A Frame of Mind: Using Statistical Models for Detection of Framing and Agenda Setting Campaigns

Oren Tsur
Harvard University
& Northeastern University
orentsur@seas.harvard.edu

Dan Calacci
Northeastern University
dcalacci@ccs.neu.edu

David Lazer
Northeastern University
& Harvard University
d.lazer@neu.edu

Abstract

Framing is a sophisticated form of discourse in which the speaker tries to induce a cognitive bias through consistent linkage between a topic and a specific context (frame). We build on political science and communication theory and use probabilistic topic models combined with time series regression analysis (autoregressive distributed-lag models) to gain insights about the language dynamics in the political processes. Processing four years of public statements issued by members of the U.S. Congress, our results provide a glimpse into the complex dynamic processes of framing, attention shifts and agenda setting, commonly known as 'spin'. We further provide new evidence for the divergence in party discipline in U.S. politics.

1 Introduction

Language is one of the main tools used by politicians to promote their agenda, gain popularity, win elections and drive societal change (Luntz, 2007). The growing availability of online archives of political data such as public statements, bill proposals, floor speeches, interviews or social media streams allows computational analysis of many aspects of the political process. The analysis performed can increase transparency, facilitate a better educated constituency and improve understanding of the political process.

In this paper we propose a framework for automatic analysis of a large collection of political texts. Specifically, we demonstrate how the use of Bayesian methods and time series analysis captures the different ways in which political parties control the political discourse. We show that topic ownership and framing strategies can be inferred using topic models. Moreover, we demonstrate how the models learned are used to construct time series of expressed agendas. These time series are fitted using autoregressive distributive-lag models in order to learn the partisan temporal relations between topics and expressed agendas.

This framework could also be applied in other domains such as ideology divergence in online forums of radical groups or for measuring the changes in public sentiment toward commercial brands.

Contribution (i) To the best of our knowledge this is the first work to analyze framing strategies on large scale in an unsupervised manner[1]. (ii) we combine topic models with regression analysis in recovering longitudinal trends. (iii) We further provide evidence for the dynamics of framing campaigns, commonly known as 'political spin[2]'. Finally, (iv) we show how this framework can shed new light on the broad scholarship on the divergence of party discipline.

2 Related Work

2.1 Political Communication Theory

Some of the theoretical constructs employed by Political Science scholars to describe features of the political communication mechanism include: *topic ownership*, *framing*, and *agenda setting*. Understanding these theoretical concepts is necessary in laying the ground for our computational approach. This subsection provides the key definitions and a brief survey of the relevant literature.

Topic/Issue Ownership We say that a candidate, a representative or a party *owns* a topic if this topic, set of ideas or the competence in handling specific issues are strongly associated with her/the

---

[1]We do use some meta data such as the speaker's party and its timestamp for the time series analysis.

[2]'Political spin' may also refer to fact twisting and factual distractions promoted using various media outlets. We do not refer to these types of spin in this work.

party (Petrocik, 1991; Petrocik, 1996; Damore, 2004). For example, environmental issues are traditionally associated with specific parties and not others (e.g. in U.S. politics, environmental issues are mostly associated with the Democratic party (Dunlap et al., 2001)).

Framing Framing is the psychological schema we use in order to organize and process our experiences. Politicians can use different contextual *frames* when referring to a specific topic, giving the public very different views on the topic at hand (Goffman, 1974; Gamson, 1989; Entman, 1993; Chong and Druckman, 2007). A notable example is the divisive partisan rhetoric used by U.S. politicians when referring to the legality of abortion. Democratic and Republican positions, framed as 'pro choice' and 'pro life', respectively, spin the abortion discourse as an issue of values of individual freedom (pro-choice) or validating the sanctity of life (pro-life). Similarly, Republicans refer to the inheritance tax by the overwhelmingly negative coinage 'death tax', while Democrats use 'estate tax'.

Framing strategies, however, go beyond the use of fixed phrases such as 'death tax' and 'prochoice'. The Affordable Care Act (ACA) and the debate over raising the minimum wage can be framed as an issue of social justice or in the context of the economic burden it incurs on tax payers and by potential job loss.

Agenda Setting and shifting Agenda setting is achieved by framing and by increased or decreased *attention* (attention shifts) in order to set or change the political, media or public agenda (McCombs and Shaw, 1972; Scheufele and Tewksbury, 2007). Some examples of agenda setting campaigns are the repeated comments about the importance of child vaccination, highlighting the need for equal pay in the 2015 State of the Union Presidential Address, or, more broadly, repeatedly addressing the need for affordable healthcare.

2.2 Computational Analysis of Political Data

The availability of archives and streams of political data is driving a growing number of computational works that address a wide array of Political Science questions. Methods vary from simple word matching to more sophisticated Bayesian models and deep learning techniques.

