JACOB JENSEN
Stanford University

ETHAN KAPLAN
University of Maryland at College Park

SURESH NAIDU
Columbia University

LAURENCE WILSE-SAMSON
Columbia University

Political Polarization and the Dynamics of Political Language: Evidence from 130 Years of Partisan Speech

ABSTRACT We use the digitized *Congressional Record* and the Google Ngrams corpus to study the polarization of political discourse and the diffusion of political language since 1873. We statistically identify highly partisan phrases from the *Congressional Record* and then use these to impute partisanship and political polarization to the Google Books corpus between 1873 and 2000. We find that although political discourse expressed in books did become more polarized in the late 1990s, polarization remained low relative to the late 19th and much of the 20th century. We also find that polarization of discourse in books predicts legislative gridlock, but polarization of congressional language does not. Using a dynamic panel data set of phrases, we find that polarized phrases increase in frequency in Google Books before their use increases in congressional speech. Our evidence is consistent with an autonomous effect of elite discourse on congressional speech and legislative gridlock, but this effect is not large enough to drive the recent increase in congressional polarization.

"Beliefs are themselves material forces."
Antonio Gramsci, *The Prison Notebooks*

"Mr. President, real leaders don't follow polls, real leaders change polls."
Gov. Chris Christie, speech at the 2012 Republican National Convention

In recent decades numerous scholars have documented the rising partisan polarization among politicians in the United States (Fiorina, Abrams, and Pope 2005, McCarty, Poole, and Rosenthal 2005). What is less well

known is how deeply this polarization has filtered into political discourse more broadly and whence it originates. Political discourse is not confined to legislatures but is instead generated and encountered by private citizens in books, newspapers, and everyday political argument, and it reflects partisan differences both inside and outside the official institutions of government. Where these partisan ideas come from, how they are propagated, and whether they matter in shaping behavior and policy are important questions for students of politics.

In this paper we use the newly digitized *Congressional Record,* matching phrases spoken on the floor of the House of Representatives to their speakers, to identify the partisanship of political phrases in each Congress since the Reconstruction era of the 1870s. Following recent literature in the statistical analysis of political language (for example, Gentzkow and Shapiro 2010 and Groseclose and Milyo 2005), we impute both the partisanship (association with left- or right-wing ideology) and the polarization (distance from the ideological center) of phrases by correlating their frequency of use with the political party of the speaker. After validating our identification of partisan phrases and checking that our method for computing partisanship does, in fact, correlate with other measures of political ideology, we use these correlations to compute an aggregate linguistic-based measure of polarization in Congress dating back to 1873. This aggregate measure computes the degree to which the use of specific phrases overlaps across the two major parties and is similar to voting-based measures of polarization, such as the DW-NOMINATE measure of Nolan McCarty, Keith Poole, and Howard Rosenthal (1997), which are based on overlap in voting patterns. In the post-1930 era, our measure correlates very highly with DW-NOMINATE, which is based on roll-call voting in Congress and is the most frequently used measure of polarization. In the pre-1930 era, we find a substantially higher degree of polarization in the late 19th century, consistent with historical narratives of Reconstruction and the Gilded Age.

We then use the imputed partisanship of phrases used in Congress, together with the frequencies of phrases used in Google Books, a database of all words from over 2 million books published in the United States since 1873. We use this database assembled by Google to construct a time-series measure of political ideology in a large sample of the printed word over 130 years of U.S. history. Our linguistic measure has two main advantages over measures based on polls, the main alternative for measurement of ideology among the population. First, the linguistic measure reflects the ideology embedded in everyday language (as expressed by an elite subset of the population, namely, the authors of

books), independent of priming by pollsters on particular topics. Second, it provides a method of scoring the implicit political ideology in a very large sample of digitized text.

The late-1990s increase in political polarization appears clearly in DW-NOMINATE, in our aggregate measure of congressional polarization, and in our aggregate measure of polarization in Google Books. However, the increase in polarization in Google Books is less pronounced throughout the second half of the 20th century than that in congressional speech, which begins in the late 1960s. If Congress is extraordinarily polarized, it has not come with a similarly large increase in the polarization of political discourse as reflected in published books until 2000.

We then use our aggregate measure of polarization in Google Books to search for important national political phenomena that correlate in time with the polarization in political discourse. Domestic political instability, driven by the violent strikes and lynchings of the late 19th and early 20th century, is robustly and strongly correlated with polarization. However, if today's political discourse is polarized, at least it is not associated with the violent conflicts of a century ago. We also find that polarization of discourse as reflected in the Google corpus is a very strong negative predictor of legislative efficiency, even more so than congressional polarization measured either by language or by roll-call votes, suggesting that ideological polarization outside of Congress does indeed get reflected in congressional behavior. Thus, we find that it is polarization in elite political discourse, rather than congressional polarization, that is most negatively correlated with legislative efficiency, suggesting that when the public is truly at ideological loggerheads, policymaking becomes very difficult.

Finally, we analyze the dynamics of individual phrases. We find strong evidence that there is substantial momentum in word use from one Congress to the next. A phrase that is used more frequently than average in one Congress is likely to be used more frequently than average for the next few Congresses. We also find robust evidence that increases in the frequency of certain phrases in congressional speech precede increases in the frequency of their appearance in the Google corpus, but only for relatively nonpartisan, nonpolarized phrases.

We follow the literature, particularly Matthew Gentzkow and Jesse Shapiro (2010), for our preferred method for choosing phrases, and some of our results are sensitive to this choice. Using this methodology, we find that increases in the use of very polarized phrases in Congress precede declines in the use of those phrases in Google Books. We then find that increases in

the frequency of partisan language in the Google corpus precede increases in the frequency of the same phrases in Congress. The role of books in developing somewhat polarizing phrases was greater in the mid- to late 20th century than in the pre–World War II period and is somewhat greater for economic issues than for social issues. However, these results are tentative and not robust to different ways of choosing phrases. Appendix B presents results for alternative methodologies.

Although causal interpretation is beyond the scope of this paper, and much more research is needed, we see our evidence as consistent with an autonomous effect of elite political discourse on congressional speech and legislative gridlock. However, we also do not see this effect as being quantitatively large enough to drive the recent increase in congressional polarization. We also cannot rule out that some unobserved factor, such as the influence of interest groups, is driving the larger political discourse as well as, with a lag, congressional speech and the policymaking process. In historical terms, the sum of our results suggests that although polarization in congressional behavior may be at an all-time high, polarization of underlying political ideology, and the attendant social conflict and legislative dysfunction, may not be.

Section I of this paper gives some background on the history of polarization in the United States as well as on the computational linguistics methods that we use. In section II we define what we mean by partisanship and polarization and show how we construct measures for these concepts. In section III we validate our measures and show that they do, in fact, capture political polarization. Section IV uses the scoring of political phrases in congressional speech to show trends in polarization in the Google Books database. We find that polarization has increased in recent decades, but not to unprecedented levels. We also document a historical association between political violence and polarization. We then turn in section V to an exploration of the diffusion of political language across domains, documenting the increase in frequency of phrases in Congress following their increased use in books, and vice versa. Section VI discusses some limitations of our analysis. We conclude in section VII with some speculations and directions for future work.

I. Background

In this section we review the measurement of historical political polarization in Congress as developed by political scientists. We also discuss theories of polarization. Finally, we discuss the text-as-data methods we use to

develop new measures and shed new light onto the debate over the origins of political polarization.

I.A. Historical Precedent

Both public commentators and academics argue that recent U.S. politics has been marked by unusual ideological extremism, but to what degree has that extremism been confined to the national legislature over the long sweep of U.S. history? The predominant view in political science is that the current polarization in Congress has not diffused much into the citizenry (Fiorina and others 2005). Other periods of political division, although associated with lower scores on voting-based measures of congressional polarization, have been arguably even more contentious, ranging from the pre–Civil War debates on slavery to the Great Depression to the struggles over civil rights in the 1960s. Unfortunately, no polls exist from those earlier periods that might allow measurement of mass ideological polarization over a long swath of history. Drawing on qualitative observation, Richard Hofstadter, in his famous 1964 essay "The Paranoid Style in American Politics," asserted that "American politics has often been an arena for angry minds" (p. 77). Hofstadter argued that what he saw as the extremist politics of Barry Goldwater's presidential campaign was not unusual, but instead just the latest incarnation of a recurrent style in American politics that "is a confrontation of opposed interests which are (or are felt to be) totally irreconcilable, and thus by nature not susceptible to the normal political processes of bargain and compromise" (p. 86). The idea that U.S. politics is necessarily polarized, owing to the intrinsic diversity and size of the country, goes back at least to James Madison and the divergence between Hamiltonian and Jeffersonian economic philosophies.

McCarty, Poole, and Rosenthal (2005) have performed valuable work in the quantitative study of congressional polarization by constructing measures of unobserved ideology from observed roll-call votes. Their score capturing ideology, DW-NOMINATE, suggests that, indeed, most of U.S. political history can be summarized on a single axis representing differing economic philosophies, but for some periods a second dimension of politics is needed. Writes Poole (2008, p. 7), "For most of American history only two dimensions are required to account for the fourteen million choices of the twelve thousand members who served in Congress. In fact, one dimension suffices except in two periods, roughly 1829–1851 and 1937–1970, when race-related issues introduced a second dimension." Furthermore, in terms of congressional DW-NOMINATE scores,

McCarty and coauthors find that in fact U.S. politics has been unusually polarized in the last 30 years. In particular, Republicans are now much more likely to vote with each other in roll-call votes rather than cross the aisle.

Despite this historical predisposition toward some degree of polarization, some recent commentators (Bartels 2010, McCarty, Poole, and Rosenthal 2005) have held that the past few decades have been historical outliers, outside the bounds of normal U.S. political debate, with parties unable to achieve agreement even on basic functions of government, and citizens at partisan loggerheads over economic and social policy. Evaluating these claims rigorously requires measures of ideology that are comparable across time, and assumptions about exactly which observable data reveal the complicated character of political ideology. The existing empirical strategies (Fiorina and others 2005, Gelman and others 2008, McCarty, Poole, and Rosenthal 2005) have generally focused on voting patterns or data from opinion polls; although informative, these do not uniquely and definitively measure political ideology. The three benefits of linguistic measures relative to these other measures are that they are largely free of the priming effects that are common in polls, that historical texts from which data can be constructed are available going back hundreds of years, and that the language use of a large number of special subgroups (such as intellectuals, church pastors, labor leaders, and business executives) of the population can be quantified and measured.

I.B. Factors Potentially Influencing Political Polarization

A host of explanations for increased political polarization have been offered, from inequality to procedural norms to race to regional realignment.[1] However, quantitatively evaluating many of the proposed factors is difficult given their limited time-series variation. And in general, ideological polarization is amenable to many (by no means exclusive) different definitions, which raises numerous measurement issues.

Furthermore, political divisions may originate not in the legislature but in the ideas disseminated by, for example, public intellectuals, churches, or the media. Yet some have noted also that the dramatic increase in measured polarization in Congress is not matched by an increase in the measured polarization of the voting population (Fiorina and others 2005).

1. See McCarty and others (2005), Glaeser and others (2004), Fiorina and others (2005), Bartels (2010), and Campante and Do (2008).

Others have pointed out that polarization has increased among the politically active and aware, but not among the electorate in general (Gelman and others 2008). The limited independent variation due to the shortness of panels in survey data, however, means that it is difficult to evaluate the long-run changes in political ideology among nonpoliticians. In addition, survey questions often lag the introduction of new issues, perhaps because survey design itself responds to political discourse and the emergence of politically sensitive issues. By constructing a phrase-level data set on the use of political language, we can better trace the ideological dynamics of political divisions, in part by bringing to bear substantially greater statistical power.

I.C. The Role of Elite Discourse

Our source for "discourse" in this paper is the Google Books database of n-grams (Google Ngrams), the construction of which we describe in more detail in appendix A. Google Books is an online collection of some 5 million digitized books, many of which were obtained from university libraries. We narrow our sample to the American Google corpus, a selection of books published in the United States, and then further to the sample between 1873 and 2000, leaving slightly fewer than 2,100,000 books in our sample. Following the advice of the authors of the Google Books corpus, we stop our analysis in 2000, when the way that texts were selected into the corpus was changed. We therefore consider Google Ngrams a measure of elite discourse, whose importance is widely stressed in the political science literature. We understand "elite" here as the intellectual elite, or people who exert an impact on public opinion through their writing or other public communication. For example, in an attempt to understand the formation of mass opinion, John Zaller (1992, p. 39) considers "three broad classes of variables: Aggregate-level variation in the information carried in elite discourse, including elite cues about how new information should be evaluated, individual-level differences in attention to this discourse, and individual-level differences in political values."

A more historical take on the role of ideology and intellectuals in U.S. politics is that of Hans Noel (2006, 2012). Examining the reversal of the two major parties' positions on racial issues in the 20th century, Noel finds that realignment by political intellectuals preceded congressional roll-call realignment by at least 20 years, which he suggests "is consistent with the view that ideology shapes party coalitions" (2012, p. 156). This ideology is expressed primarily among elites, by which Noel means

political leaders and other active participants in politics, who are the most invested. Noel and his research assistants construct the underlying data by hand-coding the positions of various pundits opining in the major contemporaneous political publications. Of particular interest is his analysis of the organization of ideology in the middle part of the 19th century around, and in response to, free labor ideology, which "formed a long coalition, between abolitionists and manufacturers (and others)" (Noel 2006, p. 21). The ascent of slavery as an ideological issue, he argues, resulted in the collapse of the prevailing party system. He also finds that "ideological polarization anticipated the political polarization of the early twenty-first century" (Noel 2006, p. xv). Using his characterization, he finds that the "conservative ideology" that underpins today's Republican Party is a mix of economic libertarianism, anti-Communist sentiment, and religiosity.

Given our data sources, we think our results are particularly descriptive of elite political discourse, in the sense of the words produced and consumed by intellectuals, as opposed to the general public discourse. However, these elites may be the segment of the population that generates ideas and promotes them to the larger population, cementing political coalitions among groups and groups of ideas.

I.D. Congressional Text-as-Data

Our approach is to develop partisanship measures for individual phrases within the historical record of congressional debate, so as to examine the phenomenon of polarization at the level of the phrase, rather than of the individual. This will allow us to project partisanship measures onto other, similar textual sources so as to understand the variation of partisan language and ideological polarization over time and space.

The use of quantitative measures of text for political analysis is by no means new. The "text-as-data" literature is still relatively young, but analysis of congressional speech has become fairly common, with methodologies varying from simple word counts to more sophisticated Bayesian models of partisan text generation. Our predecessor within the economics literature is Gentzkow and Shapiro (2010), who use congressional speech in 2005 to score partisan slant in a cross section of newspapers. Their work extends the original work by Tim Groseclose and Jeff Milyo (2005), who used partisanship measures of members of Congress to impute partisanship measures onto think tanks and the media. Justin Grimmer and Brandon Stewart (2012) provide a survey of automated text analysis in political science, and Burt Monroe, Michael Colaresi,

and Kevin Quinn (2008) discuss methods for extracting partisan phrases from the *Congressional Record.* Jean-Baptiste Michel and others (2011) introduce the Google Books corpus and use it to quantitatively analyze cultural and linguistic changes. What we contribute is a long-run analysis of the dynamics of political language, allowing an evaluation of the current moment relative to other contentious periods of U.S. political history. Our paper thus sits at the intersection of the literature on long-run patterns of partisan ideology in the United States and the literature on scaling political ideology in text to produce time-varying partisan scalings of phrases over 130 years of congressional speech.

II. Definitions and Measures

We are interested in identifying political ideas. As a first pass, we attempt to capture an idea by restricting attention to phrases of three consecutive words, or trigrams. To construct these trigrams, we strip phrases of common conjunctions, prepositions, auxiliary verbs, pronouns, and articles such as "and" and "the." We also restrict ourselves to collecting the roots of individual words. For example, "tax," "taxes," "taxing," and "taxation" all have the common root "tax," and so we count the number of times the root of a word within a phrase is "tax."[2] The sources of our data and the cleaning process are described in appendix A, but for example, one trigram we are left with is "capit.gain.tax." For each Congress from the 43rd (1873–75) through the 110th (2007–09), we collect the number of times each trigram was spoken by each member of the House of Representatives, and the party of each member. We limit our focus to the House of Representatives in order to avoid issues of weighting speech between the Senate and the House. We also drop all independent and third-party members of the House. For example, Bernie Sanders of Vermont is dropped during his tenure in the House from the 102nd through the 109th Congresses because he ran as an independent. Additionally, we drop any speech in the *Congressional Record* of less than 10 lines and any trigrams that appear fewer than 2,000 times over all years of Google Ngrams. Finally, some trigrams contain numerals or symbols, usually because of budget numbers, dates, or

2. Similarly, "increase," "increases," and "increasing" all have the common root "increas," and so we collect the number of times the root of a word within a phrase is "increas." This is done using a commonly available program called a Porter stemmer. It was also used by Gentzkow and Shapiro (2010) and is commonly used in most of the papers written within the text-as-data literature.

peculiarities related to how the *Congressional Record* is printed. We drop all trigrams containing a numeral or a symbol. Our remaining data record the frequency of use of each of the remaining trigrams for each member of the House.

Even after these filters, our potential sample of phrases is enormous (the corpus has almost 211 million individual words), so we begin by limiting our sample to the 10,000 most polarized trigrams per Congress. Since our data cover 69 Congresses, that leaves us with 690,000 phrase-Congresses. However, many phrases appear multiple times over many Congresses, so the total number of unique phrases in our data set is 56,211. Following Gentzkow and Shapiro (2010), we order the phrases using Pearson's χ^2 statistic, which is given by

$$\chi^2_{pc} = \frac{\left(f_{pcr}f^-_{pcd} - f_{pcd}f^-_{pcr}\right)^2}{\left(f_{pcr} + f_{pcd}\right)\left(f_{pcr} + f^-_{pcr}\right)\left(f_{pcd} + f^-_{pcd}\right)\left(f^-_{pcr} + f^-_{pcd}\right)},$$

where f_{pck} is the frequency of phrase (trigram) p in Congress c used by a member of party k, and f^-_{pck} is the frequency of all phrases used in Congress c by party k excluding phrase p. We use this measure to choose our phrases because the χ^2 statistic picks out the phrases that have the highest probability of being partisan. The χ^2 statistic essentially balances frequency of use with partisanship of use. For example, if Congressman Paul Ryan (R-Wisc.) mentions his daughter's full name once in the *Congressional Record,* it will be scored as a very partisan phrase because it will have been used only by Republicans—namely, Paul Ryan. However, it will have a low probability of being included in our restricted sample because it was said only once and thus does not have a very high probability of being Republican.

In a few of the very early Congresses, there are fewer than 10,000 phrases with $\chi^2_{pc} > 0$ (after computational digit limits), so in order to reach 10,000 we choose randomly from among the other phrases after exhausting the χ^2_{pc} ranking. Appendix B describes alternative restrictions of the phrases and the corresponding results. Using our 690,000 phrase-Congress observations, we then apply our correlation-based method of imputing partisanship scores onto phrases. We use a very simple method. For each Congress and each phrase, we first normalize the frequency of a phrase to have a mean of zero and a variance of 1 across speakers. This both weights all phrases equally and allows for easier numerical interpretation of our phrase partisanship measures and regression results. We then

calculate the correlation between the party of the speaker and the normalized frequency,

$$\beta_{pc} = \sum_{h} \widetilde{\text{PARTY}}_{hc} \widetilde{f}_{phc},$$

where $\widetilde{f}_{phc}$ is the normalized frequency of phrase p used by House member h in Congress c, and $\widetilde{\text{PARTY}}_{hc}$ is a similarly normalized version of PARTY_{hc}, a dummy variable that takes on the value 1 if the member is a Republican and −1 if the member is a Democrat.

Thus, β_{pc} is the correlation of phrase use with party of speaker—in other words, our measure of a phrase's *partisanship.* A correlation coefficient of $\beta_{pc} = 0$ will be given to a phrase that is equally used by Democrats and Republicans in a given Congress, a positive coefficient to a phrase used more frequently by Republicans, and a negative coefficient to a phrase that is more popular with Democrats. This measure is very similar to the measure of slant used by Gentzkow and Shapiro but does not involve running phrase-level regressions. In addition, we call a phrase highly *polarized* if the absolute value of its correlation with party, $|\beta_{pc}|$, is large.

Phrases will score high on a polarization measure to the extent either that members of Congress are talking about different things (for example, Republicans are more likely than Democrats to discuss taxes, whereas Democrats are more likely than Republicans to discuss voting rights), or alternatively, to the extent that members are talking about the same thing but using different words (for example, Republicans might discuss guns in terms of "second amendment rights," and Democrats in terms of "gun control legislation"). For a particularly verbose politician to drive the polarization measure, that politician's speech would have to use phrases different from the average speech in his or her party.

As a robustness check, we have also implemented all of our results using DW-NOMINATE in lieu of political party. The correlation between phrase partisanship calculated with the simple binary party variable β_{pc}, and phrase partisanship based on frequency correlations with the more sophisticated and informative DW-NOMINATE score, is slightly above 0.8, and so, for ease of exposition, we use the party-based measure. Another reason for this choice is that we want to compare the time series of our polarization measure with that of DW-NOMINATE, and so we use partisan affiliation to calculate phrase ideologies in order to avoid an induced mechanical relationship.

Our aggregate measures will be the average partisanship Ψ_c, Ψ_c^G, and polarization Φ_c, Φ_c^G, of a phrase, where the absence of a superscript indicates the measure calculated from the *Congressional Record* (the default), and superscript *G* indicates the measure calculated from the Google Books corpus during each Congress. To compute aggregate partisanship in our two corpuses, the *Congressional Record* and Google Books, we calculate the frequency-weighted sum of β_{pc} for all phrases p in each; to compute aggregate polarization, we do the same but substitute $|\beta_{pc}|$ for β_{pc}:

$$\Psi_c = \frac{\sum_p f_{pc}\beta_{pc}}{\sum_p f_{pc}} \text{ and } \Psi_c^G = \frac{\sum_p f_{pc}^G\beta_{pc}}{\sum_p f_{pc}^G}$$

$$\Phi_c = \frac{\sum_p f_{pc}|\beta_{pc}|}{\sum_p f_{pc}^G} \text{ and } \Phi_c^G = \frac{\sum_p f_{pc}^G|\beta_{pc}|}{\sum_p f_{pc}^G},$$

where f_{pc}^G is the frequency with which phrase p is mentioned in books published during the years of Congress c that appear in Google Books. These measures can be interpreted as the average percent deviation in use across parties of the average phrase in the samples we extract from the *Congressional Record* and Google Books, respectively. In other words, we develop measures of partisanship and polarization based upon differences across parties in speech patterns.

III. Validating the Methodology

In this section we show that the phrases we extract from the *Congressional Record* yield meaningful measures of polarization. We show our most polarized phrases and validate that they do, in fact, correspond to our intuitive notions of Democratic and Republican ideology. We also show that our measures of partisanship and polarization correlate with other measures of ideology and polarization.

