Correia, Rion B., K. N. Chan, & L.M. Rocha [2015]. "Polarization in the US Congress." *The 8th Annual Conference of the Comparative Agendas Project (CAP)*. Lisbon, Portugal, June 23-24, 2015.

Discourse Polarization in the US Congress

Rion Brattig Correia[1,3], Kwan Nok Chan[4,5], Luis M. Rocha[1,2*]

1 School of Informatics & Computing, Indiana University, Bloomington, IN, USA
2 Instituto Gulbenkian de Ciência, Portugal
3 CAPES Foundation, Ministry of Education of Brazil, Brasília DF 70040-020, Brazil
4 School of Public & Environmental Affairs, Indiana University, Bloomington, IN, USA
5 Dept of Politics and Public Administration, The University of Hong Kong
*** E-mail: rocha@indiana.edu**

Background. *Political polarization* can be defined generally as "movement away from the center toward the extremes" in policy preferences [3, p. 567]. There is general scholarly agreement that lawmakers in the United States Congress (USC) increasingly "appear to represent relatively extreme support coalitions rather than the interests of middle-of-the-road voters" [6, p. 1061]. While this definition of is broadly accepted, in practice, congressional polarization is commonly estimated from roll call votes and bill cosponsorship data. Such data, however, cannot uncover whether members across the aisle are in disagreement over programmatic details of bills or something more fundamental, such as policy agendas and values.

We begin to address this important distinction via analysis of congressional floor speeches. Our working assumption is that polarization extends to legislative activities other than voting and bill cosponsoring. In particular, we expect the discourse utilized by congress members to reflect their different policy agendas and values, something that roll call votes and cosponsorship data cannot directly reveal. To examine this proposition, we employ machine learning methods to characterize discourse polarization in legislator speeches. The advantage of our data-driven approach is that no assumptions ar required about policy differences between political parties. By capturing the (less studied) discourse aspects of polarization, our computational approach contributes a novel empirical measurement apparatus to expand current theory on legislative politics, towards a better understanding of one of the key institutions of the American democracy

If polarization exists in a substantive manner at the level of policy agendas and values—not only in voting and cosponsorship patterns—then the discourse utilized by members of both parties should be distinguishable. In other words, adopting more partisan agendas should lead to more distinctive discourse and subsequently more accurate text classification. Also, it is well-known that issues are framed with distinct phraseology to suit different political agendas [2], including political "dog whistles" [4]. This leads to two main questions: 1) how does the automatic classification of speeches by party affiliation vary in time? 2) can specific verbal features be identified to characterize legislative agendas?

Methods. Legislator speeches from 1994 to 2012 were collected from the US *Congressional Record*. They were subsequently hand-curated (e.g. for speech, topics and party affiliation) and stemmed. As textual units (documents) we considered both individual speeches and legislators (by concatenating the speeches of every legislator); this allows us to infer how distinctive the text of individual speeches is, as well as how distinctive all the text associated with a legislator is. As textual features for classifiers we used n-word-grams, with $n = 1, 2, 3$. We identified $3,000$ features per year; $1,000$ of each type. For feature selection a score $S(w) = |p_D(w)| - |p_R(w)|$ was computed, where $p_D(w)$ ($p_R(w)$) is the probability that textual feature w appears in a document from the Democratic (Republican) party set of speeches or legislators. Top features were then selected by the rank product between the rank of $S(w)$ and the rank of $D(w) = |\{d : w \in d\}|$, the number of documents (speeches or legislators) where feature w appears. Therefore, the top features maximize class discrimination and document frequency. Classifiers used were: *Variable Trigonometric Threshold* (VTT) [1,5]; *Support Vector Machine* (SVM); *Logistic Regression* (LR); and *Naive Bayes* (NB). Performance was evaluated across years using Area Under the Curve (AUC), Matthews Correlation Coefficient (MCC), Balanced F-Score, and Accuracy. Performance is reported as the fold mean using stratified 4-fold cross-validation.