Slant in news articles has been modeled by (Gentzkow and Shapiro, 2010) and (Lee, 2013), comparing word tokens and n-grams to predefined lists extracted from labeled data. Hidden Markov Models are used by (Sim et al., 2013) in order to measure ideological proportions in political speech, and (Iyyer et al., 2014) use recursive neural networks for a similar task.

Topic models have been used to detect connections between contributions and political agendas as expressed in microblogging platforms (Yano et al., 2013) and for reconstructing voting patterns based on the language in congressional bills (Gerrish and Blei, 2012). The flow of policy ideas has been modeled via measuring text reuse in different versions of bill proposals (Wilkerson et al., 2013).

Nguyen et al. (2013) use supervised hierarchical topic regression to improve prediction of political affiliation and sentiment.

Expressed agendas in press releases issued by U.S. Senators have been modeled by Grimmer using author topic models (Grimmer, 2010). It is important to point to some key differences between our work and Grimmer's work. While the model used by Grimmer allows attribution of a single topic per document, we are interested in a mixed membership model as we hypothesize possible correspondence between topics and frames. Moreover, while we are interested in partisan dynamics, Grimmer is interested in the expressed agendas of individuals thus focusing on an authorship model. Finally, unlike Grimmer, we also introduce autoregressive distributed-lag models in order to capture temporal dynamics between topics and parties as reflected in the data.

Another line of work can be found in the more traditional Political Science scholarship. The success of framing strategies is studied by the analysis of real time reactions to political debates (Boydstun et al., 2014). Autoregressive models are used for analyzing adjustment of issue positions with respect to news items during the Dutch national election campaign of 2006 (Kleinnijenhuis and de Nooy, 2013). This approach is based on manual annotation of data.

Logistic regression on manually coded campaign advertisements is used in order to learn the dynamics of issue ownership by individual candidates (Damore, 2004).

While some of the works above address related research questions (agenda setting, topic ownership) or use similar computational approaches (topic models, regression models), our work is the

```
Why Can't We Crack Down On Fraud Without A Big Government Takeover Of Health Care?
Congressman John Boehner (R-West Chester) released the following statement today on
President Obama's health care speech in St. Louis, Missouri, in which he will
announce a new plan to use private contractors to investigate Medicare and Medicaid
fraud. While the President is visiting America's heartland today, I hope he will
take the opportunity to finally listen to the American people, who are shouting,
"stop" at the top of their lungs. They just don't want out-of-touch Washington
Democrats' job-killing government takeover of health care. They don't want more
than $500 billion in tax hikes. They don't want nearly $500 billion in Medicare
cuts. They don't want these outrageous kickbacks, payoffs, and sweetheart backroom
deals. We need to scrap this bill, and start with a clean piece of paper on real,
step-by-step, commonsense reforms to lower health care costs. I support rooting out
waste, fraud, and abuse in government programs like Medicare and Medicaid. This new
proposal from the President may be worthy of bipartisan support, but why can't we
crack down on fraud without a big-government takeover of health care?
```

Figure 1: Examples of a public statement released on March 10, 2010 by Republican minority leader – Congressman John Boehner (now speaker of the U.S. House of Representatives). The highlighted sequences illustrate the different topics/frames used - health care (green), economy/budget (yellow) and corruption (orange).

first to offer a complete framework for automatic detection of topic ownership and attention shifting on a large scale. Additionally, our partisan analysis provides a model for longitudinal partisan communication strategies without the need for encoding of external events and specific campaigns.

3 Data

A brief overview of the U.S. Congress The American political system is a bicameral legislature composed of the Senate (100 senators, two from each state) and the House of Representatives (435 voting members plus 6 non-voting representatives, number depends on the population of each state). Election is held every two years, in which one third of the Senators and all members of the House face reelection. Members are typically affiliated with either the Democratic Party or the Republican Party. Congressional election and Presidential election coincide every four years.

The Corpus We use a corpus of public statements released by members of Congress in both the Senate and The House of Representatives, collected by Project Vote Smart[3]. An example of a public statement is presented in Figure 1.

In this work we use all individual statements and press releases in a span of four years (2010-2013), a total of 134000 statements made by 641 representatives. This time span encompasses two Congressional elections (November 2010 and 2012). Table 1 gives the number of Democratic and Republican representatives in the three Congress terms (111-113) covered in our data.

