Combining Digital Trace Data and Survey Data

Jonathan Nagler, jonathan.nagler@nyu.edu
Center for Social Media and Politics, New York University

Slides Prepared for `Linking Digital Footprint and Survey Data for Open Research' Workshop, University of Manchester, February, 2025. [SLIDES INTENDED ONLY FOR WORKSHOP PARTICIPANTS.].

Work described here has been funded by a variety of funders, including the National Science Foundation, the Knight Foundation and Craig Newmark Philanthropies. General operating support for CSMaP has been provided by the Hewlett Foundation, the Charles Koch Foundation, and the Siegel Family Endowment.

- Describe Data We Collected From 2 Sources:
  - YouGov Respondents (2016, 2020, 2022, 2024)
  - CSMaP Bilingual Election Monitor (2022)
- Try to Include Notes on Take-Up (Donation) Rates
- Describe Aggregates We have Tried (and/or aspired) to compute

Issues

- What is the right amount to pay people?
  - **WE** clearly should be experimenting.
  - Maybe others have?
- Aggregates to merge to survey data:

2016 - start of panel - Facebook via App

- 2016:
  - 3500 YouGov Respondents
  - 2711 said they use Facebook
  - 1331 agreed to share data
    - 38% of all respondents
    - 49.1% of respondents who use Facebook.
  - We linked 1191 of the Facebook accounts to respondents

2018 - Webtrak Data via YouGov

- In 2018 - we had a 1500 respondent webtrack panel
  - webtrack data provided by YouGov
  - Slightly mysterious:
    - mobile?
    - desktop?

2018 - Webtrak Data via YouGov

- We used the webtrack data to see the paths people took to get to low-quality news.
- And, we modeled the percentage of fake news in respondents' web diets as a function of individual level co-variates (demographics, ideology)

2020 - Just Tweets

- 2020:
  - No facebook app
  - Continuing Twitter Collection

2022 - Multi-Platform

- 2022:
  - Continued Twitter data collection
  - Facebook again!
  - Web-Tracking Data:
    - We asked respondents to install a web-browsing plug-in
    - For desktop use with Chrome
  - YouTube Data:
    - We asked respondents to download their YouTube watch history and send it to us
    - This can be done with google forms

Facebook

- 2022 - Facebook:
  - 722 respondents provided facebook data
    - 649 included likes
    - 222 included posts

Web Browsing

- 2022 - Web Browsing Data
  - 596 respondents
- 2024 - Web Browsing Data
  - 416 respondents
    - 87 of these turned on an extension to give us html pages of their visits to YouTube

YouTube

- 2022:
  - 506 respondents provided watch histories
  - 489 respondents provided subscription information
- 2024:
  - 520 respondents provided watch histories
  - 489 respondents provided subscription information

TikTok

- The donation process is the most cumbersome of the platforms we tried to get data for.
- Users must wait an unknown time for their download to be available.
  - And it is only available for a short window.
- Finding it on their mobile device to upload is cumbersome.
- For security reasons - we could not accept it as an email.

TikTok

- 275 respondents provided data with actual viewing history
- Total videos watched: 3,887,904
- Unique videos watched: 2,326,474
- Number of users who posted a comment: 51
- Total comments by users: 31,772 [??]

2022 - CSMaP Bilingual Election Monitor

- 2022:
  - In addition to YouGov, we recruited a panel of 3500 respondents (primarily) via Facebook ads (the CSMaP Bilingual Election Monitor)
    - 2300 hispanics (English-dominant, Spanish-dominant, Bilingual)
    - 900 non-hispanic whites
    - 344 other

2022 - CSMaP Bilingual Election Monitor

- Next slide gives donation rates for the 2022 respondents.
- Payment ranged from $5 to $10.
- Confusion about how to pay for web-track data:
  - One-time payment?
  - Monthly payment?

Data Shares by Ethnic-Languag Group							
	Whites	*Hispanics*	*Eng-Dominant*	*Bilinguals*	*Spa-Dominant*	Other Race	*TOTAL*
TWITTER (provided handle)	195	360	*144*	*88*	*100*	97	**652**
Twitter (Claim to have account)	457	1,219	399	497	323	201	**1877**
Twitter % Data	42.67	29.53	36.09	17.71	30.96	48.26%	**34.74%**
FACEBOOK (provided data)	302	441	*168*	*114*	*126*	112	**743**
Facebook (claim to have account)	818	1758	557	648	553	291	**2867**
Facebook % Data	36.92	25.09	30.16	17.59	22.78	38.49%	**25.92%**
WEB BROWSING (provided data)	100	150	*84*	*37*	*29*	45	**250**
(Claim Chrome is their primary browser)							**0**
YOUTUBE (provided data)	211	335	*134*	*75*	*89*	97	**546**
Youtube (Claim to have account)	664	1,649	469	631	549	299	**2612**
Youtbe % Data	31.78	20.32	28.57	11.89	16.21	32.44%	**20.90%**