Results. The automatic binary classification of speeches achieves very decent performance (Fig. 1, shows AUC and MCC results for the SVM classifier). It is possible to correctly predict if a speech was delivered by a Republican or Democratic legislator, solely by considering the occurrence statistics of n-gram features in the speeches. Classifiers that take into account feature covariance (LR & SVM) were significantly better on most performance metrics than those that do not (VTT & NB), even though performance is very similar. When legislative speech is concatenated and analyzed at the legislator level, the performance is even higher (Fig. 1, middle).

An important salient result is that classification performance clearly improves over the years, as the linear regressions in Figure 1 demonstrate. The regression lines are significant ($p < 0.001$) with high values of the coefficient of determination and clear positive slopes for all classifiers and performance measures. This demonstrates that speeches and lawmakers have become easier to automatically classify in more recent times, which in turn indicates that lawmakers increasingly used party-distinctive terms—a sign of growing polarization.

Several observations about polarization arise from the performance of the classifiers. For instance, on legislator classification (Fig. 1, middle), except for the year when the House majority changed from the Republican to the Democratic party (2007), every year of the G.W. Bush presidency observes a classification performance below the regression line. In contrast, most years of Democratic presidency (Clinton and Obama) are above the regression line—especially in the case of the Democratic party House majority in 2009 and 2010. This suggests, at least in this period, higher polarization in the House when a Democrat controlled the administration.