---
[3]http://votesmart.org/

Chamber	Party	Congress Term		
		111th	112th	113th
Senate	DEM	57	51	53
	REP	41	47	45
House	DEM	257	193	199
	REP	178	242	234

Table 1: Majority shifts in the House in the 111-113 Congress terms. Independent representatives are omitted.
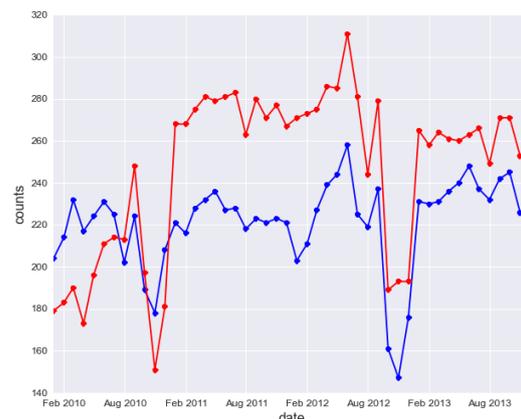

Figure 2: Monthly average number of statements by party.

While the administration was Democratic during all four years of our data, notice the Democratic loss of majority in the 112th Congress. We focus on the years 2010-2013 since Project Vote Smart has better coverage of the political discourse after 2009.

It is interesting to note that while the total number of statements per month reflects the change of majority in the November 2010 and 2012 elections (Table 1 and Figure 2), accounting for the number of seats per party it appears that the average Democrat is consistently more 'productive' with $\mu = 6.24$, $\sigma^2 = 2.1$ (Dem) and $\mu = 5.5$,

$\sigma^2 = 1.5$ (Rep) statements per month. We hence report all results after normalization by the number of seats each party posses at each timestamp.

4 Computational Framework

In order to automatically discover the correlated dynamics of attention shifts, we take a layered approach, consisting of the stages described below.

4.1 Topic Inference

In the first stage, we use topic models in order to learn topic distribution over words and identify the set of topics addressed in the corpus. Topic Modeling describes a general algorithmic framework for unsupervised discovery of a set of topics expressed in a collection of documents. The framework is based on the assumption that documents are generated by mixtures of k topics. It is therefore assumed that documents are generated by the following process: for each word in a document, we choose a topic from a given topic distribution, then choose a word from the distribution over words that the chosen topic specifies. Latent Dirichlet Allocation (LDA), the framework employed here, assumes that the distribution over topics has a Dirichlet prior. In practice, we assume a Dirichlet prior on topics and use variational Bayes (VB) optimization to infer topic distributions over words (Blei et al., 2003). In order to considerably improve efficiency, we use an online variational Bayes inference algorithm, shown to perform similarly to batch LDA (Hoffman et al., 2010). It is important to note that our goals and assumptions about the data do not lend themselvse to the use of dynamic or correlated topic models (Blei and Lafferty, 2006a; Blei and Lafferty, 2006b)[4].

4.2 Topic Assignment and Unification

The distribution of ranked topics over documents presents a "long tailed" distribution in which a few topics achieve a significant coverage of a document. This is a result of the mixed membership "generative" approach and the bag-of-words assumption. In a more realistic setting the number of topics per document is restricted. We wish to restrict the number of topics per document while still conforming to the mixture model assumption. We

therefore reassign topics to each document (statement) d in the following manner:

1. Assign a topic to each word based on distribution of topics over words infferred in the previous stage.
2. Find a set T' of k' topics ($k' < k$) that cover $q\%$ of the document in a greedy way. The topic assingment for document d will then be $d \rightarrow T'$.

4.3 Data Slicing

We slice the data according to four parameters: topic (or topical cluster), time, party and document (statement). These slicing parameters allow us the flexibility required to thoroughly analyze the data. In the time parameter, we have four settings: no slicing (all data is treated as if it were produced simultaneously), monthly slicing, weekly slicing and daily slicing, each gives different granularity of ownership patterns.

4.4 Autoregressive-Distributed-Lag Models

A linear function $b + w^T X = b + \sum_j w_j^T X^j$ is a simple yet robust method for testing dependency between X and Y. Ordinary least square regression finds the coefficients that minimize the mean square error of $Y = b + \sum_j w_j^T X^j$ given (X, Y). In our case (X, Y) are time series. We argue that a lagged dependency between two time series suggests a framing or attention shifting campaign.

Regression analysis of time series assumes independence between error terms. This key statistical property is often violated in real world data as y_t often depends on y_{t-1} thus the time series residuals tend to correlate. The consequences of violating the independence of errors are threefold: i) Statistical tests of significance are uninformative and cannot be used to prove dependency between the model parameters, ii) The coefficients learned lose accuracy, and iii) error terms are correlated, and hence contain information that is lost in analysis instead of used to leverage the prediction power of the model. The importance of controlling for autoregressive properties and for seasonality effects was recently demonstrated in the error analysis of the Google Flu Trends algorithm (Lazer et al., 2014).

