Kaitlyn Vana, Jiani Zhao
PPOL 6801: Text as Data
Report: Replicating "Yellin' at Yellen"

Introduction

"Yellin' at Yellen: Hostile Sexism in the Federal Reserve Congressional Hearings" by James Bisbee, Nicolò Fraccaroli, and Andreas Kern seeks to answer the question, 'how prevalent is hostile sexism among US politicians?' The authors investigate the prevalence of hostile sexism in U.S. congressional oversight hearings by exploiting the unique case of Janet Yellen, the first female Chair of the Federal Reserve (Fed). Covering all hearings of the Fed Chair before Congress from 2001–2020, the authors compare patterns of interruptions and aggressive verbal interaction during Yellen's tenure with those of her male predecessors, Alan Greenspan, Ben Bernanke (who acts as the reference category in their specifications) and her male successor, Jerome Powell. They investigate whether Yellen is engaged with more hostilely than male chairs, after accounting for topic, tone, and individual legislator fixed‑effects. Their paper ultimately finds that legislators who interacted with both Yellen and one or more male Fed Chairs interrupted Yellen significantly more frequently and used more aggressive language in her hearings. This held true across gender and party, although it was seen more in Republicans and men. These effects remain robust after controlling for the topic of the hearing, via two NLP measures, and Yellen's own speaking tone. The implications from the authors is that the systematic hostility faced by Yellen indicates that gender norms and bias have the capacity to undermine democratic oversight and accountability mechanisms.

Additionally, the authors exploit the "daughter presence" of legislators as a quasi‑random moderator. They find that legislators with daughters exhibited lower levels of hostile behavior toward Yellen, consistent with prior literature showing that having daughters is associated with reduced gender bias.

Methods

The authors' corpus is comprised of all congressional oversight hearings where the Fed Chair appears between 2001–2020 downloaded from govinfo. This results in the corpus capturing 79 hearings (40 House, 39 Senate) with 23,119 "utterances" spoken by 242 legislators, four Fed chairs (Greenspan, Bernanke, Yellen, Powell), plus a few experts. Each row in their main dataset becomes an utterance–hearing, with information on who is speaking, who they're talking to, chamber, date, party, etc. Their main outcome is interruptions. The transcripts mark interruptions with "--" at the end of a line; they use that notation and the sequential structure of the text to identify (a) that an interruption occurred, (b) who was interrupted, and (c) who did the interrupting.

They then construct, for each utterance, a binary indicator for "this utterance interrupts the Fed chair" versus "this is a normal turn," and aggregate up to get interruption rates per legislator–hearing. As a second, complementary outcome they use NLP-based "tone" measures:

they apply off-the-shelf text analysis tools to classify the sentiment/hostility of legislators' language toward the chair, treating more negative and aggressive tone as an additional manifestation of hostile sexism. The research design is essentially a within-legislator comparison, and so Yellen is best understood in this way as a bundled treatment, not an outcome.

Results and Differences

Our replication generated the same results and conclusions as the paper. We found that, overall, Fed Chairs were most often interrupted, and within those four Yellen was more often interrupted and having to interrupt legislators. We followed the authors' process controlling for amount of speaking, topic, etc. In our coefficient plot summarizing the difference between Bernanke's propensity to be interrupted and a subset of other speakers, we similarly found that Yellen had a higher propensity to be interrupted. Accounting for interactions and topics, we generated concurring results that Yellen was being interrupted more often regardless of topic. We encountered no substantive differences. The only differences lay in figure-generating decisions, as seen below.