Aggregates or Other Variables to Compute from Digital Trace Data (Twitter) to merge to Survey Data

- Following of specific accounts (Fox News, CNN, MSNBC, NY Times, etc)
- Following of specific politicians (Trump, any member of Congress, Governor)
- Number of political accounts followed, total number of accounts followed, number of media accounts followed
- Ideology Measures of accounts followed:
    - mean ideology of (all accounts followed, media accounts followed, political accounts followed, non-elite accounts followed)
    - variance of ideology of (all accounts followed, media accounts followed, political accounts followed, non-elite accounts followed)

Aggregates or Other Variables to Compute from Digital Trace Data (Facebook) to merge to Survey Data

- Number of Political Pages Liked
- Number of Pages Liked
- Number of low quality media pages liked
- Number of United States facebook pages liked
- Number of Latin American facebook pages liked
- Number of (! US, ! Latin American) facebook pages liked

Aggregates or Other Variables to Compute from Digital Trace Data (web-track) to merge to Survey Data

- Number of visits to specific websites: foxnews, cnn, etc
- Number of visits to political news websites (list of approx 5000 web domains)
-

Issues

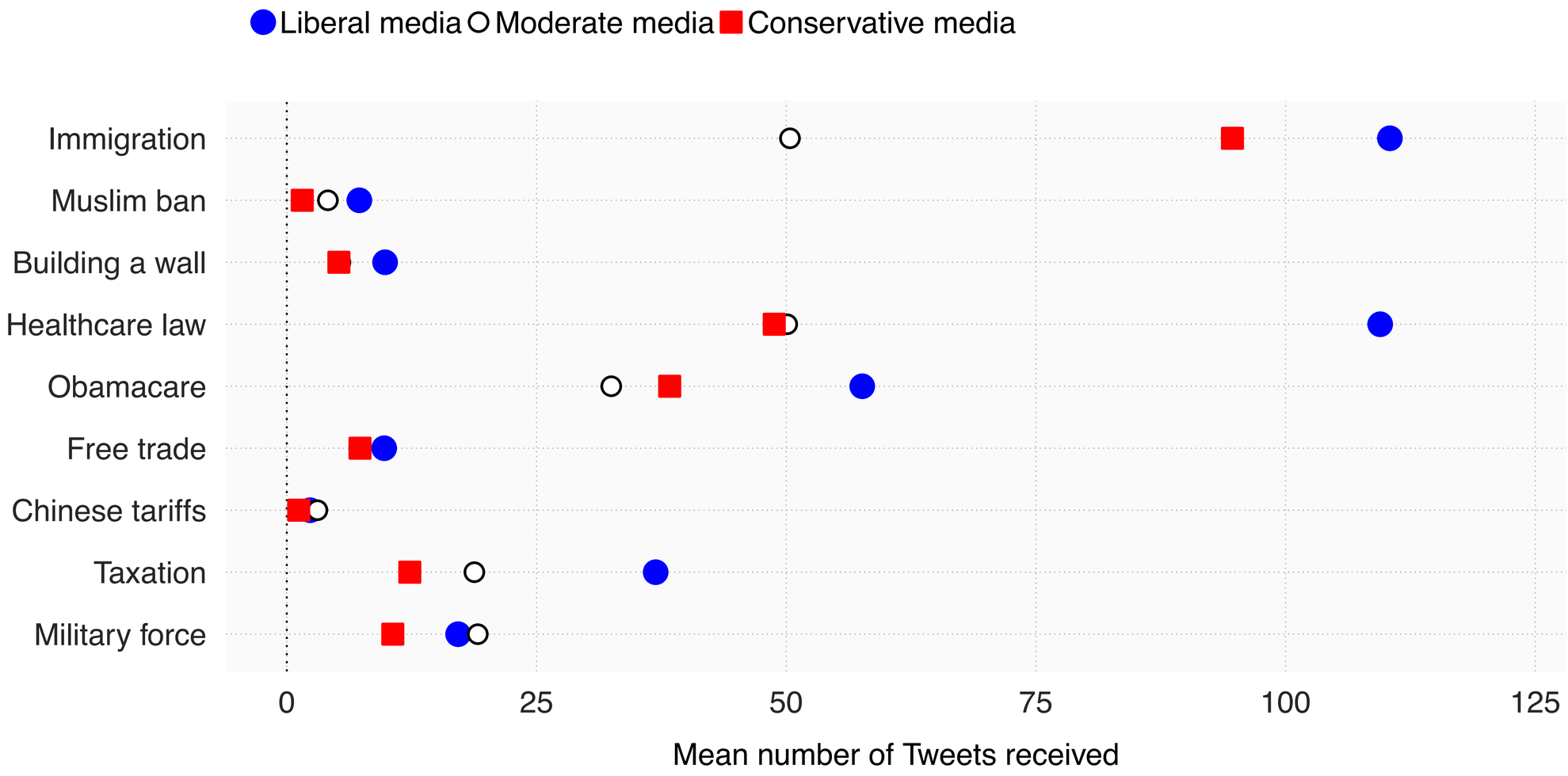
- We had a very hard time identifying the country of youtube videos
- Identifying low-quality news sites - hard
  - ESPECIALLY in spanish

