

Effects of Partisan Rhetoric on Attention and Influence in the Political Blogosphere

**Chris Brown
Devra Moehler
Elizabeth Roodhouse**

University of Pennsylvania

Do blogs that use strident partisan rhetoric garner more attention and influence than those using mixed or neutral language? Existing studies on partisanship in the blogosphere categorize political blogs as left, neutral, or right. We argue that this categorical classification system obscures one of the main features of the new media system: the greater potential for ideologically extreme rhetoric, or what we refer to as strident partisan discourse. To examine the effects of stridency, we construct a new measure of blog partisanship that uses computer-based learning to identify terms and phrases that are unique to Democratic or Republican legislators, as documented in the *Congressional Record*. This measure of stridency reflects the degree to which a blog uses partisan language by generating a continuous measure ranging from extreme Democratic to extreme Republican rhetoric, rather than coding each blog's partisanship as a categorical variable. We then use this new continuous measure to test the hypotheses that rhetorical extremity in the blogosphere increases both audience traffic and influence amongst the online public. The paper provides a methodological contribution by introducing a new measure of partisanship in the blogosphere, and an empirical contribution by analyzing the effects of understudied dimensions of the new media environment. It thus serves as a complement to extant studies of partisan media effects in the online context.

This paper presents evidence from a pilot study. Please do not cite without permission.

Prepared for delivery at the APSA Political Communication Pre-Conference Meeting
Baton Rouge, LA,
August 29, 2012

Effects of Partisan Rhetoric on Attention and Influence in the Political Blogosphere

Since the mid-1990s, both academics (e.g. Sunstein 2001) and journalists (e.g. Koppel 2010) have lamented the rise of partisan commentary on cable television and the internet. Such accounts prognosticate that opinionated coverage of politics and public affairs will polarize consumers, with “liberals watching and reading mostly or only liberals, moderates, moderates; conservatives, conservatives; neo-Nazis, neo-Nazis” (Sunstein, 2001, p. 4).¹ Recent surveys indicate that the American public concurs with this view; according to the Pew Internet and American Life Project, more than half of Americans believe that the internet has “increase[d] the influence of those with extreme political views” (55%) while making it harder to discriminate between truths and falsehoods (56%) (Smith 2011, pp. 25-27). In short, the rise of partisan media in the new information environment, and its potential polarizing effects, has become a topic of mass as well as elite concern. In these discussions, fingers are often pointed at the political blogosphere.

Existing empirical evidence indicates that concerns about a polarized, and polarizing, political blogosphere are valid, though perhaps not as acutely or uniformly as many feared. Link analysis shows that prominent political blogs are far more likely to reference other blogs with similar partisan leanings suggesting that the political blogosphere is insular. Audience research indicates that political blog readers are strong partisans who gravitate towards blogs consistent with their political views and hold more polarized political attitudes than consumers of other media (Gentzkow & Shapiro 2011; Lawrence, Sides, & Farrell 2010). However, this same research hints at diversity of linking behavior within the Democratic and Republican blogospheres and diversity of reading behavior within political blog audiences. While political bloggers are more likely to link to others with similar partisan leanings, some do link to opposing perspectives in a manner that substantively engages alternative arguments (Hargittai, Gallo, & Kane 2008).² Although most blog reads are “carnivores” gravitating to blogs that match their partisan preferences, some are “omnivores” reading blogs on the left and right (Lawrence et al. 2010). The contours of this diversity within the left and right blogospheres warrant greater attention (Shaw & Benkler 2012).

These initial studies of the polarizing potential of political blogs are very useful additions to what had been a theoretically or ideologically driven discourse. However, the field is still in its infancy and these studies miss important features of political blogs that matter for assessing the danger of polarization. This study adds to existing research by testing whether blogs with more extreme partisan rhetoric attract larger audiences and garner greater influence than more neutral or mixed-language blogs. To do so we extend upon the methodology used in Gentzkow and Shapiro (2010) to create a new continuous measure of partisan rhetoric in blogs. We record the frequency of phrases used by Democratic and Republican representatives in the 2012 *Congressional Record* and then model the degree to which each blog replicate the more biased rhetoric of partisans in Congress. We then test whether political blogs with more extreme partisan rhetoric garner more audience traffic and influence.

¹ An alternative view celebrates rather than denounces the emergence of opinionated commentary and niche publics in the new media environment for their ability to accurately represent diversity of opinions and their potential mobilizing effects (Shaw & Benkler 2012). This paper is written from the perspective of those who see greater partisan extremity of rhetoric in the blogosphere as normatively problematic, but the research findings apply equally well for those who see it as beneficial.

² Furthermore, Shaw and Benkler (2012) argue that analysis that takes blog as the unit of analysis overlooks much of the discussion that is taking place between individuals, potentially of different political persuasions, within each blog – at least with respect to the Democratic blogosphere.

The novel methodological approach in this paper advances the study of partisanship in the political blogosphere in five key ways. First, prominent quantitative studies of partisanship in the political blogosphere employ link analysis (e.g. Hargittai et al. 2008), audience characteristics (e.g. Lawrence et al. 2010), or technological practices (e.g. Shaw & Benkler 2012). Results from these studies are often understood as proxies for partisan content. These creative measurement techniques are well-suited for measuring networks, readers, and procedures, but these blog traits may or may not correlate with partisan content. Given the salience of party labels in the American polity, we might not be surprised that Democrats read, and link to, blogs associated with the Democratic Party and Republicans with the Republican Party, but we should not assume that these blog discussions are best characterized as “red meat”³ promoting a clear partisan agenda at the expense of other perspectives. We measure partisan content of blogs directly.

Second, existing research categorizes partisanship on a discrete scale (Democratic/Republican, or Liberal/Conservative), which obscures variation within each category.⁴ The rise of more extreme partisan voices is one of the key features of the new media environment, and polarization is thought to result from exposure to extreme views and not just from ignoring opposing views. We seek to measure the *degree* of partisan rhetoric in each blog (a continuous measure from neutral to extreme) and to record whether the variation within the Democratic and Republican blogosphere is consequential for attracting readers and exerting influence. In short, we hope to add evidence about extremity of blogs to what has heretofore been a conversation largely about insularity of blogs.

Third, extant quantitative analysis typically includes only blogs that are associated with a partisan perspective and excludes neutral, mixed, or independent political blogs; discrete classification of top blogs precludes most quantitative analysis of the small number of neutral, but still popular, political blogs (Shaw & Benkler 2012). Evidence of polarization in the blogosphere may appear greater than it actually is due to these selection effects. This study includes partisan, non-partisan, and mixed political blogs.