In order to control for error dependency we add the auto regressing component $\gamma^T Y^n$ to the ordinary regression, as shown in Equation 1:

$$y_t = \alpha + \beta^T X^m + \gamma^T Y^n + \epsilon_t \qquad (1)$$

---

[4] We are interested in the change of the proportions of topics over time and not in the change of the word distribution within topics and we don't assume inherent correlation of topics.

Cluster	Topic ID	Top Words
Health	30	health care law will obamacare insurance repeal affordable americans costs new reform people president healthcare act coverage mandate american obama
	51	medicare seniors program social medicaid benefits fraud payments security programs cost services costs billion payment beneficiaries waste year savings million
Energy	38	project pipeline president obama keystone jobs climate energy xl construction state change permit administration approval oil will canada environmental create
	69	oil alaska gulf coast spill drilling offshore bp murkowski begich markey resources noaa said industry moratorium mexico gas administration sen
Security	34	day nation country today americans us american war world people america lives will honor families years men many th attacks
	89	nuclear united iran international israel foreign president states security weapons people world syria nations sanctions regime must government peace
Economy	68	budget spending debt president cuts fiscal government deficit will plan trillion obama house congress year federal cut economy washington billion
	88	jobs small businesses business job economy economic create will new growth work american america help creation act manufacturing can sector

Table 2: Top twenty words in selected topics in four topical clusters.

where, $\beta^T X^m$ indicates the distributed-lag terms:

$$\beta^T X^m = \sum_{i=0}^{m} \beta_i x_{t-i} \qquad (2)$$

and $\gamma^T Y^n$ indicates the autoregressive component described by:

$$\gamma^T Y^n = \sum_{j=1}^{n} \gamma_j y_{t-j} \qquad (3)$$

for some $n \leqslant t$ (notice that i ranges from 0 while j ranges from 1).

In order to control for seasonality (such as holidays and recess') we add a set of categorical variables indicating the weekday and the week-in-year of a statement, so the autoregressive model is:

$$y_t = \alpha + \beta^T X^m + \gamma^T Y^n + \sum_l W_l^T I^l(t) + \epsilon_t \quad (4)$$

Where $l \in \{day, week\}$ thus $I^l(t)$ is the identity matrix with the dimension of the seasonal granularity, in our case $I^{day} = I_{7 \times 7}$ for each day of the week and $I^{week} = I_{52 \times 52}$ for the week of the year. $I^l_{i,i} = 1$ iff t timestamp falls in the i-th day-of-week/week-in-year.

Finally, in practice it is usually sufficient restrict the autoregressive term to one parameter with $j = 1$ (accounting to the y value at the previous time stamp), this is consistent with the 24 hours news cycle reported by (Leskovec et al., 2009) among others. Since our goal is to find correlated attention shifts we can substitute the summation distributed-lag term by a single lagged term. Thus, we aim to minimize the MSE in the following model:

$$y_t = \alpha + \beta x_{t'} + \gamma y_{t-1}^n + \sum_l W_l^T I^l(t) + \epsilon_t \quad (5)$$

Where $t' = t - i$ and $i \in \{0, 1, 2, ..., 28\}$ indicating no lag, one day lag, 2 days lag, a week's lag, etc.

5 Results

5.1 Topical Ownership and Framing

5.1.1 Inferred Topics

As an input for the topic modeling module (stage 1 of the system) we use a lexicon of the 10000 most frequent words in the corpus. We use $k = 100$ as the number of topics. Experiments with $k \in \{30, 50, 500, 1000\}$ produced topics that were either too general or too incoherent. Once the topic-word distributions were inferred, topics were validated by two annotators examining the top 50 words for each topic. Annotators used hierarchical labels – an energy related topic t_i could be annotated *energy / clean-tech*, while another topic t_j could be annotated *energy / economy / keystone-xl*. Annotations were consolidated to unify the coding[5]. After consolidation annotators agreed on all topic annotations. Some examples of topics labels are 'health', 'energy', 'economy', 'boiler-plate', 'political process', 'local' and a few 'random' topics.

After topic assignment as described in Section 4.2 each document is associated with only 2–6 topics. In this work we focus on the 14 most salient (concise, general and frequent) topics in the corpus. These 14 topics fall under four topical clusters - Health, Energy, Army/Security and Economy/Budget. Table 2 contains examples of top words and labels for some of the topics from four topical clusters.