III.A. Partisan Phrases

Table 1 lists our single most partisan phrases, for both Democrats and Republicans, by Congress. We do this as a first step toward validating their ideological content. However, the identification of these most partisan phrases over 140 years of U.S. history is interesting in itself. Although

Table 1. Most Partisan Phrases, by Party and Congress, 1873–2009[a]

	Most Democratic phrases			*Most Republican phrases*		
Congress	*Trigram*	Ψ_c	*SE*	*Trigram*	Ψ_c	*SE*
43 (1873–75)	state.legisl.power	–0.208	0.229	fifti.year.ago	0.137	0.120
44 (1875–77)	establish.unit.state	–0.165	0.105	unit.state.ii	0.222	0.181
45 (1877–79)	unit.state.judg	–0.178	0.072	methodist.episcop.church	0.232	0.041
46 (1879–81)	unit.state.requir	–0.205	0.054	unit.state.give	0.186	0.058
47 (1881–83)	hundr.fifti.million	–0.186	0.023	unit.state.labor	0.172	0.086
48 (1883–85)	public.interest.requir	–0.155	0.073	hundr.mile.distant	0.190	0.232
49 (1885–87)	year.end.june	–0.168	0.003	commerci.sulphur.acid	0.207	0.131
50 (1887–89)	time.unit.state	–0.201	0.016	industri.unit.state	0.228	0.014
51 (1889–91)	congress.pass.law	–0.278	0.025	earliest.practic.moment	0.169	0.135
52 (1891–93)	establish.unit.state	–0.123	0.081	unit.state.afford	0.242	0.177
53 (1893–95)	unit.state.resid	–0.167	0.067	protect.american.industri	0.222	0.020
54 (1895–97)	legislatur.pass.law	–0.195	0.231	south.east.west	0.137	0.074
55 (1897–99)	singl.gold.standard	–0.231	0.010	fiscal.year.end	0.136	0.004
56 (1899–1901)	singl.gold.standard	–0.272	0.009	unit.state.possess	0.196	0.034
57 (1901–03)	high.protect.tariff	–0.191	0.052	hundr.fifti.thousand	0.153	0.053
58 (1903–05)	high.protect.tariff	–0.168	0.059	past.ten.year	0.168	0.059
59 (1905–07)	exercis.polic.power	–0.245	0.021	unit.state.secretari	0.152	0.034
60 (1907–09)	unit.state.senat	–0.174	0.007	bona.fide.purchas	0.162	0.021
61 (1909–11)	high.protect.tariff	–0.284	0.013	protect.american.industri	0.195	0.012
62 (1911–13)	high.protect.tariff	–0.188	0.013	protect.american.industri	0.209	0.018
63 (1913–15)	make.ampl.provis	–0.106	0.089	unit.state.steel	0.244	0.005
64 (1915–17)	unit.state.divid	–0.131	0.088	unit.state.command	0.179	0.067
65 (1917–19)	unit.state.recogn	–0.140	0.053	time.call.attent	0.170	0.018
66 (1919–21)	great.republican.parti	–0.171	0.118	unit.state.oblig	0.142	0.096

(*continued*)

Table 1. Most Partisan Phrases, by Party and Congress, 1873–2009[a] (*Continued*)

	Most Democratic phrases			*Most Republican phrases*		
Congress	*Trigram*	Ψ_c	*SE*	*Trigram*	Ψ_c	*SE*
67 (1921–23)	high.protect.tariff	–0.229	0.017	committe.report.favor	0.113	0.075
68 (1923–25)	high.protect.tariff	–0.175	0.025	unit.state.sell	0.149	0.096
69 (1925–27)	reduc.freight.rate	–0.185	0.032	unit.state.civil	0.136	0.080
70 (1927–29)	unit.state.read	–0.154	0.048	state.great.britain	0.142	0.021
71 (1929–31)	men.women.children	–0.178	0.009	unit.state.world	0.151	0.021
72 (1931–33)	men.women.children	–0.151	0.008	rememb.year.ago	0.182	0.067
73 (1933–35)	unit.state.washington	–0.104	0.082	fundament.principl.govern	0.215	0.065
74 (1935–37)	ten.thousand.dollar	–0.110	0.070	reciproc.trade.agreement	0.276	0.005
75 (1937–39)	unit.state.applic	–0.109	0.065	nation.debt.increas	0.259	0.031
76 (1939–41)	public.work.administr	–0.164	0.002	neutral.unit.state	0.208	0.012
77 (1941–43)	lower.freight.rate	–0.116	0.088	american.expeditionari.forc	0.214	0.005
78 (1943–45)	hous.confer.report	–0.183	0.032	republican.nation.convent	0.176	0.043
79 (1945–47)	franklin.delano.roosevelt	–0.206	0.004	offic.build.washington	0.214	0.006
80 (1947–49)	provid.feder.aid	–0.160	0.073	made.unit.state	0.180	0.010
81 (1949–51)	unit.state.chamber	–0.135	0.013	compulsori.health.insur	0.220	0.005
82 (1951–53)	hundr.year.ago	–0.131	0.044	state.dean.acheson	0.205	0.013
83 (1953–55)	rural.electr.cooper	–0.169	0.018	fiscal.year.end	0.195	0.002
84 (1955–57)	unit.state.territori	–0.181	0.015	senat.agricultur.committe	0.149	0.037
85 (1957–59)	gener.unit.state	–0.159	0.003	republican.nation.convent	0.155	0.040
86 (1959–61)	unit.state.transmit	–0.159	0.006	task.forc.studi	0.235	0.059
87 (1961–63)	act.approv.juli	–0.149	0.024	nation.secur.council	0.229	0.018

88 (1963–65)	john.fitzgerald.kennedi	−0.223	0.002	outstand.young.men	0.251	0.014
89 (1965–67)	unit.state.transmit	−0.204	0.002	washington.even.star	0.233	0.011
90 (1967–69)	unit.state.transmit	−0.206	0.004	republican.nation.convent	0.248	0.018
91 (1969–71)	sleep.car.porter	−0.253	0.019	balanc.feder.budget	0.203	0.048
92 (1971–73)	provid.feder.fund	−0.203	0.021	gener.revenu.share	0.171	0.006
93 (1973–75)	act.assist.secretari	−0.193	0.012	higher.tax.bracket	0.169	0.105
94 (1975–77)	unit.state.transmit	−0.196	0.005	free.enterpris.system	0.259	0.002
95 (1977–79)	unit.state.transmit	−0.198	0.002	folk.back.home	0.245	0.035
96 (1979–81)	unit.state.transmit	−0.211	0.002	free.enterpris.system	0.241	0.002
97 (1981–83)	reagan.administr.propos	−0.228	0.007	higher.tax.bracket	0.235	0.012
98 (1983–85)	nuclear.arm.race	−0.185	0.003	unit.state.requir	0.232	0.008
99 (1985–87)	martin.luther.king	−0.208	0.002	hous.republican.confer	0.350	0.015
100 (1987–89)	martin.luther.king	−0.211	0.003	lower.tax.rate	0.207	0.042
101 (1989–91)	black.histori.month	−0.208	0.004	social.secur.earn	0.195	0.010
102 (1991–93)	defend.saudi.arabia	−0.166	0.026	social.secur.earn	0.224	0.005
103 (1993–95)	earn.incom.tax	−0.208	0.003	clinton.health.care	0.304	0.004
104 (1995–97)	billion.tax.cut	−0.321	0.002	balanc.budget.act	0.297	0.002
105 (1997–99)	black.histori.month	−0.265	0.005	chief.law.enforc	0.254	0.019
106 (1999–2001)	civil.right.movement	−0.243	0.005	peopl.pai.tax	0.206	0.014
107 (2001–03)	medicar.trust.fund	−0.280	0.002	saddam.hussein.regim	0.185	0.016
108 (2003–05)	privat.sector.job	−0.282	0.008	gener.appropri.bill	0.144	0.018
109 (2005–07)	martin.luther.king	−0.321	0.001	million.illeg.alien	0.167	0.021
110 (2007–09)	educ.health.care	−0.269	0.005	domest.energi.product	0.196	0.020

Source: Authors' calculations using data from the digitized *Congressional Record.*

a. Ψ_c is a measure of the partisanship of phrases that ranges from −1 to 1, where −1 is a highly Democratic phrase and 1 is a highly Republican phrase. See the text for details of the construction of the measure. SE = standard error.

in some Congresses the most partisan phrases are uninformative, often there are clear partisan divisions, from the Democratic opposition to high tariffs in the early 20th century (1901, 1903, 1909, 1911) to the Republican focus on business and taxation in roughly the same period (1893, 1909, 1911).[3] As Poole (2008, p. 8) observes, "through most of this period Democrats tended to be pro-agrarian [and] also anti-tariff," while "Republicans [were] pro-business, pro-capitalist." (We note that Poole's definition of "capitalist" here includes a preference for trade protection for American industry.)

Table 2 shows the 50 most partisan phrases for each party from the most recent Congress for which we had complete data, the 110th (2007–09). The most Democratic phrases are educ.health.care and mental.health.servic, while the most Republican phrases are domestic.energi.product and wall.street.journal. Democrats also frequently refer to a number of health-related topics. Unsurprisingly, the trigrams global.climat.chang and reduc.global.warm also appear among the most partisan Democratic phrases. On the Republican side, a number of tax-related topics appear in addition to the trigrams privat.properti.right and umbil.cord.blood (a phrase related to the debate over the medical use of embryonic stem cells). Some of the phrases that our methods select as partisan are not, in fact, obvious partisan phrases. These include some of the top Democratic phrases in the 1960s such as unit.state.transmit and sleep.car.porter and Republican phrases such as made.unit.state and time.call.atten. This fact could be due either to imperfections in our methods or to some of the phrases being truly partisan but less obviously so to a 21st-century audience. It is somewhat reassuring that almost all of the most partisan phrases in the 110th Congress, are in fact both recognizable and recognizably partisan. Nevertheless, improved validation of the methods, especially in the historical context, is surely needed.

III.B. Model Fit: Cross-Validation

We also attempt to validate our estimation by predicting the party of each member of the House of Representatives from his or her language use. We take the estimated partisanship coefficient for each phrase in a given

3. See the online appendix for word clouds showing the 500 most polarized phrases in each Congress. Online appendixes and replication files for the papers in this volume may be accessed on the *Brookings Papers* website, www.brookings.edu/about/projects/bpea, under "Past Editions."

Table 2. The 50 Most Partisan Phrases, by Party, in the 110th Congress, 2007–09[a]

Most Democratic phrases			*Most Republican phrases*		
Trigram	Ψ_c	*SE*	*Trigram*	Ψ_c	*SE*
educ.health.care	–0.269	0.005	domest.energi.product	0.196	0.020
mental.health.servic	–0.246	0.005	wall.street.journal	0.191	0.004
cut.interest.rate	–0.241	0.007	nuclear.power.plant	0.167	0.009
develop.block.grant	–0.225	0.007	increas.tax.burden	0.153	0.079
luther.king.jr	–0.224	0.002	trade.center.bomb	0.148	0.043
mental.health.care	–0.220	0.006	feder.govern.involv	0.148	0.029
traumat.brain.injuri	–0.208	0.003	al.qaeda.leader	0.146	0.035
middl.class.famili	–0.208	0.008	peopl.pai.tax	0.144	0.020
greenhous.ga.emiss	–0.204	0.004	umbil.cord.blood	0.144	0.030
martin.luther.king	–0.196	0.002	counti.sheriff.offic	0.142	0.016
million.peopl.live	–0.195	0.014	margin.tax.rate	0.142	0.040
health.care.educ	–0.191	0.007	govern.spend.monei	0.141	0.020
increas.public.awar	–0.191	0.015	privat.properti.right	0.140	0.013
earli.childhood.educ	–0.188	0.008	air.forc.research	0.138	0.053
dr.martin.luther	–0.187	0.003	local.nurs.home	0.138	0.137
african.american.woman	–0.186	0.012	increas.domest.product	0.136	0.029
bush.administr.propos	–0.186	0.047	adult.stem.cell	0.136	0.006
increas.minimum.wage	–0.186	0.003	lo.angel.time	0.136	0.011
global.climat.chang	–0.186	0.006	control.health.care	0.136	0.015
big.oil.compani	–0.185	0.005	capit.gain.tax	0.134	0.012
job.train.program	–0.180	0.021	outer.continent.shelf	0.134	0.002
public.health.servic	–0.179	0.009	tax.incom.tax	0.134	0.057
minimum.wage.worker	–0.178	0.015	higher.energi.cost	0.134	0.057
million.american.children	–0.177	0.012	state.law.requir	0.132	0.088
make.great.stride	–0.177	0.027	state.constitut.prohibit	0.132	0.064
adequ.health.care	–0.176	0.028	develop.nuclear.energi	0.132	0.121

(*continued*)

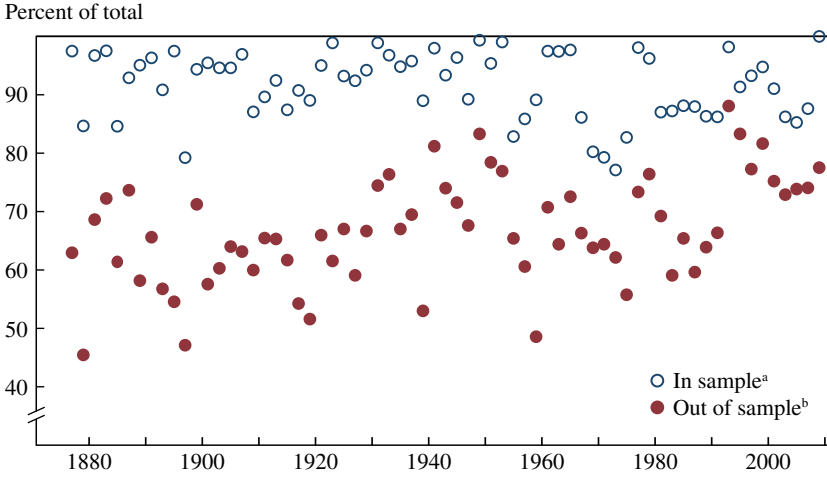
Table 2. The 50 Most Partisan Phrases, by Party, in the 110th Congress, 2007–09[a] (*Continued*)

Most Democratic phrases			*Most Republican phrases*		
Trigram	Ψ_c	*SE*	*Trigram*	Ψ_c	*SE*
expand.health.care	−0.176	0.020	special.interest.group	0.130	0.025
civil.right.movement	−0.176	0.005	american.combat.forc	0.129	0.048
health.care.crisi	−0.175	0.020	state.air.forc	0.129	0.003
make.end.meet	−0.175	0.004	feder.govern.creat	0.129	0.060
civil.right.organ	−0.175	0.024	natur.ga.resourc	0.129	0.064
carbon.dioxid.emiss	−0.173	0.019	natur.ga.reserv	0.128	0.033
reduc.global.warm	−0.173	0.020	unit.state.air	0.126	0.003
mental.health.center	−0.172	0.043	altern.fuel.sourc	0.125	0.075
california.state.assembl	−0.171	0.032	marin.corp.air	0.124	0.054
express.strong.support	−0.171	0.008	percent.incom.tax	0.123	0.088
home.energi.assist	−0.171	0.015	natur.ga.product	0.123	0.030
fuel.effici.standard	−0.171	0.021	republican.polici.committe	0.122	0.097
prevent.health.care	−0.170	0.018	unit.state.back	0.122	0.040
econom.stimulu.packag	−0.169	0.007	unit.state.treasuri	0.122	0.051
unit.state.code	−0.169	0.002	partial.birth.abort	0.122	0.061
mental.health.problem	−0.167	0.021	natur.ga.explor	0.122	0.057
energi.assist.program	−0.167	0.017	innoc.human.life	0.122	0.065
posttraumat.stress.disord	−0.166	0.017	long.long.time	0.122	0.015
work.full.time	−0.166	0.019	reduc.tax.rate	0.121	0.139
commun.develop.block	−0.165	0.010	current.tax.system	0.121	0.139
hedg.fund.manag	−0.164	0.026	lowest.unemploy.rate	0.121	0.139
student.financi.assist	−0.282	0.008	nation.honor.societi	0.144	0.018
improv.work.condit	−0.321	0.001	purchas.health.insur	0.167	0.021
health.servic.act	−0.269	0.005	air.quality.plan	0.196	0.020

Source: Authors' calculations using data from the digitized *Congressional Record.*

a. Ψ_c is a measure of the partisanship of phrases that ranges from −1 to 1, where −1 is a highly Democratic phrase and 1 is a highly Republican phrase. See the text for details of the construction of the measure. SE = standard error.

Figure 1. Party Affiliations of House Members Correctly Predicted from Their Language Use, by Congress, 1873–2007



Source: Authors' calculations using data from the digitized *Congressional Record*.

a. Share of House members whose party affiliation was correctly predicted by their frequency-weighted partisanship score described in the text.

b. The partisanship coefficient estimated from a random sample of three quarters of House members in each biennial Congress was used to predict the party affiliation of the remaining quarter.

Congress and compute a score for each House member in that Congress as the frequency-weighted sum of the member's use of partisan phrases:

$$\beta_c^h = \sum_p \beta_{pc} \widetilde{f_{phc}}.$$

If our imputed measure of partisanship for a member of the House is weakly greater than zero, we predict that member to be a Republican, and if otherwise, a Democrat. We then compare our predictions with the members' actual party affiliations and graph the percentages predicted correctly for each Congress (figure 1).

In addition, we repeat the exercise out of sample: we take a random sample of three quarters of House members in each Congress, estimate an out-of-sample β_{pc} for each phrase used at least once in that sample, and then make an out-of-sample prediction of party for the remaining members. Again, we compute and plot in figure 1 the percentage of correctly predicted party affiliations. For the in-sample predictions, our correctly predicted shares range from approximately 77 percent to approximately 99.5 percent. For the out-of-sample prediction, our results are not quite as good: the correctly

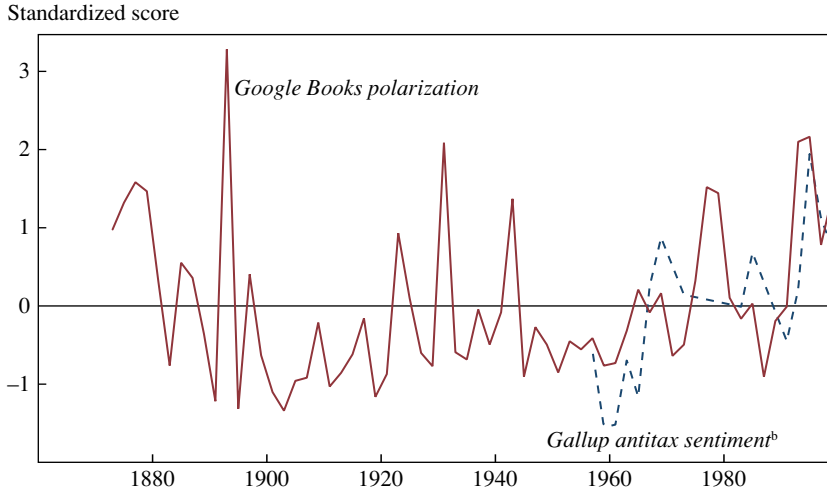
predicted shares range from just over 40 percent to slightly over 80 percent. However, the out-of-sample prediction improves steadily over time. This is to be expected, as the *Congressional Record* is shorter and measurement error due to transcription (which is done using optical character recognition) is much higher in earlier years than later. However, the trend toward increased correlation over time could also be due to the fact that the underlying variance in phrase use is higher in the more polarized 19th century or to the decreased presence over time of third-party politicians (who, as noted in section II, are dropped). We have experimented with a potentially better estimator of phrases developed by Matt Taddy (2012), which does in fact have much higher out-of-sample prediction accuracy; however, we kept the correlation measure because of its intuitiveness and ease of exposition. Of course, if we made out-of-sample predictions based on all language use excluding the individual House member whose party is being predicted, we would come very close to our in-sample prediction rates. In general, average partisanship of language does a good job of predicting party.

We also compute the average imputed partisanship of members of the House by party for each Congress. We do this by summing up across phrases, weighted by frequency of use, for each member, thus obtaining an imputed score for each member. Then we compute the mean score across all Democrats and all Republicans separately by party for each Congress. We plot these over time in appendix figure A.1. There are significant differences between our measure and DW-NOMINATE. The time series on the gap between the parties is similar, but the party that is seen as polarizing is somewhat different. We see Republican polarization in the early 1910s, when the Constitution was amended to allow for a federal income tax, and again in the 1930s. We also see Democrats moving to the left in the 2000s. It may be that in times of Republican majorities, those Democrats who remain in power may be the more left-wing Democrats, and similarly for Republicans. If so, this might partly explain the patterns in our aggregate time series. Also, a shift of everyone in the Congress to the right or the left will not show up as a mean shift in our measure. This is a drawback of our time-series measure of aggregate partisanship, although not of our average polarization measure.

III.C. Imputed Polarization and Its Relation to Voter Ideology

Our third check of our construction of partisan language is through relating shifts in voter ideology, as measured in Gallup polls, to the time trend in topic-specific polarization in the broader political discourse. We selected

Figure 2. Polarization of Tax-Related Phrases in Google Books, 1873–2007, and Antitax Sentiment in Gallup Polling, 1956–2007[a]



Source: Authors' calculations.

a. Both variables are normalized to have a mean of zero and a variance of 1. Higher absolute values indicate greater polarization.

b. Difference between the percentage of Gallup poll respondents who felt that taxes were too high and the percentage who felt they were too low.

the 6,139 phrases from among the top 10,000 in every Congress that were consistently either Republican or Democratic in each Congress within at least one decade. We then hand-coded these phrases into the following topics: narrow economic issues (the economy and taxes), broad economic issues (narrow economic issues, the environment, and health) and social issues (race, religion, sexuality, and gender issues).

We observe that the polarized topics on which Congress spends most of its time are taxes, the economy, and wars. The polarized phrases alone from each of these three subtopics account for 5 to 10 percent of all persistently polarized phrases, whereas those relating to social issues represent well fewer than 1 percent.

Figure 2 graphs the time series of our computed polarization measure for language in Google Books about taxation along with a measure of public opinion on whether or not taxes were too high, taken from Gallup polls (available only after 1957; specifically, the figure graphs the percentage difference between those who felt that taxes were too high and those who felt they were too low). To make the two series comparable, we normalize them both to have zero mean and unit variance. The graph suggests that

our measure of imputed partisanship of political text is indeed informative. Although our sample size is too small to allow a formal test of whether the opinions of politicians lead or lag voter sentiment or whether the two are dynamically correlated, it seems as though politicians' use of phrases leads to ideological change in the citizenry. A number of important periods in the history of U.S. taxation also emerge in the phrase time series. For example, the ratification of the 16th amendment in 1913, which empowered the federal government to levy an income tax, is associated in time with an increase in polarizing talk about taxes. Taxes were a polarizing topic earlier, however. For example, the People's Party (better known today as the Populists) argued for a graduated income tax in the 1890s. Closer to our own period, we see polarizing talk about taxation before Ronald Reagan's presidential victory, and a smaller bump in polarization around taxation before George W. Bush's.

