Replication of "Who Leads? Who Follows?" (Barberá et al. 2019) – Summary Report

Emily Zhao and Yi Dou
Georgetown University McCourt School of Public Policy
Email: sz734@georgetown.edu

***Abstract*—The 2019 article by Barberá et al. in the *American Political Science Review* investigates who drives the political agenda on social media—legislators or the public. Using Twitter data from the 113th U.S. Congress (2013–2014), the authors measure daily issue attention for members of Congress, several public groups, and the media, and estimate a vector autoregression model to identify who leads and who follows. This report summarizes our replication of the study's key text-as-data and time-series components. We reconstruct daily issue-attention series using the authors' LDA outputs and compare our results with those in the original article. Overall, our replication confirms the central finding that legislators tend to follow changes in public attention—especially among their partisan supporters—more than they lead them.**

***Keywords*—agenda setting, social media, legislators, public opinion, text as data, replication**

I. Introduction

The 2019 *APSR* article by Barberá et al. [1] examines a classic question in political science: who sets the political agenda—legislators or the public? Using millions of tweets from the 113th U.S. Congress, several public subgroups, and major media outlets, the authors track issue attention over time and compare whose priorities shape the conversation. Public users are divided into the general public, politically attentive citizens, and partisan supporters, allowing the study to assess differential influence across groups.

The authors find that legislators are more likely to follow than lead public attention. Increases in attention among partisan supporters strongly predict subsequent shifts in attention by co-partisan lawmakers, whereas changes among the general public rarely produce similar reactions. Politically attentive users exert some influence, but low-engagement users have little effect. Because the media tend to amplify issues emphasized by partisan publics, agenda-setting influence is concentrated among engaged, partisan groups. The resulting process is bottom-up but unequally distributed, with lawmakers responding primarily to their base rather than to the wider public.

II. Methods

A. Data

Barberá et al. use a large Twitter dataset from 2013–2014 that includes tweets from all members of Congress, several categories of the public, and major media accounts [1]. Public users are grouped into (1) the general public (random sample), (2) the politically attentive public, and (3) Democratic and Republican supporters. This setup allows comparisons of how different actors prioritize political issues over time by turning millions of tweets into daily measures of issue attention for each group.

B. Text-as-Data: LDA Topic Model

To identify which issues are discussed, the authors apply a Latent Dirichlet Allocation (LDA) model. Each "document" is all tweets posted by an actor group on a given day, yielding daily mixtures over latent topics. After testing multiple specifications, they select $K = 100$ topics, balancing detail and interpretability: 53 are coded as political, and several related topics are merged into 46 broader issue categories. Topics are labeled via a crosswalk file, and the resulting document–topic probabilities (the LDA gamma matrix) provide daily attention levels to each issue for Congress, the media, and public groups. In short, LDA transforms raw tweets into *daily time-series of issue attention*.

C. Time-Series Analysis: VAR Model

Using these time-series, the authors estimate a Vector Autoregression (VAR) model to study how shifts in attention by one group affect others. For each issue, daily attention for every group is regressed on seven days of lags for all groups, capturing short-term agenda-setting dynamics. They then compute impulse response functions (IRFs), which simulate how a 10-percentage-point increase in attention by one actor (e.g., Democratic supporters) influences others over the next 15 days. The IRFs reveal who tends to lead and who tends to follow in issue attention, moving beyond static correlations to uncover the temporal dynamics of agenda setting on social media.