Fourth, blogs follow a power-law distribution of traffic and links, so that some blogs receive dramatically more attention than others (Adamic & Huberman 2000; Farrell & Drezner 2008; Hindman 2008; Shirky 2003). While many studies acknowledge inequality in the blogosphere to justify selection of a limited number of top blogs, they do not adequately address inequality amongst the blogs selected when drawing conclusions about the polarizing potential of political blogs.⁵ We incorporate measures of audience and influence into the analysis of blog content, thus providing insights into the likely effects of partisanship in the blogosphere for majority versus minority audiences. Furthermore, the policy effect of extreme partisan blogs is likely to depend on whether extreme blogs are influential or obscure. The current analysis takes seriously the popularity and power of blogs when drawing conclusions about the polarizing potential of the blogosphere.

Fifth, we hope that the method we use for measuring partisanship in blogs facilitates replication over

³ Lawrence et. al. (2010) write: “Liberals and/or Democrats will read mostly leftwing blogs, and conservatives and/or Republicans will read mostly rightwing blogs. We dub these readers “carnivores”: they read blogs because these blogs provide them with “red meat” that accords with their partisan or ideological predilections” (144)

⁴ Gentzkow and Shapiro (2011) create a continuous measure of partisanship of internet sites based on audience data, but they devote little space to describing blogs, except to say that they attract far more lopsided partisan audiences and less traffic than news sites.

⁵ For example, Shaw and Benkler (2012) do more than most to address the power-law distribution issue by conducting a robustness check with a subsample of 65 of their initial 155 blogs. However, this approach still does not deal adequately with the fact that a very small number of blogs garner far more attention than others.

time, as well as across polities or platforms. Whereas existing means for determining partisanship of blogs rely on explicit self-identification of bloggers, US-specific or time-specific criteria for assignment, or external curators, our measure is based on language used by a legislative body. The method can be employed whenever and wherever legislative transcripts and media transcripts are concurrently available, thus facilitating comparative research.

By addressing each of these issues, we hope to provide a more accurate understanding of the political blogosphere so as to facilitate effective policies in the public interest. In addition, a better understanding of the relationship between partisan rhetoric and audience or influence can also help inform larger theoretical debates about the motivations and behaviors of media audiences and producers. The next section discusses the theories and evidence to develop predictions about the popularity of extreme partisan rhetoric in blogs. The third section describes the data, measures and methodology we employ to test the hypotheses. The fourth section discusses the results and we conclude with a discussion of implications and future research.

Partisan Rhetoric and Popularity

There are good reasons to expect a positive relationship between the extremity of partisan rhetoric and audience size, which is partly why observers have expressed such concern about the polarizing potential of the blogosphere. To predict the relationship between partisanship and web traffic, we need to examine the motivations and behaviors of political blog readers. The theory of cognitive dissonance argues that people have a motivational drive to seek out discussions that support their pre-existing attitudes, and eschew opposing perspectives, in order to avoid the discomfort of conflicting ideas (Festinger, 1957). Empirical evidence on media choice suggests that individuals selectively expose themselves to media that accords with their political views (see Hart 2009 for a comprehensive overview; Iyengar & Hahn 2009; Iyengar, Hahn, Krosnick, & Walker 2008; Knobloch-Westerwick & Jingbo Meng 2009; Stroud 2008; Valentino, Banks, Hutchings, & Davis 2009).⁶ Research also shows that political blog readers are strong partisans with relatively extreme political attitudes (Lawrence et al. 2010). Furthermore, readers of political blogs on the left and right are “more polarized than either non-blog-readers or consumers of various television news programs, and roughly as polarized as US senators” (Lawrence et al. 2010, pp. 141), and political blogs garner more distinct partisan audiences than other web-sites (Gentzkow & Shapiro 2011). Readers might turn to the web to satisfy their cravings for strong partisan perspectives that are less easy to find in mainstream media. In sum, research on selective exposure leads us to expect that political blog popularity will reflect the distribution of political attitudes among readers; given that political blog readers tend to be strong partisans and ideologues, we expect that more extreme partisan blogs will garner more attention than more neutral blogs.

H1: Extremity of rhetoric in political blogs is positively associated with audience traffic.

While we postulate a positive relationship, there are reasons to expect no relationship or even a negative relationship. Political blog readers are also more politically interested, active, educated, and elite than the general population (Lawrence et al. 2010). These knowledgeable and engaged readers might have a preference for more informative and less biased sources and they may seek out neutral or mixed platforms. If political blog readers are primarily motivated to avoid dissonant perspectives, then

⁶ Importantly, the hypothesis about selective avoidance of counter-attitudinal content has met with less success than the hypothesis about selective exposure (Garrett 2009; Knobloch-Westerwick & Jingbo Meng 2009).

the relationship will be positive. If, however, political blog readers are primarily motivated by a deliberate search for information, then the relationship will be negative. A mixture of the two motivations could result in a null relationship.

With respect to our second outcome of interest, we also expect a positive relationship between the extremity of rhetoric in political blogs and authority/influence. First, we might expect influence to follow audience traffic. If more partisan political blogs garner the most traffic then they are more likely to attract the attention of other internet content producers. Second, we expect blogs that present a distinct and strident point of view will garner more attention from other internet producers than blogs containing mixed perspectives and temperate dialogue. Producers of internet content can use links to strident blogs as a way of establishing their own credentials as a partisan spokesperson, or a group insider. Or, producers seeking controversy may cite blog writers associated with opposite ends of the spectrum. Either way, more extreme partisan blogs have an advantage over blogs where partisanship is ambiguous.

H2: Extremity of rhetoric in political blogs is positively associated with influence.

Importantly, there are also reasons to expect the opposite relationship of that posited in our second hypothesis. Mainstream web content producers may be more concerned with establishing credibility by providing quality information and balanced coverage than they are with establishing themselves as members of a partisan in-group. Credibility-seeking content providers may avoid linking to more extreme partisan blogs in favor of more balance or analytic discussions. Extant research indicates that mainstream news organizations dominate the online political environment (Hindman 2008; Tewksbury 2003), and that bloggers depend upon traditional news organizations for both content (Reese, Rutigliano, Hyun, & Jeong 2007) and credibility (Tsui 2008). From this perspective, we might expect less strident political blogs to garner more influence.

In sum, we hypothesize that blogs with more extreme partisan rhetoric will garner more audience traffic and influence, though there are also reasons to expect that evidence will prove otherwise. If we find empirical support for our hypotheses, then there are additional reasons to be concerned about the polarizing potential of political blogs. We now describe our methodological strategy for testing these hypotheses.