IV. Historical Patterns of Partisan and Polarized Political Discourse

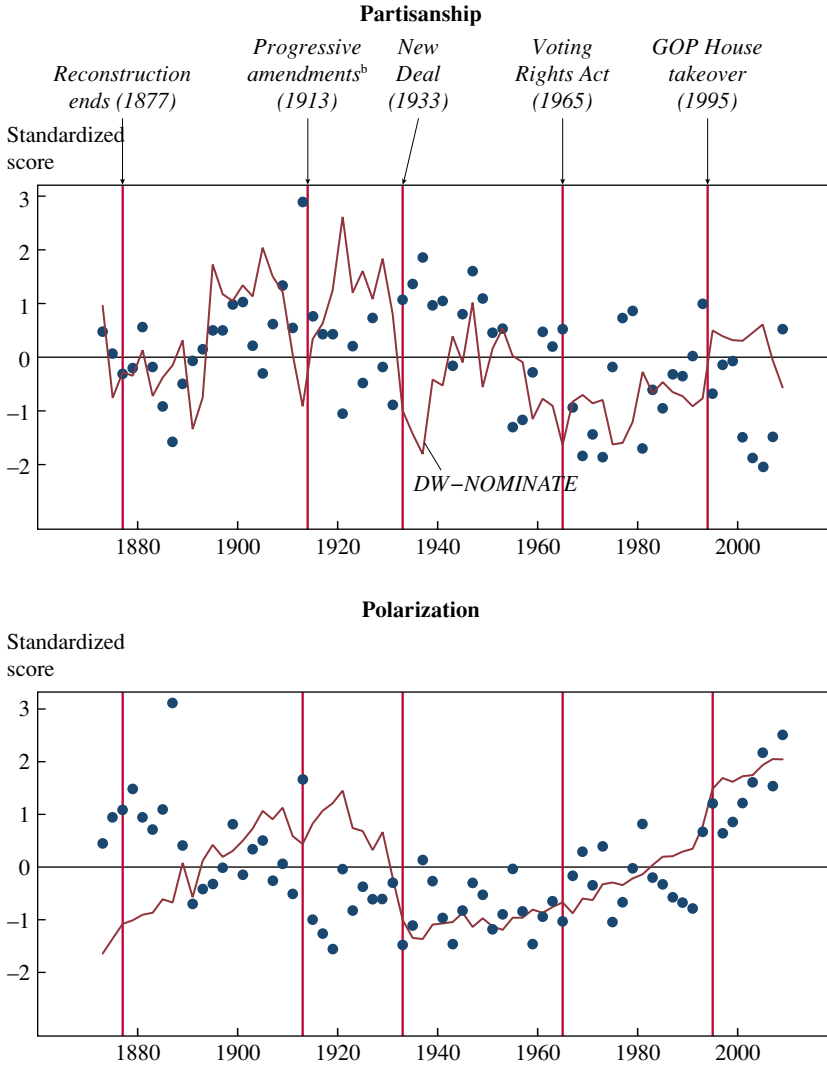
In this section we use our scoring of political phrases in congressional speech to identify trends in political polarization in the Google Books database and explore some other variables that may correlate with polarization over time. We also investigate what impact, if any, polarization may have had on legislative efficiency.

IV.A. Aggregate Time Series

Figure 3 plots over time our measures of partisanship and polarization in Congress, as well as the corresponding DW-NOMINATE measure; figure 4 does the same for Google Books and DW-NOMINATE. All the plotted variables are standardized to have a zero mean and unit variance. Five key events in U.S. political history are also indicated. The DW-NOMINATE measure of partisanship is the weighted sum of the party-level means of the DW-NOMINATE-1 score, where the weights are the fractions of seats held by each party in Congress. Similarly, the DW-NOMINATE measure of polarization is the party-seat-weighted sum of the absolute value of the DW-NOMINATE-1 score. Appendix A provides further detail.

The top panels of figures 3 and 4 suggest that our measures of partisanship are not strongly correlated with DW-NOMINATE: the correlation between our congressional partisanship measure and DW-NOMINATE is –0.02, and that between the Google Ngrams measure of partisanship and DW-NOMINATE is 0.08. The time series also reveal what appears to be

Figure 3. Partisanship and Polarization as Measured in the *Congressional Record* and by DW-NOMINATE, 1873–2007[a]

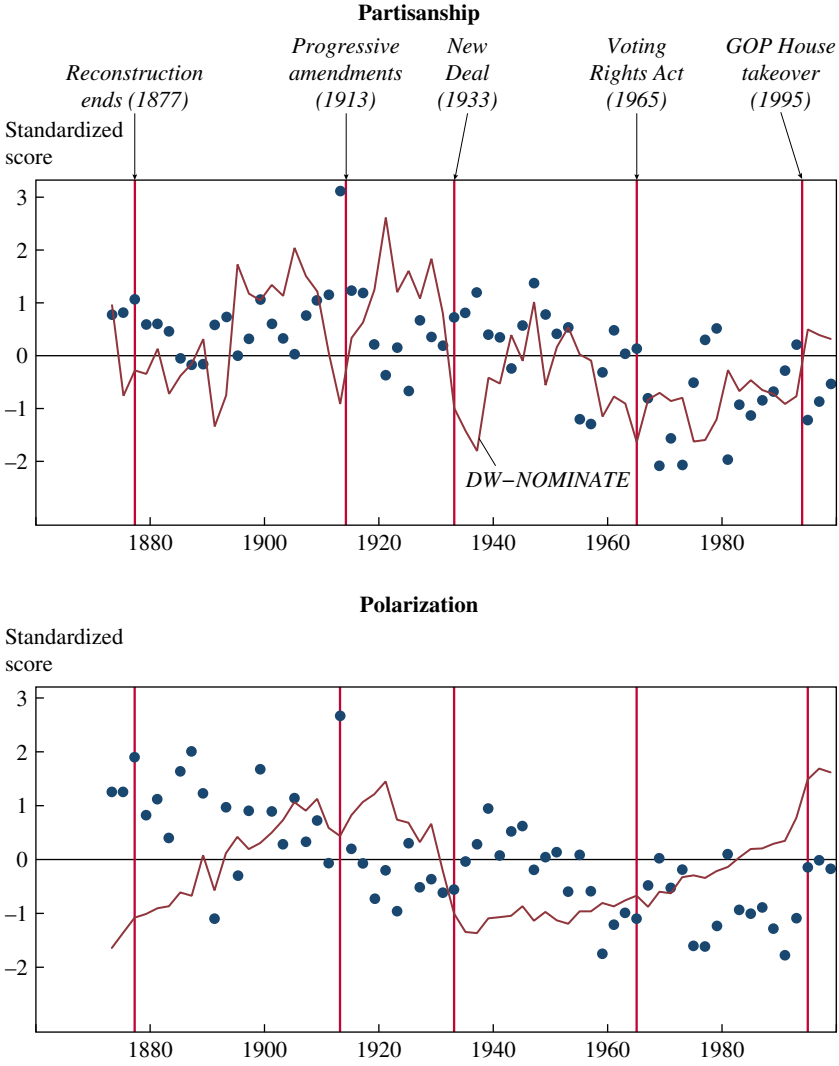


Sources: Authors' calculations using data from the digitized *Congressional Record* and the legislator estimates on voteview.com/dwnomin.htm.

a. All measures are standardized to have a mean of zero and a variance of 1.

b. The 16th (income tax) and 17th (direct election of senators) Amendments to the U.S. Constitution.

Figure 4. Partisanship and Polarization as Measured in Google Books and by DW-NOMINATE, 1873–1999



Sources: Authors' calculations using data from Google Books and the legislator estimates on voteview.com/dwnomin.htm.

a. All measures are standardized to have a mean of zero and a variance of 1.

a curious pattern in our partisanship measure: the partisanship of language tends to switch when House control switches, but in the direction of the new minority party. We formally tested the hypothesis that there was no shift in measured linguistic partisanship with changes in partisan control of the House, and were unable to reject it, but the limited power of such a test warrants caution. The pattern could arise because minority parties talk more and use more-partisan language in an effort to slow the enactment of policies they oppose. By contrast, the correlation between the Google Ngrams and *Congressional Record* measures of partisanship is a very high 0.85, suggesting that the changing correlations with party use in Congress are driving much of the change in Google Ngrams.

The broad pattern of congressional polarization lines up well with the DW-NOMINATE measure (bottom panel of figure 3), although the correlation coefficient is only 0.08. The Google Ngrams polarization measure (bottom panel of figure 4) has a correlation of –0.078 with the corresponding DW-NOMINATE measure. But the correlation between congressional polarization and Google Ngrams polarization is a modest 0.64, suggesting that much more of the variation is coming from the Google Ngrams frequency of partisan phrases, rather than from changes in the *Congressional Record.*

The bottom panels of figures 3 and 4 echo the common claim that political polarization has increased starting in the mid-1990s. Unfortunately, the composition of the Google Books corpus changes in 2001, so we cannot consistently examine our measure of polarization in political discourse in the last decade.[4] But the increase is more pronounced and historically anomalous in congressional speech than in the Google Books corpus. By 2000, political polarization as measured in Google Ngrams (figure 4) indicates an increase in the polarization of political discourse to levels last seen around the 1950s and the civil rights era. This is well below the measured polarization in the Google corpus though most of the 1930s and most of the pre-1920 period. In contrast, measured polarization in Congress is, by 2000, close to its historic peak (with the exception of 1877, which marks the end of the Reconstruction era). This suggests that although Congress's polarization may have begun increasing in the late 1980s, polarization of political discourse as a whole may not have increased to the same degree. However, we will have to await the development of comparable post-2000 Google corpora to compare these trends comprehensively.

4. When we estimate polarization in the 2001–07 Google Books, we do see large increases in 2005 and 2007, but our measure still remains below the highest levels in the late 19th century. However, following Michel and others (2011) and the recommendations in Google Books, we limit our further analysis to the pre-2001 corpus.

Our measures of polarization start higher than DW-NOMINATE in the 1870s, consistent with the sharply polarized environment following the Civil War. Our measures thus fall earlier than does DW-NOMINATE (although they spike upward around 1913, when the progressive 16th and 17th Amendments to the Constitution were ratified), before rising following the New Deal, a period when DW-NOMINATE polarization falls.

Our measures are perhaps more consistent with the high degree of political tension voiced during the first administration of Franklin Roosevelt. Later, our measures spike at various places during the civil rights era of the 1960s and the Watergate era of the 1970s, while the increase in DW-NOMINATE polarization is secular. We find these differences instructive: they occur at times when Congress was unified in passing legislation but the polarization of public discourse was intensified.

IV.B. Time-Series Correlates of Polarized Political Discourse

We next examine a small list of potential correlates of polarization in political discourse. We consider fatalities from domestic political violence, military casualties, and GDP growth, all averaged to the biennial (congressional) level.[5] All the data sets are described in appendix A. We chose these variables on the basis of data availability and the hypothesis that political discourse becomes more polarized during periods of economic or political change. We run simple ordinary least squares regressions of the form

$$\Psi_c^G = \gamma^1 \text{ GDP growth}_c + \gamma^2 \text{ Political violence}_c + \gamma^3 \text{ Military casualties}_c + \gamma X_c + \varepsilon_c,$$

where X is a vector of control variables that includes the number of Democratically held seats in the House, the DW-NOMINATE measure of polarization, and the level of polarization in the House Ψ_c in order to control for congressional characteristics. We also report specifications with aggregate partisanship in Google Φ_c^G as the dependent variable. As the results in the first four columns of table 3 show, political violence and political polarization are correlated. Moreover, figure 5 shows that both are higher

5. We also considered inequality, as measured by the share of income accruing to the top 1 percent of the population. This measure is a robust predictor of polarization in Google Ngrams, even when conditioned on various covariates (including DW-NOMINATE). However, because the inequality measure is available only back to 1913, we decided to focus exclusively on variables we could measure over the full 1873–2009 period.

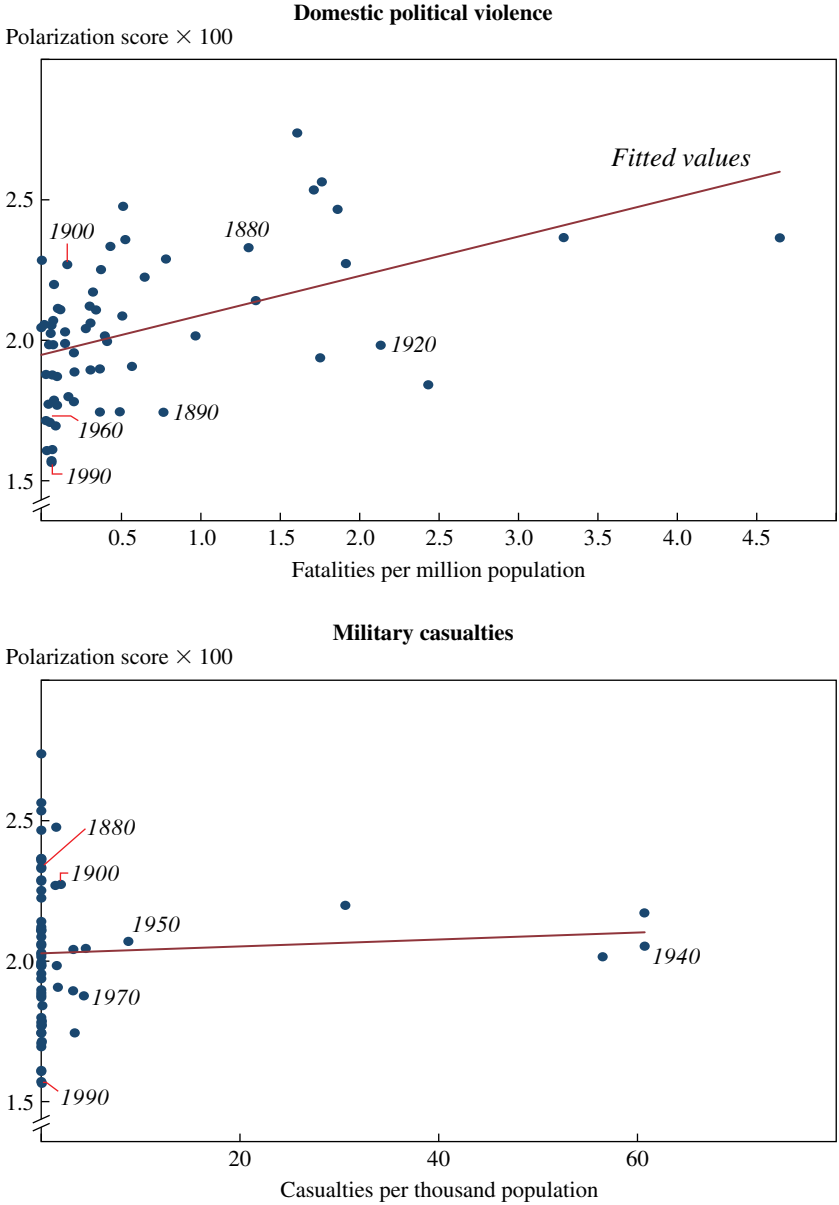
Table 3. Regressions Explaining Political Polarization and Partisanship as Measured in Google Books[a]

	Dependent variable: polarization Φ_c^G				*Dependent variable: partisanship* Ψ_c^G			
Independent variable	*3-1*	*3-2*	*3-3*	*3-4*	*3-5*	*3-6*	*3-7*	*3-8*
Fatalities from domestic political violence (per million population)	186.8**	104.7**	79.71	−5.841	200.7	234.1***	167.7**	84.50
	(69.37)	(34.53)	(44.27)	(32.57)	(102.6)	(61.80)	(62.66)	(53.94)
Military casualties (per thousand population)	0.0161	4.384*	4.199**	4.533**	4.396	2.059	3.936	2.553
	(1.589)	(1.656)	(1.498)	(1.678)	(3.119)	(2.188)	(2.273)	(1.977)
Real GDP growth (percent per year)	358.3*	182.1	116.5	59.59	−102.0	−293.3	−488.8*	−434.8*
	(158.0)	(148.5)	(142.6)	(108.9)	(349.0)	(224.0)	(234.5)	(193.9)
Polarization in House Φ_c		24465.7***	22210.7***	23816.2***				
		(4159.4)	(5143.6)	(4080.1)				
No. of Democratic seats in House			−1.277	0.378			−3.203*	0.946
			(0.777)	(0.572)			(1.303)	(0.642)
DW-NOMINATE polarization			−58.92*					
			(27.08)					
Year trend				−4.022***				−5.766***
				(0.721)				(0.818)
Partisanship in House Ψ_c						24415.0***	25823.6***	21420.9***
						(1312.9)	(1444.3)	(1270.6)
DW-NOMINATE partisanship							−119.1*	
							(51.15)	
Constant	1871.5***	626.2**	1046.5*	8445.7***	−219.4*	−140.0**	630.0	10912.3***
	(39.82)	(216.9)	(409.7)	(1334.9)	(95.10)	(50.29)	(329.3)	(1558.5)

Source: Authors' regressions using data from the digitized *Congressional Record,* Google Books, and other sources.

a. Data are biennial and correspond to years of Congresses from 1873–75 (43rd Congress) to 2007–09 (110th Congress), 62 observations in all. All partisanship and polarization variables are standardized to have a mean of zero and variance of 1. Robust standard errors are in parentheses. Asterisks indicate statistical significance at the *0.05, **0.01, and ***0.001 level.

Figure 5. Domestic Political Violence, Military Casualties, and Polarization as Measured in Google Books



Sources: Authors' calculations using data from Google Books, Turchin (2012), and the Correlates of War project (www.correlatesofwar.org).

in the 19th century than in the 20th. The top panel of figure 5 also shows that this effect is driven not by outliers but largely by the violent strikes and racial violence of the late 19th and early 20th centuries, particularly during Reconstruction, a period of U.S. political history when legal institutions were widely regarded as weak. The regression reported in column 3-3 of table 3 includes the full set of congressional controls X_c, and although the effect of political violence falls by 50 percent, it remains large and close to statistically significant at the 10 percent level. However, including a simple linear trend (column 3-4) turns the effect from positive to negative, reflecting the fact that violence, particularly political violence, has declined secularly within the United States since the 1870s. We also run these regressions omitting the end of the Reconstruction Era (1873 and 1875, results not reported) and still find a significant correlation between political violence and polarization of political discourse.

Another candidate predictor of polarization is military casualties, reflecting the hypothesis that wars create periods of national unity and thus of lower political polarization. We find the opposite result, but this effect is driven completely by World War II, as can be seen from the bottom panel of figure 5. (In results not reported, omitting the war years renders the casualties variable insignificant in virtually all specifications, but the political violence variable remains significant.) Perhaps surprisingly, economic growth has no correlation with our measure of polarization.

The last four columns in table 3 report the same regressions but with partisanship, rather than polarization, as the outcome variable. Again we find that political violence, and little else, is a robust predictor of right-wing partisanship in political discourse. Much of the variation is captured in the congressional variables included in X_c, particularly congressional partisanship. This suggests that right-wing phrases become more prevalent in Google Books during periods when political violence is high, but we do not push this interpretation.

Although comparisons between political polarization today and in the Gilded Age of the late 19th century are possible (Bartels 2010), our results, subject to our methodological caveats, suggest that the period was both much more polarized in its political discourse and more politically violent. Recent political polarization may be high relative to the 1970s, but it is a far cry from the open violence of the late 19th century. We make no attempt to say anything about the causal relationship between these variables, as both political violence and polarized political discourse could drive each other or be generated simultaneously by other variables. However, we have attempted to control for measures of polarization in Congress, and the

results suggest that polarized political discourse varies independently of the polarization of politicians.

IV.C. Polarization and Legislative Efficiency

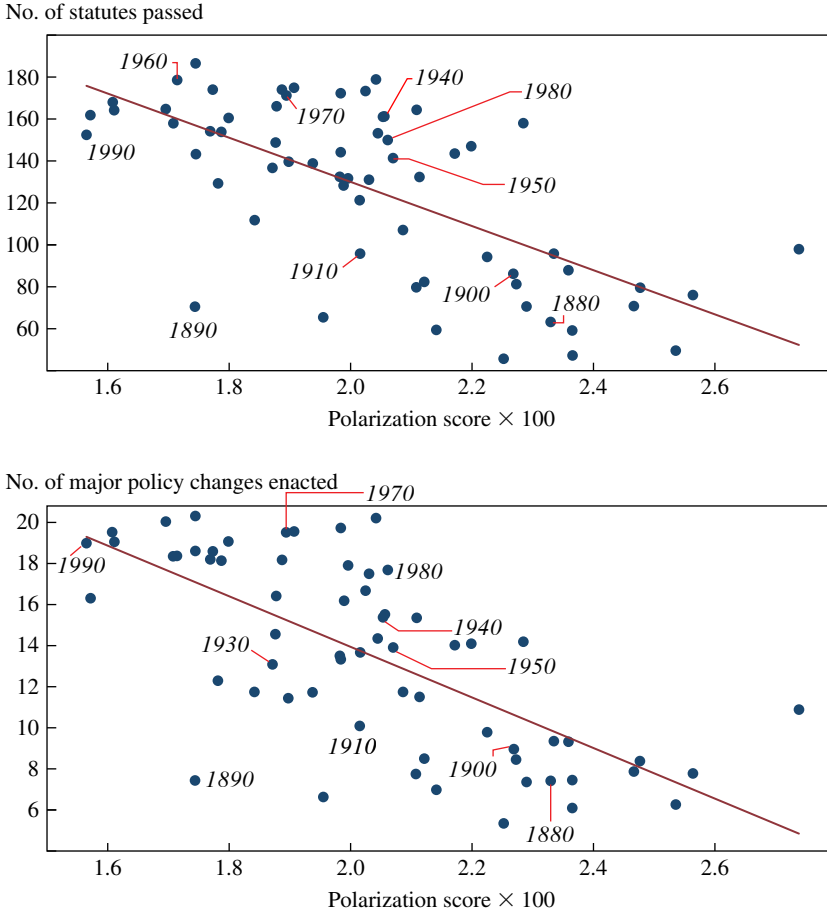
We now turn to simple estimates of the effect of political polarization on legislative efficiency. We consider two measures of legislative efficiency from Tobin Grant and Nathan Kelly (2008), who aggregate a number of partially overlapping series of legislative productivity to arrive at a time series that spans the entire history of Congress. Their first measure, the legislative productivity index (LPI), combines a number of series including the raw number of laws passed by each Congress and a number of "major enactments" and "key votes" as identified by the *Congressional Quarterly.* Their second measure, the major legislation index (MLI), excludes the raw number of laws passed from the aggregate. The two series are very highly correlated, with a correlation coefficient of 0.93. Figure 6 shows clearly that polarization in political discourse is a strong negative correlate of legislative productivity. Of course, this finding could be due to any of a number of omitted variables, so we next estimate a set of simple time-series regressions of the form

$$LE_c = \gamma^G \Psi_c^G + \gamma^C \Psi_c^C + \gamma X_c + \varepsilon_c,$$

where LE_c is either the LPI or the MLI. Columns 4-1 and 4-2 in table 4 show results of a simple bivariate regression of legislative efficiency on our polarization measure, confirming the strong negative relationship seen in figure 6. Surprisingly, columns 4-3 and 4-4 show that polarization of political discourse remains a strong negative correlate of legislative productivity even when measures of congressional polarization, either from text (our measure) or from roll-call votes (DW-NOMINATE), are included as predictors. This remains true when the same time-series correlates used in the previous subsection are included as controls, in columns 4-5 and 4-6.