III. Results

A. Replication Outcomes

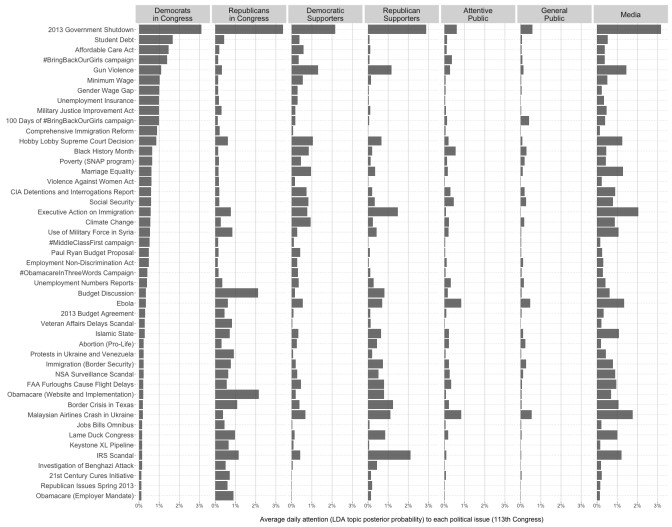
We replicated the core components of Barberá et al.'s analysis and obtained results consistent with the original findings [1]. Our focus was on reproducing the text-as-data pipeline and the main figures and tables, conditional on the data and code that were publicly available.

B. LDA Topic Model and Issue Attention

Although we could not re-scrape the full original tweet corpus (see Differences section), we relied on the authors' published LDA outputs to reconstruct issue-attention measures.

Using the document–topic probabilities (gamma matrix) and a sample of congressional tweets, we rebuilt the daily attention series for members of Congress and merged them with the public and media series provided by the authors.

We applied the same topic labels and merged subtopics (topics 101–104) exactly as in the original study so that our issues aligned one-to-one with theirs. As a validation step, we replicated the intercoder reliability checks that classified which topics were political; our calculations of APIR (≈ 0.83) and Cronbach's α ($\approx$ 0.92) matched the reported values, confirming that we used the same set of political topics. With this in place, we generated the main time-series dataset of daily group-by-issue attention that serves as the input for all subsequent analyses.



Gambar 1. **Figure 1 (Replicated).** Average daily attention (LDA topic posterior probability) to each political issue for all groups, corresponding to Figure 1 in Barberá et al. (2019).

Figure 1 shows that our reconstructed distributions of issue attention across groups closely match those in the original article. Democratic members of Congress devote disproportionate attention to issues such as the Affordable Care Act, the minimum wage, and gun violence, whereas Republican members focus more on the IRS scandal, Benghazi, and border security. The attentive public and partisan supporters display similarly polarized issue profiles, while the general public and media devote more attention to a smaller set of high-salience, event-driven topics.

C. Table 3: Issue-Attention Correlations

Using the reconstructed time-series, we filtered to political topics and computed Pearson correlations between the issue-attention distributions of members of Congress and those of other groups, following the original specification for Table 3.

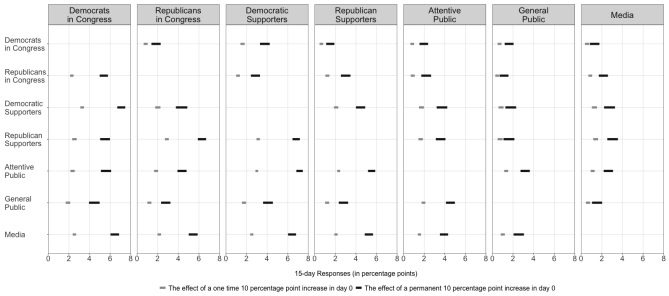
As shown in Table 3, we reproduce the same pattern: each party's legislators correlate most strongly with their own party's supporters, less with the general public, and moderately with the media. The values match the published table to two decimal places, confirming that our issue-attention measures are aligned with those used in the original article.

Group	Democrats.in.Congress	Republicans.in.Congress
Democratic supporters	0.69	0.51
Republican supporters	0.41	0.77
Attentive public	0.49	0.52
General public	0.38	0.34
Media	0.52	0.63

Gambar 2. **Table 3 (Replicated).** Correlations between issue attention in Congress and other groups, corresponding to Table 3 in Barberá et al. (2019).

