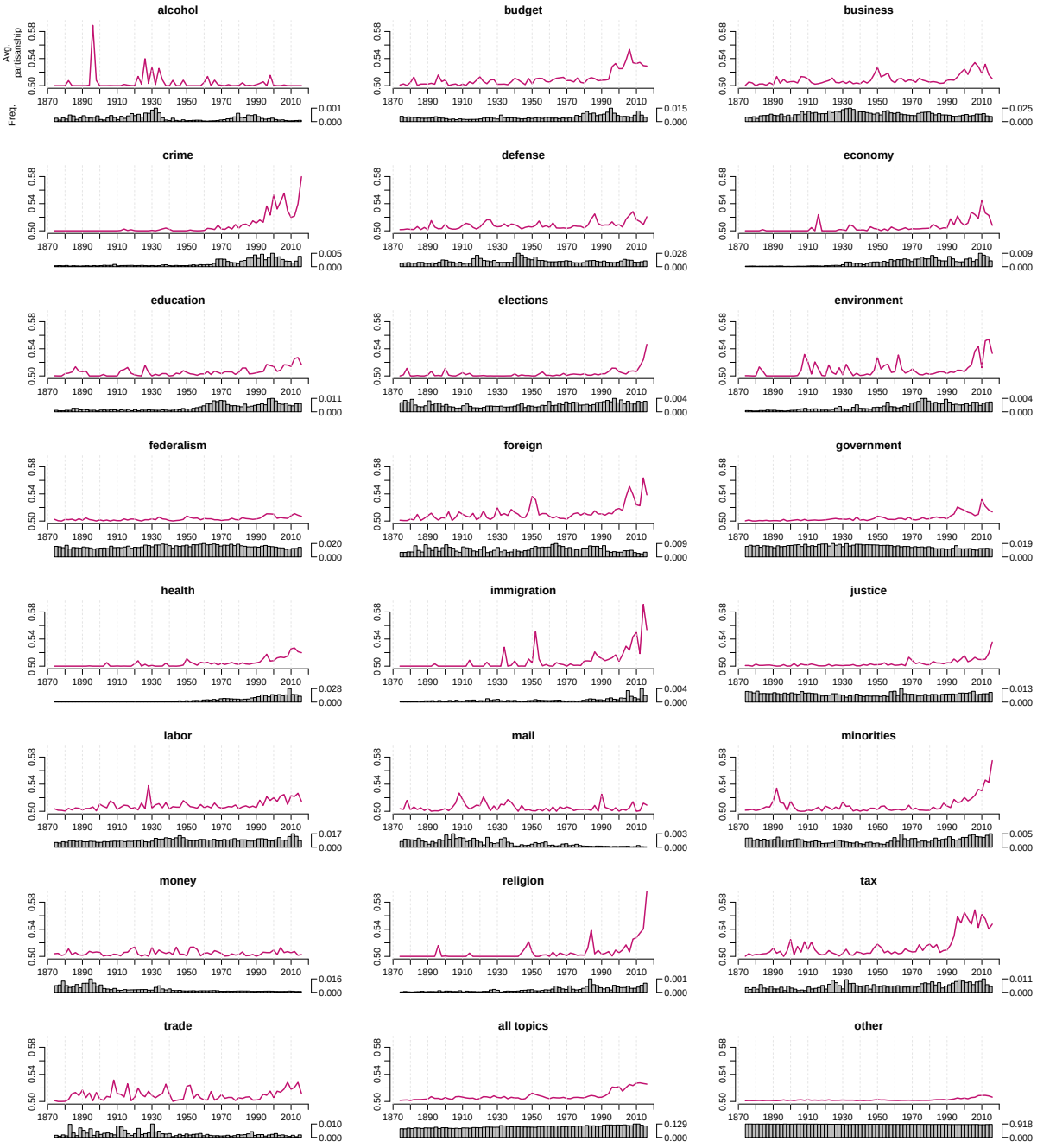
Notes: Let $\hat{\boldsymbol{\rho}}_t(\mathcal{S})$, $\hat{\mathbf{q}}_t^R(\mathcal{S})$, and $\hat{\mathbf{q}}_t^D(\mathcal{S})$ be functions estimated using the maximum likelihood estimator on a sample of speakers $\mathcal{S}$. Let $\hat{\boldsymbol{\rho}}_t^*(\mathcal{S})$, $\hat{\mathbf{q}}_t^{*R}(\mathcal{S})$, and $\hat{\mathbf{q}}_t^{*D}(\mathcal{S})$ be functions estimated using our preferred penalized estimator on sample $\mathcal{S}$ and evaluated at the sample mean of the covariates in session t and sample $\mathcal{S}$. Let $\overline{\mathcal{S}} = \cup_t (R_t \cup D_t)$ be the full sample of speakers and let $\overline{\mathcal{S}}_k$ for $k = 1,...,K$ denote $K = 5$ mutually exclusive partitions ("folds") of $\overline{\mathcal{S}}$, with $\overline{\mathcal{S}}_{-k} = \overline{\mathcal{S}} \setminus \overline{\mathcal{S}}_k$ denoting the sample excluding the k^{th} fold. For $P \in \{R,D\}$, denote $P_{k,t} = \overline{\mathcal{S}}_k \cap P_t$ and $\hat{\mathbf{q}}_t^P(\overline{\mathcal{S}}_k) = \frac{1}{|P_{k,t}|} \sum_{i \in P_{k,t}} \hat{\mathbf{q}}_{i,t}(\overline{\mathcal{S}}_k)$. The lines labeled "in-sample" in Panels A, B, and C present the in-sample estimated partisanship using the maximum likelihood estimator, leave-out estimator, and our preferred penalized estimator. These are the same as in Figure 1 and Figure 2. The line labeled "out-of-sample" in Panel A presents the