Although these results are still subject to many concerns about measurement and identification, we view them as provocative evidence that it is underlying ideological polarization among political elites that is the true obstacle to legislative efficiency, rather than the polarization of Congress. The rhetoric in Congress may be extremely heated, but the personal relationships among representatives may be sufficiently strong to allow for substantial legislation to get passed, despite the obstacles imposed by the opposing party. However, if party activists and intellectuals are truly polarized in their beliefs, then politicians may have no leeway to compromise,

Figure 6. Legislative Efficiency and Polarization as Measured in Google Books



Sources: Authors' calculations using data from Google Books and Grant and Kelly (2007).

because they are constrained by reelection concerns or other intraparty ideological constraints.

V. The Diffusion of Political Language

We now turn to the micro-level diffusion of partisan language between Congress and the domain of books. We are interested to see whether increases in the use of a phrase by members of Congress precede or follow an increase of its use in books, and whether the patterns differ depending

Table 4. Regressions Explaining Congressional Legislative Efficiency with Political Polarization and Other Variables[a]

	Dependent variable: major legislation index			*Dependent variable: legislative productivity index*		
Independent variable	*4-1*	*4-2*	*4-3*[b]	*4-4*	*4-5*	*4-6*[b]
Polarization in Google Books Φ_c^G	−3.253***	−3.646***	−2.387***	−27.80***	−24.90***	−13.14*
	(0.408)	(0.504)	(0.511)	(3.685)	(4.570)	(5.296)
Polarization in Congress Φ_c		0.615	1.281**		−6.078	−1.205
		(0.569)	(0.444)		(5.015)	(4.351)
DW-NOMINATE polarization		−0.546	0.566		−8.622*	−1.350
		(0.460)	(0.412)		(4.305)	(3.900)
Fatalities from domestic political violence (per million population)			−0.782			−8.933
			(0.624)			(6.947)
Real GDP growth (percent per year)			−0.827			−11.05
			(1.984)			(17.76)
Military casualties (per thousand population)			0.0499*			0.308
			(0.0199)			(0.174)
No. of Democratic seats in House			0.0487***			0.368***
			(0.00701)			(0.0658)
Constant	13.53***	13.53***	3.454*	126.5***	124.3***	51.43**
	(0.417)	(0.443)	(1.708)	(3.856)	(3.787)	(16.16)

Source: Authors' regressions using data from the digitized *Congressional Record,* Google Books, Grant and Kelly (2008), and other sources.

a. Data are biennial and correspond to years of Congresses from 1873–75 (43rd Congress) to 2007–09 (110th Congress), 64 observations in all except where noted otherwise. All partisanship and polarization variables are standardized to have a mean of zero and variance of 1. Robust standard errors are in parentheses. Asterisks indicate statistical significance at the *0.05, **0.01, and ***0.001 level.

b. Sample consists of 62 observations.

Table 5. Summary Statistics for the Phrase Sample[a]

Variable	*Mean*	*Standard deviation*[b]
Mean polarization (25th percentile)	0.046	0.015
Mean polarization (50th percentile)	0.053	0.015
Mean polarization (75th percentile)	0.061	0.015
Google Books frequency	0.00001	0.00011
Congressional Record frequency	0.00002	0.00021
Phrases relating to social issues	0.0014	0.0380
Phrases relating to narrow economic issues	0.0087	0.0927
Phrases relating to broad economic issues	0.02	0.14
Pre-1941	0.47	0.50

Source: Authors' calculations.
a. The sample contains a total of 2,741,178 observations.
b. The standard deviation reported in the first three rows is that of the polarization variable.

upon the degree of polarization of the phrase. We form the balanced panel of phrases across our 10,000-per-year sample of phrases, imputing a value of zero to the congressional frequency if the phrase did not appear in our top 10,000 using the χ^2 metric. Table 5 reports the summary statistics for the sample of phrases we use for this analysis.

We divide the frequency counts by the total number of phrases spoken in the Congress to obtain frequency shares. We denote the frequency shares by $\hat{f}^G_{pc}$ for Google Books and $\hat{f}^C_{pc}$ for Congress. We estimate dynamic panel equations at the phrase-Congress level of the form

$$\hat{f}^G_{pc} = \sum_{j=1}^{4}\sum_{k=1}^{3} LGMPol^{jh} \times MeanPolQtile_{pj} \times \hat{f}^C_{pc-h} + \sum_{k=1}^{3} LG^k \hat{f}^G_{pc-h} + \gamma_c + \gamma_p + \varepsilon_{pc}$$

$$\hat{f}^C_{pc} = \sum_{j=1}^{4}\sum_{k=1}^{3} LCMPol^{jk} \times MeanPolQtile_{pj} \times \hat{f}^G_{pc-h} + \sum_{k=1}^{3} LG^k \hat{f}^C_{pc-h} + \gamma_c + \gamma_p + \varepsilon_{pc},$$

where LG^k and LC^k are the corresponding coefficients on lag k. $LGMPol^{jk}$ and $LCMPol^{jk}$ are the coefficients on the interaction of the kth lag of the Google Books and congressional frequencies with the jth quartile of average phrase polarization in the sample. That is, we average the absolute value of the correlation of the phrase with party over all years that it appears in the sample and then construct dummies ($MeanPolQtile_{pj}$) for the quartiles of this mean polarization variable. The specification includes phrase and Congress fixed effects, denoted by γ_p and γ_c respectively. Standard errors are clustered at the phrase level. The fixed effects imply that we are estimating the relationship between changes in the frequency of phrase use in one domain (Congress or Google Books) and

changes in the frequency of phrase use in the other. These fixed effects induce the well-known Nickell bias in panel regressions. However, because we have a relatively long panel, this should not be a major concern, and in any case, standard generalized method of moments (GMM) solutions (for example, Arellano-Bond) to this problem are computationally infeasible. We interpret significant coefficients on the lagged frequencies as evidence that increases in the frequency of the phrase in one domain precede increases in its frequency in the other. Tables 6 and 7 report the results; the cumulative sums of the lags for the mean and each quartile are separately reported at the bottom of each table.

Table 6 shows the partial correlations of lags of congressional frequencies with Google Ngram frequencies. One very strong and robust finding is that there is substantial momentum in language use. Increased use of a phrase in Google Books persists for at least 6 years and is very robust to other ways of selecting phrases, as we show and discuss in appendix B. Another qualitatively robust finding is that increased use of a phrase in Congress anticipates its use in Google, but only for the most bipartisan phrases. Given lags in writing and publication, this finding may also be consistent with low-partisanship phrases simultaneously emerging in the minds of authors writing books and in the mouths of politicians. The leads are stronger in the pre-1941 era: a doubling in the frequency with which a phrase is used in Congress is associated with an increase of 5.6 percent in the use of that phrase in Google Books. In the post-1941 period that number drops to around 1.4 percent and loses statistical significance at conventional levels. Narrow economic phrases (phrases on economic or tax-related topics) represent 10 to 15 percent of our persistently polarized phrases. There are no significant instances where the use of polarized economic phrases in Congress leads their use in Google Books. High correlations are found between the use of social phrases in Congress and future use in Google Books, but these are statistically significant only at the 10 percent level. Again, a doubling in use of bipartisan phrases in Congress is associated with an almost 10 percent increase in use in Google Books within 2 years. For more-polarized social phrases, increased use in Congress seems to be negatively correlated with future use in books, but the coefficients do not reach statistical significance at even the 10 percent level.

Table 7 shows the partial correlation of lags of Google Ngram frequencies with congressional frequencies. Again there is a very strong and robust momentum in congressional phrase use. The coefficients are similar in size to those for the correlations of phrase use in Google Books with lagged Google Books phrase use. Bipartisan language does not seem to flow from

Table 6. Regressions Estimating Diffusion of Polarized Phrases from Congress to Google Books[a]

	Dependent variable: frequency of phrase in Google Books at time t						
Independent variable	*Straight panel*	*Panel with lagged dependent variable*	*Pre-1941*	*1941 and after*	*Narrow economic issues*	*Broad economic issues*	*Social issues*
Congressional frequency$_{t-1}$	0.0630***	0.0311***	0.0458*	0.0270***	0.0140	0.0205	0.0949**
	(0.0117)	(0.00716)	(0.0194)	(0.00673)	(0.0151)	(0.0162)	(0.0302)
Congressional frequency$_{t-2}$	0.0283**	–0.00299	0.00772	–0.00372	–0.000393	0.0152	0.00830
	(0.00986)	(0.00406)	(0.00960)	(0.00365)	(0.00411)	(0.0121)	(0.0800)
Congressional frequency$_{t-3}$	0.0424***	–0.00815	0.00203	–0.00882	–0.00127	–0.00594	0.00111
	(0.00864)	(0.00553)	(0.0144)	(0.00544)	(0.00277)	(0.00805)	(0.0474)
(Congressional frequency × phrase in 25th–50th percentile)$_{t-1}$	–0.00652	–0.00284	–0.0193	0.00215	–0.000398	0.0352	–0.0877**
	(0.0248)	(0.0124)	(0.0235)	(0.0135)	(0.0194)	(0.0450)	(0.0304)
(Congressional frequency × phrase in 25th–50th percentile)$_{t-2}$	–0.00765	–0.00228	–0.0124	0.00109	–0.00365	–0.0274*	0.00310
	(0.0154)	(0.00506)	(0.0105)	(0.00480)	(0.00690)	(0.0137)	(0.0809)
(Congressional frequency × phrase in 25th–50th percentile)$_{t-3}$	–0.0124	0.00142	–0.00405	–0.000177	0.00100	0.00903	–0.0156
	(0.0138)	(0.00637)	(0.0149)	(0.00609)	(0.00414)	(0.0155)	(0.0485)
(Congressional frequency × phrase in 50th–75th percentile)$_{t-1}$	–0.0155	–0.0153	–0.0264	–0.0117	–0.00665	–0.00305	–0.0755*
	(0.0175)	(0.00858)	(0.0208)	(0.00732)	(0.0167)	(0.0195)	(0.0310)
(Congressional frequency × phrase in 50th–75th percentile)$_{t-2}$	0.00262	0.00252	–0.00474	0.00327	–0.00445	0.00962	–0.0186
	(0.0154)	(0.00673)	(0.0129)	(0.00452)	(0.00821)	(0.0282)	(0.0802)
(Congressional frequency × phrase in 50th–75th percentile)$_{t-3}$	–0.000878	0.00430	0.00457	–0.00201	0.00907	–0.0234	–0.0117
	(0.0159)	(0.00622)	(0.0150)	(0.00549)	(0.00701)	(0.0258)	(0.0487)
(Congressional frequency × phrase in 75th–99th percentile)$_{t-1}$	–0.0407**	–0.0178*	–0.0320	–0.0122	–0.00477	–0.0103	–0.0778*
	(0.0128)	(0.00781)	(0.0203)	(0.00746)	(0.0162)	(0.0174)	(0.0325)
(Congressional frequency × phrase in 75th–99th percentile)$_{t-2}$	–0.0176	0.00202	–0.00528	–0.00341	–0.00213	–0.0253	0.0133
	(0.0110)	(0.00567)	(0.0106)	(0.00518)	(0.00564)	(0.0143)	(0.0811)
(Congressional frequency × phrase in 75th–99th percentile)$_{t-3}$	–0.0387***	–0.000900	–0.0107	0.00502	0.000518	0.0134	–0.0315
	(0.00941)	(0.00643)	(0.0146)	(0.00557)	(0.00353)	(0.0101)	(0.0490)

(*continued*)

Table 6. Regressions Estimating Diffusion of Polarized Phrases from Congress to Google Books[a] (*Continued*)

	Dependent variable: frequency of phrase in Google Books at time t						
Independent variable	*Straight panel*	*Panel with lagged dependent variable*	*Pre-1941*	*1941 and after*	*Narrow economic issues*	*Broad economic issues*	*Social issues*
Google Books frequency$_{t-1}$		0.449***	0.359***	0.456***	0.364***	0.518***	0.494***
		(0.0388)	(0.0457)	(0.0614)	(0.0552)	(0.0584)	(0.0337)
Google Books frequency$_{t-2}$		0.290***	0.233***	0.281***	0.298***	0.302***	0.395***
		(0.0234)	(0.0204)	(0.0296)	(0.0403)	(0.0577)	(0.0781)
Google Books frequency$_{t-3}$		0.143***	0.0977***	0.154***	0.0429	0.0681	–0.0138
		(0.0283)	(0.0273)	(0.0225)	(0.0723)	(0.0348)	(0.0102)
No. of observations	2,533,513	2,533,513	1,287,523	1,245,990	30,805	65,392	16,165
R^2	0.014	0.643	0.352	0.605	0.407	0.698	0.689
Sum of coefficients for congressional frequency, all lags							
0–25th percentiles	0.134***	0.0200***	0.0555***	0.0144	0.0123	0.0297	0.104*
	0.0275	0.00773	0.0137	0.00878	0.0125	0.0270	0.0630
25th–50th percentiles	–0.0266	–0.00369	–0.0357**	0.00306	–0.00304	0.0168	–0.100
	0.0512	0.00885	0.0154	0.0124	0.0151	0.0418	0.0625
50th–75th percentiles	–0.0138	–0.00852	–0.0266	–0.0104*	–0.00202	–0.0168	–0.106*
	0.0458	0.00668	0.0163	0.00564	0.0135	0.0287	0.0624
75th–99th percentiles	–0.0970***	–0.0167**	–0.0480***	–0.0106	–0.00639	–0.0222	–0.0960
	0.0294	0.00669	0.0140	0.00731	0.0132	0.0274	0.0617

Source: Authors' regressions using data from the digitized *Congressional Record* and Google Books.

a. Data are from a balanced biennial phrase-level panel covering 1879–1999. All specifications include time and phrase fixed effects. The first and second columns use the full panel, the third and fourth divide the panel by period at 1941, and the last three columns use only phrases related to the indicated set of issues. Phrase-clustered standard errors are in parentheses. Asterisks indicate statistical significance at the *0.05, **0.01, and ***0.001 level.

Table 7. Regressions Estimating Diffusion of Polarized Phrases from Google Books to Congress[a]

	Dependent variable: frequency of phrase in the digitized Congressional Record at time t						
Independent variable	*Straight panel*	*Panel with lagged dependent variable*	*Pre-1941*	*1941 and after*	*Narrow economic issues*	*Broad economic issues*	*Social issues*
Google Books frequency$_{t-1}$	0.0420***	0.000920	0.00623*	–0.00593	0.866	0.0604	–0.0338
	(0.00735)	(0.00351)	(0.00312)	(0.0101)	(1.009)	(0.0643)	(0.0458)
Google Books frequency$_{t-2}$	0.0159***	–0.00539	–0.00498	0.00217	–0.281	–0.00671	0.0643
	(0.00333)	(0.00348)	(0.00439)	(0.00524)	(1.068)	(0.115)	(0.0591)
Google Books frequency$_{t-3}$	0.00498	–0.00163	–0.00356*	0.00345	–0.480	–0.0208	0.0998
	(0.00338)	(0.00262)	(0.00165)	(0.00911)	(0.511)	(0.0963)	(0.0717)
(Google Books frequency × phrase in 25th–50th percentile)$_{t-1}$	–0.00251	–0.00464	–0.0167	0.0127	–0.906	–0.0206	0.272
	(0.0224)	(0.00878)	(0.0132)	(0.0141)	(1.020)	(0.0716)	(0.145)
(Google Books frequency × phrase in 25th–50th percentile)$_{t-2}$	0.00152	0.00592	0.000998	0.00914	0.262	0.0343	–0.185
	(0.0129)	(0.00524)	(0.00702)	(0.0121)	(1.083)	(0.123)	(0.199)
(Google Books frequency × phrase in 25th–50th percentile)$_{t-3}$	–0.00286	0.00834	0.00947*	0.00227	0.475	0.0688	–0.138
	(0.00769)	(0.00636)	(0.00477)	(0.0146)	(0.520)	(0.0995)	(0.131)
(Google Books frequency × phrase in 50th–75th percentile)$_{t-1}$	0.0661	0.00558	0.0147	0.0178	–0.878	–0.191*	0.000562
	(0.0365)	(0.0101)	(0.0128)	(0.0148)	(1.010)	(0.0862)	(0.0673)
(Google Books frequency × phrase in 50th–75th percentile)$_{t-2}$	0.0700*	0.0203	0.0369*	–0.0146	0.403	0.135	0.207
	(0.0276)	(0.0140)	(0.0188)	(0.0116)	(1.078)	(0.215)	(0.108)
(Google Books frequency × phrase in 50th–75th percentile)$_{t-3}$	0.0667**	0.00621	0.0140	0.00480	0.546	0.00217	–0.254**
	(0.0246)	(0.0106)	(0.0120)	(0.0127)	(0.524)	(0.174)	(0.0822)
(Google Books frequency × phrase in 75th–99th percentile)$_{t-1}$	0.0652**	0.0105	0.00194	0.0270	–0.675	0.00972	0.0553
	(0.0248)	(0.0131)	(0.0166)	(0.0167)	(1.015)	(0.0694)	(0.0610)
(Google Books frequency × phrase in 75th–99th percentile)$_{t-2}$	0.0265	–0.00560	–0.0116	–0.00207	0.214	–0.0459	–0.0593
	(0.0253)	(0.0137)	(0.0163)	(0.0155)	(1.071)	(0.120)	(0.0654)
(Google Books frequency × phrase in 75th–99th percentile)$_{t-3}$	0.0254	0.0182*	0.0185*	–0.00158	0.643	0.0467	–0.0946
	(0.0160)	(0.00787)	(0.00783)	(0.0138)	(0.517)	(0.101)	(0.0717)

(*continued*)

Table 7. Regressions Estimating Diffusion of Polarized Phrases from Google Books to Congress[a] (*Continued*)

	Dependent variable: frequency of phrase in the digitized Congressional Record at time t						
Independent variable	*Straight panel*	*Panel with lagged dependent variable*	*Pre-1941*	*1941 and after*	*Narrow economic issues*	*Broad economic issues*	*Social issues*
Congressional frequency$_{t-1}$		0.513***	0.435***	0.503***	0.559***	0.540***	0.483***
		(0.0327)	(0.0417)	(0.0434)	(0.0591)	(0.0668)	(0.0881)
Congressional frequency$_{t-2}$		0.170***	0.147***	0.111***	0.132*	0.227***	0.134**
		(0.0196)	(0.0246)	(0.0292)	(0.0648)	(0.0525)	(0.0457)
Congressional frequency$_{t-3}$		0.121***	0.103***	0.0368	0.0782*	0.0468	0.141
		(0.0242)	(0.0284)	(0.0323)	(0.0384)	(0.0319)	(0.0837)
No. of observations	2,533,513	2,533,513	1,287,523	1,245,990	30,805	65,392	16,165
R^2	0.011	0.550	0.353	0.369	0.515	0.570	0.470
Sum of coefficients for Google Books frequency, all lags							
0–25th percentiles	0.0628***	–0.00610*	–0.00231	–0.000310	0.104	0.0329***	0.130***
	0.00957	0.00355	0.00469	0.00662	0.182	0.00923	0.0313
25th–50th percentiles	–0.00385	0.00962	–0.00619	0.0241	–0.169	0.0825*	–0.0501
	0.0390	0.00756	0.0140	0.0269	0.184	0.0463	0.109
50th–75th percentiles	0.203**	0.0321**	0.0657***	0.00798	0.0711	–0.0545	–0.0464
	0.0809	0.0134	0.0193	0.00949	0.200	0.0391	0.0441
75th–99th percentiles	0.117**	0.0231***	0.00882	0.0234*	0.183	0.0105	–0.0986***
	0.0518	0.00885	0.0102	0.0130	0.184	0.0179	0.0322

Source: Authors' regressions using data from the digitized *Congressional Record* and Google Books.

a. Data are from a balanced biennial phrase-level panel covering 1879–1999. All specifications include time and phrase fixed effects. The first and second columns use the full panel, the third and fourth divide the panel by period at 1941, and the last three columns use only phrases related to the indicated set of issues. Phrase-clustered standard errors are in parentheses. Asterisks indicate statistical significance at the *0.05, **0.01, and ***0.001 level.

books to Congress. However, there is evidence that very polarized language does flow from books to Congress. A doubling of phrase use in the top quartile of phrase polarization is associated with a subsequent rise in congressional use of a little over 2 percent.

Together our results suggest that intellectuals are more likely to be an autonomous engine of polarization in Congress than Congress is an engine for polarization in books. This is particularly true in economic matters. However, even for narrow economic issues, where the coefficients are larger, their magnitudes still seem incapable of explaining the much stronger increase in congressional polarization than in polarization in Google Books in recent years. Most important, our results should be taken as suggestive and preliminary rather than definitive.

VI. Limitations

It is quite possible that our "distant reading" of congressional speech is missing important changes in the data-generating process over time. Two such changes that quickly come to mind are institutional changes to House rules or party practices that may have altered the instrumental use of political speech on the floor of the House, and changes in media coverage of politics. As an example of the effect of the institutional structure on congressional speech, consider the issue of "unconstrained floor time" allowed politicians in Congress. Forrest Maltzman and Lee Sigelman (1996, p. 828) find that

> special orders and short speeches serve as potential tools of policy influence within the House . . . that party leaders, ideological extremists and minority party members resort to such speeches suggests that structured floor debate inadequately serves many members' policy goals. Unstructured floor speeches provide those members whose views are largely ignored in a majoritarian institution such as the House the opportunity to participate, and may thus serve as something of an institutional safety valve.

The media available and the coverage given national politics may also alter both members' incentives for speech and the mechanisms for its diffusion. Since the advent of extensive television coverage of Congress by C-SPAN in 1979, congressional debate may have transmuted into congressional performance, with members speaking into the camera for the benefit of domestic constituents rather than to each other for purposes of persuasion. In fact, Jonathan Morris (2001, p. 102), in an analysis of 1-minute unconstrained speeches, finds that "even though the practice of granting one minute on the floor began long before live television coverage of Congress had started, its use has increased significantly since

1979, when cameras were placed in the House chamber." It is an open empirical question whether similar changes in the production of speech accompanied changes in other media such as newspapers and radio.

There are also important limitations to our data. We truncate our data to a maximum of 10,000 phrases per year, chosen by rank based on a χ^2 measure of phrase partisanship. As we describe in appendix B, we find that some of our results are not robust to the measure we use to select phrases. Moreover, our results may also be sensitive to the number of phrases we select. We also have relied in this paper largely on trigrams, which may not be optimal for identifying meaningful patterns from plain text. An important next step is to examine the robustness of our results to incorporation of different-length n-grams (such as unigrams and bigrams) or skipgrams (lists of nonconsecutive words). Other, more sophisticated methods of transforming language into data, such as parts-of-speech tagging, might also be useful. In particular, capturing whether a phrase is used in a positive or a negative connotation might dramatically reduce measurement error in the labeling of phrases as partisan. We have also used only a single criterion, the party of House members, to score the partisanship of phrases. Using other measures, including presidential vote share by congressional district, ratings by groups such as the Americans for Democratic Action and the American Conservative Union, and DW-NOMINATE scores is an important future step. Bringing party platforms and Senate and presidential speech into the analysis is another potential extension.

