June 2021

README for "Social Networks Shape Beliefs and Behavior" by Michael Bailey, Drew Johnston, Martin Koenen, Theresa Kuchler, Dominic Russel, and Johannes Stroebel

The data used in this paper is described in two places: in Section 1, we describe the data from Facebook on which we conduct most of our analyses, while in Appendix B we describe the (publicly available) aggregated data that we use in that section to validate some of our core results. We include in this replication package the R and Stata code that we use to produce the figures and tables in the paper using these two datasets.
Given the sensitive nature of the data described in Section 1, we are unable to share this data as a part of this replication package. However, replication is possible for any researcher who obtains authorized access to the data directly from Facebook. Researchers interested in working with the de-identified confidential Facebook data can send inquiries to Michael Bailey (mcbailey@meta.com). Facebook works with outside researchers on a case-by-case basis through formal collaborations or internships after a satisfactory data use agreement has been agreed. In the past years, such collaborations were initiated with researchers at a number of universities, including LSE, Stanford, MIT, Harvard, NYU, Northwestern, Duke, and Boston College. The onboarding process may take several months, and researchers should expect to be subject to extensive background checks. The authors will assist with reasonable replication attempts for two years following publication.

In addition to the regression files and logs that are included in the replication package, some data preparation files are integrated into the Facebook codebase; researchers who have obtained internal access through Facebook will find these pipelines in the following folders:

- dataswarm-pipelines/tasks_adhoc/datascience/covid_academic/
- dataswarm-pipelines/tasks_adhoc/measurementsystems/covid_academic/

Facebook is unable to provide a "stable" version of the raw data that does not change over time. As a result, each re-run of these data construction pipelines will produce a slightly different sample. However, the effects of this change in the sample on final estimates are likely to be negligible.

Steps for internal data construction and replication

In this section, we will summarize the steps necessary to replicate our analyses using individual-level Facebook data.

1. **Run dataswarm-pipelines/tasks_adhoc/datascience/covid_academic/prep.py**

   This step creates basic information about the movement of users in our sample

2. **Run dataswarm-pipelines/tasks_adhoc/measurementsystems/covid_academic/public_post_sentiment.py**

   This step gathers information about users' public posts.

3. **Run dataswarm-pipelines/tasks_adhoc/datascience/covid_academic/main.py (up to make_R_output)**

   This step performs the bulk of the preprocessing necessary to gather our analysis data set. In the middle of the process, it will output training and test data needed to compute the conditional probabilities.

4. **From a Facebook-provided devserver, run the following files using Stata, which are all in the directory dataswarm-pipelines/tasks_adhoc/datascience/covid_academic/stata/. These files are provided publicly as a part of this repository.**
   - **exposure_groups_regressions.do**
   - **exposure_mvmt_regressions.do**
   - **exposure_mvmt_timing_regressions.do**

- **exposure_posts_regressions.do**

  These steps include all of our Stata-based analyses, producing figures and summary stat tables. These files are included in the replication package (in the "internal analysis code") folder, but they cannot be run as the data used in them is proprietary.

5. **Run "make_R_output" in dataswarm-pipelines/tasks_adhoc/datascience/covid_academic/main.py**

   These steps include all of our R-based analyses, producing figures and summary stat tables.

6. **Run dataswarm-pipelines/tasks_adhoc/datascience/revisions.py**

   This performs various small tasks that gather data used to add figures and numbers added in response to feedback we received during the revision process.

7. **From a Facebook-provided devserver, run the following files using R, which are all in the directory dataswarm-pipelines/tasks_adhoc/datascience/covid_academic/revisions_R/. These files are provided publicly as a part of this repository.**
   a. **fixest_robustness_regs.R**
   b. **make_exposure_balance_table.R**
   c. **make_weights.R**
      tablet_pullout_stat.R

   These files use the output of the above pipeline to generate several figures and numbers added in the revision process
   fixest_robustness_regs.R is included in the replication package, in the "internal analysis code" folder. The other three files make use of internal Facebook tools and cannot be included in the replication package.

Steps for public data construction and replication

1. Download the replication package and install all necessary software listed below
2. Acquire the Safegraph data used in the replication links, by following the instructions included alongside the list of the files that are needed
3. In Stata, set the working directory to the code directory in the replication package, and run "do metafile_zip_lvl_analysis.do". This will run all of the Stata analyses and produce the following figures:
   a. Table A24
   b. Figure A14
   c. Figure A15
4. In R, set the working directory to the code directory, then run "run_all.R". This will produce the following figures:
   a. Figure A3

Software Dependencies (Public Data)

There are several external dependencies we rely on.