---

[5]For example, if topic t_i was labeled *energy, cleantech* by one annotator and *energy, green* by the other, the annotators would agree to use either *cleantech* or *green* consistently.
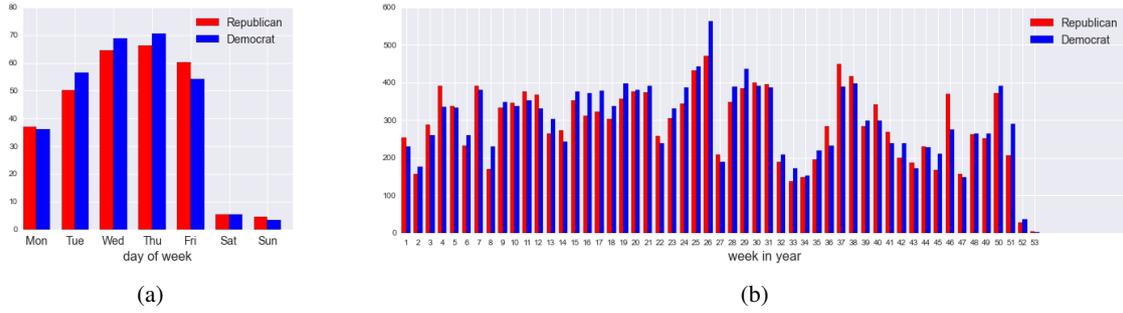
Figure 3: Seasonality effect: average number of statements issued per day of week (a) and per week in year (b).

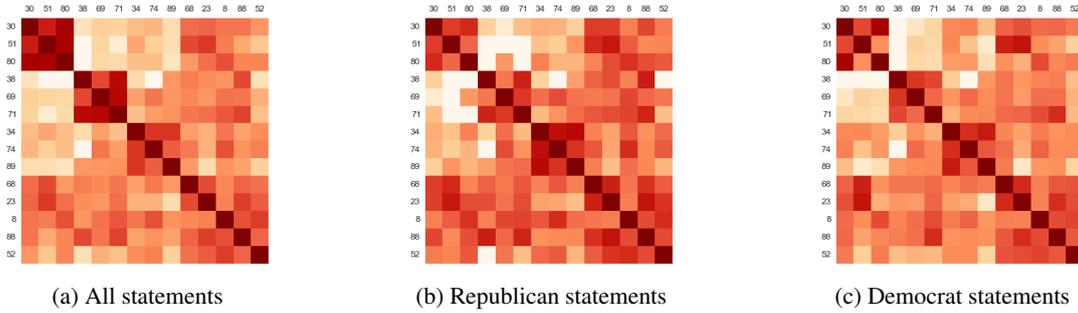(a) All statements (b) Republican statements (c) Democrat statements

Figure 4: Normalized Pointwise Mutual Information (PMI) of topic cooccurrence of 14 topics of four topical clusters Health (30, 51, 80), Energy (38, 69,71), securtity (34, 74, 89) and Budget & Economy (68, 23, 8, 88, 52)

Cluster	Topic	DEM	REP	DEM	REP
Health	30	1679	4622		
	51	746	233	3169	5386
	80	898	437		
Energy	38	128	255		
	69	1102	948	4042	3415
	71	2859	2119		
Security	34	6239	5121		
	74	3875	3071	12393	11140
	89	3807	4138		
Economy	68	12260	19916		
	23	5221	3742		
	8	6981	2456	31604	31706
	88	12845	11139		
	52	3479	1154		

Table 3: Total number of statements by party in four topical clusters. DEM indicates the Democrat party, REP indicates the Republican party.

Cluster	Topic	DEM	REP	DEM*	REP*
Health	30	46	154	2	79
	51	151	22	34	1
	80	157	27	37	2
Energy	38	47	43	2	6
	69	114	56	18	5
	71	144	52	43	5
Security	34	141	63	52	9
	74	144	46	34	8
	89	80	113	10	22
Economy	68	32	174	7	127
	23	151	49	60	9
	8	205	2	165	0
	88	137	68	63	17
	52	190	12	123	0

Table 4: Number of *weeks* each party "owned" a topic by issuing more statements (DEM, REP) and number of weeks the party owned the topic with statistical significance $p < 0.05$ (DEM*, REP*).

5.1.2 Partisan Topic Ownership

Table 3 shows the partisan ownership by providing the number of statements issued by each party on each topic and for topical clusters. It also illustrates that different topical granularities portray different ownership patterns. For example, while it seems like the health cluster is owned by the Republican party (Table 3, cluster level), a closer look at specific topics in the cluster reveals a more complex picture – the Republicans actually own only topic 30, which turns to be the most dominant topic in the cluster. Similarly, while the statement

counts in the Economy cluster are quite balanced (31604 vs. 31706), the counts of the individual topics in the cluster are polarized. Remember that these topical classes were all inferred by the LDA in an unsupervised way. These partisan ownership patterns were also confirmed by domain experts.

Longevity is a crucial factor in topic ownership. A weekly ownership of a topic is achieved by a party Q if it issued more statements on the topic than party R in that particular week. We compute the significance of the ownership assuming a null hypothesis that statements are issued by the parties

by two Bernoulli processes with the same parameters. Table 4 provides the number of weeks each party owned each topic and the number of weeks it had a significant ownership ($p < 0.05$)[6].