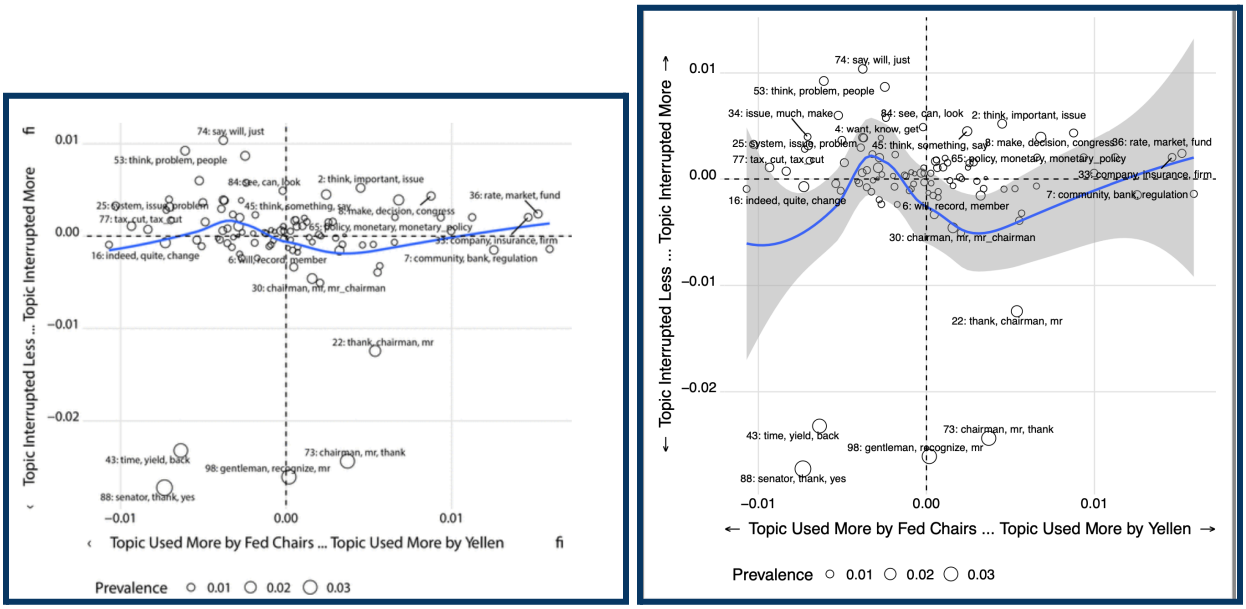

Figure 1. Scatter plot of topics by whether they were used more by Yellen or male Fed chairs (x-axis) and whether they were more or less interrupted (y-axis). Left: from "Yellin' at Yellen" (Figure 9.) Right: generated from replication materials

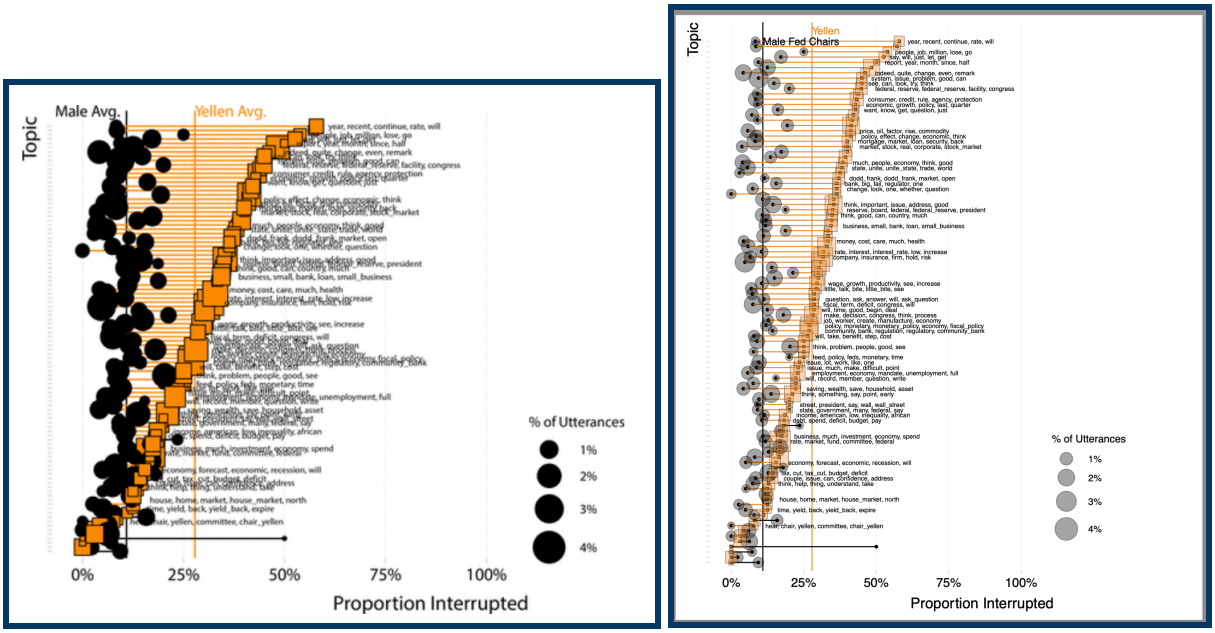

Figure 2. Each utterance is assigned to its highest-scored topic (y-axis) and then aggregated to the speaker. The x-axis indicates the proportion of each speaker's utterances that is interrupted, broken out by the utterance's highest-scoring topic. Points are sized by the proportion of all utterances that are assigned to each topic. Left: from "Yellin' at Yellen" (Figure 10) Right: generated from replication materials

In Figure 1, we incorporated a confidence band around the smoother, representing uncertainty in the estimated trend, in order to more coherently analyze the contributions by topic. We also made aesthetic changes, as seen in the second set of figures (Figure 2), for readability.