average, across folds, of the out-of-sample estimated partisanship using the maximum likelihood estimator: $\frac{1}{K}\sum_{k=1}^K \left[\frac{1}{2}\hat{\mathbf{q}}_t^R(\overline{\mathcal{S}}_k) \cdot \hat{\boldsymbol{\rho}}_t(\overline{\mathcal{S}}_{-k}) + \frac{1}{2}\hat{\mathbf{q}}_t^D(\overline{\mathcal{S}}_k) \cdot (1 - \hat{\boldsymbol{\rho}}_t(\overline{\mathcal{S}}_{-k}))\right]$. The line labeled "out-of-sample" in Panel B presents the average, across folds, of the out-of-sample estimated partisanship using the leave-out estimator: $\frac{1}{K}\sum_{k=1}^K \left[\frac{1}{2}\hat{\mathbf{q}}_t^R(\overline{\mathcal{S}}_k) \cdot \hat{\boldsymbol{\rho}}_t(\overline{\mathcal{S}}_{-k}) + \frac{1}{2}\hat{\mathbf{q}}_t^D(\overline{\mathcal{S}}_k) \cdot (1 - \hat{\boldsymbol{\rho}}_t(\overline{\mathcal{S}}_{-k}))\right]$, which is derived by replacing $\hat{\boldsymbol{\rho}}_{-i,t}(\overline{\mathcal{S}})$ in the in-sample leave-out with its counterpart calculated on the sample excluding the k^{th} fold. The line labeled "out-of-sample" in Panel C presents the average, across folds, of the out-of-sample estimated partisanship using our preferred penalized estimator: $\frac{1}{K}\sum_{k=1}^K \left[\frac{1}{2}\hat{\mathbf{q}}_t^R(\overline{\mathcal{S}}_k) \cdot \hat{\boldsymbol{\rho}}_t^*(\overline{\mathcal{S}}_{-k}) + \frac{1}{2}\hat{\mathbf{q}}_t^D(\overline{\mathcal{S}}_k) \cdot (1 - \hat{\boldsymbol{\rho}}_t^*(\overline{\mathcal{S}}_{-k}))\right]$.