Data, Measurement, and Methodology

Our measure of blog partisanship assumes that staunch partisan blogs, as compared to neutral blogs, include more rhetorical phrases associated with a like-minded party, and fewer phrases associated with an opposing party. We use the *Congressional Record* to identify partisan phrases, and then rank blogs according to how similar their rhetoric is to that of a Congressional Democrat or Congressional Republican. At its heart, measuring partisanship using speech falls under the rubric of opinion mining or sentiment analysis in natural language processing and computational linguistics.⁷ Generally, these areas of research are motivated by a “desire to provide tools for information analysts in government, commercial, and political domains, who want to automatically track attitudes” (Wiebe 2005, 165). This is precisely the task undertaken below when we develop a way to measure the underlying attitudes, or degree of partisanship, of political blogs through an analysis of its text.

⁷ For a thorough overview of the field of sentiment analysis and opinion mining see Pang and Lee (2008).

In computer science there is a rich literature devoted to identifying latent opinion or sentiment using any number of the features of text or speech; furthermore, a number of articles in economics, political and computer science have previously attempted to measure political bias or partisanship using fully- or semi-autonomous methods to classify speech or text, and (Balahur, Kozareva & Montoyo, 2009; Diermeier, Godbout, Yu & Kaufmann, 2012; Gentzkow & Shapiro, 2010; Grimmer, 2010; Laver, Benoit & Garry, 2003; Pang & Lee, 2008; Wiebe, Wilson & Cardie, 2005; Yu, Kaufmann & Diermeier, 2008). This paper builds on this previous research, adopting primarily the method put forth by Gentzkow and Shapiro (2010) to identify partisan rhetoric in newspapers and place them on a continuous left-to-right scale. We replicate with some modifications the method used to identify partisan rhetoric from the *Congressional Record*, update the evidence to the current 112th Congress, and apply it for the first time to rank political blogs on a continuous left-to-right scale.

This section describes the data gathered to build a model designed to measure the partisan rhetoric of political blogs. It then provides an explanation of how partisan phrases were identified and how a model was built to measure partisanship from phrase counts. Next the section describes the independent and dependent variables of interest to this study. Finally the section describes the methodology used to determine the relationship between the independent variable, extreme rhetoric, and the dependent variables, audience and influence in the political blogosphere.

Data Sources

The analysis carried out in this paper relies on two sets of newly collected data the *Congressional Record* and a newly collected corpus of 76,172 blogposts from a sample of 48 blogs for the first six months of 2012.⁸

First, we use the *Congressional Record* to identify partisan rhetoric. Several existing studies have taken advantage of the *Congressional Record* as a source of easily identifiable and attributable partisan language (Gentzkow & Shapiro, 2010; Pang & Lee, 2008; Quinn *et al.*, 2010; Yu *et al.*, 2008). For example Diermeier *et al.* (2012) used Senate speeches from the 101st to 108th Congresses to develop a classifier for parties using a Support Vector Machine (SVM). The *Congressional Record* has one nice, notable feature; text/speech is already classified according to partisanship as each speech can be attributed to a member of the House of Representative or Senate.⁹ With recourse to existing measures of ideology derived from roll call votes (such as Clinton, Jackman, and Rivers' ideal point estimates) it is much easier to develop a more fine-tuned model to predict ideology from text (Clinton, Jackman & Rivers, 2004). We extend the research of these methods by combining several different methods and updating rhetoric to the current 112th Congress.

The sample of the *Congressional Record* collected covers the 112th Congress up until June 2012. Speeches were collected and downloaded from the Library of Congress' website.¹⁰ The *Congressional Record* itself is divided into separate sections that include all floor actions in the House of Representatives and the Senate, as well as a separate section called "Extensions of Remarks". In the "Extension of Remarks" members of the House can enter text they were unable to say on the floor due to time constraints or other extenuating circumstances,¹¹ and members may also ask to enter news

8 See Tables in the Appendix for a full list of blogs selected

9 This effectively provides researchers with a set of already annotated speech which with to train or build a model for classification purposes, often one of the most time-consuming aspects of a supervised-learning task.

10 <http://thomas.loc.gov/home/LegislativeData.php?&n=Record>

11 Speeches that run over time or for whatever reason cannot be finished in the House are included in the Extensions of Remarks.

articles, editorials, letters, or other text into the record. We started with all three sections of the *Congressional Record* in their entirety generating a total of 33,759 documents.

Besides the existing research that establishes the partisan content of floor speeches in Congress which the *Congressional Record* contains, the parties themselves pay particular attention to the issues that their members raise. For instance, in the much more restrictive House, many opportunities to speak on the floor are controlled by party leadership (Harris, 2005). During the 1990s, both Republicans and Democrats began monitoring what their members said, keeping close tabs on which members stayed on topic and rewarding those who did so. Since then party control over content has only increased. As an example, in September 2005 the Republican “Theme Team” kept track of statistics for message discipline and found that Republicans stayed on message 70 percent of the time compared to 56 percent for Democrats over the same time period. In addition to keeping track of the statistics, Republicans also created daily mailings that detailed which members were speaking most often and how many stayed on message (O'Connor, 2005).

Second, the sample of political blogs collected is listed in the Appendix at table 1 with the total number of blogposts collected as well for the 6-month period (January – June 2012). The sample itself was drawn from the highest rated blogs on the list of top 100 political blogs as rated by Technorati. A few blogs could not be collected either due to technical difficulties or due to the inaccessibility of archives, and these blogs are not included in the sample. For instance, in the case of a site like Business Insider, the number of posts during the sample period vastly outnumbers that of other blogs in the sample. A total of 48 blogs were collected from the top 100 list of political blogs.¹²

A total of 76,172 blogposts were downloaded from the list of sites. The blogs themselves were downloaded using custom Python scripts and frameworks.¹³ Once downloaded, html mark-up was stripped either by using a port of Arc90's Readability algorithm for Python which is designed to only display article text, stripping unnecessary mark-up.¹⁴ When the algorithm failed, the specific tags that contained relevant text were identified and relevant text was separated from the rest of the webpage.

Identifying Partisan Phrases

Text data is notoriously sparse and building any model to measure partisanship with text first requires pre-processing and filtering to determine the most relevant phrases that identify partisanship.

The pre-processing involved a few steps. First, text from both the *Congressional Record* and the political blogs downloaded had stopwords removed,¹⁵ words stemmed using the Porter stemming algorithm, and grouping words into bigrams and trigrams(2- and 3- word phrases/tokens). Stopwords were removed because they are unlikely to provide any information on partisanship. Stemming is done so phrases that are likely to convey similar sentiment but differ because one phrase has a singular word as opposed to a plural word in another phrase should be counted similarly. This reduces dimensionality

12 The plan is to collect more political blogs over a longer period of time, though the sample will ultimately be limited by the availability of traffic and authority data of the blogs as well.