Google Books, although more comprehensive than any other database we know of, does not provide any information about the authors, publishers and their locations, or subjects of the books underlying the phrase counts it generates in any given year. Moreover, we know little about the process by which books are selected for inclusion in the corpus. We hope to eventually use historical text databases with more such information to look at subtler patterns of diffusion of partisan ideas. For example, electronic databases of newspapers dating back to the 19th century exist that we could use to track the geographical origins of certain phrases. We are also in the process of obtaining corpora of historical books with associated author data, which would allow us to look directly at the patterns of phrase diffusion among political public intellectuals, Congress, and the public at large.

In the interest of space and clarity, our methodology here has been extremely simple, ignoring many features of text that may allow for finer estimates of phrase partisanship. We have experimented with other methods that yield different results, and more research is needed to validate these different estimators substantively. Among the things we experimented

with were a sparsity-augmented model developed by Taddy (2010), which forces near-zero correlations to be exactly zero, eliminating some noise in the measure. We also attempted to distinguish partisan subjects of speech from partisan discussions of a bipartisan subject. We did this by estimating a topic model (Blei and Li 2010), which statistically infers the topic of speech from the *Congressional Record,* as well as a simple sentiment analysis algorithm to extract the tone of speech from the recorded text (Pang and Lee 2008) and infer the sentiment (positive or negative) with which a phrase is spoken or written. For brevity and simplicity of exposition, we do not report results using these alternative methods. However, all of these approaches are worth further investigation, perhaps with modifications to take into account the particular conventions in congressional speech.

Finally, we have reported only correlations. The entire project relies on an assumption that the correlation between party and phrase use is driven by some underlying ideological preference or belief. Although this is plausible, it is also possible that ideology is a very poor predictor of speech patterns, and that in fact language use reflects other variables that we have not incorporated into the analysis, such as the geographical origin, race, or sex of the speaker or the characteristics of the audience. At another level, in our estimates of the relationships in both the aggregate time-series and the phrase-level panel, we make no pretense of using only identifying variation. It could be that latent changes in language, driven for example by broad cultural changes or waves of immigration, are driving all of our results. We hope to revisit this question in the future, but we believe more progress will be made using a combination of the data we have constructed and the kinds of panel regressions we explored in the last section, perhaps at higher frequency and combined with more convincingly exogenous variation.

VII. Conclusion

In this paper we have extended and combined the literature on text analysis of congressional speech with measures of the partisanship of phrases to extract partisan and polarizing language from the *Congressional Record* since Reconstruction. We have used these partisan phrases together with the Google Books corpus to construct a measure of polarization of political discourse, which we then used to study the dynamics of political language. This measure is complementary to those based on roll-call voting and other measures, as it allows ideology to be measured across many domains of the printed word as well as over long periods of time. We find that although

political discourse became substantially more polarized in the late 1990s, the increase was smaller than the increase of polarization in Congress over the same time period.

We explore the diffusion of political phrases in a limited way by estimating the dynamic relationship between the frequency of a phrase in the Google Books corpus and its polarization in Congress. An intriguing hypothesis is that Congress is itself driving polarization in political ideas outside of Congress. This has potentially interesting implications for the endogeneity of public opinion and political discourse to politics itself. What we find, however, is that, at least in the case of the House of Representatives, although Congress might lead the printed word in the use of bipartisan phrases, the increased use of polarized phrases in Congress is followed by declines, not increases, in the use of those phrases in books. Moreover, there is some evidence that intellectual political discourse, as measured in Google Books, anticipates the language used in Congress, particularly in the latter half of our sample period. An interesting but statistically weak result is that this effect seems particularly strong for economic phrases, and weak for social phrases, suggesting that although Congress may take its economic language from public intellectuals, it does not adopt its language on social issues from the same sources. Although polarized political discourse may influence congressional speech and legislative gridlock, this effect seems not to be driving the recent increase in congressional polarization.

The data set we have developed allows us to project the partisanship of phrases (imputed from the things partisan politicians say) onto many other textual domains where data are available. Extensions of this methodology could explore the writings of public intellectuals, the sermons of church leaders, the work of think tank scholars, and the op-eds of media pundits. We are interested in the relative influence of these opinion leaders in part to determine whether and how much the words of scribblers matter. Here a historical approach is important, as the ideas that deeply influence policy likely gestate over a number of years, evolving from idiosyncratic intellectual bon mots to catchphrases used in Congress to sell policy. In addition, the institutions mediating the relationship between policymaking and political discourse may have changed substantially over time, and a study of this relationship in the past may provide clues to the impact of new media or think tanks on policymaking.

Why study political ideas and political writing in the first place? An old question, which we do not pretend to answer, is whether ideas or interests drive political action. For Karl Marx and some other economists, for

example, interests were primary—this is partly the point of economic materialism—and politics and ideology merely reflect economic structure and demands. But as our epigraphs suggest, this perspective is by no means universally held, and it could be that ideas and opinions are simply autonomous or endogenous to political institutions. Our own intuition is that ideas expressed in words must have some of their own momentum and power, but that there are likely important background material conditions through which groups and individuals modify these ideas and make their propagation more or less likely. We hope that future work will use linguistic measures of ideology to better identify the sources of ideological change over the long run and across different groups within society.

APPENDIX A

Data Sources

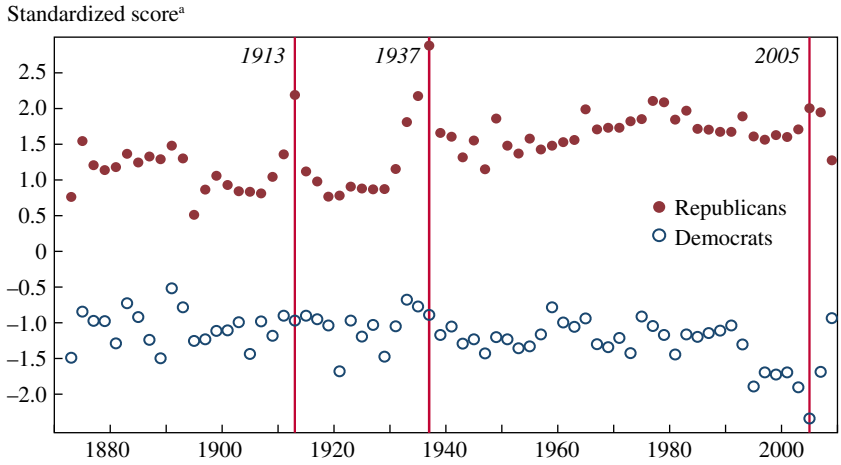
A chief contribution of this paper is the development of measures to tease out the political polarity and partisanship of text. We consider two sources of text: the *Congressional Record* from 1873 to 2011, and Google Ngrams, which begins in 1520 and ends in 2008 and is drawn from Google Books, a large corpus of books digitized by Google. We also relate the imputed polarization from these sources to measures of voter ideology (captured by Gallup polls) and to various important political events and economic trends.

A.1. The Congressional Record

We begin with plain-text versions of the *Congressional Record,* which is the official written record of the proceedings of both the House of Representatives and the Senate. The scanned series begins in 1873, although digitized versions have been made available by the U.S. government only since 1994. We downloaded plain-text versions of earlier sessions from databases accessed via Columbia University. An important caveat is that the *Congressional Record* can be amended with "nonsubstantial changes" by the office of the member making the remarks before publication at the end of the session, so the *Congressional Record* contains speech that was not necessarily actually uttered on the floor. The *Congressional Record* has also grown substantially over time, from 5,500 pages in 1873–74 to over 40,000 in the 1970s (the peak).

First we preprocess the data. We remove capitalization and then use a Porter stemmer (Porter 1980). This is a common first step in natural language processing. The stemmer converts words into common roots (suffixes

Figure A.1. Average Within-Party Partisanship, by Congress, 1873–2007



Source: Authors' calculations using data from the digitized *Congressional Record.*
a. Standardized to have a mean of zero and a variance of 1.

like "-ing" and "-ly" get removed) and strips out "stopwords" (common words such as articles and conjunctions). We then convert the text into three-word phrases called trigrams. These are recorded by speaker. We were unable to match about 2 percent of congressional speech with the speaker because of transcription errors. We also dropped any speech by a member of length less than 10 trigrams, and any phrase that did not appear in the Google corpus at least 2,000 times. We then select the top 10,000 phrases ranked by the χ^2 measure described in the text for each Congress. In a few early Congresses, there were fewer than 10,000 phrases with a positive χ^2. In those cases we included all phrases with a positive χ^2 and selected the remaining phrases at random.

Figure A.1 uses the *Congressional Record* data to score each of the major parties in each Congress by its degree of partisanship.

A.2. Google Ngrams

Our second source of textual data is Google Ngrams.[6] The details of the construction of the Google Books corpus are set out in Michel and others (2011), but we briefly describe the data here. We use the "American" corpus, which does not filter by subject or cap the number of books from

6. The data are available at books.google.com/ngrams/datasets.

any year. The corpus contains over 5 million digitized books, equivalent to over 4 percent of the total number of distinct books in print. The first year of our sample is 1873, and Google Books has 3,978 books published in that year. The year with the largest number of published books was 2008, with 149,373 books. However, the authors of the database recommend that the corpus be used primarily between 1800 and 2000, as around 2000 the inception of the Google Books project itself changed the composition of the corpus in subtle ways.[7] Thus, we restrict all of our time series using the Google Ngrams data set to between 1873 and 2000.

Google Ngrams tells us, for a set of *n* consecutive word stems, how many times that set appeared in books digitized by Google in a particular year. Punctuation (aside from apostrophes) is separated out into its own tokens, so that, for instance, the set of trigrams for the sentence "The cat sat on the mat!" would be {The.cat.sat, cat.sat.on, sat.on.the, on.the.mat, the.mat.!}.

We use the Google Ngrams corpus for two distinct purposes. First, we use it to filter the phrases in the *Congressional Record* that meet a threshold of at least 2,000 appearances in the entire Google Books corpus. Thus, for example, we exclude trigrams that are so infrequent that they appear fewer than 2,000 times in Google Books. Our view is that trigrams so idiosyncratic are likely to reflect such things as typographical errors. Second, we use the Google Ngrams as measures of the penetration of the partisan and polarized phrases in political discourse among intellectual elites.

A.3. Voter Ideology

To proxy for public ideology, we rely on a Gallup poll covering taxation that dates back to 1957. The question asked is, "Do you consider the amount of federal income tax you have to pay as too high, about right, or too low?" and the data report percentage shares for each of these responses.

A.4. Macrocorrelates

For our macroeconomic data we use the methodology outlined by Joseph Davis (2004) for combining the series compiled by Jeffrey Miron and Christina Romer (1990), the Federal Reserve, and Davis himself to construct a measure of real GDP growth during each Congress. We have also used data from *Historical Statistics of the United States* (HSUS) *Millenial Edition Online* (Carter and others 2006) as a robustness check. The

7. See "Culturomics" (www.culturomics.org/Resources/A-users-guide-to-culturomics).

HSUS also reports data on real GDP from 1872 to 2002, which we combine with recent data from the Bureau of Economic Analysis to calculate annual real GDP growth rates.

Statistics on U.S. military casualties are taken from the Correlates of War (COW) project. We use the COW wars v4.0 (1873–2007), summing deaths from the Intra-State War Data, the Inter-State War Data, and the Extra-State War Data, to generate a time series of U.S. war deaths. Since the COW project reports deaths by conflict and not by year, we assign deaths equally over each year of a conflict. We normalize the data by the total resident population of the United States so that our variable for war deaths is military casualties per 100,000 people.

The data on political violence are drawn from Peter Turchin's U.S. Political Violence database (Turchin 2012). We sum for each year the fatalities from riots, lynchings, and terrorism, again normalizing by population.

APPENDIX B

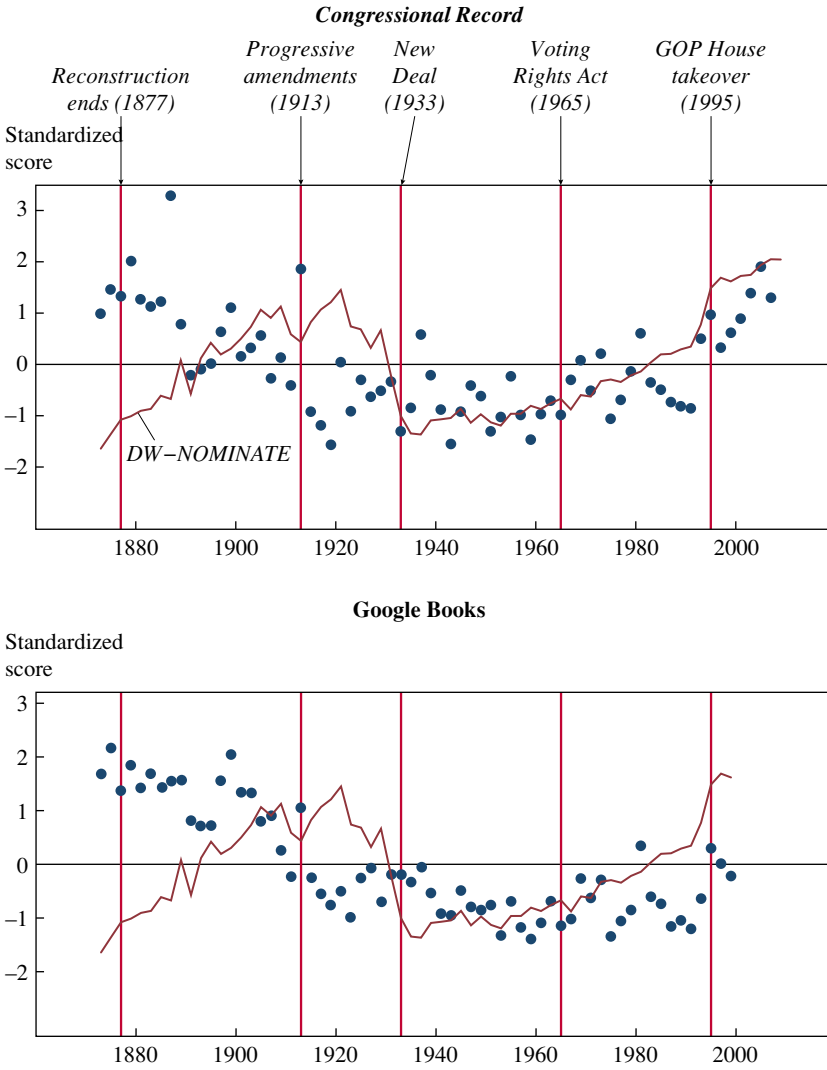
Alternative Phrase Selection Methods

In the main analysis, we followed Gentzkow and Shapiro (2010) in selecting the top 10,000 phrases in each year according to a χ^2 statistic. This method selects for phrases that are both frequent and disproportionately used by one party. It has the virtue of being computationally simple, so that very few calculations are required for each phrase. In this appendix we report results from two alternative ways of selecting the top 10,000 phrases: according to frequency and according to the t statistic of the correlation. We do this to illustrate the sensitivity of our results to this prima facie arbitrary choice. Although some of our results are robust, we take this appendix as showing the necessity of further research on methods.

Appendix figure B.1 replicates the aggregate time-series graphs for the sample of phrases selected by restricting the sample to the top 10,000 by frequency (the frequency-restricted sample); appendix figure B.2 replicates the aggregate time-series graphs for the sample of phrases selected by restricting the sample to the top 10,000 by t statistic (which we call the t-stat-restricted sample). The t statistic is calculated by dividing the correlation coefficient of a phrase by its standard error.

The patterns in the time series calculated from the frequency-restricted sample are broadly similar to those generated from the χ^2-restricted sample,

Figure B.1. Polarization Measured Using the Frequency-Based Threshold and by DW-NOMINATE, 1873–2007[a]

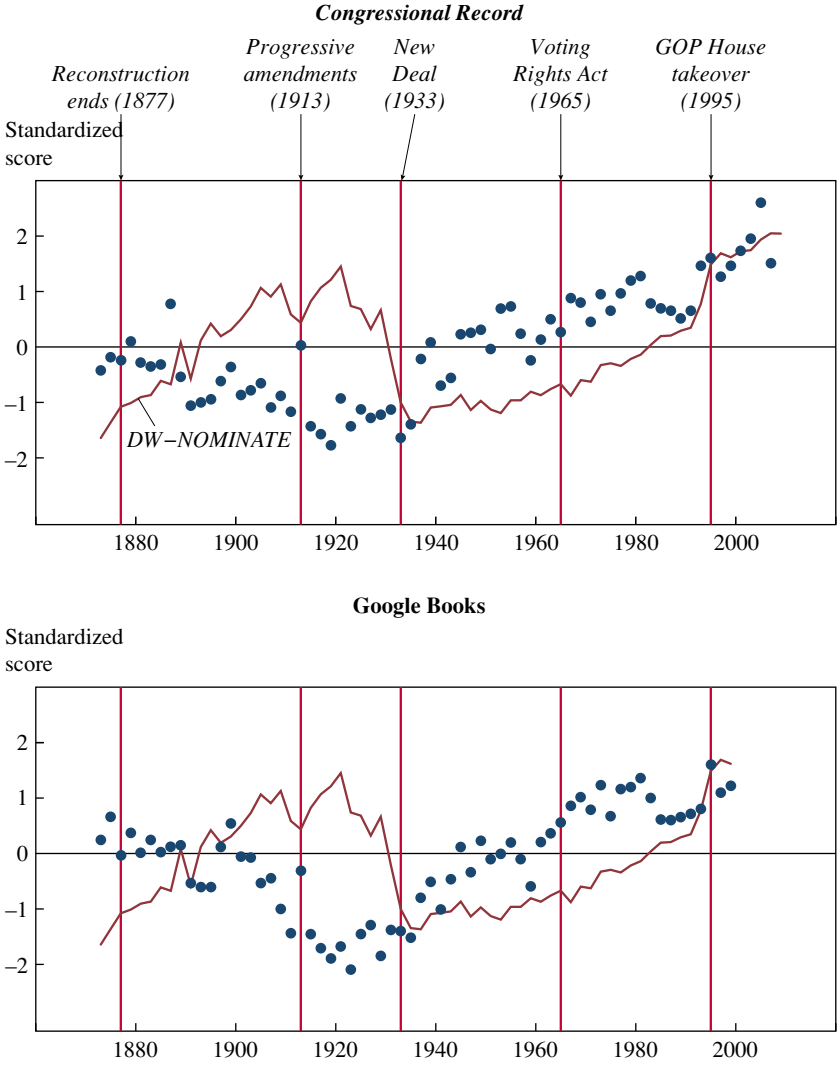


Sources: Authors' calculations using data from the digitized *Congressional Record,* Google Books, and the legislator estimates on voteview.com/dwnomin.htm.

a. All measures are standardized to have a mean of zero and a variance of 1.

b. The 16th (income tax) and 17th (direct election of senators) Amendments to the U.S. Constitution.

Figure B.2. Polarization Measured Using *t*-Statistic-Based Threshold and by DW-NOMINATE, 1873–2007[a]



Sources: Authors' calculations using data from the digitized *Congressional Record,* Google Books, and the legislator estimates on voteview.com/dwnomin.htm.

a. All measures are standardized to have a mean of zero and a variance of 1.

with both the congressional and the Google Books measures showing their peaks in the late 19th century, falling toward the middle of the 20th century, and then increasing again, with the recent increase much more muted in Google Books phrases than in the congressional phrases.

However, the patterns in the time series calculated from the *t*-stat-restricted sample are noticeably different. The congressional and the Google Books time series look much more similar, and both show a decreasing trend between Reconstruction and World War I and a monotonic increase from the New Deal to the end of the series. Some of the differences between the *t*-stat-restricted series and the two others are likely to be mechanical. As noted in the text, over time there has been a dramatic increase in the size of the *Congressional Record,* because of which the number of phrases surviving our filters quintupled. As a consequence, the polarization of the 10,000th phrase has increased, and thus the average polarization has risen. This is less likely to be true of the frequency cut, which does not select on polarization, or of the χ^2 cut, which does so to a lesser degree than the *t* statistic cut. Perhaps a better dynamic estimation strategy would be to take the top 50 percent of polarized phrases across years rather than the top 10,000 phrases. We leave this for future research.

We next show replications of the panel regressions in tables 6 and 7 for these alternative phrase samples. Appendix tables B.1 and B.2 replicate tables 6 and 7 for the frequency cut sample, and tables B.3 and B.4 do so for the *t*-stat-restricted sample. Our results for the relationship between congressional language use and subsequent use in books are generally qualitatively although not statistically robust. In particular, although our results showing that Congress leads Google Books for bipartisan phrases are robust across both the *t*-stat-restricted and the frequency cut samples, the coefficients are not always statistically significant. The same is true for the result showing that increased use of the top quartile of polarized phrases in Congress is correlated with a subsequent decline in use in Google Books.

