New Thinking About When Social Media and Survey Responses May Align

Frederick G. Conrad[1]
Michael F. Schober[2]
Johann A. Gagnon-Bartsch[1]
Robyn Ferg[3]
Mao Li[1]
Paul C. Beatty[4]

1 UNIVERSITY OF MICHIGAN

2 THE NEW SCHOOL FOR SOCIAL RESEARCH

3 Westat

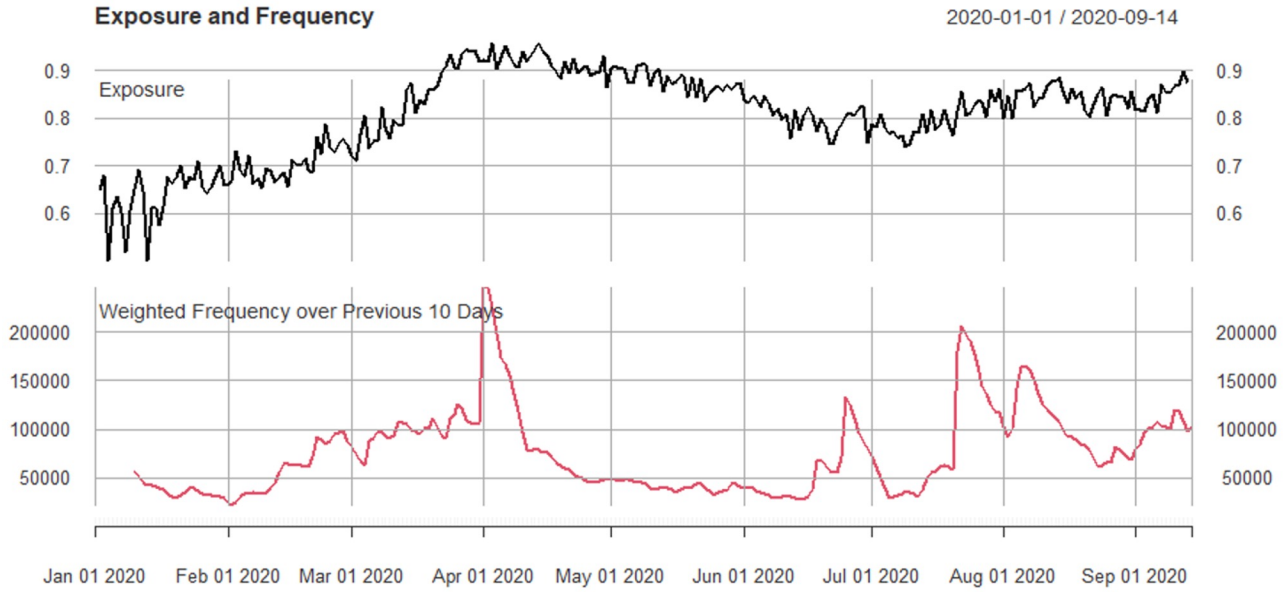
4 United States Census Bureau

Views expressed are those of the authors, and not those of the US Census Bureau

Can social media tell the same story, more or less, that a survey tells?

- Do the patterns of answers to a survey question mirror the patterns of social media posts during the same time period?
- If so, the two data sources are “aligned”

Have you heard or read that the census helps inform how public funds are distributed to communities for things such as schools, roads, and health clinics, or have you not heard or read that?



Why Does Alignment Matter?

- If the two data sources align, social media could be more timely and cheaper than survey data for some research questions
- Despite early evidence of alignment, relationships have not held up
  - e.g., O'Connor et al. (2010); Conrad et al. (2015), Conrad et al. (2021), Pasek et al. (2018)
- But the approach seems too promising to abandon (yet)

Current work: when is alignment more and less likely?