13 Upon publication these tools and documentation will be available for download so they can be used by the community for other research projects

14 For a description of the methodology of the algorithm, see <http://lab.arc90.com/2009/03/02/readability/>. To download the port for Python, see <https://github.com/buriy/python-readability>

15 For a full list, see the Appendix. This is the default list under the Natural Language Toolkit available for Python (<http://nltk.org/>) - a collection of tools used by many researchers doing computational linguistics and natural language processing.

of the data without sacrificing the information left in the data. Lastly, words are grouped into bigrams and trigrams in recognition of the fact that many important political phrases are 2-3 words long¹⁶ (for instance, ``Affordable Care Act'', ``Keystone XL Pipeline'', ``Tax Cuts'', and ``Tax Breaks''). Three examples of the pre-processing of text are presented in the Appendix.

After pre-processing text, the next step is determining which phrases are relevant for identifying partisanship. The process used to do so follows from Gentzkow and Shapiro (2010). First, all the bigrams and trigrams used in the blog dataset are counted for each blog post in which they occur. Bigrams and trigrams that occur in less than .01 percent of blog posts are excluded. This exclusion is necessary for computational reasons since most phrases occur only once or twice and this would have vastly increased the number of phrases to search for in the *Congressional Record* data. Additionally, phrases that occur only a few times are unlikely to be useful in distinguishing partisanship between blogs as well.

Next, after gathering the set of phrases from blogs we count the number of times each phrase is used by both Republicans and Democrats in the 112th Congress. The 33,759 documents downloaded from the *Congressional Record* were then separated by speaker and party.¹⁷ To further limit the most useful phrases that identify partisanship chi-squared values for each phrase were calculated using the method detailed in Gentzkow and Shapiro (2010).

After calculating chi-squared values, the top 250 bigrams and trigrams for both Republicans and Democrats were kept for the next stage of estimating weights for each set of words.¹⁸ One of the benefits of using the *Congressional Record* as a source of partisan phrases is that estimates of partisanship and ideology have already been estimated by political scientists using roll call votes. In this case we use ideal point scores obtained using the Bayesian methods outlined in Clinton, Jackman, and Rivers (2004).¹⁹ This means that there is a readily available source of data to use to calculate the `weights' for our bigrams and trigrams. Letting the known ideology²⁰ of legislators be y_i and the vector of phrase counts be x_i , a simple approach would simply regress y_i on x_i to recover the partisan weight for each phrase then estimate the *unknown* partisanship of blogs using those weights for phrase counts in the blog. However, a simple OLS model with ideology regressed on the counts of bigrams and trigrams is not possible as the number of independent variables greatly outnumbers the number of observations. An alternative method proposed by Taddy (nd) is to use multinomial logit inverse regression while normalizing word counts by dividing each phrase count by the total number of phrases. Taddy's technique of inverse regression reduces the dimensionality of text and recovering the weights of particular phrases. For a full explanation and discussion of the method see Taddy (nd).²¹

16 Unigrams could also be used, though robustness checks show that little benefit is obtained.

17 Many times a single speech or entry in the *Congressional Record* included multiple speakers. To deal with this a Perl script was used to separate and identify the speakers for attribution.

18 For a sample of Democratic and Republican phrases see the Appendix

19 Other methods include Poole and Rosenthal's DW-NOMINATE; however, at the time of writing estimates were not available for the 112th Congress. Nevertheless, an extensive literature on the robustness and similarity of the two measures is available as well.

20 Ideology in this case is the ideal point scores calculated from roll call votes using the method found in Clinton, Jackman, and Rivers (2004). Data is available at <http://pscl.stanford.edu/blog/>. This contrasts with Gentzkow and Shapiro's (2010) use of Bush vote-share for districts as a proxy for ideology. We choose to use this measure as opposed to vote-share from the 2008 election to eliminate as much noise from the measure of ideology as possible since it is a more direct measure of ideology rather than simply a proxy, though we expect our results to hold using either measure of ideology.

21 The *textir* package available through the Comprehensive R Archive Network (CRAN) provides additional details as well (Taddy,).

How well does this method perform at predicting ideology? An easy test-case is to extract phrase-weights using inverse regression then predict ideology using the phrase weights. Figure 1 displays the predicted ideology to the original ideal point scores. The scores are correlated at .83 with only a few notable outliers. For instance, Congresswoman Gabrielle Giffords, understandably did not give many speeches relative to many of the other members of Congress due to the unfortunate shooting in December 2011 and her long recovery. As such, with limited data available to predict partisanship with speech and few roll-call votes given, the ideal point estimates using both roll call votes and speech are not near one another

Measuring Blog Partisanship

After building a model to predict partisanship using speeches from the *Congressional Record*, the next task is using the model to measure blog-partisanship using data from blogposts. Using the weights recovered from the inverse-regression, counts of bigrams and trigrams were used to measure partisanship. Table 3 in the Appendix displays the results.

Upon inspection, the measure seems to work fairly well as the measure places blogs where one would reasonably expect them in relation to one another. One of the most liberal blogs according to our measure is ThinkProgress, a well-known, progressive think tank. In comparison, one of the most conservative blogs is The Foundry, the blog of the conservative think tank The Heritage Foundation.

Independent Variable – Extreme Rhetoric in Political Blogs

Our key question is whether strident partisan rhetoric in blogs garners more influence and traffic. To answer this question we first change our partisanship measure slightly by creating our key independent variable of extremity by calculating the absolute difference between the partisanship measure and 0. Our expectations make a claim about the effect of strident rhetoric, but no claims about the effect of partisanship. Our extremity measure then removes partisanship such that the most extreme liberal and conservative blogs receive similar scores.

Dependent Variable – Audience of Political Blogs

For audience we used the number of visitors for the site during May 2012 available from Compete and the number of visitors over a 3-month period from Alexa. Unfortunately, it was difficult to obtain detailed information on the number of visitors to the blogs in our sample, though we do not have reason to believe that May 2012 was an outlier that would restrict the generalizability of any of our claims in the paper. There were no major events in May 2012 that we believe would bias results such that Conservative (or Liberal) blogs received an increase in traffic unique to that month. The issue here is that if for some reason May 2012 is unique and traffic increased for either Conservative or Liberal blogs, but not both, in May then this would bias our results, as would unusual levels of traffic on a handful of blogs that are either all extreme or all moderate. Ideally, we could look at average or total traffic statistics for a longer period of time to deal with this issue.

Dependent Variable - Authority/Influence of Political Blogs

There are a couple measures we used to assess the authority and influence of blogs in our sample. We present here 2 measures of authority that we used – the number of sites linking in through Alexa. The second measure we use is the WebArbiter Global Rank (obtained on 06/16/2012) for the blogs in the

sample.²²

Methodology for Analysis

To test the effect of extreme rhetoric on audience and authority we used a simple univariate ordinary least-squares regression with a single independent variable of extremity and dependent variables. The results are summarized in figure 2 and in table 4. Note, dependent variables were transformed using a log-transformation due to heavily right-skewed data in our sample for these variables.