Figure 5: Partisanship within and between Topics



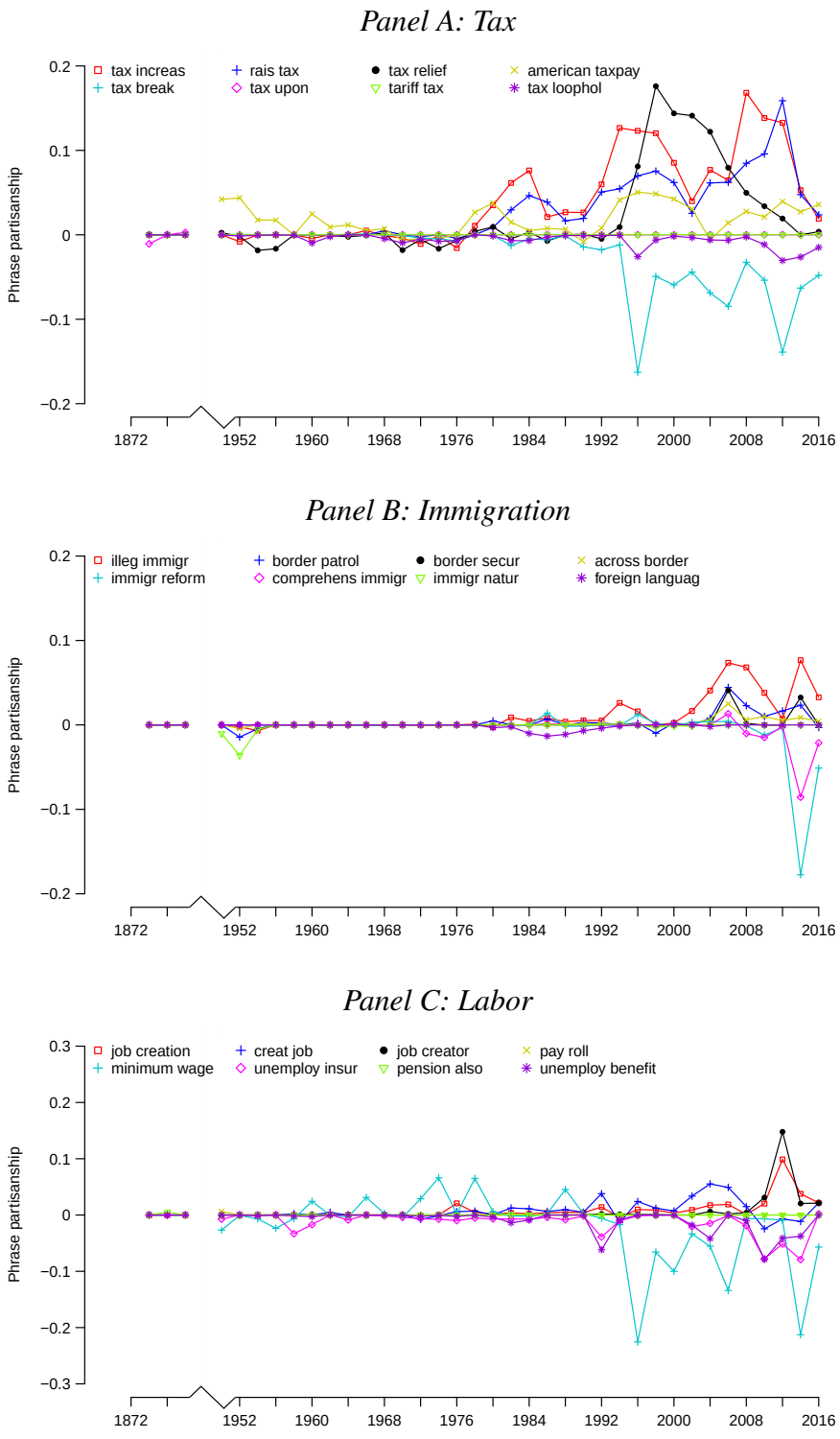
Notes: "Overall" average partisanship is from our preferred specification in Panel B of Figure 2. The other two series are based on the same parameter estimates and use the vocabulary of phrases contained in one of our manually defined topics. Between-topic average partisanship is defined as the expected 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.