- We explore alignment between responses to the Census Tracking Survey and Tweets containing Census-related words from the same time period
  - Census Tracking Survey: 23 questions, daily web survey, nonprob., n=76,919, conducted from Jan 2 to September 13, 2020
  - Tweets: ~3.5 million from Jan 1 to September 14, 2020, contained terms such as “Census” or ”American Community Survey” listed in large filter (query) designed by Census Bureau and implemented in Sprinklr platform
- We measure alignment for each of the 23 questions from the Tracking Survey using *comovement*:
  - fraction of time that two time series move in the same direction (Fechner, 1897; Moore & Wallis, 1943; Goodman & Grunfeld, 1961)
  - Less sensitive to outliers than Pearson correlation

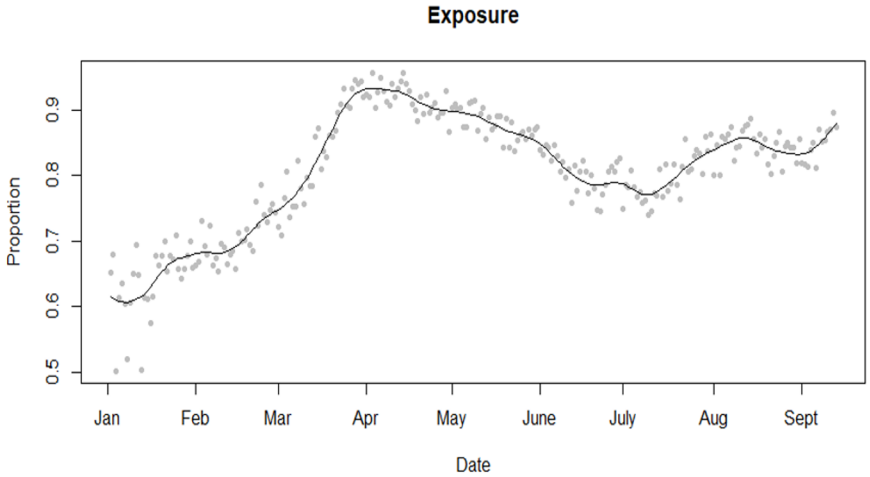
We propose that alignment should be more likely when

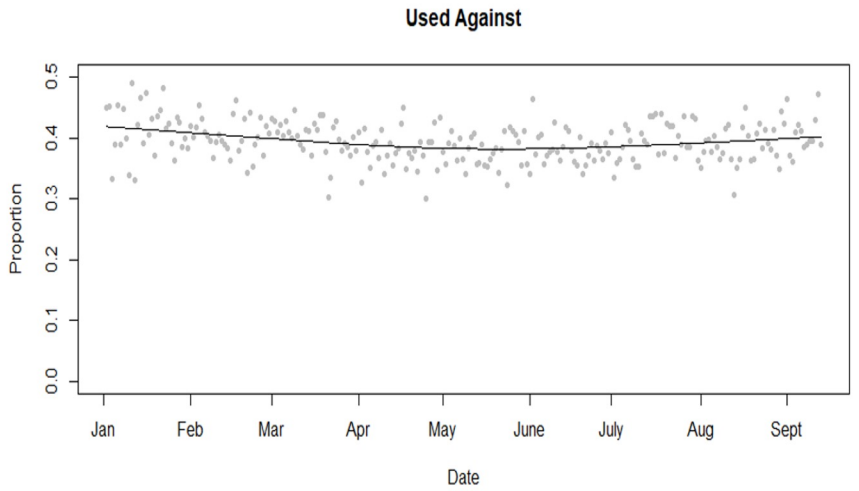
1. survey responses have a relatively high signal-to-noise ratio (i.e., there is something for social media to align with)
    - Signal: Responses actually change over the time frame
    - Noise: Variability is not too high

Questions with high (0.83) and low (0.14) signal-to-noise ratio:

How much have you seen or heard recently – within the last week or so – about the 2020 Census?

How concerned are you, if at all, that the answers you provide to the 2020 Census will be used against you?