Discussion of Results

Initial results show no effect of extreme rhetoric on audience or authority for our sample. Regardless of the outcome measure used, there is no significant relationship between the extremity of partisan rhetoric and either traffic or influence. However, it is unclear whether these results would necessarily hold with better measurement of the outcomes, and with a larger more representative sample of blogs and blogposts.

Our measures of authority and audience may be too noisy and our findings of null effects may be driven by measurement error. We only had audience data available for a single month and for the sites linking in it seems Alexa only updates their information every 3 months. Future iterations of this paper will experiment with different measures. One potential alternative is the Blogosphere Authority Index. An advantage of the BAI is that it is a more direct measure of authority amongst political blogs that combines a few different metrics such as the number of times a blog occurs in another blog's "blogroll" and how active a community is based on the number of comments for the blog (Karpf, 2008). Additionally, we hope to gather blogposts over a longer period of time to reduce error that might result from unusual events or irregularities in measurement for a given period of time. Furthermore, we plan to collect a larger sample of blogs to increase our power and possibly also the variation on both our independent and dependent variables. Finally, we may gain additional inferential leverage by controlling for blog characteristics, such as type of blog partisanship (Democratic or Republican), and number of blogposts.

Conclusion

This paper provides a new measure of partisanship for the political blogosphere to test whether or not strident rhetoric affects the audience of authority of a blog. The measure is derived from the *Congressional Record* by identifying partisan phrases and developing a model that takes counts of two-word and three-word phrases as the inputs and produces a continuous measure of partisanship of rhetoric as the output. We've shown that the measure of rhetorical phrases can be used to accurately predict the partisanship of both members of Congress and political blogs.

Notably, initial results that test the relationships between strident rhetoric and blog traffic, or strident rhetoric and blog authority, show that there are no relationships. We did not find empirical support for either hypothesis based on conventional wisdom. If these preliminary findings are correct, they suggests that readers are not necessarily drawn to strident partisan rhetoric online despite the fact that

22 Rankings from Alexa (<http://www.alexa.com/>) and WebArbiter ((<http://www.webarbiter.com/en/>)) are available at their respective websites.

political blog readers tend to be strong partisans. Readers do not seem to be seeking out extreme views, and blogs may be less polarizing than suggested by pessimistic observers of the internet. Furthermore, bloggers that act as mouthpieces for political parties do not seem to garner more authority in the political blogosphere. Apparently, influence is based on criteria other than partisan rhetoric. While the results may help temper the fears of the pessimists, the optimists will find little comfort in our results. Even though polarized rhetoric does not automatically lead to greater success, blog readers do not seem to avoid strident voices and even staunch ideologues can become influential bloggers.

Future iterations of this paper will test the robustness of the results. First, we want to assess whether there is an interaction between the direction of partisanship and the extremity of rhetoric. It could be that the effects of strident rhetoric are different for conservative and liberal blogs, though it is unclear why that would be the case. Second, the measures used in this paper for our dependent variables likely suffer from a great deal of measurement error, obscuring any effects that actually do exist. In the future using a more robust measure that combines several different metrics like the BAI may help us identify the effects, if any. Lastly, we would like to expand the sample of data collected by increasing the number of blogs and the time periods under examination. Because our sample covers a contentious period of primary elections for Republicans and the ramp-up to a general election, the effects of strident rhetoric may be swamped due to the time period examined. A longitudinal analysis that looks at a longer time period of analysis could help deal with this issue.

References

- Adamic, Lada A., & Huberman, Bernardo A. (2000). Power-Law Distribution of the World Wide Web. *Science*, 287(5461), 2115.
- Balahur, A., Kozareva, Z. & Montoyo, A. (2009). Determining the polarity and source of opinions expressed in political debates. *Computational Linguistics and Intelligent Text Processing*, , 468-480.
- Clinton, J., Jackman, S. & Rivers, D. (2004). The Statistical Analysis of Roll Call Data. *American Political Science Review*, 98, 355-370.
- Diermeier, D., Godbout, J., Yu, B. & Kaufmann, S. (2012). Language and ideology in Congress. *British Journal of Political Science*, 42, 31-55.
- Farrell, Henry, & Drezner, Daniel. (2008). The Power and Politics of Blogs. *Public Choice*, 134(1), 15-30.
- Garrett, R. Kelly. (2009). Politically Motivated Reinforcement Seeking: Reframing the Selective Exposure Debate. *Journal of Communication*, 59(4), 676-699.
- Gentzkow, M. & Shapiro, J.M. (2010). What Drives Media Slant? Evidence from U.S. Daily Newspapers. , 78, 35-71.
- Gentzkow, Matthew, & Shapiro, Jesse. (2011). Ideological Segregation Online and Offline. *The Quarterly Journal of Economics*, 126(4), 1799-1839.
- Grimmer, J. (2010). A Bayesian hierarchical topic model for political texts: Measuring expressed agendas in Senate press releases. *Political Analysis*, 18, 1.
- Hargittai, Eszter, Gallo, Jason, & Kane, Matthew. (2008). Cross-Ideological Discussions among Conservative and Liberal Bloggers. *Public Choice*, 134(1), 67-86.
- Harris, D. (2005). Orchestrating Party Talk: A Party-Based View of One-Minute Speeches in the House of Representatives. *Legislative Studies Quarterly*, 30, 127-141.
- Hart, William; Albarracín, Dolores; Eagly, Alice H.; Brechan, Inge; Lindberg, Matthew J.; Merrill, Lisa. (2009). Feeling Validated Versus Being Correct: A Meta-Analysis of Selective Exposure to Information. *Psychological Bulletin*, 135(4), 555-588.
- Hindman, Matthew. (2008). *The Myth of Digital Democracy*. Princeton, NJ: Princeton University Press.
- Iyengar, Shanto, & Hahn, Kyu S. (2009). Red Media, Blue Media: Evidence of Ideological Selectivity in Media Use. *Journal of Communication*, 59(1), 19-39.
- Iyengar, Shanto, Hahn, Kyu S., Krosnick, Jon A., & Walker, John. (2008). Selective Exposure to Campaign Communication: The Role of Anticipated Agreement and Issue Public Membership. *The Journal of Politics*, 70(01), 186-200.
- Karpf, D. (2008). Measuring Influence in the Political Blogosphere. *Politics Technology Review, Institute for Politics, Democracy & the Internet*, , 33-41.
- Knobloch-Westerwick, Silvia, & Jingbo Meng. (2009). Looking the Other Way. *Communication Research*, 36(3), 426-448.
- Koppel, Ted. (2010, November 14). Ted Koppel: Olbermann, O'reilly and the Death of Real News. *The Washington Post*.
- Laver, M., Benoit, K. & Garry, J. (2003). Extracting policy positions from political texts using words as data. *American Political Science Review*, 97, 311-331.
- Lawrence, Eric, Sides, John, & Farrell, Henry. (2010). Self-Segregation or Deliberation? Blog Readership, Participation, and Polarization in American Politics. *Perspectives on Politics*, 8(01), 141-157.
- O'Connor, P. (2005). Special orders and one-minute speeches hone parties' discipline. *The Hill*, , 4.
- Pang, B. & Lee, L. (2008). Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval*, 2, 1-135.
- Quinn, K., Monroe, B., Colaresi, M., Crespin, M. & Radev, D. (2010). How to analyze political attention with minimal assumptions and costs. *American Journal of Political Science*, 54, 209-