Figure 6: Partisanship by Topic



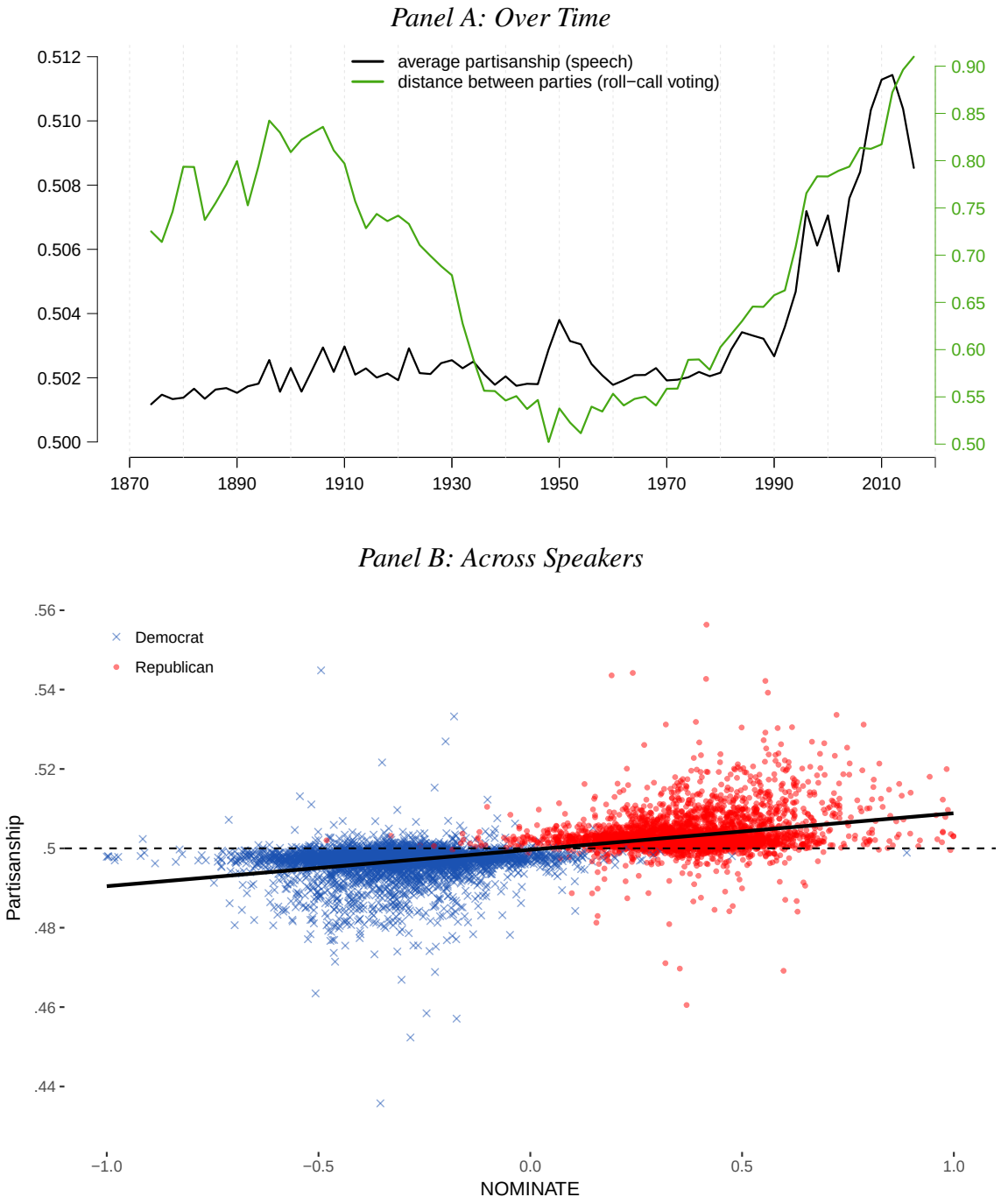
Notes: Calculations are based on our preferred specification in Panel B of Figure 2. Each panel corresponds to a topic. In each panel, for each session the top (line) plot shows estimated average partisanship for the given topic, and the bottom (bar) plot shows the share of all speech that is accounted for by phrases in the given topic. 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 7: Partisanship over Time for Phrases within Topics



Notes: Calculations are based on our preferred specification in Panel B of Figure 2. Panel A shows 1,000 times the estimated value of phrase partisanship ζ_{jt}, as defined in Section 6.1, for the four Republican (Democratic) phrases in the "tax" topic that have the highest (lowest) average phrase partisanship across all sessions. The legend lists phrases in descending order of the magnitude of average phrase partisanship across all sessions. Panels B and C show the same for the "immigration" and "labor" topics.