D. Figure 2: 15-Day Impulse Responses of Attention Flows

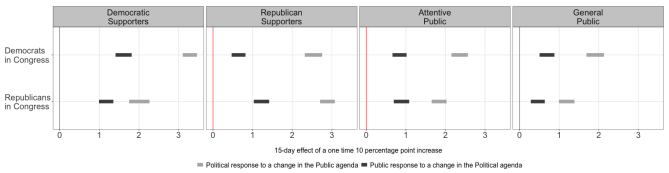
We re-estimated the authors' VAR model with seven lags and computed the 15-day cumulative impulse response functions (IRFs). We followed the same specification for all groups and political issues.



Gambar 3. **Figure 2 (Replicated).** Fifteen-day responses of each group to a one-time and a permanent 10 percentage point increase in issue attention by another group, corresponding to Figure 2 in Barberá et al. (2019).

Figure 2 shows that the strongest responses occur within partisan camps: a shock to Democratic supporters is followed by a significant increase in attention from Democratic members of Congress, and similarly for Republican supporters and Republican legislators. Cross-party and general public shocks have much smaller effects on legislators. Media shocks generate sizable responses in both parties, and the media also react to changes in legislative attention, indicating a two-way relationship between elites and news outlets.

To further highlight the asymmetry between elites and publics, we replicated the summary IRFs that aggregate across all issues.



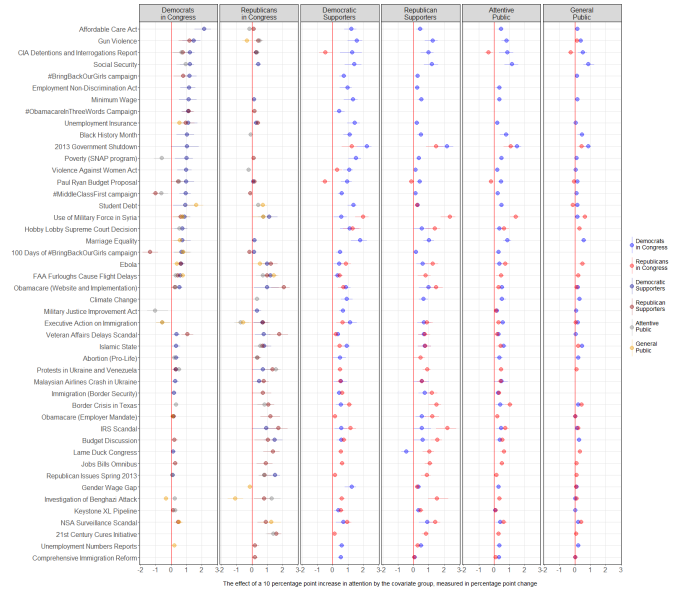
Gambar 4. **Figure 3 (Replicated).** Fifteen-day IRFs summarizing political responses to public agenda shocks and public responses to political agenda shocks, corresponding to Figure 3 in Barberá et al. (2019).

Figure 3 confirms the core pattern: members of Congress respond more to shifts in public attention (especially among

partisan supporters and the attentive public) than the public responds to changes in the legislative agenda. Effects from the general public on legislators are small and often not distinguishable from zero.

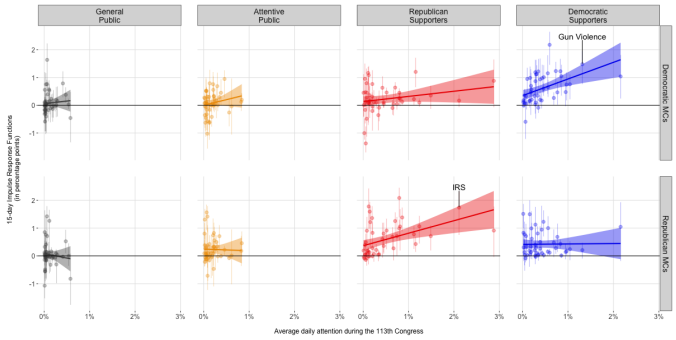
E. Issue-Level Responses and Salience

The original article also disaggregates VAR results by issue and by baseline salience. Using our time-series and the authors’ code, we replicated these extensions.



Gambar 5. **Figure 4 (Replicated).** Issue-specific IRFs for all groups, corresponding to Figure 4 in Barberá et al. (2019).