- Reese, Stephen D., Rutigliano, Lou, Hyun, Kideuk, & Jeong, Jaekwan. (2007). Mapping the Blogosphere: Professional and Citizen-Based Media in the Global News Arena. *Journalism*, 8(3), 235-261.
- Shaw, Aaron, & Benkler, Yochai. (2012). A Tale of Two Blogospheres. *American Behavioral Scientist*, 56(4), 459-487.
- Shirky, Clay. (2003). *Powerlaws, Weblogs and Inequality*. Retrieved August 20, 2012, from http://shirky.com/writings/powerlaw_weblog.html
- Smith, A. (2011). *The Internet and Campaign 2010*. Washington, D.C.: Pew Internet and American Life Project.
- Stroud, Natalie. (2008). Media Use and Political Predispositions: Revisiting the Concept of Selective Exposure. *Political Behavior*, 30(3), 341-366.
- Sunstein, Cass. (2001). *Republic.Com*. Princeton: Princeton University Press.
- Taddy, M. (). Inverse Regression for Analysis of Sentiment in Text. , , .
- Tewksbury, David. (2003). What Do Americans Really Want to Know? Tracking the Behavior of News Readers on the Internet. *Journal of Communication*, 53(4), 694-710.
- Tsui, Lokman. (2008). The Hyperlink in Newspapers and Blogs. In J. Turow & L. Tsui (Eds.), *The Hyperlinked Society* (pp. 70-83). Ann Arbor: University of Michigan Press.
- Valentino, Nicholas A., Banks, Antoine J., Hutchings, Vincent L., & Davis, Anne K. (2009). Selective Exposure in the Internet Age: The Interaction between Anxiety and Information Utility. *Political Psychology*, 30(4), 591-613.
- Wiebe, J., Wilson, T. & Cardie, C. (2005). Annotating expressions of opinions and emotions in language. *Language Resources and Evaluation*, 39, 165-210.
- Yu, B., Kaufmann, S. & Diermeier, D. (2008). Classifying party affiliation from political speech. *Journal of Information Technology & Politics*, 5, 33-48.

Appendix

Blog	URL	Total Posts (January- June 2012)
Althouse	althouse.blogspot.com	2192
American Power	http://americanpowerblog.blogspot.com/	1838
American Thinker	http://americanthinker.com	4049
Balloon Juice	http://www.balloon-juice.com/	3012
Ballot Access News (KOREA)	http://www.ballot-access.org	1297
Bookworm Room	http://www.bookwormroom.com	675
CNA Daily News	http://www.catholicnewsagency.com	1183
Cato @ Liberty	cato-at-liberty.org	999
Danger Room	http://www.wired.com/dangerroom/	450
Doug Ross @ Journal	directorblue.blogspot.com	1031
Politics K-12	http://blogs.edweek.org/edweek/campaign-k-12	151
Gay Patriot	http://www.gaypatriot.net	778
Hot Air	http://hotair.com	691
Hotline on Call	hotlineoncall.nationaljournal.com	1178
Informed Comment	http://www.juancole.com	508
JOSHUAPUNDIT	http://joshuapundit.blogspot.com	498
Legal Insurrection	http://legalinsurrection.com/	1401
The Lonely Conservative	http://lonelyconservative.com	3046
The Long War Journal	http://www.longwarjournal.org	254
Marathon Pundit	http://marathonpundit.blogspot.com	784
Mediaite	mediaite.com	5810
Media Matters	http://mediamatters.org	1082
Michelle Malkin	michellemalkin.com	611
Naked Capitalism	http://www.nakedcapitalism.com	828
Nice Deb	http://nicedeb.wordpress.com	500
No More Mister Nice Blog	http://nomoremister.blogspot.com	639
OpenMarket	http://www.openmarket.org	614
Pink News (UK)	http://www.pinknews.co.uk	1505
The American Prospect	http://prospect.org	514
Hit & Run	reason.com/blog	3564
Rhymes with Right	http://rhymeswithright.mu.nu	534
Riehl World View	http://www.riehlworldview.com/carnivorous_conserval	575
Right Wing Watch	http://rightwingwatch.org	1202
SCOTUSblog	http://www.scotusblog.com	778
Shadow Government	http://shadow.foreignpolicy.com/	162
Sister Toldjah	http://sistertoldjah.com	1215
Via Meadia	blogs.the-american-interest.com/wrm	1477
The Blaze	theblaze.com	7381
The Foundry (Heritage Inst)	blog.heritage.org	1700
The Inquisitr	http://InQuisitr.com	3495
Think Progress	thinkprogress.org	1931
-Towleroad News	http://www.towleroad.com	3331
Western Journalism	westernjournalism.com	1998
Whitehouse.gov	http://www.whitehouse.gov/blog	881
White House Dossier	whitehousedossier.com	784
Wake up America	http://www.wakeupamericans-spree.blogspot.com	746
ZeroHedge	http://www.zerohedge.com/	6296

Table 1: List of Blogs Collected and Total Posts for Sample Period

Stopwords

i	are	below
me	was	to
my	were	from
myself	be	up
we	been	down
our	being	in
ours	have	out
ourselves	has	on
you	had	off
your	having	over
yours	do	under
yourself	does	again
yourselves	did	further
he	doing	then
him	a	once
his	an	here
himself	the	there
she	and	when
her	but	where
hers	if	why
herself	or	how
it	because	all
its	as	any
itself	until	both
they	while	each
them	of	few
their	at	more
theirs	by	most
themselves	for	other
what	with	some
which	about	such
who	against	no
whom	between	nor
this	into	not
that	through	only
these	during	own
those	before	same
am	after	so
is	above	than
don	can	too
should	will	very
now	just	

Table 2: Stopwords Removed during Pre-Processing

Example of Pre-Processing

From American Thinker, 02/04/2012 (http://www.americanthinker.com/printpage/?url=http://www.americanthinker.com/articles/./2012/02/tax_reality_vs_liberal_propaganda.html)

Original

"Want to help make a huge step towards balancing the budget, make the tax code fair and equitable, and improve the economy? According to President Obama, Massachusetts Senate candidate Elizabeth Warren, and other liberals, the way to do this is to raise taxes on the wealthy: From a CBS partial transcript of the 2012 State of the Union address: Tax reform should follow the Buffett Rule. If you make more than \$1 million a year, you should not pay less than 30 percent in taxes."

