# Structural Topic Models

Giovanni Maya

2025-05-15

# Agenda

- **Paper Overview**
  - Topic + Research Question
  - Data
  - Theory
  - Findings
- **Replication Progress**
  - My Progress
  - Bottlenecks
  - Future Timeline
- **Extension Considerations**
  - Methodological Angle for innovation
  - Preliminary ideas

# Paper Overview

- **Research Question**: What is the importance of incorporating individual level covariates on topic prevalence and content?
- **Topic**: This paper's novel approach is an STM, Structual Topic Model. In simple terms, it is an extension of LDA where it finds the topics prevalent in documents, but now, also incorporates document-level metadata to see how it influences said prevalence.
- **Data sources**:
  - ANES 2008–2009 Panel Study
    - Open-ended responses to the *Most Important Problem* (MIP) question
  - Gadarian & Albertson Experimental Study
    - Participants primed with fear (or not), then asked to explain views on immigration
  - Rand, Greene, and Nowak (2012)
    - Responses on how intuitive vs reflective reasoning affects cooperation in public goods games

# Mathematical Theory

- **Structural Topic Model (STM)**
  - Extension of Latent Dirichlet Allocation (LDA)
  - Allows inclusion of covariates to influence:
    - Topic prevalence: how much each topic appears in a document
    - Topic content: how the words used within a topic vary
- **Key components**:
  - $\theta_d \sim \text{LogisticNormal}(X_d \cdot \gamma, \Sigma)$
    *(Document-topic proportions depend on covariates $X_d$)*
  - $z_{d,n} \sim \text{Multinomial}(\theta_d)$
    *(Topic assignment for word $n$ in document $d$)*
  - $w_{d,n} \sim \text{Multinomial}(\beta_{z_{d,n}})$
    *(Word drawn from the topic-specific word distribution)*
- **Estimation**:
  - Performed using **variational EM** to approximate posterior distributions

# Takeaways

- **ANES 2008–2009 Panel Study (Most Important Problem responses):**
  - STM recovered major topics like "War," "Economy," and "Unemployment" that aligned well with ANES hand-coding.
  - Topic prevalence varied by education and partisanship:
    - War topics discussed more by lower-education Democrats.
    - Economic topics more common among Republicans.
  - STM provided a nuanced account: individual responses could belong to multiple topics.
- **Gadarian & Albertson Immigration Experiment:**
  - Participants primed with fear used more moral, emotional language:
    - Words like "feel," "believe," "hope," "god."
  - Control group used more policy-driven, calculated terms:
    - Words like "money," "risk," "figure."
  - STM detected distinct topic shifts by treatment group in strategy explanations.

# Takeaways (cont)

- **Rand, Greene, and Nowak Public Goods Experiment:**
  - Intuition-primed participants used more emotional, instinctive language:
    - "feel," "good," "believe," "chance," "god"
  - Reflection-primed participants used more analytical, self-focused terms:
    - "money," "myself," "keep," "gain," "figure"
  - STM uncovered topic differences aligned with experimental treatments (intuition vs. reflection, fast vs. slow decision-making)
- Now onto my progress!

# Replication Progress

- **ANES 2008–2009 Panel Study:**
  - Successfully replicated STM on open-ended MIP responses.
  - Identified top topics including War/Iraq and Economy.
  - Reproduced Figure 8: Interaction between education and partisanship on the War topic.
    - Found that Democrats with higher education were less likely to discuss War.
- **Gadarian & Albertson Immigration Experiment:**
  - Replicated STM treatment effect analysis (Figure 12).
    - Participants in the fear condition more likely to use emotional/moral language.
  - Produced representative quotes using `findThoughts()` as shown in Figure 10.
  - Created interaction plot of party ID $\times$ treatment on Topic 2 (emotion-focused strategy).
    - Reproduced visual structure of Figure 8 in the paper.