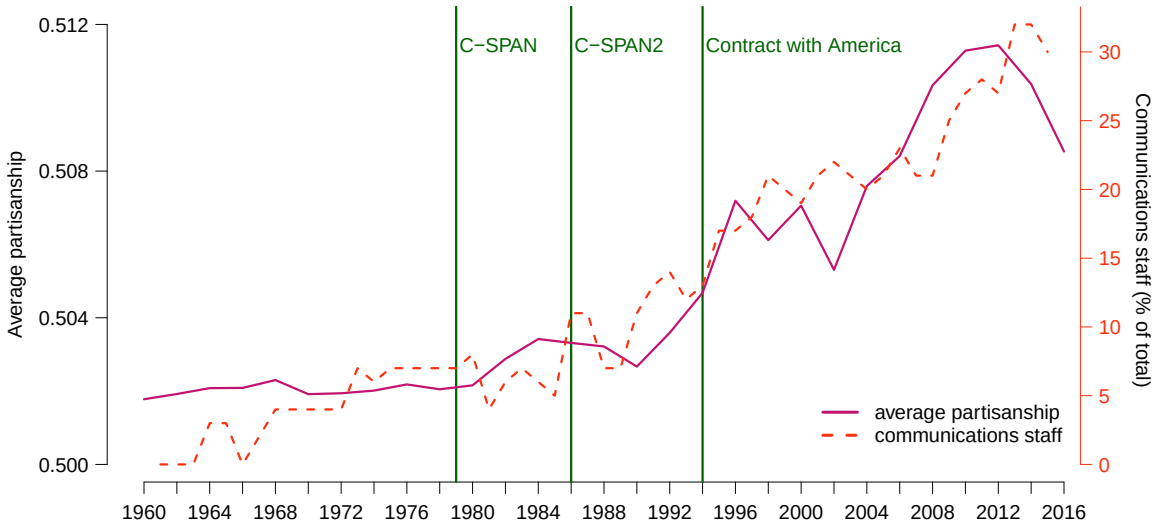
Figure 8: Partisanship vs. Roll-Call Voting



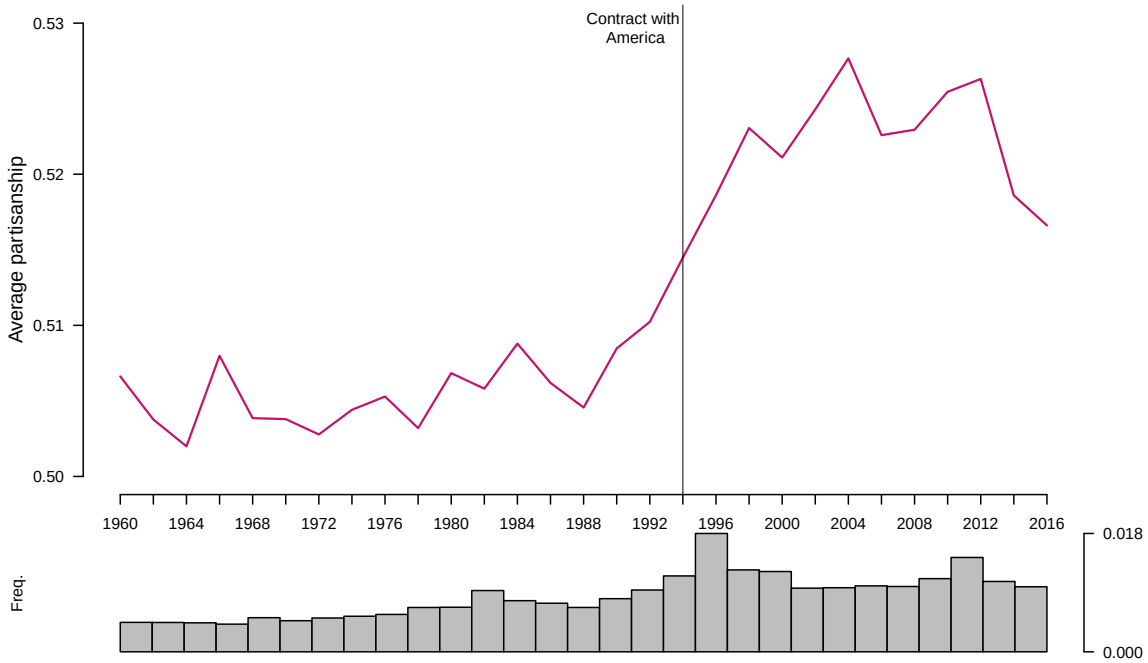
Notes: Panel A shows our preferred estimate of average partisanship from Panel B of Figure 2 and the difference between the average Republican and the average Democrat in the first dimension of the Common Space DW-NOMINATE score from McCarty et al. (2015). Panel B plots each speaker's posterior probability $\hat{\rho}_i$ of being Republican based on speech against the first dimension of the Common Space DW-NOMINATE score (McCarty et al. 2015). We drop observations for which we cannot match a DW-NOMINATE score to the speaker. 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. Nine outliers are excluded from the plot. The solid black line denotes the linear best fit among the points plotted.