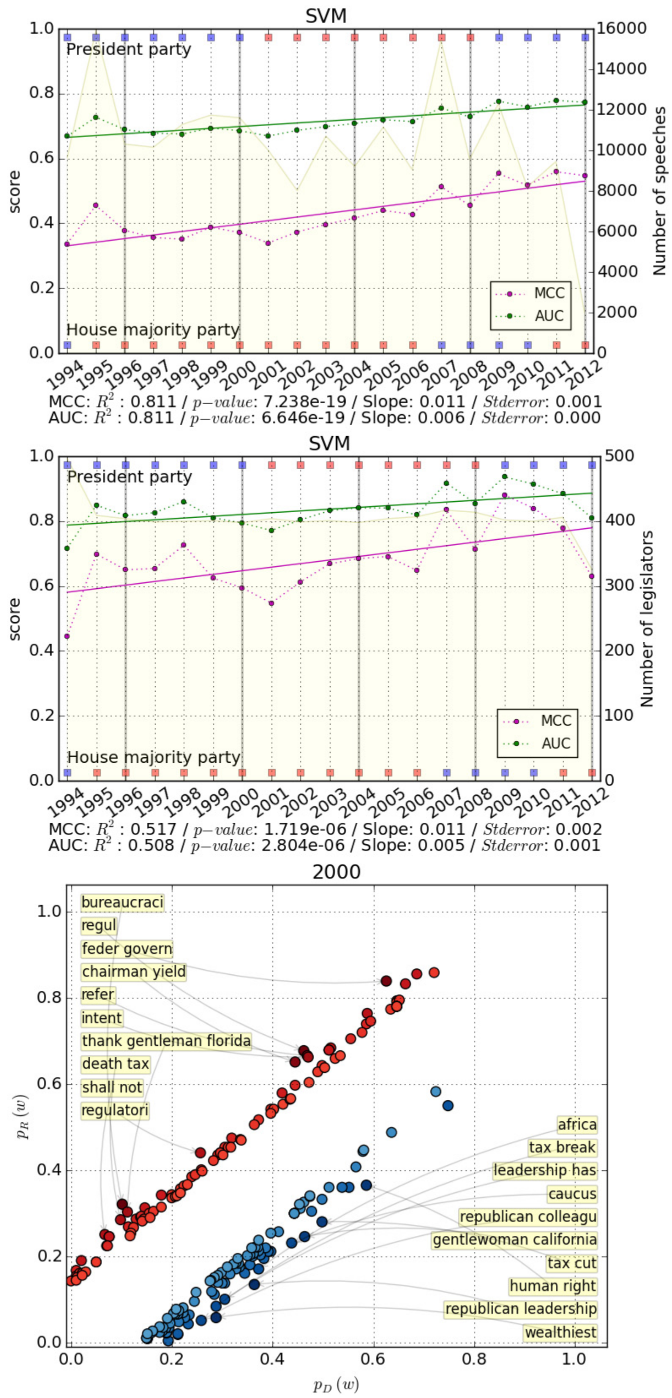

Figure 1. SVM classification for speech (top) and lawmakers (middle). MCC (magenta) and AUC (green) mean fold scores across years (values in left vertical axis). Linear regression shown for each performance measure, with R^2, p-value and slope values underneath the plot. Top and bottom squares denote party of president and house majority, respectively; blue for Democratic, red for Republican. Shaded yellow area denotes the number of speeches or lawmakers per year available (right axis). (bottom) Top 100 features for the year 2000; top 10 features for each party are annotated. Blue (red) circles are features more likely to occur in Democratic (Republican) speeches; darker colors correspond to higher $S(w)$ scores.

The distinguishing textual features (e.g. for year 2000 in Fig. 1, bottom) reveal broad patterns of partisan rhetoric. While some of the terms are domain-specific references to chamber proceedings (e.g. 'chairman yield') and political entities and personalities (e.g. 'thank gentleman Florida', 'gentlewoman California'), other features offer agenda characterizations that comport with the general policy positions of the two parties. Democrats appear to have a distinctive interest on education, health, and wealth inequality for example ("human right", "wealthiest"), whereas Republicans are are distinguished by interest in fiscal sustainability, efficiency, debt, and the free market ("regulation", "bureaucracy"). Moreover, features in particular years indicate use of distinct phraseology for issue framing. Some examples include "death tax" (2000) and "african american" (1994).

We repeated the analysis for subsets of speeches annotated with a top-level, human-annotated topic such "budget", "energy", "tax" and "security". Interestingly, classification performance significantly increases longitudinally for topics such as "energy" and "security", but shows no significant increase for topics such as "tax" and "budget". This suggests that some topics have become more polarized, but others have not. Computing the classification performance of speeches per topic for all years reveals that some topics are much easier to classify overall than others. For instance, the AUC for topics like "budget", "medicare", "tax", "social security" is highest, reaching values within $[0.7, 0.8]$, whereas topics like "gun", "terror", "Israel" and "security" lead to an AUC below 0.6. This shows that the discourse between the parties is much more distinct in the former set of topics than the latter. The variation in intensity indicates that polarization could be contained to those issues where partisan rhetoric is viewed favorably. Lawmakers might adopt a more "bipartisan" tone over issues where moderate views prevail.

Conclusion. Our analysis demonstrates that text mining methods can be used to reveal policy agendas and issue frames in House deliberations of the U.S. Congress. The ability to distinguish the discourse of legislators from each party provides a useful means to study the semantic roots of polarization and how it unfolds in time, under specific policy issues or topics. From 1994 to 2012 the ability to classify party affiliation from the text of speeches delivered by lawmakers has systematically increased; moreover, this result is robust to choice of classifier. This means that the discourse utilized by each party is increasingly distinct, suggesting increased *discourse polarization* in this period. The polarizing discourse features we extracted across years, and the per-topic analysis of polarization, should facilitate future analyses of the use of framing devices in political communication such as "dog whistles".

[1] Alaa Abi-Haidar, Jasleen Kaur, Ana Gabriela Maguitman, Predrag Radivojac, Andreas Rechtsteiner, Karin Verspoor, Zhiping Wang, and Luis Mateus Rocha. Uncovering protein interaction in abstracts and text using a novel linear model and word proximity networks. *Genome Biology 9*, Suppl 2(S11), 2008.

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

[3] Morris P Fiorina, Samuel A Abrams, and Jeremy C Pope. Polarization in the american public: Misconceptions and misreadings. *The Journal of Politics*, 70(02):556–560, 2008.

[4] Robert E. Goodin and Michael Saward. Dog whistles and democratic mandates. *The Political Quarterly*, 76(4):471–476, 2005.

[5] Artemy Kolchinsky, Anália Lourenço, Heng-Yi Wu, Lang Li, and Luis Mateus Rocha. Extraction of pharmacokinetic evidence of drug-drug interactions from the literature. *PLoS ONE*, 10(5):e0122199, 2015. doi: 10.1371/journal.pone.0122199.

[6] Keith T Poole and Howard Rosenthal. The polarization of american politics. *The Journal of Politics*, 46(04):1061–1079, 1984.