Pre-Processed

['want', 'help', 'make', 'huge', 'step', 'toward', 'balanc', 'budget', 'make', 'tax', 'code', 'fair', 'equit', 'improv', 'economy', 'accord', 'presid', 'obama', 'massachusett', 'senat', 'candid', 'elizabeth', 'warren', 'liber', 'way', 'rais', 'tax', 'wealthi', 'cb', 'partial', 'transcript', 'state', 'union', 'address', 'tax', 'reform', 'follow', 'buffett', 'make', 'million', 'year', 'pay', 'less', 'percent', 'tax']

From ZeroHedge, 04/30/2012 (<http://www.zerohedge.com/news/rosenberg-takes-student-loan-bubble-wagon-and-1937-38-collapse-summarizes-big-picture>)

Original

Few have been as steadfast in their correct call that the US economy sugar high of the first quarter was nothing but a liquidity-driven, hot weather-facilitated uptick in the economy, which has now ended with a thud, as seen by the recent epic collapse in all high-frequency economic indicators, which have not translated into a market crash simply because the market is absolutely convinced that the worse things get, the more likely the Fed is to come in with another round of nominal value dilution.

['steadfast', 'correct', 'call', 'us', 'economy', 'sugar', 'high', 'first', 'quarter', 'noth', 'hot', 'uptick', 'economy', 'end', 'thud', 'seen', 'recent', 'epic', 'collaps', 'economy', 'indic', 'translat', 'market', 'crash', 'simpli', 'market', 'absolut', 'convinc', 'wors', 'thing', 'get', 'like', 'fed', 'come', 'anoth', 'round', 'nomin', 'valu', 'dilut']

From The Corner, 02/09/2012 (<http://www.nationalreview.com/corner/290608/contraceptives-vs-health-ashley-mcguire>)

Original

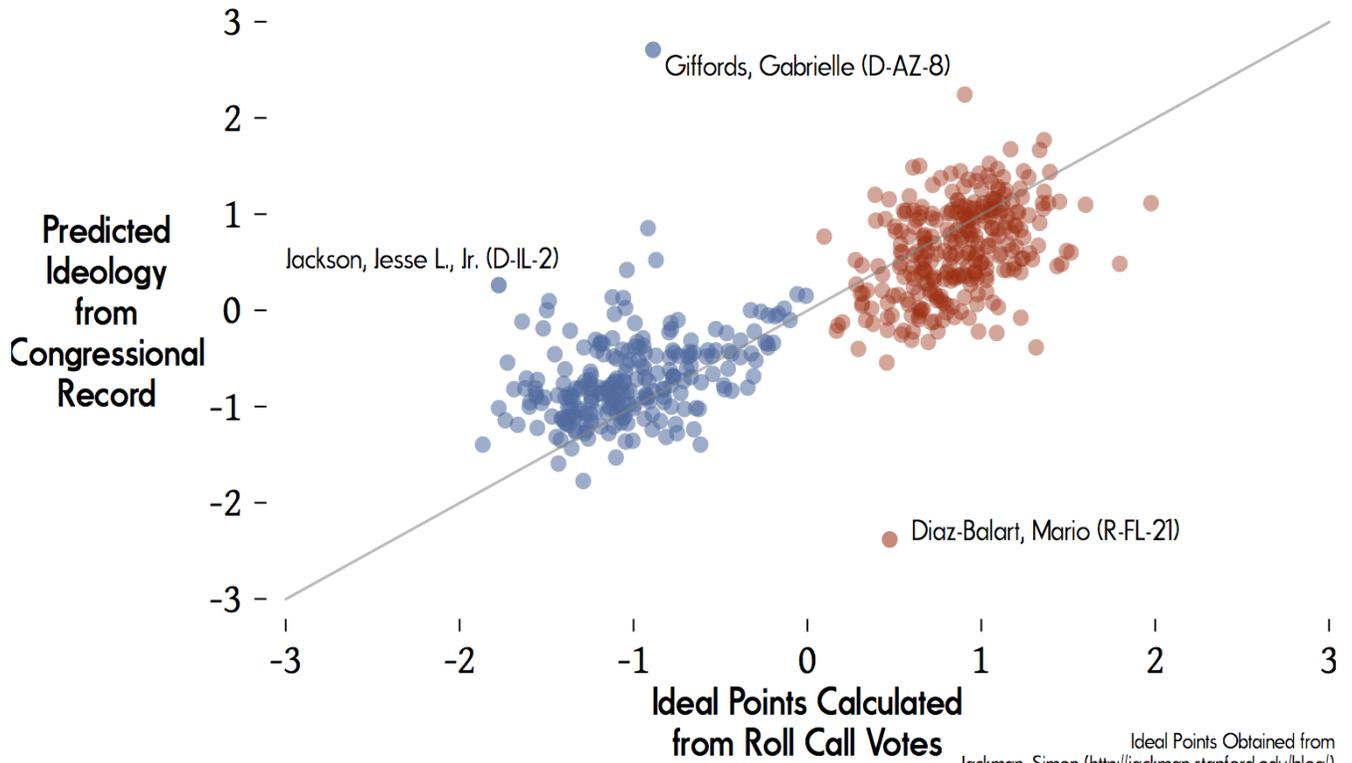
The debate surrounding the Health and Human Services Department's contraception mandate has too often been framed as one between essential women's health and the scruples of some hierarchical old priests. But Catholics of all stripes are coming to the defense of the hierarchical old priests, recognizing that this debate is, in reality, about religious freedom on the most basic of levels. This view even passed the E. J. Dionne test: Dionne wrote that the administration "utterly botched" the handling of contraception with the new health-care law.

Pre-Processed

['debat', 'surround', 'health', 'human', 'servic', 'depart', 'contracept', 'mandat', 'often', 'frame', 'one', 'essenti', 'women', 'health', 'scrupl', 'hierarch', 'old', 'cathol', 'stripe', 'come', 'defens', 'hierarch', 'old', 'priest', 'recogn', 'debat', 'realiti', 'religi', 'freedom', 'basic', 'view', 'even', 'pass', 'dionn', 'test', 'dionn', 'wrote', 'administr', 'utterli', 'botch', 'handl', 'contracept', 'new', 'health', 'care', 'law']

Ideology of the 112th Congress

Predicted vs. Actual



Ideal Points Obtained from
Jackman, Simon (<http://jackman.stanford.edu/blog/>)
(See Clinton, Jackman, and Rivers 2004)