Topic 30 illustrates the different perspectives. The total statement count (see Table 3) reveals a clear ownership by the Republican party, issuing 73% of the statements. While turning to weekly ownership (Table 4) we get similar number (Republicans control 77% of the weeks); assuming only significance ownership, Republicans significantly own the discourse for 79 weeks while Democrats have significant ownership in only 2 weeks which means the Republicans own 97% of the significantly owned weeks.

5.1.3 Topic Cooccurrence

Topic cooccurrence could approximate the way topics are framed. A heatmap of *within* statement topic cooccurrence based on Pointwise Mutual Information (PMI) (Church and Hanks, 1990) is presented in Figure 4. The topical clusters are characterized by blocks along the diagonal. The blocks structure is to be expected due to the inherent topical similarity within clusters. It is interesting to see the inter-cluster differences in PMI between the two parties. At the cluster level, Republicans tend to use the Budget & Economy topics with topics in all other topical clusters, evident by the stronger colors in the five bottom (left) rows (columns) in 4b comparing to 4c.

A notable example is the way Republicans frame the controversial Keystone XL project (Energy, topic 38) with the impact on the job market and small businesses (Budget & Economy, topic 88), a topic traditionally owned by Democrats (see top topical words in Table 2 and topic ownership at Table 4).

5.2 Partisan Discipline

Party discipline is of great interest for political scientists (Crook and Hibbing, 1985; Bowler et al., 1999; Krehbiel, 2000; McCarty, 2001) . Typically, party discipline is examined by analysis of roll call votes on bills. Discipline, however can be also measured by adherence to party lines in talking points and agenda setting campaigns. Discipline, therefore, can be captured by conformity of lan-

---

[6]The numbers do not necessarily add up to 208 (the number of weeks in four years) due to weeks with no significant ownership , e.g. the parties issued a similar number of statements (usually zero) on that topic.
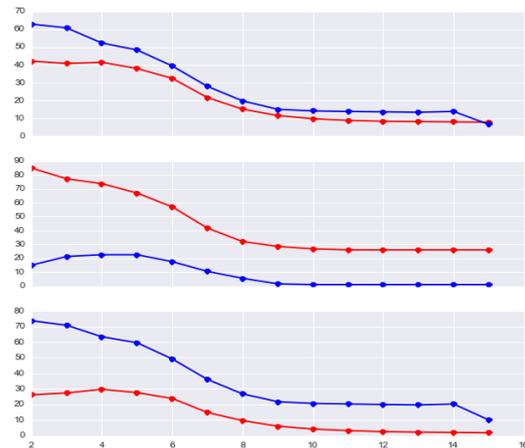

Figure 5: Average number of n-grams owned by each party on all topics (top), in Republican owned topics (middle) and in Democrat owned topics (bottom).

guage in public statements. While it is "common knowledge" among political scientists that Republicans are more adherent to "talking points" – to the best of our knowledge there are no large scale studies that support (or refute) that.

In the absence of official lists of "talking points", repeated use of similar phrases (n-grams) can provide an indication for the level of party discipline. In each topic, we looked at all n-grams ($n \in \{2, ..., 14\}$) that were used by more than five members of the Congress. For example, the tri-gram "the American people" (topic 38) appears in 81 statements made by 54 members of congress, only two of them were Democrats. Similarly, the tri-gram "social security benefits" (topic 51) appears in 123 statements, issued by 89 members, 71 of which were Democrats. Examining "ownership" of n-grams (per n-gram, per topic) reveals that that Republicans do tend to stick to talking points more than Democrats do.

Figure 5 provides the average number of n-grams owned by each party over all topics (top), over Republican owned topics (middle) and over Democratic owned topics (bottom). While on average Democrats own more n-grams than Republicans (Figure 5, top), the difference is marginal and is attributed to the fact that Democrats own more topics than the Republicans (10 vs. 4, see Table 3). Comparison between n-gram ownership in Democratic owned topics and Republican owned topics (Figure 5, middle and bottom) shows that while each party owns more n-grams in the *topics it owns*, Republicans present stronger ownership over the n-grams in their owned topics than Democrats in their respective owned topics. Moreover, Republicans present relative discipline even

in Democratic owned topics.

Manually looking at some sampled n-grams it appears that mid-length n-grams are shared "talking points" and longer n-grams are full citations from bill proposals and committee reports. These findings are in line with textual sharing semantics (Lin et al., 2015).