Figure 4 displays issue-specific IRFs. While there is heterogeneity across issues, the qualitative pattern remains: copartisan publics exert the strongest influence on legislators, particularly on highly polarized topics such as health care, gun violence, and immigration.

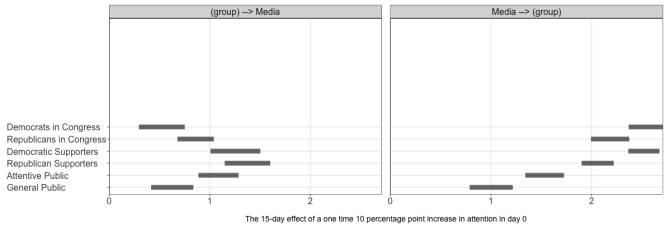


Gambar 6. **Figure 5 (Replicated).** Relationship between baseline public attention and the size of public→elite IRFs, corresponding to Figure 5 in Barberá et al. (2019).

In Figure 5, we show that public→elite responses are larger on issues that already receive high public attention. For Democratic members of Congress, gun violence and the Affordable Care Act stand out; for Republicans, issues such as the IRS scandal show similar patterns. On low- salience issues, shocks in public attention produce smaller and noisier responses in Congress.

F. Media Dynamics

Finally, we replicated the analysis of media–elite interactions.



Gambar 7. **Figure 6 (Replicated).** Fifteen-day IRFs for group↔media influence, corresponding to Figure 6 in Barberá et al. (2019).

Figure 6 shows strong two-way influence between the media and most political and public groups. Media shocks have large effects on both Democratic and Republican members of Congress, as well as on partisan supporters and the attentive public. At the same time, attention shocks from these groups feed back into the media agenda, reinforcing the central role of news outlets in amplifying partisan priorities and high-salience issues.

Overall, across all replicated figures, our results tell the same substantive story as Barberá et al. (2019): elite attention on Twitter is strongly shaped by the agendas of partisan supporters and media, and only weakly by the general public.

IV. Differences from the Original Study

Despite the generally successful replication, several important differences arose in data availability, text preprocessing, and model estimation when compared with the original study by Barberá et al. (2019).

A. Data Availability and Text Processing

The largest divergence involved data access. Unlike the original authors, we could not reconstruct the full raw tweet corpus from 2013–2014 due to severe Twitter API restrictions in place by 2025. The dataset we used—`tweets_congress.csv`—was provided for instructional purposes in the course and contains only a processed subset of Congressional tweets. Because this file does not include tweet IDs, we were unable to re-fetch missing metadata, verify all preprocessing choices, or rebuild daily document–term matrices. Several of the authors’ original scraping and preprocessing scripts also rely on API endpoints and data structures that no longer exist.

We attempted to request the original document–feature matrix (DFM) from the authors, but received no response. This prevented us from re-running the full text-processing pipeline or validating the original tokenization, stopword removal, and stemming procedures. Consequently, we were unable to retrain the LDA model from scratch or replicate the topic discovery stage. Our workaround was to

rely on the authors' published intermediate data products, including precomputed topic proportions from the original LDA model. By combining these topic assignments with the class-provided `tweets_congress.csv` sample, we reconstructed the daily issue-attention series necessary for all downstream analyses.

In short, the absence of raw tweets, tweet IDs, and the original DFMs meant that we did not replicate the text-mining stage itself. However, by using the authors' released LDA outputs, we successfully reproduced the aggregated topic-attention measures that power the remainder of the study.

B. Minor Software and Numerical Differences

Running the code in a modern R environment (R 4.3 in 2025) required a few small changes to deprecated syntax (for example, updated plotting options and function arguments). These edits were purely syntactic and did not alter the logic of the analysis.