2020 Census Tracking Survey Data from the Integrated Communications Contract Campaign Monitoring and Optimization

We propose that alignment should be more likely when

2. social media corpus includes content semantically related to survey question

Selecting tweets semantically related to survey question

- SBERT* is NLP tool that calculates semantic distance between two texts
- For each of the 23 survey questions,
  - SBERT assigned distance score to each tweet in corpus (n~ 3.5 million)
  - We ordered the tweets by distance from the question and established a cutoff above which tweets were related to the question and below which they were unrelated – based on human judgment
  - Created corpus of “relevant” tweets all of which were above the cutoff – much smaller than full corpus (e.g., range from ~56K to ~500K)

*Sentence-Bidirectional Encoder Representations from Transformers

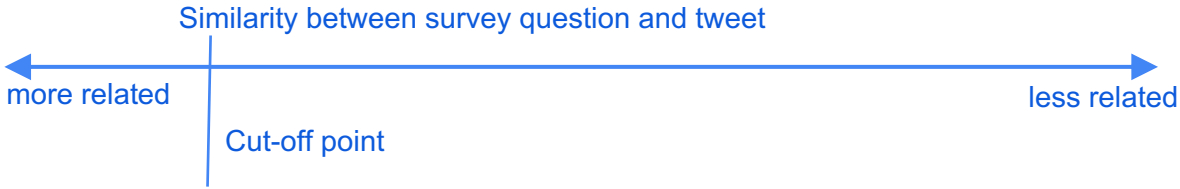
SBERT quantifies semantic distance between tweets and survey question, e.g.

"Do you think the 2020 Census questionnaire will or will not ask which people living in your household are U.S. citizens?"

"The 2020 Census questionnaire will NOT ask which people in their households are citizens."

@seanhannity can she get dumber? the census is for AMERICAN CITIZENS not illegal aliens. geez

"RT @GUslavery Dr. Young was one of four doctors credited with founding the Medical Department at the college in 1850. Two other department founders, Flodoaro Howard and Johnson Elliot were also listed as slaveholders on the 1860 census."



We propose that alignment should be more likely when

3. the attributes of tweets that are counted “make sense” for survey question’s content

Attributes of tweets that might affect alignment

- Volume of tweets could comove with a question like
  - *"Have you heard or seen the message 'Shape your future. Start here' about the 2020 Census, or have you not heard or seen that?"*
  - because the more people who have heard or seen the message, the more Twitter users are likely to post about the Census
- Sentiment of tweets (positive vs. negative) could comove with a question like
  - *"How concerned are you, if at all, that the Census Bureau will share individuals' answers to the 2020 Census with other government agencies?"*
  - because as more respondents indicate concern more Twitter users may post negative tweets – especially if they are semantically related to the survey question (i.e., in the relevant corpus)
- Predicted alignment (yes or no) for each question based on this kind of rationale
- Accuracy: we predict alignment and it is observed (hit) or if we predict no alignment and no alignment is observed (correct rejection)

Results: Signal to Noise Ratio

Tweet attribute, Corpus	Correlation with Comovement
Volume, Full corpus	0.568824
Volume, Relevant corpus	0.694120
Sentiment, Full corpus	0.353126
Sentiment, Relevant corpus	- 0.125321

Results: Relevant versus All Tweets

Tweet attribute, Corpus	# Questions (out of 23) for which comovement* observed
Volume, Full corpus	7
Volume, Relevant corpus	9
Sentiment, Full corpus	7
Sentiment, Relevant corpus	6

- We are encouraged by the number of questions for which we observe comovement
- But not clear that restricting tweets to those that are related to the survey question is revealing much “hidden” comovement

*significant or marginally significant

Relevance versus Stance

- “Semantically related” means tweet concerns same topic as question, is not sensitive to user’s opinion (e.g., “will ask” or “will not ask”) expressed in tweet
  - If most of the tweets in a relevant corpus express one position, e.g., “will ask,” then comovement -- if present -- stands a relatively good chance of being observed
  - But if the corpus contains equal proportions of tweets expressing either side of an issue (“will ask” and “will not ask”) – a big mix of opinion -- then no reason these tweets should comove with “will ask” responses
- Currently engaged in “stance” detection, i.e., discriminating tweets by the opinion they express