5.3 Time Series Analysis

To this end we create two time series for each topic $c \in T$: $S_c^{C_D}$ – daily normalized counts for Democrats and $S_c^{C_R}$ – daily normalized counts for Republicans. Normalization of counts is needed in order to account for the bias introduced by the difference in the number of seats each party holds and the changes in that number in the different terms as apparent from Table 1.

Our data exhibit two strong seasonality effects: a weekly cycle with the lowest point on the weekend and peaking on Thursday (Figure 3a), and a yearly cycle with low points at the weeks of 4th of July, Thanksgiving, August recess and Christmas (Figure 3b). These seasonality effects are captured by the added terms in Equation 4.

After time series are constructed we apply first-difference detrending (Enders, 2008) in order to transform the time series to stationary series and avoid trend-incurred correlations.

We fit autoregressive-distributed-lag models for all pairs in $\{X = S_{c,l}, Y = S_{c'}\}$, where $c, c' \in T$ (topics), $l \in \{0, 1, 2, 3, ..., 7, 14, ..., 28\}$.

In this setting we fit 5153 pairs of time series of which 718 pairs had a significant coefficient for X ($p < 0.05$). Artificial significance due to abundance of fitted models was accounted to by applying the strict Bonferroni correction (Dunn, 1961) on the significance level. The correction resulted in 103 significant correlations, most of them with lag of up to 3 days. Table 5 gives the number of intra/inter significant correlation for lags $l \in 0, 1, 2$.

One example for such correlation is the Republicans "responding" to Democratic topic 88 with with topic 8 (intra-cluster) in one and two days lag. We interpret this as a different spin on the budget issue. Another example is the Democratic party corresponds to Republican topic 30 with topic 88 (inter-cluster) on the same day (no lag). We interpret this as a way to place the Acordable Care Act in a specific economic frame. We note that while the lagged correlated time series do not imply a responsive pattern, a significance of lagged

Cluster	Dependent	Significant Correlations		
		$l = 0$	$l = 1$	$l = 2$
Intra-cluster	DEM	28	1	1
	REP	26	4	5
Inter-cluster	DEM	15	2	0
	REP	17	0	0

Table 5: Number of statistically significant ($p < .05$, Bonferroni corrected) daily lagged correlations between cross-party time series.

correlation may suggest such a pattern. We provide some evidence in the qualitative analysis in the next section.

5.4 Discussion and Qualitative Analysis

Inter and intra-cluster correlations can be interpreted as manifestations of different types of framing strategies and campaigns for attention shifts. A detailed analysis of the interplay between the different frames is beyond the scope of this paper and is left for political scientists.

The majority of the significant correlations were found with no lag. It is important to note that these correlations are found significant even after accounting to autoregressive patterns. Zero-lag correlations could be interpreted in a number of ways. Two probable interpretations are (i) daily time series are too crude to model lag patterns, and (ii) the parties respond to some external event at the same time. While we cannot address (i) due to sparseness and noise[7], we can sample statements and examine them manually. Manual examination reveals a strong responsive pattern in peaking trends. One typical example is the Republican spike in topic 30 on March 10. The statement at Figure 1 is very illustrative as it *explicitly* refers to a statement by President Obama. Explicit references to statements made by the other side are found more frequently in Republican statements and reveal a clear responsive pattern that also suggest a strong party discipline, in line with the results in Section 5.2. This small scale qualitative analysis complements the quantitative results reported in Section 5.3 and provide evidence for a responsive pattern even in zero lag series.

6 Conclusion

We presented a statistical framework for the analysis of framing strategies and agenda setting campaigns in the political sphere. Combining topic models and time series analysis, we modeled topic

---

[7]The exact timestamp is sometimes missing, set to midnight or affected by external factors.

ownership and party discipline and analyzed responsive patterns in an unsupervised way and with no prior knowledge of the political system. Our work draws from political science theory, validating some theoretical constructs and shedding new light on others. The proposed framework and the results could be further used and interpreted by political scientists and communication scholars.

7 Acknowledgments

References

David Blei and John Lafferty. 2006a. Correlated topic models. *Advances in neural information processing systems*, 18:147.

David M Blei and John D Lafferty. 2006b. Dynamic topic models. In *Proceedings of the 23rd international conference on Machine learning*, pages 113–120. ACM.

David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent dirichlet allocation. *the Journal of machine Learning research*, 3:993–1022.

Shaun Bowler, David M Farrell, and Richard S Katz. 1999. Party cohesion, party discipline, and parliaments. *Party discipline and parliamentary government*, pages 3–22.

Amber E Boydstun, Rebecca A Glazier, Matthew T Pietryka, and Philip Resnik. 2014. Real-time reactions to a 2012 presidential debate a method for understanding which messages matter. *Public Opinion Quarterly*, 78(S1):330–343.