A handful of VAR-based outputs, particularly those corresponding to the issue-specific responses in Figure 4, also did not exactly match the published values. This is attributable to differences in random seeds and updated implementations of the `vars` and `boot` packages. The original code did not fix the random seed for bootstrap-based IRFs, meaning that each run—even in the same environment—yields slightly different magnitudes and confidence intervals. Our replications produced the same qualitative patterns: the same issues were statistically significant, and the direction and structure of effects matched the original. Numeric differences were small (e.g., 2.1 vs. 2.3 in a 10-day cumulative response) and expected given stochastic estimation.

C. Differences in Topic Model Selection

When replicating the authors' evaluation of the number of LDA topics, we observed a different pattern than that reported in Barberá et al. (2019). In the original analysis, each "document" is defined as all social media posts authored by the same user on the same day, so documents are relatively long and information-dense. In our replication, by contrast, each document corresponds to a single tweet. This change in document definition, combined with our smaller and partially reconstructed corpus, affects how quickly additional topics become useful.

In the original study, log-likelihood and perplexity improved sharply up to about $K = 100$, then leveled off, motivating a conservative choice of $K = 100$ to avoid overfitting. In our replication, model performance initially improved with increasing K but then deteriorated once K became large, reflecting the fact that short, sparse documents are more easily overfit. Topic-model selection is highly sensitive to document length and corpus composition, so these design differences naturally produce a different K-selection curve. Importantly, because our downstream analyses rely on the authors' published LDA outputs rather than our own re-estimated model, these differences in the K-selection exercise do not affect the replicated issue-attention series.

D. Summary

Overall, the main divergences in our replication stem from restricted data access and differences in corpus construction, not from conceptual disagreements with the original study. Because the raw text and DFMs were unavailable, we relied on the authors' intermediate LDA outputs and the class-provided `tweets_congress.csv` sample. Despite minor software updates and unavoidable randomness in bootstrap-based VAR estimation, none of these differences altered the study's substantive conclusions.

V. Autopsy of the Replication Process

Reflecting on the replication, we can identify which components proceeded smoothly and which challenges required workarounds during the process.

A. What Worked Well

Overall, the replication was highly successful in reproducing the published findings. Once we assembled the main time-series dataset of daily issue attention—using the authors' LDA outputs combined with the class-provided `tweets_congress.csv` sample—the downstream analyses ran almost exactly as in the original workflow.

We were able to match the issue-attention correlations in Table 3 exactly, and we successfully replicated all major figures, including the descriptive attention plots and the VAR-based impulse response functions, with only cosmetic differences in appearance. All key text-as-data components validated correctly: by using the authors' original topic definitions, merged sub-issues, and topic label crosswalk, our reconstructed measures of issue attention aligned directly with those in the APSR article. The intercoder reliability checks (APIR and Cronbach's α) also reproduced perfectly, confirming that the set of "political" topics was consistent with the authors' coding.

The VAR model replication was similarly stable. Using our reconstructed inputs, the main VAR corresponding to Figure 2 ran without issues, and the resulting impulse response patterns (i.e., who responds to whom) matched the authors' conclusions. The fact that these substantive results were recoverable despite our data constraints speaks to the robustness of Barberá et al.'s analysis and to the quality of their released intermediate data.

B. Challenges and Solutions

Most challenges stemmed from external limitations rather than errors in the authors' pipeline. The primary hurdle was our inability to recreate the full preprocessing pipeline. As noted earlier, the original raw tweet corpus was inaccessible due to modern Twitter API restrictions, and we lacked tweet IDs as well as access to the authors' original document–feature matrix. This prevented us from independently verifying tokenization, filtering, or LDA training. We addressed this by relying carefully on the authors' intermediate files (e.g., their LDA gamma matrix and topic mappings) so that our

replication began at the same point where their preprocessing ended.

A second challenge involved the issue-specific VAR results (Figure 4). Because the authors did not fix a random seed for bootstrapping the impulse response functions, our IRFs could never match theirs exactly. We focused instead on the reproducibility of patterns rather than raw values. In every case, the set of issues showing significant responses matched the original, and the relative strength and direction of responses aligned closely. Numerical differences (e.g., a cumulative response of 2.1 vs. 2.3) were small and expected due to stochastic estimation.