The results showing anticipation of future language use in Congress from language use in Google Books is generally less robust. Across samples, we still find that an increase in use of very polarized phrases in Google Books is strongly correlated with future use in Congress. However, the results, despite relatively sizable coefficients, are not statistically significant. The same correlation for the entire sample is neither qualitatively nor statistically robust across the sample. The results on polarized economic language in Google Books showing up in Congress with a lag are

Table B.1 Regressions Estimating the Diffusion of Polarized Phrases from Congress to Google Books, Using a Frequency-Based Threshold[a]

	Dependent variable: frequency of phrase in Google Books at time t						
Independent variable	*Straight panel*	*Panel with lagged dependent variable*	*Pre-1941*	*1941 and after*	*Narrow economic issues*	*Broad economic issues*	*Social issues*
Congressional frequency$_{t-1}$	0.0592*	0.0292***	0.0356	0.0231***	0.0415*	0.00641	0.0462*
	(0.0259)	(0.00865)	(0.0198)	(0.00420)	(0.0210)	(0.00368)	(0.0199)
Congressional frequency$_{t-2}$	0.0276	–0.00340	0.00426	–0.00705*	0.0165	–0.0139	0.0443
	(0.0201)	(0.00639)	(0.0152)	(0.00288)	(0.0115)	(0.00747)	(0.0330)
Congressional frequency$_{t-3}$	0.0469	–0.00733	–0.00297	–0.00847*	–0.0194	0.00903	–0.0462
	(0.0304)	(0.00436)	(0.00694)	(0.00381)	(0.0115)	(0.00688)	(0.0262)
(Congressional frequency × phrase in 25th–50th percentile)$_{t-1}$	–0.0115	–0.0133	–0.0171	–0.00690	–0.0353	0.0253	–0.0294
	(0.0296)	(0.0104)	(0.0217)	(0.00542)	(0.0219)	(0.0163)	(0.0258)
(Congressional frequency × phrase in 25th–50th percentile)$_{t-2}$	0.00352	0.00548	0.00192	0.00600	–0.0208	0.0545	–0.0589
	(0.0239)	(0.00894)	(0.0181)	(0.00470)	(0.0128)	(0.0371)	(0.0387)
(Congressional frequency × phrase in 25th–50th percentile)$_{t-3}$	–0.00704	–0.000109	0.00557	–0.00251	0.0254*	–0.0541	0.0490
	(0.0337)	(0.00565)	(0.00841)	(0.00455)	(0.0125)	(0.0318)	(0.0264)
(Congressional frequency × phrase in 50th–75th percentile)$_{t-1}$	–0.0274	–0.0141	–0.0194	–0.00646	–0.0336	–0.000736	–0.0414*
	(0.0272)	(0.00984)	(0.0208)	(0.00582)	(0.0223)	(0.00515)	(0.0201)
(Congressional frequency × phrase in 50th–75th percentile)$_{t-2}$	–0.0144	–0.00260	–0.00615	–0.00369	–0.0222	0.00739	–0.0436
	(0.0206)	(0.00692)	(0.0155)	(0.00704)	(0.0136)	(0.00951)	(0.0334)
(Congressional frequency × phrase in 50th–75th percentile)$_{t-3}$	–0.0329	0.00305	0.000374	0.00539	0.0198	–0.00436	0.0380
	(0.0307)	(0.00544)	(0.00754)	(0.00434)	(0.0114)	(0.00821)	(0.0264)
(Congressional frequency × phrase in 75th–99th percentile)$_{t-1}$	–0.0305	–0.00986	–0.0162	–0.00238	–0.0185	0.0346	–0.0231
	(0.0277)	(0.00997)	(0.0208)	(0.00899)	(0.0231)	(0.0242)	(0.0251)

(Congressional frequency × phrase in 75th–99th percentile)$_{t-2}$	−0.0157	0.00171	−0.00393	0.00690	−0.0158	0.00398	−0.0210
	(0.0215)	(0.00912)	(0.0173)	(0.00391)	(0.0117)	(0.0176)	(0.0375)
(Congressional frequency × phrase in 75th–99th percentile)$_{t-3}$	−0.0441	−0.00527	−0.00694	−0.00297	0.0180	0.00403	0.00987
	(0.0309)	(0.00549)	(0.00820)	(0.00581)	(0.0124)	(0.0173)	(0.0336)
Google Books frequency$_{t-1}$		0.451***	0.361***	0.460***	0.360***	0.515***	0.494***
		(0.0393)	(0.0461)	(0.0623)	(0.0549)	(0.0587)	(0.0336)
Google Books frequency$_{t-2}$		0.290***	0.233***	0.282***	0.298***	0.301***	0.395***
		(0.0237)	(0.0206)	(0.0299)	(0.0406)	(0.0581)	(0.0781)
Google Books frequency$_{t-3}$		0.142***	0.0978***	0.152***	0.0440	0.0688	−0.0139
		(0.0287)	(0.0277)	(0.0230)	(0.0721)	(0.0352)	(0.0103)
No. of observations	2,395,104	2,395,104	1,217,184	1,177,920	30,805	65,392	16,165
R^2	0.014	0.647	0.355	0.611	0.410	0.700	0.689
Sum of coefficients for Congressional frequency, all lags							
0–25th percentiles	0.134*	0.0185	0.0369	0.00753	0.0386*	0.00151	0.0442
	(0.0754)	(0.0129)	(0.0332)	(0.00600)	(0.0229)	(0.00101)	(0.0306)
25th–50th percentiles	−0.0150	−0.00789	−0.00966	−0.00341	−0.0307	0.0257	−0.0393
	(0.0852)	(0.0121)	(0.0347)	(0.00336)	(0.0232)	(0.0209)	(0.0304)
50th–75th percentiles	−0.0747	−0.0137	−0.0252	−0.00476	−0.0360	0.00230	−0.0469
	(0.0770)	(0.0121)	(0.0336)	(0.00671)	(0.0231)	(0.00276)	(0.0304)
75th–99th percentiles	−0.0903	−0.0134	−0.0271	0.00155	−0.0163	0.0426**	−0.0342
	(0.0775)	(0.0123)	(0.0334)	(0.00545)	(0.0243)	(0.0199)	(0.0309)

Source: Authors' regressions using data from the digitized *Congressional Record* and Google Books.

a. Data are from a balanced biennial phrase-level panel covering 1879–1999. Phrases are selected for inclusion according to frequency rather than by the χ^2 method used in tables 6 and 7 in the text. All specifications include time and phrase fixed effects. The first and second columns use the full panel, the third and fourth divide the panel into periods at 1941, and the last three columns use only phrases related to the indicated set of issues. Phrase-clustered standard errors are in parentheses. Asterisks indicate statistical significance at the *0.05, **0.01, and ***0.001 level.

Table B.2. Regressions Estimating the Diffusion of Polarized Phrases from Google Books to Congress, Using a Frequency-Based Threshold[a]

	Dependent variable: frequency of phrase in the digitized Congressional Record at time t						
Independent variable	*Straight panel*	*Panel with lagged dependent variable*	*Pre-1941*	*1941 and after*	*Narrow economic issues*	*Broad economic issues*	*Social issues*
Google Books frequency$_{t-1}$	0.151**	0.0122	0.0167	0.0150	0.208	0.0477***	0.141
	(0.0477)	(0.00780)	(0.00863)	(0.0131)	(0.159)	(0.0139)	(0.0905)
Google Books frequency$_{t-2}$	0.0887*	–0.0133	–0.0102	–0.0111	–0.197	–0.0217	0.0197
	(0.0452)	(0.00713)	(0.00780)	(0.0114)	(0.161)	(0.0271)	(0.149)
Google Books frequency$_{t-3}$	0.0875	0.0245	0.0246	0.000651	0.0432	–0.00167	–0.136
	(0.0626)	(0.0232)	(0.0247)	(0.0115)	(0.0566)	(0.0291)	(0.0915)
(Google Books frequency × phrase in 25th–50th percentile)$_{t-1}$	–0.0848	–0.0263*	–0.0263	–0.00933	–0.232	–0.178**	–0.0551
	(0.0542)	(0.0124)	(0.0165)	(0.0148)	(0.165)	(0.0585)	(0.170)
(Google Books frequency × phrase in 25th–50th percentile)$_{t-2}$	–0.0101	0.0451**	0.0591**	0.0110	0.372	0.184	–0.0138
	(0.0517)	(0.0152)	(0.0206)	(0.0149)	(0.215)	(0.175)	(0.153)
(Google Books frequency × phrase in 25th–50th percentile)$_{t-3}$	–0.0419	–0.0257	–0.0220	0.00236	–0.0570	–0.0559	0.146
	(0.0658)	(0.0245)	(0.0265)	(0.0136)	(0.139)	(0.141)	(0.117)
(Google Books frequency × phrase in 50th–75th percentile)$_{t-1}$	–0.106*	–0.00206	–0.00324	–0.00401	0.000457	–0.0559	0.00783
	(0.0532)	(0.0107)	(0.0114)	(0.0156)	(0.172)	(0.0482)	(0.113)
(Google Books frequency × phrase in 50th–75th percentile)$_{t-2}$	–0.0739	0.000433	–0.00348	0.00426	0.105	0.0220	–0.0968
	(0.0477)	(0.0114)	(0.0146)	(0.0137)	(0.168)	(0.0480)	(0.178)
(Google Books frequency × phrase in 50th–75th percentile)$_{t-3}$	–0.0499	–0.0101	–0.00849	0.000573	0.118	0.0820	0.337*
	(0.0642)	(0.0236)	(0.0256)	(0.0127)	(0.118)	(0.0555)	(0.136)
(Google Books frequency × phrase in 75th–99th percentile)$_{t-1}$	–0.0380	0.00202	–0.00608	0.00831	–0.272	0.0325	–0.149
	(0.0561)	(0.0144)	(0.0185)	(0.0230)	(0.240)	(0.0322)	(0.0970)

(Google Books frequency × phrase in 75th–99th percentile)$_{t-2}$	–0.0509	0.00243	–0.00997	0.0307	0.0616	–0.0342	0.0233
	(0.0478)	(0.0123)	(0.0153)	(0.0220)	(0.227)	(0.0561)	(0.155)
(Google Books frequency × phrase in 75th–99th percentile)$_{t-3}$	–0.0827	–0.0190	–0.0237	0.0123	0.269	0.0657	0.137
	(0.0630)	(0.0241)	(0.0255)	(0.0180)	(0.205)	(0.0475)	(0.0921)
Congressional frequency$_{t-1}$		0.522***	0.446***	0.507***	0.570***	0.554***	0.483***
		(0.0337)	(0.0435)	(0.0437)	(0.0624)	(0.0677)	(0.0876)
Congressional frequency$_{t-2}$		0.167***	0.145***	0.110***	0.128	0.221***	0.135**
		(0.0207)	(0.0262)	(0.0296)	(0.0707)	(0.0568)	(0.0463)
Congressional frequency$_{t-3}$		0.118***	0.0992***	0.0362	0.0734*	0.0372	0.138
		(0.0250)	(0.0298)	(0.0331)	(0.0364)	(0.0293)	(0.0840)
No. of observations	2,395,104	2,395,104	1,217,184	1,177,920	30,805	65,392	16,165
R^2	0.010	0.556	0.360	0.374	0.520	0.574	0.471
Sum of coefficients for Google Books frequency, all lags							
0–25th percentiles	0.328**	0.0234	0.0311	0.00448	0.0537	0.0243***	0.0241
	(0.153)	(0.0227)	(0.0242)	(0.00860)	(0.0795)	(0.00798)	(0.0833)
25th–50th percentiles	–0.137	–0.00700	0.0108	0.00400	0.0838	–0.0501	0.0768
	(0.166)	(0.0240)	(0.0301)	(0.0108)	(0.115)	(0.0377)	(0.121)
50th–75th percentiles	–0.230	–0.0117	–0.0152	0.000820	0.223***	0.0481	0.248**
	(0.160)	(0.0229)	(0.0256)	(0.0100)	(0.0586)	(0.0369)	(0.101)
75th–99th percentiles	–0.172	–0.0145	–0.0398	0.0513	0.0579	0.0640*	0.0117
	(0.159)	(0.0239)	(0.0277)	(0.0392)	(0.129)	(0.0381)	(0.0802)

Source: Authors' regressions using data from the digitized *Congressional Record* and Google Books.
a. See table B.1 for details of the estimation.

Table B.3. Regressions Estimating the Diffusion of Polarized Phrases from Congress to Google Books, Using a *t*-Statistic-Based Threshold[a]

	Dependent variable: frequency of phrase in Google Books at time t						
Independent variable	*Straight panel*	*Panel with lagged dependent variable*	*Pre-1941*	*1941 and after*	*Narrow economic issues*	*Broad economic issues*	*Social issues*
Congressional frequency$_{t-1}$	0.0423**	0.0210**	0.0251**	0.0136**	0.0293*	0.00560	0.00147
	(0.0147)	(0.00661)	(0.00906)	(0.00513)	(0.0140)	(0.00398)	(0.00250)
Congressional frequency$_{t-2}$	0.0274	−0.00273	0.0000572	−0.00182	−0.0151	−0.00879	0.00916
	(0.0141)	(0.00441)	(0.00633)	(0.00343)	(0.0146)	(0.00622)	(0.00616)
Congressional frequency$_{t-3}$	0.0307*	−0.00631	−0.000785	−0.00875*	0.0117	0.00343	−0.0115*
	(0.0155)	(0.00349)	(0.00382)	(0.00404)	(0.0140)	(0.00320)	(0.00463)
(Congressional frequency × phrase in 25th–50th percentile)$_{t-1}$	−0.00630	−0.0161*	−0.0144	−0.0136	−0.0302*	0.0219	0.0414*
	(0.0214)	(0.00747)	(0.0106)	(0.00719)	(0.0142)	(0.0175)	(0.0197)
(Congressional frequency × phrase in 25th–50th percentile)$_{t-2}$	0.00264	0.00817	0.00706	0.00660	0.0155	0.0358	−0.0256
	(0.0196)	(0.00619)	(0.0102)	(0.00371)	(0.0144)	(0.0242)	(0.0192)
(Congressional frequency × phrase in 25th–50th percentile)$_{t-3}$	0.00489	0.00503	0.00503	0.00522	−0.00968	−0.0298	−0.0312
	(0.0227)	(0.00471)	(0.00625)	(0.00388)	(0.0140)	(0.0190)	(0.0357)
(Congressional frequency × phrase in 50th–75th percentile)$_{t-1}$	−0.0156	−0.00977	0.00400	−0.00713	−0.0148	0.000194	0.00832
	(0.0163)	(0.00764)	(0.0181)	(0.00540)	(0.0163)	(0.00526)	(0.00615)
(Congressional frequency × phrase in 50th–75th percentile)$_{t-2}$	−0.00978	0.00133	−0.00813	0.00299	0.0142	0.00261	−0.00558
	(0.0152)	(0.00530)	(0.0154)	(0.00303)	(0.0146)	(0.00875)	(0.0104)
(Congressional frequency × phrase in 50th–75th percentile)$_{t-3}$	−0.0101	0.00237	−0.00266	0.00716	−0.0126	0.00115	0.0242*
	(0.0167)	(0.00346)	(0.00637)	(0.00370)	(0.0142)	(0.00658)	(0.0104)
(Congressional frequency × phrase in 75th–99th percentile)$_{t-1}$	−0.0238	−0.0122	−0.00363	−0.00740	−0.0159	0.0108	0.00570
	(0.0156)	(0.00709)	(0.0131)	(0.00544)	(0.0143)	(0.0105)	(0.00525)

(Congressional frequency × phrase in 75th–99th percentile)$_{t-2}$	−0.0169	0.00373	0.00413	0.00271	0.0177	0.0105	0.00473
	(0.0147)	(0.00468)	(0.00938)	(0.00340)	(0.0145)	(0.00827)	(0.0141)
(Congressional frequency × phrase in 75th–99th percentile)$_{t-3}$	−0.0215	0.00121	−0.0132*	0.00626	−0.0154	−0.000834	−0.000684
	(0.0160)	(0.00415)	(0.00538)	(0.00407)	(0.0165)	(0.00706)	(0.0118)
Google Books frequency$_{t-1}$		0.451***	0.361***	0.458***	0.359***	0.518***	0.495***
		(0.0390)	(0.0458)	(0.0616)	(0.0544)	(0.0604)	(0.0343)
Google Books frequency$_{t-2}$		0.290***	0.233***	0.281***	0.298***	0.305***	0.395***
		(0.0236)	(0.0205)	(0.0296)	(0.0401)	(0.0574)	(0.0782)
Google Books frequency$_{t-3}$		0.142***	0.0979***	0.153***	0.0422	0.0682	−0.0141
		(0.0285)	(0.0276)	(0.0226)	(0.0716)	(0.0349)	(0.00996)
No. of observations	2,487,031	2,487,031	1,263,901	1,223,130	30,805	65,392	16,165
R^2	0.012	0.645	0.354	0.606	0.411	0.699	0.689
Sum of coefficients for Congressional frequency, all lags							
0–25th percentiles	0.100**	0.0119	0.0244*	0.00299	0.0259**	0.000237	−0.000903
	(0.0432)	(0.00776)	(0.0128)	(0.00686)	(0.0125)	(0.00124)	(0.00259)
25th–50th percentiles	0.00122	−0.00284	−0.00228	−0.00175	−0.0244**	0.0278	−0.0153
	(0.0624)	(0.00787)	(0.0167)	(0.00735)	(0.0124)	(0.0240)	(0.0106)
50th–75th percentiles	−0.0355	−0.00606	−0.00679	0.00302	−0.0132	0.00396	0.0270
	(0.0467)	(0.00681)	(0.0141)	(0.00517)	(0.0149)	(0.00317)	(0.0172)
75th–99th percentiles	−0.0622	−0.00729	−0.0127	0.00157	−0.0136	0.0204*	0.00975*
	(0.0448)	(0.00720)	(0.0147)	(0.00546)	(0.0153)	(0.0105)	(0.00564)

Source: Authors' regressions using data from the digitized *Congressional Record* and Google Books.

a. Phrases are selected for inclusion according to *t*-statistic, calculated by dividing the correlation coefficient of a phrase by its standard error, rather than by the χ^2 method used in tables 6 and 7 in the text. See table B.1 for other details of the estimation.

Table B.4. Regressions Estimating the Diffusion of Polarized Phrases from Google Books to Congress, Using a *t*-Statistic-Based Threshold[a]

	Dependent variable: frequency of phrase in the digitized Congressional Record at time t						
Independent variable	*Straight panel*	*Panel with lagged dependent variable*	*Pre-1941*	*1941 and after*	*Narrow economic issues*	*Broad economic issues*	*Social issues*
Google Books frequency$_{t-1}$	0.0869**	0.00989	0.00219	0.0185	–0.0335	0.0196	0.0893
	(0.0321)	(0.00903)	(0.0114)	(0.0166)	(0.0463)	(0.0324)	(0.0563)
Google Books frequency$_{t-2}$	0.0537*	–0.000874	–0.00494	–0.00442	0.230*	–0.0610**	–0.130
	(0.0264)	(0.00726)	(0.00769)	(0.0115)	(0.105)	(0.0208)	(0.111)
Google Books frequency$_{t-3}$	0.0586	0.0236	0.0170	0.00446	–0.0574	0.0885	0.263*
	(0.0331)	(0.0151)	(0.0105)	(0.0135)	(0.107)	(0.0564)	(0.134)
(Google Books frequency × phrase in 25th–50th percentile)$_{t-1}$	–0.0647	–0.0141	0.00487	0.00136	0.164	–0.190*	–0.0905
	(0.0445)	(0.0175)	(0.0192)	(0.0267)	(0.155)	(0.0784)	(0.0823)
(Google Books frequency × phrase in 25th–50th percentile)$_{t-2}$	–0.00223	0.0232	0.0496*	–0.0299	0.0245	0.305	0.523***
	(0.0384)	(0.0210)	(0.0244)	(0.0254)	(0.306)	(0.217)	(0.113)
(Google Books frequency × phrase in 25th–50th percentile)$_{t-3}$	–0.00891	–0.0148	–0.0150	0.0265	0.473***	–0.163	–0.407**
	(0.0420)	(0.0188)	(0.0151)	(0.0325)	(0.118)	(0.191)	(0.142)
(Google Books frequency × phrase in 50th–75th percentile)$_{t-1}$	–0.0306	–0.00913	–0.00751	0.00348	–0.272	0.0996	–0.499*
	(0.0340)	(0.0119)	(0.0163)	(0.0205)	(0.374)	(0.0588)	(0.238)
(Google Books frequency × phrase in 50th–75th percentile)$_{t-2}$	–0.0142	0.00760	0.00905	0.0210	–0.665*	0.143	0.477**
	(0.0285)	(0.0114)	(0.0112)	(0.0166)	(0.285)	(0.0954)	(0.183)
(Google Books frequency × phrase in 50th–75th percentile)$_{t-3}$	–0.0191	–0.0133	–0.00667	–0.00360	0.640	–0.176*	–0.160
	(0.0350)	(0.0159)	(0.0118)	(0.0159)	(0.384)	(0.0823)	(0.176)
(Google Books frequency × phrase in 75th–99th percentile)$_{t-1}$	0.0843	0.0385*	0.0264	0.0938**	0.0883	–0.0598	0.0469
	(0.0535)	(0.0195)	(0.0215)	(0.0346)	(0.0959)	(0.147)	(0.135)

(Google Books frequency × phrase in 75th–99th percentile)$_{t-2}$	−0.0121	−0.00976	−0.0349	0.0601	0.133	0.0835	0.0623
	(0.0466)	(0.0241)	(0.0186)	(0.0549)	(0.261)	(0.0935)	(0.142)
(Google Books frequency × phrase in 75th–99th percentile)$_{t-3}$	−0.0547	−0.0208	−0.00925	−0.0518	−0.379	0.0999	−0.271*
	(0.0373)	(0.0234)	(0.0133)	(0.0883)	(0.343)	(0.167)	(0.136)
Congressional frequency$_{t-1}$		0.412***	0.438***	0.263***	0.465***	0.429***	0.401***
		(0.0282)	(0.0436)	(0.0298)	(0.0505)	(0.0408)	(0.0655)
Congressional frequency$_{t-2}$		0.125**	0.139***	0.0170	0.139**	0.139**	0.182*
		(0.0459)	(0.0266)	(0.0632)	(0.0453)	(0.0516)	(0.0796)
Congressional frequency$_{t-3}$		0.172***	0.0992**	0.0845***	0.113**	0.125**	0.107
		(0.0226)	(0.0315)	(0.0174)	(0.0410)	(0.0386)	(0.0549)
No. of observations	2,487,031	2,487,031	1,263,901	1,223,130	30,805	65,392	16,165
R^2	0.007	0.365	0.346	0.088	0.402	0.345	0.361
Sum of coefficients for Google Books frequency, all lags							
0–25th percentiles	0.199**	0.0326	0.0143	0.0186**	0.140**	0.0471*	0.222
	(0.0897)	(0.0215)	(0.0144)	(0.00787)	(0.0700)	(0.0247)	(0.150)
25th–50th percentiles	−0.0759	−0.00572	0.0395	−0.00204	0.662	−0.0473	0.0251
	(0.108)	(0.0256)	(0.0246)	(0.0139)	(0.423)	(0.0594)	(0.187)
50th–75th percentiles	−0.0639	−0.0149	−0.00512	0.0209	−0.297	0.0666	−0.182
	(0.0936)	(0.0213)	(0.0153)	(0.0151)	(0.208)	(0.0413)	(0.155)
75th–99th percentiles	0.0175	0.00792	−0.0178	0.102	−0.157	0.124	−0.162
	(0.121)	(0.0315)	(0.0176)	(0.0709)	(0.135)	(0.0817)	(0.152)

Source: Authors' regressions using data from the digitized *Congressional Record* and Google Books.

a. Phrases are selected for inclusion according to *t*-statistic, calculated by dividing the correlation coefficient of a phrase by its standard error, rather than by the χ^2 method used in tables 6 and 7 in the text. See table B.1 for other details of the estimation.

decently robust in the frequency cut sample but not in the *t*-stat-restricted sample.[8] Future research is needed to determine appropriate ways to construct phrase samples.

ACKNOWLEDGMENTS We thank David Gergen, Arthur Spirling, and the editors for extensive and very helpful feedback. We also thank the participants at the Brookings Panel conference, and Matt Connelly, Ellora Derenoncourt, Jean-Baptiste Michel, and Sébastien Turban for helpful comments.

8. Note that we did not generate a new set of persistently polarized phrases and hand-code topics for the *t*-stat-restricted and frequency-selected samples. Instead, we used the persistently polarized phrases from the χ^2 cut.

References

Bartels, Larry M. 2010. *Unequal Democracy: The Political Economy of the New Gilded Age.* Princeton University Press.

Blei, David M., and Jon D. McAuliffe. "Supervised Topic Models." Princeton University and University of Pennsylvania. arxiv.org/pdf/1003.0783.pdf.

Campante, F. R., and Q. A. Do. 2008. "Inequality, Redistribution and Population." In *American Law and Economics Association Annual Meetings,* p. 103. Berkeley, Calif.: bepress.