Additional Variables to Extract/Label/Merge

- We labeled all tweets by accounts followed for selected topics:
  - immigration
  - build a wall
  - healthcare law
  - free trade
  - progressive taxation
  - use of military force
  - Covid-19
  - Abortion

Additional Variables to Extract/Label/Merge

- We assume stance for tweets sent by politicians and media based on ideology of the source


Liberal media
Moderate media
Conservative media
Immigration
Muslim ban
Building a wall
Healthcare law
Obamacare
Free trade
Chinese tariffs
Taxation
Military force
0
25
50
75
100
125
Mean number of Tweets received

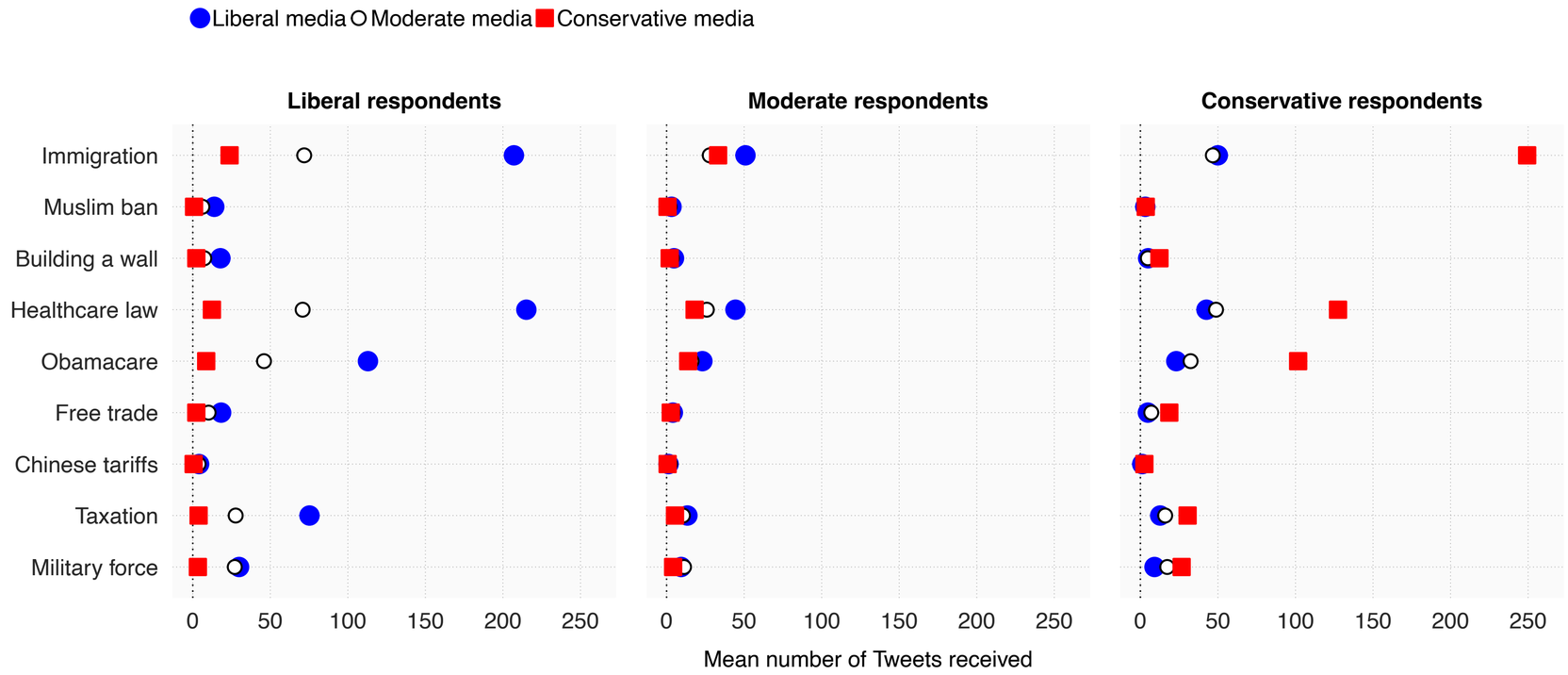


Figure 2: Average number of Tweets on campaign-related topics received by liberal, moderate, and conservative respondents. (The sample consists of respondents whose Twitter timeline was reconstructed based on the information about the accounts followed by participants. The tweets are disaggregated by the ideology of the news source.)

Effects of Media-tweets seen on Issue Placements