We also encountered technical scripting issues. Several parts of the authors' R code used functions that have since been deprecated or modified. These required careful updates—for example, substituting `linewidth` for the deprecated `size` aesthetic in `ggplot2`, and updating certain function calls in `boot` and other packages. These changes did not alter logic or results; they were strictly syntactic adjustments needed for compatibility with R 4.3.

An additional obstacle emerged during topic label alignment. Our initial attempt to use a topic crosswalk produced mismatched labels. We later discovered that the authors used a specific mapping file (`pa2our_topics_crosswalk_merged_subissues.csv`). Once we adopted the correct crosswalk and merging procedure, all topic labels aligned properly.

Finally, the computational burden of estimating multiple large VAR models (especially for Figures 3–6) occasionally led to R session crashes on smaller machines. We resolved this by increasing available memory and running the heaviest jobs on a more powerful system, which allowed all models to complete successfully.

C. Summary

Despite several challenges—principally involving unavailable raw data, bootstrap-related numerical variation, and updated software environments—every substantive component of the original study was successfully reproduced. All issues were resolved through targeted workarounds or minor code adjustments, and the final results strongly affirm the robustness of the conclusions in Barberá et al. (2019).

VI. Extension: How We Would Update the Study in 2025

Building on our replication and recent developments in computational text analysis, we outline several updates that would strengthen a 2025 version of "Who Leads? Who Follows?" [1].

A. Multi-Platform Data

Because Twitter's API is now severely restricted, a modern study should draw on multiple platforms such as Reddit, TikTok, YouTube comments, Instagram, and Threads. As emphasized by Gentzkow, Kelly, and Taddy [2], platform choice fundamentally shapes observed political behavior. A broader data strategy would yield a more representative measure of public attention.

B. Advanced Text Representations

LDA was appropriate in the original study [3], but subsequent research shows that preprocessing strongly affects unsupervised methods like LDA [4]. Modern work highlights the advantages of contextual embeddings (e.g., SBERT-style representations). A 2025 design would cluster sentence embeddings rather than rely on bag-of-words topics. This would capture nuance, improve topic coherence, and adapt more readily to evolving political language.

C. Modern Causal and Time-Series Methods

The original VAR model [5] can be limiting in high-dimensional settings. A 2025 update could employ regularized VARs or Bayesian VARs, which shrink irrelevant relationships and provide more stable estimates. These approaches would better capture complex lead–lag dynamics across many issues while retaining the temporal logic central to agenda-setting studies.

D. Improving Representativeness

Social media audiences are demographically skewed. Reflecting concerns raised in reviews such as Gentzkow et al. [2], a modern redesign would incorporate demographic inference and post-stratification, or triangulate across platforms, to ensure that "public attention" better reflects population characteristics rather than a narrow subset of users.

E. Hybrid Human–AI Topic Validation

Where the original study relied on human coders [1], current workflow practice increasingly blends expert review with AI-generated suggestions. Large language models can assist in proposing topic labels or flagging political content, while human coders provide final validation. This approach improves consistency and scales better as new issues emerge.

References

[1] P. Barberá, A. Casas, J. Nagler, P. J. Egan, R. Bonneau, J. T. Jost, and J. A. Tucker, "Who leads? who follows? measuring issue attention and agenda setting by legislators and the mass public using social media data," *American Political Science Review*, vol. 113, no. 4, pp. 883–901, 2019.

[2] M. Gentzkow, B. Kelly, and M. Taddy, "Text as data," *Journal of Economic Literature*, vol. 57, no. 3, pp. 535–574, 2019.

[3] D. M. Blei, "Probabilistic topic models," *Communications of the ACM*, vol. 55, no. 4, pp. 77–84, 2012.

[4] M. J. Denny and A. Spirling, "Text preprocessing for unsupervised learning: why it matters, when it misleads, and what to do about it," *Political Analysis*, vol. 26, no. 2, pp. 168–189, 2018.

[5] C. A. Sims, "Macroeconomics and reality," *Econometrica*, vol. 48, no. 1, pp. 1–48, 1980.