Replication materials

Replication materials for "Who Leads? Who Follows? Measuring Issue Attention and Agenda Setting by Legislators and the Mass Public Using Social Media Data", forthcoming in APSR

Abstract: Are legislators responsive to the priorities of the public? Research demonstrates a strong correspondence between the issues about which the public cares and the issues addressed by politicians, but conclusive evidence about who leads whom in setting the political agenda has yet to be uncovered. We answer this question with fine-grained temporal analyses of Twitter messages by legislators and the public during the 113th U.S. Congress. After employing an unsupervised method that classifies tweets sent by legislators and citizens into topics, we use VAR models to explore whose priorities more strongly predict the relationship between citizens and politicians. We find that legislators are more likely to follow, than to lead, discussion of public issues, results that hold even after controlling for the agenda-setting effects of the media. We also find, however, that legislators are more likely to be responsive to their supporters than to the general public.

This README file provides an overview of the replications materials for the article. The Data section describes the main dataset required to reproduce the tables and figures in the paper. The Analysis section summarizes the purpose of each R or python script. The var sections contains the output (Impulse Response Functions, IRFs) of the main VAR model used in the study (as well as the issue-level VAR models). Finally, the Dashboard and Topics section offers instructions on how to replicate the topic models in the paper.

Note: In compliance with Twitter's Terms of Service, we cannot publicly share the raw files containing the full text of the tweets. Instead, we provide the document-feature matrices required to replicate the topic models in the paper. They can be found in Dataverse.

Data

- `data/main-time-series.csv`: dataset with information about the proportion of attention that each group under analysis paid to each issue each day. It contains the following variables:
  - **date**: dates in numeric format
  - **topic**: topic codes, range from 1 to 104, where [101,102,103,104] are topics that we merged a ad-hoc. Not all topics are of a political nature. These are the political ones used in our analysis: [3, 7, 9, 12, 14, 15, 16, 18, 20, 23, 28, 32, 33, 36, 37, 39, 41, 43, 46, 47, 48, 49, 50, 51, 53, 58, 62, 63, 64, 66, 67, 70, 75, 81, 83, 85, 88, 89, 93, 96, 97, 99, 100, 101, 102, 103, 104]
  - **dem**: the attention that Democrats in Congress paid to the given issue in that particular day
  - **rep**: the attention that Republicans in Congress paid to the given issue in that particular day
  - **public**: the attention that the Attentive Public paid to the given issue in that particular day
  - **pubdem**: the attention that Democratic Supporters paid to the given issue in that particular day
  - **pubrep**: the attention that Republican Supporters paid to the given issue in that particular day
  - **random**: the attention that the General Public paid to the given issue in that particular day
  - **random_us**: the attention that the General Public (located in the United States) paid to the given issue in that particular day
  - **media**: the attention that naional Media organizations paid to the given issue in that particular day
- `data/pa2our_topics_crosswalk_merged_subissues.csv`: dataset mapping our political topics to the major topic codes of the *Comparative Policy Agendas Project*, as well as providing human readable labels for all these political topics.

Var

- `var/var_irfs-MAIN.Rdata`: Impulse Response Functions for the main VAR model of the study (used to generate the results in Figure 2, 3, and 6).
- `var/onetime-structural-shock-irfs-results.csv`: 15-day IRFs for a one-time and a permanent 10-point increase in attention (used also to generate Figure 2, 3, and 6).
- `var/issue-level/*`: Impulse Response Functions for the issue-level VAR models of the study (used to generate the results in Figure 4 and 5).

Analysis

- 01-table3.R to replicate Figure 3 of the paper, where we show the correlation between the issue attention distribution of the different groups under analysis.
- 02-figure1.R to replicate Figure 1 of the paper, where we show the average attention paid to each topic for the whole period of analysis.
- 03-figure2.R to replicate Figure 2 of the paper, where we show the results of our main VAR model, by showing 15-day Impulse Response Functions for one-time as well as permanent 10-percentage point increases in attention.
- 04-figure3.R to replicate Figure 3 of the paper, where we explore in more detail the ability of politicians *versus* groups of the public to lead the agenda of the other; and *viceversa*.
- 05-figure4.R to replicate Figure 4 of the paper, where we show the issue-level IRFs.
- 06-figure5.R to replicate Figure 5 of the paper, where we show the correlation between issue-level responsiveness and issue salience.
- 07-table4.R to replicate Table 4 of the paper, where we show the correlation between the issue attention distribution of the media, and the political and public groups under study.
- 08-figure6.R to replicate Figure 6 of the paper, where we show the ability of the the different groups under study to lead the agenda of the media, and *viceversa*.

Dashboard

To facilitate this validation exercise we have prepared an online dashboard where we offer a visualization of each of the topics that results from our analysis. The dashboard is available in the following URL: http://www.pablobarbera.com/congress-lda.

The code and data to reproduce the dashboard is available in the `02-dashboard-code` folder in this repository.

Topics

We provide all the code required to re-run the topic models for the different sets of actors in the `01-topic-models` folder:

- `01-create-dtm.py` creates the document-feature matrices in our study. Please note that running this script requires access to the original corpus of tweets, which we cannot share publicly in order to comply with Twitter's terms of service. All other scripts can be run without running this one first, since we are providing the files required to replicate the document-feature matrices we use in the analysis.
- `02-choosing-number-topics.r` runs multiple LDA models to choose the right number of topics based on cross-validated model fit, as reported in the Appendix of the paper.
- `03-running-lda.r` runs the topic model for the different sets of tweets in our sample.

- `04-output-data-for-reg.R` extracts the topic model probabilities for each set of accounts and topic.
- `05-adding-random-US-and-merging-subissues.R` adds the topic probabilities for the random set of US users and merges the topics that we merged into combined issues. This is the file that generates our main dataset – `data/main-time-series.csv`
- `06-intercoder-reliability-stats.R` computes intercoder reliability statistics for our coding of topics into political and non-political topics, as reported in the main text of the article.