Carter, Susan B., Scott Sigmund Gartner, Michael R. Haines, Alan L. Olmstead, Richard Sutch, and Gavin Wright. 2006. *Historical Statistics of the United States, Earliest Times to the Present: Millennial Edition.* Cambridge University Press.

Davis, Joseph H. 2004. "An Annual Index of US Industrial Production, 1790–1915." *Quarterly Journal of Economics* 119, no. 4: 1177–1215.

Fiorina, Morris P., with Samuel J. Abrams and Jeremy C. Pope. 2005. *Culture War? The Myth of a Polarized America.* London: Longman.

Gelman, Andrew, David Park, Boris Shor, Joseph Bafumi, and Jeronimo Cortina. 2008. *Red State, Blue State, Rich State, Poor State: Why Americans Vote the Way They Do,* expanded edition. Princeton University Press.

Gentzkow, Matthew, and Jesse M. Shapiro. 2010. "What Drives Media Slant? Evidence from US Daily Newspapers." *Econometrica* 78, no. 1: 35–71.

Glaeser, Edward L., Giacomo A. M. Ponzetto, and Jesse M. Shapiro. 2005. "Strategic Extremism: Why Republicans and Democrats Divide on Religious Values." *Quarterly Journal of Economics* 120, no. 4: 1283–1330.

Grant, Tobin, and Nathan Kelly. 2008. "Legislative Productivity of the U.S. Congress, 1789–2004." *Political Analysis* 16 (February): 303–23.

Grimmer, Justin, and Brandon M. Stewart. 2012. "Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts." Stanford University.

Groseclose, Tim, and Jeffrey Milyo. 2005. "A Measure of Media Bias." *Quarterly Journal of Economics* 120, no. 4: 1191–1237.

Hofstadter, Richard. 1964. "The Paranoid Style in American Politics." *Harper's* (November): 77–86.

Lind, Michael. 2012. *Land of Promise: An Economic History of the United States.* New York: Harper.

Maltzman, Forrest, and Lee Sigelman. 1996. "The Politics of Talk: Unconstrained Floor Time in the U.S. House of Representatives." *Journal of Politics* 58, no. 3: 819–30.

McCarty, Nolan, Keith T. Poole, and Howard Rosenthal. 1997. *Income Redistribution and the Realignment of American Politics.* Washington: American Enterprise Institute.

———. 2005. *Polarized America: The Dance of Ideology and Unequal Riches.* MIT Press.

Michel, J.-B., Y. K. Shen, A. P. Aiden, A. Veres, M. K. Gray, J. P. Pickett, D. Hoiberg, D. Clancy, P. Norvig, J. Orwant, and others. 2011. "Quantitative Analysis of Culture Using Millions of Digitized Books." *Science* 331, no. 6014: 176.

Miron, Jeffrey A., and Christina D. Romer. 1990. "A New Monthly Index of Industrial Production, 1884–1940." *Journal of Economic History* 50, no. 02: 321–37.

Monroe, Burt L, Michael P. Colaresi, and Kevin M. Quinn. 2008. "Fightin' Words: Lexical Feature Selection and Evaluation for Identifying the Content of Political Conflict." *Political Analysis* 16: 372–403.

Morris, Jonathan S. 2001. "Reexamining the Politics of Talk: Partisan Rhetoric in the 104th House." *Legislative Studies Quarterly* 26, no. 1: 101–21.

Noel, Hans Christopher. 2006. "The Coalition Merchants: How Ideologues Shape Parties in American Politics." Ph.D. dissertation, University of California, Los Angeles.

———. 2012. "The Coalition Merchants: The Ideological Roots of the Civil Rights Realignment." *Journal of Politics* 74, no. 1: 156–73.

Pang, Bo, and Lillian Lee. 2008. "Opinion Mining and Sentiment Analysis." *Foundations and Trends in Information Retrieval* 2, no. 1–2: 1–135.

Poole, Keith T. 2008. "The Roots of the Polarization of Modern U.S. Politics." University of Georgia. ssrn.com/abstract=1276025.

Porter, M. F. 1980. "An Algorithm for Suffix Stripping." *Program* 14, no. 3: 130–37.

Sarkees, Meredith Reid, and Frank Wayman. 2007. *Resort to War: 1816–2007.* Washington: CQ Press.

Taddy, Matt. 2012. "Inverse Regression for Analysis of Sentiment in Text." University of Chicago. arxiv.org/pdf/1012.2098.pdf.

Turchin, Peter. 2012. "Dynamics of Political Instability in the United States, 1780–2010." *Journal of Peace Research* 4: 577–91.

Zaller, John R. 1992. *The Nature and Origins of Mass Opinions.* Cambridge University Press.

Comments and Discussion

COMMENT BY

DAVID GERGEN and MICHAEL ZUCKERMAN With this paper, Jacob Jensen, Ethan Kaplan, Suresh Naidu, and Laurence Wilse-Samson offer a methodologically innovative contribution toward answering a long-standing question: what, exactly, drives the national political conversation? In this comment we summarize their paper and, in a constructive spirit, recommend that future research into this topic be broadened to answer three basic questions:

—What role do the media play in driving the political conversation?

—How do historical changes in the nature of Congress affect its rhetoric (as well as its results), and to what extent is the historical *Congressional Record* both a consistent source of data and a reliable predictor of polarization and results going forward?

—To what extent are elites (such as members of Congress and published authors) responding to rather than driving the ideology espoused by more popular voices?

Jensen and coauthors probe the words behind the American political conversation, building on the ascendant "text-as-data" approach to create intriguing new measures of polarization and partisanship. They then pair these newly constructed longitudinal measures with preexisting data to seek out "important national political phenomena that correlate in time with the polarization in political discourse" and may be driving its direction. The data powering these ambitious efforts are a set of quantifiably partisan three-word phrases ("trigrams") stretching back to 1873, sifted out by the authors (using, presumably, a lot of computing power) from both

the newly digitized *Congressional Record* archives and the Google Books corpus of the English language.[1]

The paper's first contribution, then, is this new measurement of polarization and partisanship in both congressional and broad "political discourse" stretching back over the past 140 years (and limited only by digital availability of the *Congressional Record*—the Google corpus extends more than 350 years earlier). The series of data they construct is no small undertaking and a welcome addition to an area of inquiry that has previously been dominated by the DW-NOMINATE score, a measure of polarization drawn from congressional voting patterns and curated most prominently by political scientists Nolan McCarty, Keith Poole, and Howard Rosenthal. Run historically, the authors' new model correlates closely with DW-NOMINATE in the post-1930 era but diverges from it before 1930, finding high polarization toward the end of the 19th century and less in the early 20th century. What it shows—that "recent political polarization may be high relative to the 1970s, but it is a far cry from the open violence of the late 19th century"—offers some useful historical scope and context. It also builds an empirical scaffolding for some anecdotal bits from history that challenge the contention that polarization is worse now than ever before: recall that in 1856 Congressman Preston Brooks of South Carolina caned Massachusetts Senator Charles Sumner to within an inch of his life on the floor of the Senate, or that one debate in the House in the 1800s grew so rancorous that some 30 members drew their guns. Politics, to paraphrase Mr. Dooley, was never beanbag.

From this initial measurement, the authors extend their exploration in an effort to tease out correlations between polarization over time and other, broader factors. Among their more targeted observations are a robust correlation between polarization and political violence; a switch in relative polarization in the direction of the *minority* party when House control switches (the authors reasonably hypothesize that minority parties may "talk more and use more partisan language in order to slow the enactment of policies they oppose"); and, perhaps most interesting (but "statistically weak," in the authors' words), a lag between the Google Books database and Congress for polarized economic phrases only, suggesting that "although

1. As the authors themselves suggest, the paper is worth the read just for the delight of perusing tables 1 and 2, which list, respectively, the single most partisan phrase in each congressional session back to 1873 for both Democrats and Republicans, and the 50 most partisan phrases, again by party, for the 110th Congress. Both tables present engrossing time slices of the national discourse.

Congress may take its economic language from public intellectuals, it does not adopt its language on social issues from the same sources."[2]

A discussion of the statistical techniques that the authors employ lies beyond our training, but the range of their search—through Gallup polling, military casualty figures, economic data, multiple measures of legislative efficiency, major historical events, and other data—is extensive, and promising for future inquiry. Although the paper deliberately stops short of addressing causality, the authors do report that they see their "evidence as consistent with an autonomous effect of elite political discourse on congressional speech and legislative gridlock," but they deny that this effect is responsible for "the recent increase in congressional polarization." They argue instead that "polarized political discourse varies independently of the polarization of politicians," asserting that Congress is far more polarized than the broader discourse, finding evidence that polarization in the discourse diffuses into congressional speech, and tying this broader polarization of discourse—but not that of congressional language—to legislative gridlock.

This conclusion—that "it is underlying ideological polarization of political elites that is the true obstacle to legislative efficiency, rather than the polarization of Congress"—is important. Congressional polarization may be superable through personal relationships or the co-opting of a few members of the opposition, the authors reason, but "if party activists and intellectuals are truly polarized in their beliefs, then politicians may have no leeway to compromise, because they are constrained by reelection concerns or other intraparty ideological constraints."

Sharing the authors' enthusiasm (if not their methodological expertise) for addressing the underlying questions, we offer three responses. First, although the authors responsibly acknowledge that they have deliberately left the media largely outside the scope of their inquiry, this exclusion leaves significant potential insight on the table. Second, beyond what is noted in their limitations section, a firmer engagement with the changing nature of Congress (and congressional history) in recent years would deepen the analysis and shed light on possible sources of bias or inaccuracy in the congressional vein of inquiry (we claim less insight as to the shortcomings of the Google Books corpus). Lastly, although

2. The measure also, among its other capabilities, predicts with fair accuracy a member's party based on his or her words, and shows ("perhaps surprisingly," the authors note) no correlation with economic growth.

the authors are careful to offer a cautious, data-grounded theory of "the sources of ideological change" and to divorce "the autonomous effect of elite political discourse on congressional speech and legislative gridlock" from the contemporary rise in congressional polarization, we wonder if their analysis is missing important elements in the American political discourse that could account not only for the rise in congressional polarization, but for the polarization of the broader discourse and overarching political gridlock as well.

THE ROLE OF THE MEDIA The authors could substantially deepen their subsequent research by engaging to a much larger extent with the media's impact on the ideological ecosystem—particularly its more democratic forms such as television and the Internet. Although the authors note in their limitations section that "the media available and the coverage given national politics" are missing from their analytical model, even in this acknowledgment they largely focus on the rise of C-SPAN (a phenomenon that is actually more relevant to our discussion below of the changes within Congress itself). For obvious reasons, television and the Internet will not be useful for an analysis intended to stretch back to 1873, but even so, the study's focus on *Congressional Record* statements and the Google Books corpus leaves unaddressed a substantial portion of the venues through which Americans receive and refine their political information and thinking. Exploring how political ideologies are formed in the modern era without considering television news or the Internet seems a little like exploring how a production of *La Bohéme* is performed without considering any of the singers or musicians.

One instructive analysis of the media's role in driving the contemporary political conversation comes from Theda Skocpol and Vanessa Williamson's (2011) research into the Tea Party. In a chapter addressing the "media as cheerleader and megaphone," they detail the ways in which cable news networks—most prominently Fox News in the case of the Tea Party—have become "communities of meaning" and skilled message distribution venues for political thinking, as influential and autonomous as a major political party (if not more so). Using Fox News and CNN transcripts from Lexis-Nexis, Skocpol and Williamson demonstrate how Fox's coverage was able to drive its competitors' coverage, turning the Tea Party from a minor story into a national craze. "Commercial competition," they write, "means that issue-mongers can fan a supposedly scandalous sound bite into an uproar of intense coverage across many channels" (Skocpol and Williamson 2011, p. 126). In their book *Echo Chamber,* Kathleen Hall Jamieson and Joseph Capella (2008)

demonstrate how media venues on both sides of the political spectrum have created "self-protective enclaves" for their adherents that serve to disseminate, confirm, and strengthen party thinking. Their argument underscores our own observations from the multiple interviews we have conducted with senior elected officials in both parties: often these individuals have recounted being booked for cable television appearances only to find themselves summarily dropped once it became clear they were planning to express a nuanced, moderate opinion, rather than a sharply partisan screed.

The present authors, of course, are working with a very different methodology over a much longer timeline. Nevertheless, their future research could build substantially on this paper's findings by also studying the correlations in time of partisan trigrams (and, as the authors note, eventually unigrams and skip-grams) drawn from Fox News, MSNBC, and CNN transcripts, available online or via Lexis-Nexis, or from partisan blogs, message boards, and the regular e-mail distributions of partisan organizations. Because these venues are likely to affect political thinking on the ground on a far wider scale than either congressional floor statements or (perhaps sadly) most published books, a broader investigation that analyzed these elements alongside the bodies of political speech the authors have already explored may well create a fuller, richer picture of how the conversation unfolds. It would not be surprising to us if that picture were to reveal the media as playing a significant role in amplifying the recent rise in rhetorical temperature that the authors observe, both in Congress and in the broader discourse.

THE CHANGING NATURE OF CONGRESS The authors rightly note in their limitations section that potential shifts in "House rules or party practices," as well as the "advent of extensive television coverage of Congress by C-SPAN," may introduce compromising factors into their data. Although the paper's purview extends beyond congressional rhetoric alone, it is worth considering some of these factors and examining changes in Congress more closely.

The authors quote an article by Forrest Maltzman and Lee Sigelman (1996) arguing that the "special orders and short speeches" introduced as *Congressional Record* statements "serve as potential tools of policy influence within the House." This explanation seems overstated; in our experience, these statements are often simple political tactics employed by members of Congress who want either to place a marker on the record for use in messages to constituents or to indemnify themselves against future attacks by opponents.

The authors further observe that, in the C-SPAN era, "congressional debate may have shifted into congressional performance"; this notion is apt. Former Democratic Senate Majority Leader Tom Daschle, for example, has told us in an interview that, in his experience, the C-SPAN cameras had a substantial impact on members' conduct and language on the floor. As he recalled it, the Senate had the cameras turned off a few times while they were debating confidential or otherwise protected business, and each time this happened, the chemistry in the room instantly flipped, yielding some of the most emotional and thoughtful moments he ever witnessed on the Senate floor. Aside from considerations of Senate efficacy and culture, Daschle's recollections clearly support the authors' recognition of a potential skew in the data arising from the birth of C-SPAN.

On a related note, the growth of the party caucus meetings and their message distribution strategies seem to exert an amplifying influence on the frequency and unity of partisan messaging in recent Congresses. When we interviewed Rep. Lee Hamilton, a long-standing (now retired) moderate Democratic congressman from Indiana, last year, he bemoaned the Democratic caucus meetings of his later years in Congress and the level of message discipline they enforced. As he told it, at each meeting a new set of talking points would be distributed to each of the members; because the average person watches C-SPAN for only 10 to 12 minutes, he then explained, the party would send everyone out to deliver the same message over and over again in 10- to 15-minute increments. In addition to being gut-wrenchingly frustrating for an independent-minded member, such a practice could cause the authors' model to overestimate current polarization by artificially boosting the number of responsive trigrams. (At the same time, of course, the mere existence of such a practice may be a symptom of such widespread polarization that any artificial amplification is, in fact, meaningful data.)

These procedural notes point to the ultimate question of whether rhetorical data drawn from the *Congressional Record* are really the best measure of meaningful political polarization, especially over long periods (a question that can likewise be extended to DW-NOMINATE scores, which the authors' measure largely tracks post-1930). To begin with, the mechanical counting of partisan phrases tends to ignore both the context in which those phrases were uttered and the intensity of feeling behind their utterance. Unique acts of extreme partisanship—such as the "You lie!" outburst of Rep. Joe Wilson, Republican of South Carolina, during the 2009 State of the Union address—although harder to model, are arguably more telling rhetorical symptoms of congressional polarization than simple word counts.

The historical context also matters. Much of the authors' data, for example, originate in the period of institutional strengthening of the presidency during and after the New Deal, and thus of the steady erosion of Congress's influence. Scholarship on this point goes back at least as far as Arthur Schlesinger's (1973) *Imperial Presidency,* but one handy modern heuristic is simply to look at media coverage: the major media players today tend to deploy much larger White House contingents and much smaller congressional contingents than in the past, calling into question the centrality of congressional statements to the current national dialogue.

Finally, as far as legislative efficacy is concerned, history seems also to suggest that congressional rhetoric can, in certain periods, be a misleading indicator. Members of Congress in the 1950s and 1960s, for example, often talked tough to impress the folks back home or to keep in line with regional traditions, but came together on bipartisan compromises to build an interstate highway system, to pass landmark civil rights legislation, and to create Medicare and Medicaid. In the 1980s, President Ronald Reagan and House Speaker Tip O'Neill attacked each other repeatedly and vociferously in public, yet much was accomplished in Congress in those years, in part because the two men shared a political culture—and a deep affection for each other. Perhaps because members of Congress of these generations had shared the experience of World War II, they enjoyed stronger common bonds and were more inclined to place country over party than are many of today's members. Yet President Bill Clinton and House Speaker Newt Gingrich, in the 1990s, were similarly able to prevent hot rhetoric from becoming an obstacle to political achievement, in particular welfare reform and four straight balanced budgets.

All that said, the authors' investigations into the relationship between the polarization of broader discourse and legislative efficiency contain much that is useful and instructive (although it would have been interesting to see *Congressional Record* polarization graphed against legislative efficiency as well). But for their further efforts to elucidate the relationship between the broader discourse and the functionality and makeup of Congress, we would propose exploring other potential sources of Congressional data—members' campaign remarks or issue releases, for example—alongside the other laudable possibilities the authors themselves cite (Senate and presidential speech, newspaper databases, think-tank issuances). The result would be a fuller picture that might lend itself to more robust correlates and potential causal connections. It may well turn out that the polarization of rhetoric that the authors document is not what is driving the present

dysfunctionality of Congress but rather a reflection of it, and that the true causes lie deeper in the political culture.

THE PEOPLE'S ROLE IN THE CONVERSATION Finally, although we are impressed by the statistical rigor and methodological ingenuity of the authors' research, we still fear, given the elite nature of the two data sources they investigate, that their reliance on these two elite-driven sources may lead their analysis to disregard key aspects of American political life and thus skew it toward finding that elites drive the conversation. The authors are careful not to ascribe causation to the correlations they identify and to admit the high likelihood of third-party effects; they are likewise careful to point out that their findings regarding the effect of elites on congressional speech and productivity do not account for the recent increase in congressional polarization. Nevertheless, since their analysis seems to side at least modestly with a more elite-driven theory of ideology (they approvingly cite John Zaller), as well as with Morris Fiorina (whom they likewise cite), in suggesting that the broader polity is less polarized than Congress, it is worth considering the alternative view, espoused by political scientists such as Alan Abramowitz. In *The Disappearing Center* (2010), Abramowitz argues that the elites are not driving polarization of their own accord, but rather are being forced to adapt as the citizens on the extremes grow in passionate intensity and the center fails to hold. These elites (politicians and the authors of books included in the Google corpus) may well be reacting to wider trends in the public culture rather than shaping them.

Beyond the contemporary movements that caution skepticism in the face of an elite-driven conception of public opinion (such as the Tea Party, which certainly appears to be driving congressional Republicans today far more than vice versa, and which has been mowing down establishment-promoted candidates for several years now), an instructive voice on this question is Cass Sunstein, whose 2009 book, *Going to Extremes: How Like Minds Unite and Divide,* posits two different kinds of polarization: planned and spontaneous. Sunstein writes (p. 34):

> Some people act as polarization entrepreneurs: They attempt to create communities of like-minded people, and they are aware that these communities will not only harden positions but also move them to a more extreme point. But sometimes polarization arises spontaneously, though entirely voluntary choices, without the slightest kind of planning. Consider, for example, people's reading patterns, which suggest an unmistakable form of self-sorting into liberal and conservative networks. Or consider the blogosphere itself, which shows a similar kind of spontaneous sorting and polarization. Or consider simple geographical choices; like-minded people, in essential agreement on political issues, may end up living in the same area simply because that is what they want to do.

Whether polarization is truly "spontaneous" or wells up from other sources (including, but not limited to, cultural and economic experiences), the side of the equation that the elite-driven model of public opinion seems to miss is the influence over the public discourse that is seated within the people themselves (and, perhaps, within politicians' and authors' strategic desires to know enough of the people's thinking to be able to tell them what they wish to hear).

There are, certainly, politicians and authors who *do* drive the conversation and who are, in Sunstein's phrase, "polarization entrepreneurs": former Alaska Gov. Sarah Palin helped coin the term "death panels"; Ann Coulter's books can be found on Google Books, as can Keith Olbermann's. But the actions, ideas, and messages of many more individuals may still inform, if much less prominently, a more demand-driven conception of political speech, one in which the voters hold a heterogeneous and changing but at least quasi-autonomous set of political values and beliefs, such that the politicians (and writers) who are able to best approximate and communicate those views are the ones who get elected (and published). Politicians and other elite voices would still, it bears noting, play an important role as synthesizers and coherers of the disparate views and values contained within the larger republic, and the messages they issue would re-inform the larger conversation in an ongoing cycle. But they would be reacting to spontaneous polarization more than driving it. One does not need to subscribe to Joseph Schumpeter's (1962) unpalatably top-down democratic theory to accept his underlying premise that democratic politics still is, at its core, a competition for the most votes.

We are not suggesting that the authors turn to the opposite extreme and adopt the kind of stripped-down rational choice theory of democracy that the political scientist Peter Euben once lampooned as viewing the voter as "a consumer in drag" (quoted in Sabl 2002, p.123). Rather, we are suggesting that the authors, without taking their eyes off of the elites' contributions, counterbalance their future efforts with a robust look at the preferences of voters as well, not to mention how those preferences sway elite-informed data sources like Google Books and the *Congressional Record.* Temperamentally and practically, the path we are suggesting is thus closer to Andrew Sabl's (2002) "democratic constancy" theory and his treatment of what he calls "the most refined exponents of the [rational choice] school"—thinkers like Anthony Downs, William Riker, and Fiorina himself, who "merely assume that citizens have certain aims, that politicians have goals that require their staying in office, and that democracy is a process of reconciling these two realities" (Sabl 2002, p. 123).

As far as polarization goes, therefore, a mix of Sunstein's two effects—which in our thinking would offer both a supply side and a demand side to the hypothetical market for political ideas and would leave room for the findings of more bottom-up-minded political scientists like Abramowitz—seems about right. To their credit, the authors end their paper on a thoughtful and circumspect note, stating their "intuition" that "ideas must have some of their own momentum and power, but that there are likely important background material conditions through which groups and individuals modify these ideas and make their propagation more or less likely." By plumbing the distinction between their own two types of polarization—polarization of the broader discourse and *Congressional Record* polarization—and showing some of the attending correlates, the authors have made a valuable foray into understanding how this complex process plays out. As the authors continue their fresh, methodologically path-breaking exploration of a question as old as democracy itself, we would urge them to keep their eyes on both sides of the political equation: remembering that America's ultimate authority is still constitutionally seated in the votes of everyday people, and extending their study to the messages and ideas those people both receive and express as they go about their day-to-day lives.

REFERENCES FOR THE GERGEN AND ZUCKERMAN COMMENT

Abramowitz, Alan. 2010. *The Disappearing Center.* Yale University Press.