Autopsy

The repository on Harvard Database which contained replication data for this paper was extremely well-organized. The ReadMe was detailed and comprehensive, and all pre-processing steps were well-documented. The authors included all intermediary files, and comments were thorough and clear. Across the replication exercise, several components of the text-processing and topic-modeling pipeline performed strongly. The clarity on text pre-processing is likely what allowed us to exactly replicate the authors' results. On the preprocessing side, we successfully cleaned and standardized the hearing transcripts by removing headers, fixing whitespace, and harmonizing speaker attributes such as party, gender, and demographic variables. We also replicated the authors' steps enriching the dataset by merging electoral, family, and biographical information for both legislators and Fed officials, along with bill-level legislative activity measures.

For topic modeling, the baseline LDA approach worked reliably. After tokenizing the documents, removing stopwords, and creating document–term and co-occurrence matrices, we implemented a train/test split and selected k=100, based on perplexity and semantic coherence

and in agreement with the authors, then refit the final model on the full dataset. The STM model also performed well. Using textProcessor() and prepDocuments(), we maintained the same number of topics for comparability and specified covariates for prevalence, which enabled meaningful estimates of speaker-level and partisan effects. For example, we also found that female speakers were more likely to discuss financial stability, interrupted legislators gravitated toward defensive topics, and Republicans framed inflation differently than Democrats.

Robustness checks (RR1) also succeeded. We reran LDA on chunk-level documents by aggregating short utterances. The goal was to check whether the discovered topics were stable and reliable under different document definitions. For the chunk-level model, we grouped short utterances from the same speaker into larger blocks and used a 50/50 train–test split. This gave us enough test data to compute perplexity reliably. After searching over many values of K, the best K was around 70 as opposed to 100. This makes sense since the unit of text changed. We then fit LDA with K = 70 on all chunks. Next, for the speaker-level model, we aggregated all text by speaker and used an 80/20 split. This split approach was deployed since the number of documents was smaller. Speaker-level data has fewer documents, and so we needed more training data to learn coherent topics. Again, the best K was about 70.

There were two steps which did not work well in our replication process. Some of the original visualization functions from the published codebase, such as plot_cme() and marginaleffects(), are now deprecated or unavailable, preventing exact reproduction of all figures in the original paper. As mentioned in differences, this led to us updating some of the figures for readability.

Extension

Potential opportunities for extension exist in applying this research framework to different institutional contexts or executive-branch officials who are similarly required to testify before Congress, such as the Secretary of Commerce, the Secretary of Education, the CFPB Director (still applicable despite its current status), or the CBO Director. This would test the generalizability of hostile sexism in US Congress. Finance and banking is a field which traditionally serves as a male source of power. The authors mention that this context "[embodies] the types of situations where hostile sexism should be most pronounced." Expanding into these adjacent fields would help determine whether the patterns observed with Janet Yellen are unique to central banking or reflect broader gendered dynamics in congressional oversight. Applying this empirical framework to other countries, as in Sebastian Vera Vallejo's and Analia Vidal Gomez's "The Politics of Interruptions: Gendered Disruptions of Legislative Speeches," would also be a worthwhile extension.

Another extension would focus on intra-party and family-life dynamics, probing more deeply into how legislators' partisan identities, ideological leanings, or personal backgrounds, like the investigation of the presence of daughters' effect, moderate their treatment of female officials. This could clarify whether the "daughter effects" identified in the original study point to causal mechanisms or simply correlate with deeper partisan or social-identity factors.

Finally, researchers could explore alternative methodological strategies. It would be interesting to apply PreText and investigate which text pre-processing steps did or did not affect the resulting models, or introducing hard-encoding to better adjust for overall increases in congressional hostility across time. The authors acknowledged that their analysis "does not preclude the possibility that speakers simply became more hostile in 2014," but it does mean that "this shift would have had to occur within legislators, instead of reflecting an incoming class of more hostile interlocutors." These methodological refinements would help ensure that measured differences in treatment are not artifacts of shifting institutional norms or changes in discourse unrelated to gender.