Figure 9: Possible Explanations for the Rise in Partisanship

Panel A: Partisanship Post-1961 with Key Events



Panel B: Partisanship and the Contract with America



Notes: Calculations are based on our preferred specification in Panel B of Figure 2. Panel A shows average partisanship starting from 1961, the "Communications staff (% of total)" series from Lee (2016a, b) which plots (from 1961 through 2015) the share of House leadership staffers working in communications, and line markers for select events. Panel B quantifies partisanship of phrases in the *Contract with America*. 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 speech that is accounted for by phrases in the *Contract* in a given session.

Rate of Convergence of Penalized Maximum Likelihood Estimator

Let $\boldsymbol{\theta}$ be a vector that stacks the parameters $(\boldsymbol{\alpha}_t, \boldsymbol{\gamma}_t, \boldsymbol{\varphi}_t)$, and write $\boldsymbol{\theta}_0$ for its true value. Let $\mathbf{C}$ be a matrix that stacks the matrices $\mathbf{C}_t$, adding a row of zeros for speakers who are inactive in a given session. The matrix $\mathbf{C}$ then has dimension $NT \times J$, where N is the number of unique speakers, T is the number of unique sessions, and recall that J is the number of unique phrases. All limits are with respect to N.

Define the negative log likelihood for (1) and (2) as

$$\mathcal{L}(\boldsymbol{\theta}, \mathbf{C}) = \sum_{i=1}^{N} \sum_{t=1}^{T} \left[m_{it} \log \left(\sum_j \exp\left(u_{ijt}\right) \right) - \sum_j c_{ijt} u_{ijt} \right].$$

Now define $\hat{\boldsymbol{\theta}}$ to minimize the objective

$$\mathcal{L}(\boldsymbol{\theta}, \mathbf{C}) + \sum_j \lambda_{Nj} \sum_t \left| \varphi_{jt} \right|$$

where $\lambda_{Nj} \geq 0$ is a data-dependent penalty. Let

$$\mathbf{F}_N = \mathcal{L}''(\boldsymbol{\theta}_0, \mathbf{C})$$

be the matrix of second derivatives of the negative log likelihood evaluated at the true value $\boldsymbol{\theta}_0$.

Proposition 1. *If (i)* $\mathbf{F}_N / N \to \mathbf{F}$ *for some positive definite matrix* $\mathbf{F}$*, and (ii)* $\lambda_{Nj} / \sqrt{N} \xrightarrow{p} \lambda_{0j} \geq 0$ *for all* j*, then*

$$\sqrt{N}\left(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\right) \xrightarrow{d} \tilde{\theta}$$

where $\tilde{\theta}$ *is a random variable with a non-degenerate distribution.*