Hamilton, Lee. 2004. *How Congress Works.* Indiana University Press.

Jamieson, Kathleen Hall, and Joseph N. Capella. 2008. *Echo Chamber: Rush Limbaugh and the Conservative Media Establishment.* Oxford University Press.

Sabl, Andrew. 2002. *Ruling Passions.* Princeton University Press.

Schlesinger, Arthur M., Jr. 1973. *The Imperial Presidency.* Boston: Houghton Mifflin.

Schumpeter, Joseph. 1962. *Capitalism, Socialism, and Democracy.* New York: Harper Perennial.

Skocpol, Theda, and Vanessa Williamson. 2011. *The Tea Party and the Remaking of Republican Conservatism.* Oxford University Press.

Sunstein, Cass. 2009. *Going to Extremes: How Like Minds Unite and Divide.* Oxford University Press.

COMMENT BY

ARTHUR SPIRLING At the time of this writing in late 2012, the United States risks falling off a "fiscal cliff." Absent a bipartisan agreement between a Democratic president and a Republican House of Representatives, taxes will rise and public spending will be cut automatically in a bid to decrease a large budget deficit, regardless of the (seemingly baleful) consequences. The received wisdom is that reaching a legislative deal to prevent this outcome is a difficult proposition: the parties bicker and are intransigent, and they operate in a Congress that is "the most polarized since the end of Reconstruction," according to Ezra Klein, a columnist and blogger for the *Washington Post.*[1]

For social scientists, at least three research questions arise from this purported nadir of American politics and the rancor and bitterness that supposedly characterize it: First, is it true? Second, does it matter? And third, how did this state of affairs come about? Broadly, it is these queries that this paper by Jacob Jensen and coauthors seek to answer. In so doing, they collect an extraordinary new and voluminous data set that incorporates a century of congressional speech, use innovative measures of political partisanship, and compare their results with a corpus of published phrases (Google Ngrams, drawn from Google Books) to look for possibly causal relationships between what politicians say and what is said by their publics. What emerges from their efforts is an impressive, data-driven look at political polarization and its development since the Reconstruction Era. Unsurprisingly, given its sheer scope, the analysis is not without flaws; commensurately, I will comment here on possible avenues for improvement and refinement, primarily on technical grounds. In addition, as it is clear that the paper and its data will inspire future research, my concluding section will attempt to point such efforts in fruitful directions, in terms of both methods and substance.

THE PAPER'S CONTRIBUTION AND ITS MOTIVATION It is important to emphasize the wealth of text data the authors have gathered: it is, to this discussant's knowledge, unprecedented in the study of U.S. politics. In sum, it is 130 years of information from the speech of the nation's representatives in Congress, and after much reduction it still includes some 690,000 phrase observations. The authors have matched these data to the party of the speaker, which no doubt required thorough cleaning and

1. Ezra Klein, "14 Reasons Why This Is the Worst Congress Ever," *Wonkblog (Washington Post),* July 13, 2012. www.washingtonpost.com/blogs/wonkblog/wp/2012/07/13/13-reasons-why-this-is-the-worst-congress-ever/.

much careful effort. The authors' inferential task is then similarly ambitious: to track the polarization of the parties over time, and to see what effects this varying polarization might have on other important outcomes, such as political violence. They perform extremely computationally intensive operations on their text data from the *Congressional Record,* and then do the same for the Google Books corpus. All of this is as impressive as it is important, and the paper deserves to be well cited—quite apart from the fact that the data set itself will form the backbone of many future studies. The paper is candid, thoughtful, and circumspect, and it comes at a time when methods for "text-as-data" are coming to the fore in the toolkit of political science (see, for example, Quinn and others 2010, Grimmer and Stewart forthcoming), and when "polarization" is a buzzword both in popular media and in academia (McCarty, Poole, and Rosenthal 2008, Fiorina, Abrams, and Pope 2010).

However, no good deed goes unpunished, and no good paper goes uncriticized. This is most assuredly a good paper, and any harshness in the comments that follow should indicate the degree to which reading it provokes thought—approving or otherwise.

TWO PROBLEMS WITH TRIGRAMS The core of the authors' measurement strategy is the trigram, a three-word sequence. Because they both "stem" and "stop" the raw text, meaning that words are truncated to their "roots" and function words (articles, conjunctions, auxiliary verbs, prepositions, and pronouns) are removed, it is not necessarily the case that any given trigram appears as is in the speeches. For example, "capital gains tax" becomes "capit.gain.tax," and any parts of sentences containing nothing but function words, such as "what he did with them," will disappear from the counting process altogether. The authors can hardly be blamed for attempting to reduce the dimensions of the feature space: although operating at the "token" (in this case, single word) level would be more general, the estimation problem would become much more computationally difficult—perhaps prohibitively so. Since one imagines that political phrases are precisely about context and a particular relationship with the words around them, stemmed and stopped trigrams are a reasonable pragmatic choice, capturing subtleties of meaning and allowing relative ease of interpretation while retaining tractability on the statistical side. Nonetheless, social scientists might have a few concerns. First, the idea of trigrams in this context is to capture some notion of word order. That is, phrases like "capit.gain.tax" and "nation.debt.increas" relate to concerns specific to political economy in a way that these words uttered separately do not. For many classification exercises, working with such

simplifications of spoken language present almost no cost. But life may be less rosy when, as here, issues of the speaker's sentiment are at stake. To cite a crude example, one imagines that the sentences "I do not support the New Deal" and "I do support the New Deal" would be spoken by legislators of quite different ideological stripes. But depending on the specific choice of stop words, both reduce to "support.new.deal" for the purposes of the present analysis, with potentially confusing consequences for interpretation.

A second, more subtle issue when working with stopped, stemmed trigrams arises from the fact that not all partisan phrases will be treated equally. As a running example, consider two very different three-word phrases: "Martin Luther King" and "By Almighty God." Notice that the second phrase includes a noun ("God") that has many synonyms: Creator, Lord, Heavenly Father, and so on. In principle, then, members of Congress could use any of these alternatives and communicate approximately the same meaning, and each would be counted separately under the authors' scheme. This is much less true of "Martin Luther King," a phrase that refers to an obvious individual and for which there are few close substitutes. As a result, speakers who wish to make a comment about that individual have little choice but to coordinate on "Martin Luther King" as a phrase. The consequence is that even if a particular concept—such as talking of God in whatever way—is highly discriminatory (and, one might hypothesize, indicative of Republicans), the diversity of options will reduce the chances that it appears as such. Matters are even worse in this particular case, because "By Almighty God" includes a stop word, which will be removed and some other word joined to the other two, further diversifying the nature of its appearance in the texts at hand.

What to do? One obvious robustness check would be to vary the stemming and stopping rules and verify that the results are similar. Another is to be more explicit about sentence structure and word order. Here the work of Huma Lodhi and others (2002) might prove profitable, and in particular their use of string kernels, which allow the researcher to break up documents into sets of *n*-contiguous characters and then base analysis on the relative frequency of these characters. There is no stemming or stopping with such procedures, and thus the statements regarding the New Deal above would be categorized as different. In addition, future work might consider identifying synonyms, perhaps with the help of a thesaurus or its equivalent, although this would involve more human coding a priori than the authors were perhaps willing to undertake for this study.

POLARIZED VIEWS, OR PARTISAN TOPICS? The metric used in the paper to calculate a phrase's (that is, trigram's) partisanship gives extra weight to word sequences that are used frequently by one party but infrequently by the other. That is, "partisan" words are those that discriminate between Democrats and Republicans. But as the authors themselves acknowledge, this difference in use may come from two very different sources: parties may talk about different things (guns versus immigration, for example), or about the same things in different ways ("illegal aliens" versus "undocumented workers"), or perhaps some combination of the two occurs. Depending on what the researcher wants to do with the results generated, this conflation is of varying concern. The broad goal of this paper is to measure "polarization," which is usually taken to mean a difference of opinion on the same topic, such as taxes, abortion, or immigration, because ideological distance decreases the ability of a given Congress and administration to deliver public policy efficiently. That is, we care about parties' positions rather than the valence they accord to different issues. If all the authors have captured is a difference in what subjects are "important" to the parties, then they have deviated some distance from the original goal. Notice here that validating the trigrams—in the sense that they predict party membership well in a holdout sample—cannot discriminate between ideological and topical division as an organizing principle for congressional speech.

Inspection of the trigrams identified as partisan does not help on this matter. As the authors note, the most recent examples do indeed appear to capture different views on the same issues, but in many years the selected trigrams appear entirely uninformative, "fiscal.year.end" (Republicans, 1897), "unit.state.oblig" (Republicans, 1919), and "unit.state.transmit" (Democrats, 1967) being easily found examples. One way to proceed may be that described by Burt Monroe, Michael Colaresi, and Kevin Quinn (2008), who limit attention to the difference on particular topics, thus getting immediately to the estimand of interest for the current authors, and in a model-based way. Note further that "topic" in this context could refer to some exogenously imposed issue that must be discussed, such as an OPEC oil shock, rather than one endogenously introduced for the specific purpose of partisan legislating.

UNSURE ABOUT UNCERTAINTY The paper's core measure, β_{pc}, is the correlation between the frequency of use of a phrase and the party of a speaker (coded 1 if the member is a Republican, −1 if a Democrat). Thus, if β_{pc} is negative, the phrase is associated with Democrats, and if positive, with Republicans. If the correlation is large in absolute terms (the authors do not say how large), the phrase is denoted as "polarizing." Unusually for

an estimated quantity, there is no uncertainty around this metric. This is unfortunate for several reasons. First, when comparing words within a given Congress, it would presumably be helpful to know how different the use of phrases actually is. Suppose, for example, that "Franklin.Delano. Roosevelt" receives a score of –0.3, implying it is a Democratic phrase; suppose further, however, that the bounds on that correlation are (–0.7, 0.1). In that case it clearly includes zero—or perfect nonpartisanship—implying that one cannot claim it is "truly" a Democratic phrase. Second, the same logic applies over time, too: the fact that a phrase is used more in a later Congress should affect one's certainty about its status as a polarizing term, even if the relative proportion of times it is used by the different parties remains constant. This matters given that Congress says and does more and more today than in the past, and it is precisely the notion of "never been so bad" that the authors seek to tackle. How might the authors proceed? Obviously, correlations have sampling distributions, and one can place confidence intervals around them. If that is objectionable, one might proceed via a bootstrap approach, although as with the confidence interval approach, care is needed in demarcating exactly what is being sampled from and dealing with the fact that it is the normalized frequency that enters the correlation calculation (set to be zero, on average, for every Congress).

INFERENCE, TIME, AND INSTITUTIONS The authors look at several time series that they expect to be correlated with, if not causally related to, the polarization of Congress. They find, among other things, that polarization is related to political violence, but not to legislative efficiency. That is, the work of government still gets done even if the parties disagree. Of course, such claims raise obvious issues of simultaneity and endogeneity: for example, the more a party gets done (such as Obamacare), the more the other party may respond by acting in polarized fashion. The authors also find that polarizing phrases in the Google Books corpus diffuse into Congress over time, but that less polarized language diffuses from Congress into books. The authors are quite candid that making causal inferences about such time series is fraught with difficulty: to put it most crudely, it is hard to know which causes which, and getting at the mechanisms behind the causation is even harder. One interesting observation that might lead to more helpful theorizing about all these problems is given by the authors in their comments on House control: they note that partisanship of language tends to switch in the direction of the (new) minority party. The authors speculate that this may be due to a more vocal minority attempting to filibuster majority progress. An alternative possibility is that minority parties represent more of a draw from the core of their party, since moderates tend

to come from more evenly bipartisan districts and are more vulnerable to electoral forces there, so that when a party loses power, it tends disproportionately to lose its most centrist voices (see Canes-Wrone, Brady, and Cogan 2002 for a discussion of this literature).

Thinking about parties in this way introduces a more general notion of institutions (of which parties are one type) and norms of behavior. Parties are known to "whip" their members—that is, to pressure them to vote in certain ways—and it seems plausible that they would cajole them to speak in certain ways as well. In addition, the rules that Congress uses to run itself vary over time, and future work in this area should note that such changes are likely to be reflected in the debates, and the debate structure, observed in practice.

TOWARD A STRUCTURAL MODEL? As noted above, the authors have not been shy about linking their data on speeches with the historical record in books. A further project might attempt to compare and contrast like with like, at least as it pertains to national legislatures. Recent times have seen the digitization of the British Hansard House of Commons records: every speech, every member, every session (www.hansard-archive.parliament.uk). Although it would certainly be interesting to look at polarization in comparative perspective, a more compelling target for analysis is the changing nature of language across the systems. Consider, for example, the term "liberal." In the United States this adjective is typically applied to those on the political left and connotes social permissiveness combined with notions of strict regulation of industry and a relatively generous welfare state. In Europe, in contrast, and particularly in the United Kingdom, "liberal" is less likely to be used to describe such views. Indeed, as traditionally considered, liberalism refers to free trade and a more laissez-faire method of economic production. Precisely where these European and American notions diverged is of profound interest in understanding both American "exceptionalism," in the Tocquevillian sense, and the general development of European political movements—including socialism—that are curiously absent in the United States (Hartz 1955). The authors' methods provide some clues as to how one might proceed in such an inquiry. One option is to take each trigram including the word (or appropriate stem) "liberal" and take account of the words preceding and following it. It may be that in some initial historical period, before the American Revolution, the words were used identically on both sides of the Atlantic (pertaining to trade, or speech, or meetings), but that "liberty" later took on a special ideological meaning in the United States that it did not in Britain. Furthermore, the Google Books corpus has separate data bases for British and

American English publications: whether parliaments follow presses, or presses follow parliaments, is a question for both countries. There is, of course, nothing unique about terms such as "liberal," and the best approach would be—as the authors are—agnostic about what divides groups both within and outside the parliaments of their countries.

The authors of this paper have shown how political science and economics can come together fruitfully to yield insights of value to both. Further collaborative work between the disciplines on methods of measurement is surely in order, too. Put most crudely, the social sciences do not yet have a generally accepted (or perhaps even a useful) structural model of text generation that would allow researchers to connect the language choices observed in the data with a model of rational human behavior, the parameters of which can be directly interpreted in terms of quantities we care about. In this respect the contrast between analysis of congressional speech and analysis of congressional votes is stark. For the latter, the last 15 years has seen an explosion in the application of item response models to roll-call data (see, for example, Poole and Rosenthal 1997, Clinton, Jackman, and Rivers 2004). The theoretical model underpinning the techniques typically used is that of "spatial voting" (in the sense of Davis, Hinich, and Ordeshook 1970), which is based on the proposition that elected representatives compare the status quo with the outcome promised by the new bill and choose the option that offers greater utility. Of course, not every feature of the structural model is identified (in particular, one cannot obtain outcome locations without additional assumptions), but the reduced-form estimates nonetheless correspond to some "helpful"—if somewhat idealized—world of human interaction and decisionmaking. Thus, one can readily ask, in a comparative statics fashion, what is expected to happen on seeing a bill become more attractive to a member of Congress along some dimension, how ideologically cohesive a party is, or (by imposing more structure) how representatives have moved through ideological space over time.

Matters are much less clear with text. In particular, we lack a satisfying theoretical model of human behavior that describes how and why different words, or different words in combination, are selected from some possible dictionary such that they communicate a political point or maximize utility in some way. In part, this lacuna is due to the fact that the strategy space—what agents can do given the situation they face—is extremely complicated: rather than simply vote aye or nay, politicians must decide which words (out of thousands) strike the right tone, quite apart from any selection of topic to discuss. Second, although some strategic and reactive

voting certainly does occur in legislatures, ignoring this variation seems fairly harmless in the case of voting (but see Spirling and McLean 2007). It is much less innocuous in the case of speeches, which by their very nature are responses to one another: studying their words and phrases as independent observations seems a bold, and possibly disastrous, statistical choice. Although political scientists have given presumed data generating processes for documents, especially in the context of "topic models" (for example, Quinn and others 2010), they are generally vague in terms of the role of human decisionmaking. Thus, there is room for improvement in this part of political economy: writing down a (simple) structural model that could be fit to data in some reduced form should be the goal. We have plenty of data—the authors have shown us that; we now need to work together as social scientists to put this type of data to its best use.

REFERENCES FOR THE SPIRLING COMMENT

Canes-Wrone, Brandice, David Brady, and John Cogan. 2002. "Out of Step, Out of Office: Electoral Accountability and House Members Voting." *American Political Science Review* 96, no. 1: 127–40.

Clinton, Joshua, Simon Jackman, and Douglas Rivers. 2004. "The Statistical Analysis of Roll Call Data." *American Political Science Review* 98, no. 2: 355–70.

Davis, Otto, Melvin Hinich, and Peter Ordeshook. 1970. "An Expository Development of a Mathematical Model of the Electoral Process." *American Political Science Review* 64: 426–48.

Fiorina, Morris, Samuel Abrams, and Jeremy Pope. 2010. *Culture War? The Myth of a Polarized America,* 3rd ed. London: Pearson.

Grimmer, Justin, and Brandon Stewart. Forthcoming. "Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Documents." *Political Analysis.*

Hartz, Louis. 1955. *The Liberal Tradition in America: An Interpretation of American Political Thought since the Revolution.* New York: Harcourt, Brace & World.

Lodhi, Huma, Craig Saunders, John Shawe-Taylor, Nello Christianini, and Chris Watkins. 2002. "Text Classification Using String Kernels." *Journal of Machine Learning Research* 2: 419–44.

McCarty, Nolan, Keith Poole, and Howard Rosenthal. 2008. *Polarized America: The Dance of Ideology and Unequal Riches.* MIT Press.

Monroe, Burt, Michael Colaresi, and Kevin Quinn. 2008. "Fightin' Words: Lexical Feature Selection and Evaluation for Identifying the Content of Political Conflict." *Political Analysis* 16, no. 4: 372–403.

Poole, Keith, and Howard Rosenthal. 1997. *Congress: A Political-Economic History of Roll Call Voting.* Oxford University Press.

Quinn, Kevin, Burt Monroe, Michael Colaresi, Michael H. Crespin, and Dragomir Radev. 2010. "How to Analyze Political Attention with Minimal Assumptions and Costs." *American Journal of Political Science* 54: 209–28.

Spirling, Arthur, and Ian McLean. 2007. "UK OC OK? Interpreting Optimal Classification Scores for the United Kingdom House of Commons." *Political Analysis* 15, no. 1: 85–96.

GENERAL DISCUSSION Bradford DeLong praised the authors for their contribution to documenting and explaining polarization in American politics. He thought it important to differentiate between ideological polarization and partisan polarization, with the latter being much more in evidence today. To illustrate the difference, DeLong noted that a century ago Theodore Roosevelt began his political career as an ideological firebrand, yet was also very willing not only to cut deals across partisan lines but even to wreck his own party's electoral chances to promote the policies he supported. That was an example of ideological but not partisan polarization. By contrast, the current Congress demonstrates so much partisan polarization—predominantly but not overwhelmingly on the Republican side—that it cannot even enact policies on which the two parties have historically agreed.

Steven Davis found it difficult to interpret the paper's results that drew on Google Books without knowing more about the composition of the Google Books database. In particular, he wondered whether that composition had shifted over time as economic factors—changes in the pricing of books, the emergence of new media—changed the relative supply and demand for different types of books. Such shifts, for example in the relative output of serious nonfiction books versus cheap romances or sci-fi novels, could call into question whether phrase counts from Google Books provided a valid and stable measure of political discourse. Davis also hypothesized that the more widely a book is circulated, the greater its impact on polarization, and so he asked whether data were available to allow weighting of books by their sales.

David Romer said that although he agreed with Arthur Spirling that a structural model of speech would be ideal, at the very least the paper would benefit from some simple statistical baselines. For example, measuring polarization by counting trigrams might automatically lead to finding the most frequently discussed topics to be the most polarized, even when there is broad agreement on the topic. If instead the trigrams could be compared against a null data set, like that which might be generated by

the proverbial monkeys at typewriters, then the authors could control for those trigrams that are not related to an ideological topic.

Christopher Carroll commented that the paper's technology had other potential econometric applications. For example, an analysis based on searching of a newspaper database for words like "bubble" could supplement Case, Shiller, and Thompson's survey-based analysis (see their paper in this volume) of expectations regarding housing prices.

Benjamin Kay questioned the authors' use of a simple time trend to measure political violence. He suggested replacing the lagged political violence variable with variables known to correlate with the level of political violence, such as income per capita and the proportion of teenage and young adult males in the population.

Hilary Hoynes asked if it were possible to identify the context in which a trigram occurs, for example to specify whether the underlying term or concept is being discussed positively or negatively. Doing so would sharpen the authors' ability to measure ideological polarization, she suggested.

Asked by Gerald Cohen whether he agreed with DeLong's assertion that Republicans today are more partisan and less compromising, David Gergen replied in the affirmative.

Given the thorny econometric problems associated with the time-series data that several panelists had cited, David Laibson suggested that a cross-sectional approach might fruitfully address a number of interesting questions while encountering fewer methodological difficulties. For example, such an approach might investigate what kind of language is correlated with reelection.

Responding to the discussion, Suresh Naidu agreed with Laibson that a cross-sectional analysis would be interesting, and for that reason the authors had structured the data in panel form. But, Naidu cautioned, even in such an analysis, what are of interest are such things as where a word or phrase comes from, which leads back to the problem of defining shocks that can help in identification.

Replying to Hoynes, Naidu asserted that the context of statements could only be determined with the help of a structural model of language, which, as Spirling had noted, was currently lacking. In any case, Naidu said, their data consisted only of Google Ngram counts and not the raw texts, which are what would be needed to determine context, sentiment, and the like.

Naidu agreed with Carroll that the paper's methodology would be valuable in many other contexts such as the study of expectations. He mentioned, as an example, that similar methodology had been used to measure

the farsightedness of the general public, by counting Google searches that specify a year in the future.

Replying to Davis, Naidu said that the composition of the Google Books data set was something of a mystery even to Google. The books in the data set are digitized copies from university libraries and thus reflect whatever those universities chose to acquire in the past. An attempt to screen out "junk" material had been made, but some surely remained. A further issue, according to Naidu, was that a structural break will always appear in the data set 75 years before the time of inquiry, because of the expiration of copyrights: only books in the public domain are included.

Naidu accepted DeLong's and Gergen's distinction between ideological and partisan polarization and suggested that members of Congress in fact interact along three dimensions, the third being their individual interpersonal relationships. Unfortunately, the paper's model reduced all these to a single dimension, the ideological, and that as reflected only in the members' public utterances. This, Naidu conceded, could fail to capture how the legislative process really works, for example because members might think and vote ideologically on some issues but not on others.

Naidu found Gergen's point about the extent to which the media drive ideological polarization today both interesting and worthy of investigation. The wealth of media data now available could indeed reveal much about the interaction between the media and politicians in contemporary society, but such an analysis would require focusing on a much shorter time frame than that of the present paper.

Finally, Naidu said that he and his coauthors were the first to acknowledge the many methodological problems in the paper's analysis—the lack of clarity about causality, the potential measurement errors, and so on—but he viewed the paper as a starting point for using these new data sources to answer questions that previously could not have been addressed.