# ANES 2008-2009 Panel Study

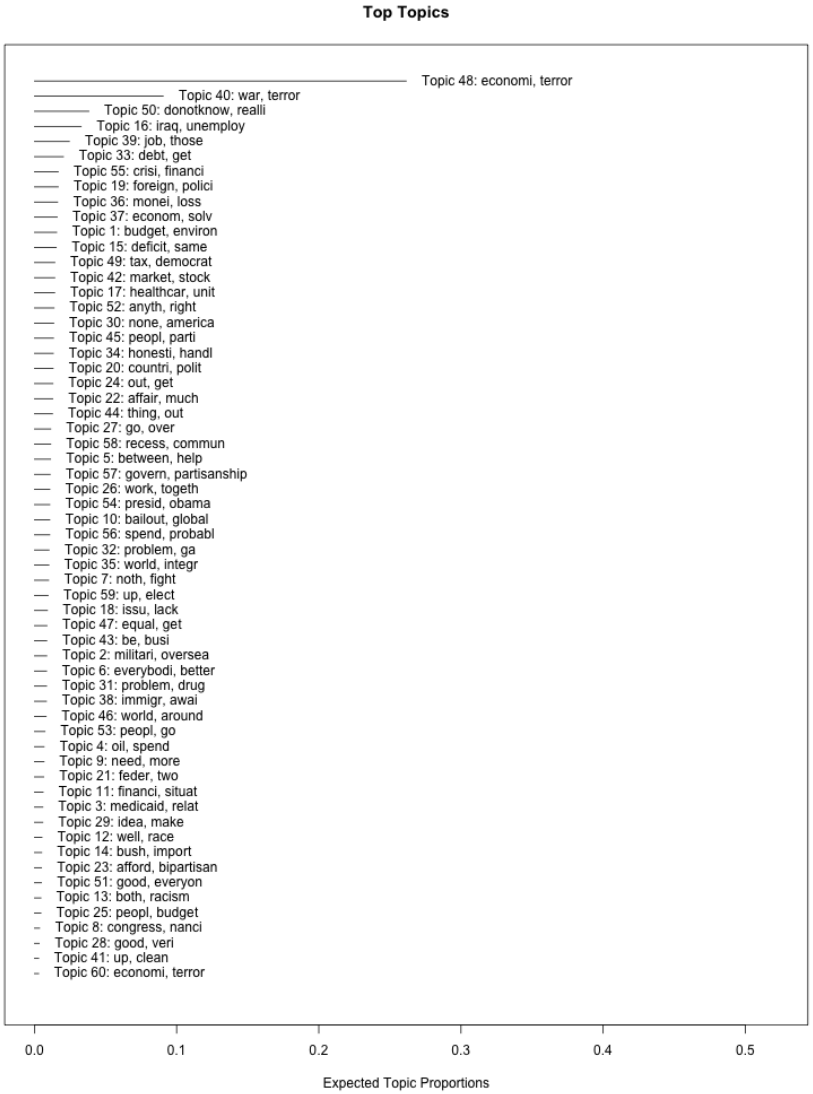


Figure 17: Top Topics

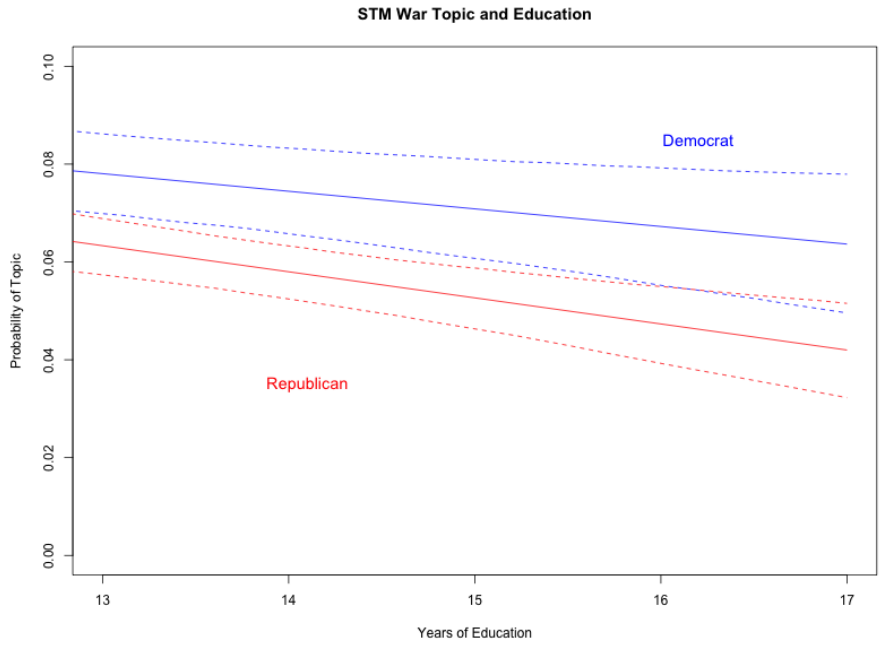


Figure 18: STM War Topic and Education

# Gadarian & Albertson Immigration Experiment

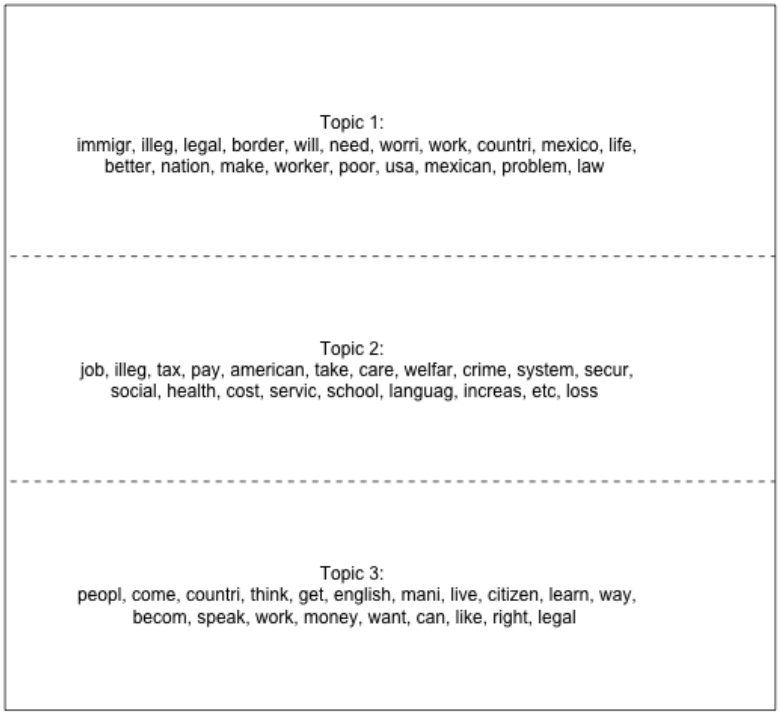


Figure 8: Vocabulary Associated with Topics 1 and 2

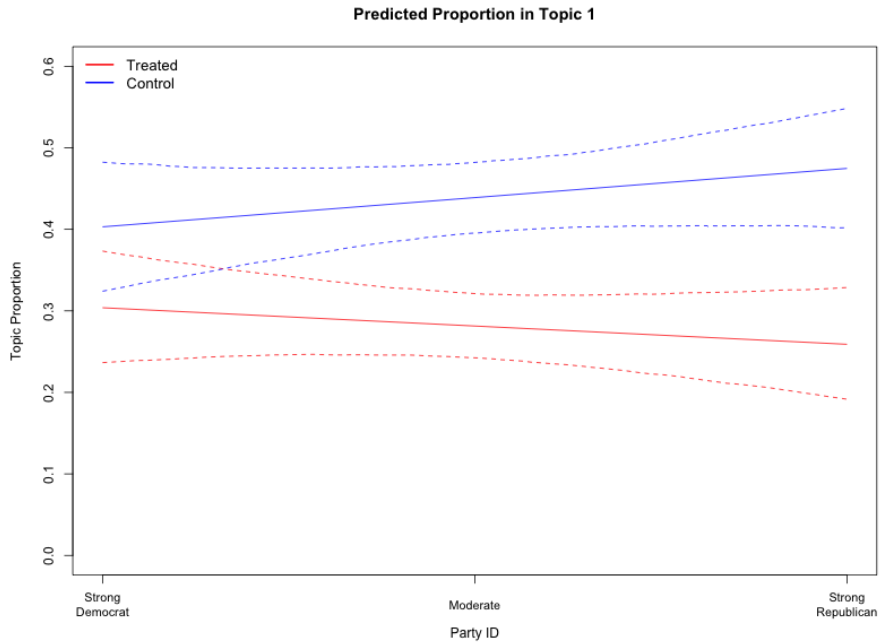


Figure 8: Party Identification, Treatment, and the Predicted Proportion in Topic 1

# Bottlenecks & Timeline

- **Current Bottlenecks:**
  - Need to resolve data inconsistencies with Rand, Greene, and Nowak (2012).
  - Debugging figure creation errors (e.g., axis labels and topic assignments).
  - Conducting additional validation to align STM output with paper expectations.
- **Planned Timeline:**
  - **This week:** Finalize replication for ANES and Gadarian studies; complete Rand implementation.
  - **Next week:** Execute and document all extensions; prepare final video presentation.

# Extension Considerations

- **STM over Time: Topic Shifts**
  - Compare topic prevalence in *Most Important Problem* (MIP) responses across two waves:
    - **ANES 2008 Panel Study** vs. **CES/ANES 2020**
  - Find and examine the differences in topic models across time (temporal aspect)
- **Embedding-Based Topic Clustering**
  - Use Sentence-BERT or similar model to embed responses
  - Apply clustering (e.g., k-means, HDBSCAN) to discover prevalent topics and see how covariates are able to affect the embeddings?
  - Validate and contrast with STM findings