Proof. The proof follows Knight and Fu (2000). Let $\mathbf{a}$ denote a vector whose dimensions match $\boldsymbol{\theta}$. We will write $a_{\varphi jt}$ to denote the element matching φ_{jt}. Now define a data-dependent function $V_N()$ with

$$\begin{aligned} V_N(\mathbf{a}) = &\left[\mathcal{L}\left(\boldsymbol{\theta}_0 + \mathbf{a}/\sqrt{N}, \mathbf{C}\right) - \mathcal{L}(\boldsymbol{\theta}_0, \mathbf{C}) \right] \\ &+ \sum_j \lambda_{Nj} \sum_t \left(\left| \varphi_{jt}^0 + a_{\varphi jt}/\sqrt{N} \right| - \left| \varphi_{jt}^0 \right| \right). \end{aligned}$$

The function $V_N(\mathbf{a})$ is minimized at

$$\hat{\mathbf{a}} = \sqrt{N}\left(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\right).$$

By (i), the first term in $V_N(\mathbf{a})$ converges in distribution to

$$\mathbf{a}'\mathbf{w} - \frac{1}{2}\mathbf{a}'\mathbf{F}\mathbf{a}$$

where

$$\mathbf{w} \sim N(0, \mathbf{F}).$$

By (ii), the second term in $V_N(\mathbf{a})$ converges in probability to

$$\sum_j \lambda_{0j} \sum_t \left[a_{\varphi jt} \operatorname{sgn}\left(\varphi_{jt}^0\right) \mathbf{1}_{\varphi_{jt}^0 \neq 0} + \left| a_{\varphi jt} \right| \mathbf{1}_{\varphi_{jt}^0 = 0} \right].$$

Therefore

$$V_N(\mathbf{a}) \xrightarrow{d} V(\mathbf{a})$$

where

$$\begin{aligned} V(\mathbf{a}) = {} & \mathbf{a}'\mathbf{w} - \frac{1}{2}\mathbf{a}'\mathbf{F}\mathbf{a} \\ & + \sum_j \lambda_{0j} \sum_t \left[a_{\varphi jt} \operatorname{sgn}\left(\varphi_{jt}^0\right) \mathbf{1}_{\varphi_{jt}^0 \neq 0} + \left| a_{\varphi jt} \right| \mathbf{1}_{\varphi_{jt}^0 = 0} \right]. \end{aligned}$$

Because $V_N()$ is convex and $V()$ has a unique minimum, we have that

$$\sqrt{N}\left(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\right) \xrightarrow{d} \arg\min_{\mathbf{a}} V(\mathbf{a})$$

which is nondegenerate as desired. □

Corollary 2. *Write average partisanship $\overline{\pi}_t(\boldsymbol{\theta})$ as a function of the parameter $\boldsymbol{\theta}$. Then under the conditions of Proposition 1, for each t*

$$\sqrt{N}\left(\overline{\pi}_t\left(\hat{\boldsymbol{\theta}}\right) - \overline{\pi}_t(\boldsymbol{\theta}_0)\right) \xrightarrow{d} \tilde{\pi}$$

where $\tilde{\pi}$ is a random variable with a non-degenerate distribution that depends on t.

Proof. First note that Proposition 1 implies that $\hat{\boldsymbol{\theta}} \xrightarrow{p} \boldsymbol{\theta}_0$. Because $\overline{\pi}_t()$ is continuous and differentiable we can write that

$$\overline{\pi}_t\left(\hat{\boldsymbol{\theta}}\right) - \overline{\pi}_t(\boldsymbol{\theta}_0) = \nabla_{\overline{\boldsymbol{\theta}}} \overline{\pi}_t\left(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\right)$$

where $\overline{\boldsymbol{\theta}}$ is a mean value. The rest follows from standard limit results. □

Remark 3. Corollary 2 implies that $\overline{\pi}_t\left(\hat{\boldsymbol{\theta}}\right)$ satisfies Assumption 2.2.1 of Politis et al. (1999), with $\tau_N = \sqrt{N}$. It then follows by Theorem 2.1.1 and Remark 2.2.1 of Politis et al. (1999) that if we

choose subsets of size $B \to \infty$ with $B/N \to 0$, subsampling-based confidence intervals on $\overline{\pi}_t\left(\hat{\boldsymbol{\theta}}\right)$ will have asymptotically correct coverage.