Dennis Chong and James N Druckman. 2007. Framing theory. *Annu. Rev. Polit. Sci.*, 10:103–126.

Kenneth Ward Church and Patrick Hanks. 1990. Word association norms, mutual information, and lexicography. *Computational linguistics*, 16(1):22–29.

Sara Brandes Crook and John R Hibbing. 1985. Congressional reform and party discipline: The effects of changes in the seniority system on party loyalty in the us house of representatives. *British Journal of Political Science*, 15(02):207–226.

David F Damore. 2004. The dynamics of issue ownership in presidential campaigns. *Political Research Quarterly*, 57(3):391–397.

Riley E Dunlap, Chenyang Xiao, and Aaron M McCright. 2001. Politics and environment in america: Partisan and ideological cleavages in public support for environmentalism. *Environmental politics*, 10(4):23–48.

Olive Jean Dunn. 1961. Multiple comparisons among means. *Journal of the American Statistical Association*, 56(293):52–64.

Walter Enders. 2008. *Applied econometric time series*. John Wiley & Sons.

Robert M Entman. 1993. Framing: Toward clarification of a fractured paradigm. *Journal of communication*, 43(4):51–58.

William A Gamson. 1989. News as framing. *American Behavioral Scientist*, 33(2):157–161.

Matthew Gentzkow and Jesse M. Shapiro. 2010. What drives media slant? evidence from u.s. daily newspapers. *Econometrica*, 78:35–71.

Sean Gerrish and David M. Blei. 2012. How they vote: Issue-adjusted models of legislative behavior. In *Neural Information Processing Systems (NIPS)*, pages 2762–2770.

Erving Goffman. 1974. *Frame analysis: An essay on the organization of experience*. Harvard University Press.

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

Matthew Hoffman, Francis R Bach, and David M Blei. 2010. Online learning for latent dirichlet allocation. In *advances in neural information processing systems*, pages 856–864.

Mohit Iyyer, Peter Enns, Jordan Boyd-Graber, and Philip Resnik. 2014. Political ideology detection using recursive neural networks. In *Proceedings of the 52nd meeting of the Association for Computational Linguistics*.

Jan Kleinnijenhuis and Wouter de Nooy. 2013. Adjustment of issue positions based on network strategies in an election campaign: A two-mode network autoregression model with cross-nested random effects. *Social Networks*, 35(2):168–177.

Keith Krehbiel. 2000. Party discipline and measures of partisanship. *American Journal of Political Science*, pages 212–227.

David M Lazer, Ryan Kennedy, Gary King, and Alessandro Vespignani. 2014. The parable of google flu: Traps in big data analysis. *Science Magazine (AAAS)*.

Han Soo Lee. 2013. Do national economic and political conditions affect ideological media slant? *Political Communication*, 30:395–418.

Jure Leskovec, Lars Backstrom, and Jon Kleinberg. 2009. Meme-tracking and the dynamics of the news cycle. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 497–506. ACM.

Y. Lin, D. Margolin, and D Lazer. 2015. Uncovering social semantics from textual traces: A theory-driven approach and evidence from public statements of u.s. members of congress. *Journal of the Association for Information Science and Technology*, page (forthcoming).

Frank I. Luntz. 2007. *Words That Work: It's Not What You Say, It's What People Hear*. Hyperion.

Nolan McCarty. 2001. The hunt for party discipline in congress. In *American Political Science Association*, volume 95, pages 673–687. Cambridge Univ Press.

Maxwell E McCombs and Donald L Shaw. 1972. The agenda-setting function of mass media. *Public opinion quarterly*, 36(2):176–187.

Viet-An Nguyen, Jordan Boyd-Graber, and Philip Resnik. 2013. Lexical and hierarchical topic regression. In *Neural Information Processing Systems*.

John R Petrocik. 1991. Divided government: Is it all in the campaigns? *The politics of divided government*, pages 13–38.

John R Petrocik. 1996. Issue ownership in presidential elections, with a 1980 case study. *American journal of political science*, pages 825–850.

Dietram A Scheufele and David Tewksbury. 2007. Framing, agenda setting, and priming: The evolution of three media effects models. *Journal of communication*, 57(1):9–20.

Yanchuan Sim, Brice Acree, Justin H Gross, and Noah A Smith. 2013. Measuring ideological proportions in political speeches. In *Proceedings of EMNLP*.

John Wilkerson, David A. Smith, and Nick Stramp. 2013. Tracing the flow of policy ideas in legislatures: A text reuse approach. In *New Directions in Analyzing Text as Data*, September.

Tae Yano, Dani Yogotama, and Noah A. Smith. 2013. A penny for your tweets: Campaign contributions and capitol hill microblogs. In *International AAAI Conference on Weblogs and Social Media (ICWSM)*, July.