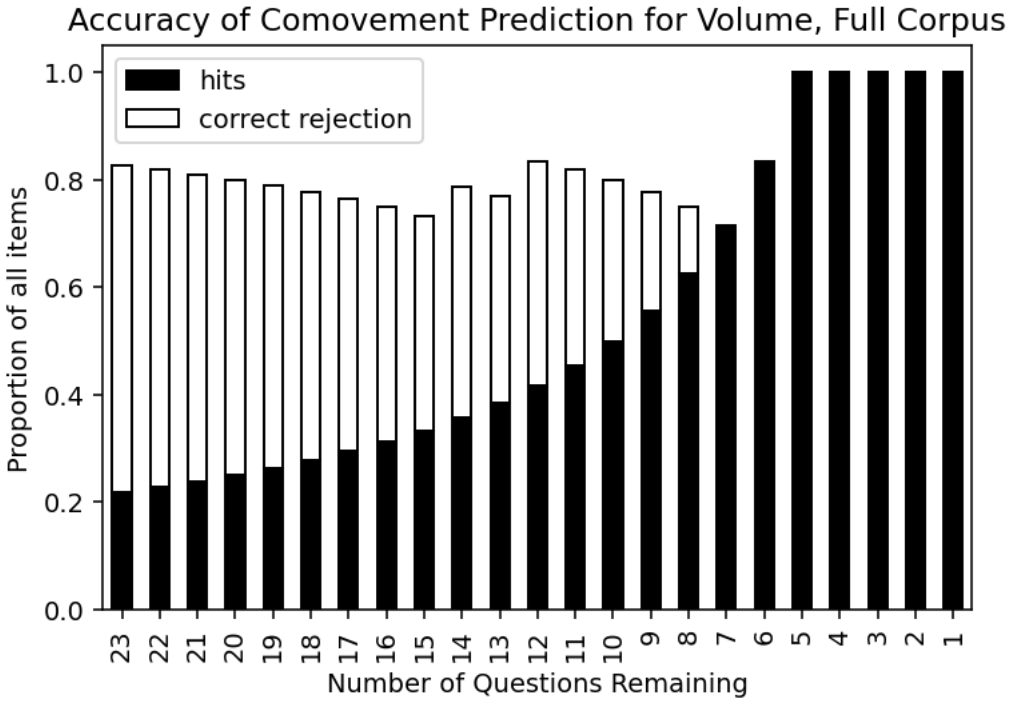
Results: Predicting Alignment

Tweet attribute, Corpus	Prediction Accuracy
Volume, Full corpus	.83
Volume, Relevant corpus	.74
Sentiment, Full corpus	.69
Sentiment, Relevant corpus	.56

- Accuracy mostly perfect if restricted to questions with high signal to noise ratio

Accuracy of predicted comovement

- Calculated Hits and Correct Rejections across all 23 survey questions
- Successively removed questions with lowest signal-to-noise ratio, recalculating Accuracy



Conclusions (so far)

- Alignment detected in several (6 to 9) of the 23 questions, and we accurately predicted when alignment would be observed
- Promising for potential use of social media to assess public opinion
- but this kind of analysis takes a lot of hard thinking, computing power, computational skill, and judgment about when responses and tweets should move together
- Restricting search for alignment to posts that are semantically close to survey question led to slightly more alignment for volume (but not for sentiment)
- Currently conducting stance detection to see if this reveals alignment hidden by the mix of opinion

Thank You

fconrad@umich.edu

We gratefully acknowledge support from the US Census Bureau through a cooperative agreement with the University of Michigan and the New School, "New Approaches to Analyzing Social Media Content for Enhancing Census Bureau Data” (award #CB20ADR0160002)