In Stata 18.0:

- ftools (version 2.37.0)
- gtools (version 1.10.1)
- reghdfe (version 5.7.3)
- binscatter (version 7.02)
- estout (verson 3.32)

We also make use of a handful of R (4.3.1) packages:

- tidyverse (2.0.0)
- ggplot2 (3.4.2)
- tigris (2.0.3)
- raster (3.6-23)
- scales (1.2.1)
- sf (1.0-14)

Data Sources (Public Data)

- ./data/raw/geo/zcta_county_rel_10.txt
    - This is a zcta->county crosswalk downloaded from the Census Bureau here
- ./data/raw/SCI/*
    - This data is the the US zip code -> US zip code Social Connectedness data, the latest version of which can be found here.
    - The version of this data included in the replication package dates back to early 2020, and so does not include (possibly endogenous) friendship formation that followed the onset of the pandemic.
- ./data/raw/cases/time_series_covid19_confirmed_US.csv
  ./data/raw/cases/time_series_covid19_deaths_US.csv
    - This data is from the Johns Hopkins Coronavirus Research Center, and can be downloaded from their GitHub page. This data is occasionally retroactively updated, so downloading the latest version of these files will give different case counts from the ones used to produce the figures in the paper. For the most accurate version, you should use the versions from this commit on the repository.
- ./data/raw/cases/us-counties.csv
    - This data is derived from the Covid-19 case reports from the New York Times, and can be downloaded from their GitHub page. Like the Johns Hopkins data, this data is retroactively updated at times. To reproduce the figures most accurately, you should use the versions from this commit.
- ./data/raw/covariates/nhgis0061_csv/nhgis0061_ds239_20185_2018_zcta.csv
  ./data/raw/covariates/nhgis0062_csv/nhgis0062_ds239_20185_2018_zcta.csv
  ./data/raw/covariates/nhgis0063_csv/nhgis0063_ds172_2010_county.csv
  ./data/raw/covariates/nhgis0063_csv/nhgis0063_ds239_20185_2018_county.csv
    - This data holds various area-level covariates, drawn from NHGIS. It can be downloaded by including the necessary column names in the data downloader, which is documented here
- ./data/raw/covariates/zip_area.csv
    - This data is derived from the 2010 TIGER zcta shapefiles. It can be recreated by dropping the geometry column from this file.
- ./data/raw/covariates/us_counties2018_data.dta
    - This data is derived from the 2018 TIGER county shapefiles. It can be recreated by dropping the geometry column from this file.
- ./data/raw/safegraph/social_distancing/2020-**-**-social-distancing.csv
    - This data is drawn from SafeGraph's social distancing data set. Researchers can request a copy of this dataset from this link, though **usage requires that researchers sign a non-commercial agreement. For this reason, we cannot include this data in the replication package**. Once you have downloaded the files, place "2020-01-01-social-distancing.csv" through to "2020-07-31-social-distancing.csv" in the directory listed above.
- ./data/raw/safegraph/places/CoreApr2020Release-CORE_POI-2020_03-2020-04-07/*.csv
    - These files (core_poi-part{1-5}.csv and brand_info.csv) contain information on the Points Of Interest that Safegraph records. They can be downloaded in the same manner as the social distancing data above.
    - **This data is not included in the replication package due to contractual conditions.**
- ./data/raw/safegraph/weekly place patterns/*
    - These files contain weekly visit information for visits to the POIs listed above, measured by Safegraph. They can be downloaded in the same way. Due to a change in the format of the data midway through the series, the csvs should be stored as follows:
        - 2020-01-06-weekly-patterns.csv to 2020-06-22-weekly-patterns.csv should be placed directly in the folder above

        - Data for subsequent weeks should be stored in the five-piece format in which they are distributed in folders named “06_24” through “08_05”, with each holding “patterns-part1.csv” through “patterns-part4.csv” for the corresponding week.
    - **This data is not included in the replication package due to contractual conditions.**
- ./data/raw/HUD-zip-tract/ZIP_TRACT_122019.xlsx
    - This file contains the 2019Q4 zip->tract crosswalk from the US Department of Housing and Urban Development. It can be downloaded here.
- ./data/raw/HUD-zip-tract/ZIP_COUNTY_032020.zlsx
    - This dataset crosswalks between zip codes and counties and is provided by the US Department of Housing and Urban Development here.

Software Dependencies (Private Data)

In the analysis code we provide with Facebook data, we use the following packages on R 4.3.1:

- fixest
- tidyverse
- httr
- anesrake
- gt
- webshot
- phantomjs
- ggplot
- ggthemes
- tigris
- raster
- scales

This software will be pre-installed on the server used to run the analysis code.

For the Stata portion of the code, we use Stata version 16.0 and the following packages:

- ftools (version 2.31.3)
- reghdfe (version 5.7.3)
- binscatter (version 7.02)
- xml_tab (version 3.35)
- estout (version 3.23)