Figure 1: Comparing Ideology Predicted from Text and Roll Call Votes

Partisanship of Blogs

Partisan Rank	Blog	Estimated Partisanship	Extremity Measure
1	Ballot Access News	-0.8176587	0.8176587
2	Pink News (UK)	-0.6504537	0.6504537
3	Think Progress	-0.5286878	0.5286878
4	Politics K-12	-0.4902978	0.4902978
5	-Toweroad News	-0.4753065	0.4753065
6	Danger Room	-0.4148405	0.4148405
7	The American Prospect	-0.3992572	0.3992572
8	SCOTUSblog	-0.3916132	0.3916132
9	Balloon Juice	-0.3087217	0.3087217
10	Naked Capitalism	-0.2750453	0.2750453
11	Media Matters	-0.2494312	0.2494312
12	Whitehouse.gov	-0.2461999	0.2461999
13	American Power	-0.2141607	0.2141607
14	CNA Daily News	-0.0615947	0.0615947
15	Althouse	-0.0533959	0.0533959
16	Via Meadia	-0.0356797	0.0356797
17	Hit & Run	-0.020306	0.020306
18	No More Mister Nice Blog	-0.0195334	0.0195334
19	Hotline on Call	-0.0032183	0.0032183
20	The Inquisitr	0.0132371	0.0132371
21	Rhymes with Right	0.0146619	0.0146619
22	Riehl World View	0.0796929	0.0796929
23	OpenMarket	0.2282366	0.2282366
24	American Thinker	0.2565257	0.2565257
25	ZeroHedge	0.2893494	0.2893494
26	The Long War Journal	0.3411107	0.3411107
27	Legal Insurrection	0.3477108	0.3477108
28	The Blaze	0.3562728	0.3562728
29	Cato @ Liberty	0.4249285	0.4249285
30	Marathon Pundit	0.4861694	0.4861694
31	The Lonely Conservative	0.4968106	0.4968106
32	Bookworm Room	0.5332581	0.5332581
33	Sister Toldjah	0.5422986	0.5422986
34	Western Jorunalism	0.5530443	0.5530443
35	Right Wing Watch	0.5954408	0.5954408
36	Informed Comment	0.6016108	0.6016108
37	Michelle Malkin	0.6234348	0.6234348
38	Gay PatrioT	0.6237066	0.6237066
39	White House Dossier	0.6372961	0.6372961
40	Mediaite	0.6386387	0.6386387
41	Shadow Government	0.6465569	0.6465569
42	JOSHUAPUNDIT	0.7162407	0.7162407
43	Wake up America	0.7502289	0.7502289
44	Hot Air	0.7521146	0.7521146
45	Doug Ross @ Journal	0.7871495	0.7871495
46	The Foundry (Heritage Inst)	0.8211926	0.8211926
47	Nice Deb	0.9281767	0.9281767

Table 3: Estimated Partisanship and Extremity

Regression Summary Tables

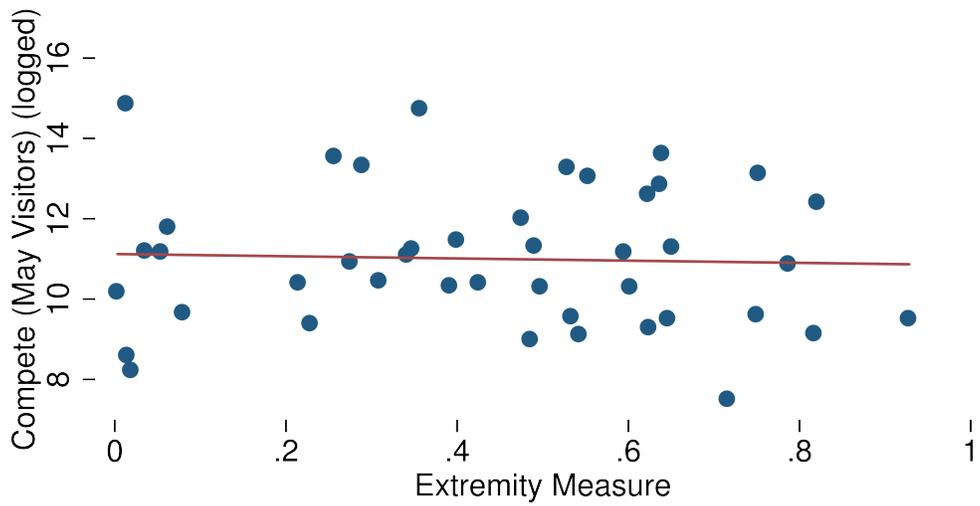
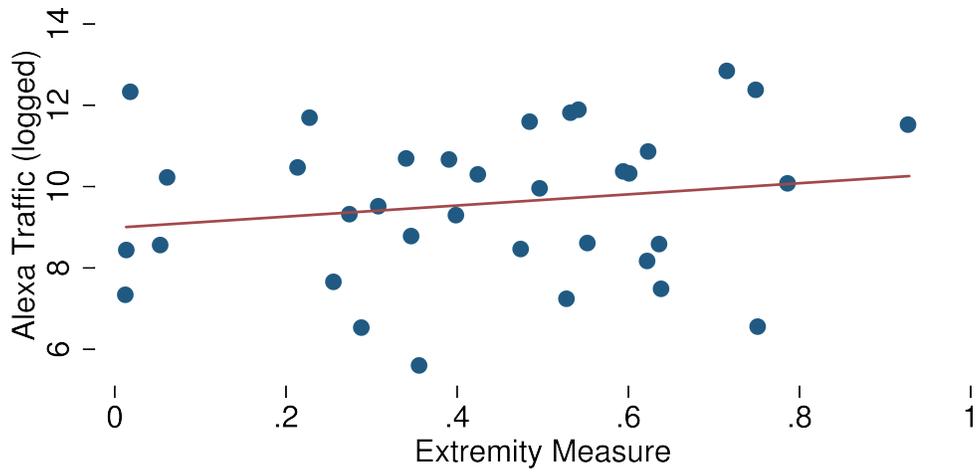
Dependent Variables	Model	Audience Models		Authority Models	
		3-Month Traffic (Alexa)	Unique Visitors May 2012 (Compete)	Sites Linking In (Alexa)	Global Rank (Web Arbiter)
Extremity Measure		1.369	-0.275	-1.141	0.35
		-1.335	-1.056	-0.77	-1.411
Constant		8.987***	11.12***	8.550***	11.01***
		-0.661	-0.534	-0.382	-0.699
Observations		35	43	35	35
R-squared		0.031	0.002	0.062	0.002

Standard errors in parentheses

*** p<0.01, ** p<0.05, * p<0.1

Note: Not all Dependent Variables were able for all sites, those sites were excluded from the analyses in those cases. Additionally, all dependent variables were transformed using a log-transformation

Regression Results – Traffic (Scatterplots with fitted lines)
Traffic vs. Extreme Rhetoric



Regression Results – Authority (Scatterplots with fitted lines)

Authority vs. Extreme Rhetoric

