AER-2019-0658.R2 "The Welfare Effects of Social Media" Validation and Replication results

You may want to consult Unofficial Verification Guidance for additional tips and criteria.

SUMMARY

I applaud the authors for providing a sophisticated and robust lab environment. Unfortunately, the archive provided is non-functional for replication purposes. Please see detailed issues below. In a nutshell,

- the setup is very complex, and could presumably be greatly simplified. In practice it failed because the key "gslab_make" custom python module is not found despite being available, and no instructions to that extent are available.
- the archive provided is a straight copy from a Git repo. Unfortunately, no Git-LFS references were resolved, and all datasets remain as Git-LFS placeholders, with no actual data.
- While the README is very detailed with respect to setup and execution, it is entirely devoid of any data description. No manifest of provided data files is available. No reference to the confidential data is provided, nor any description of access, or reasons for absence of access, are provided in manuscript or README.

I am confident the authors can remediate these issues, and look forward to the revised archive, README, and manuscript (with complete data citations).

Details follow.

Data description

> [REQUIRED] The replication archive lacks the confidential data. The README makes no mention of that. The README should specify how to access the confidential data (if possible), and any accomodations that were made to account for the absence of confidential data. A template README file can be found on the AEA Data Editor Github site.

Data sources:

Author-collected survey data

- The authors collected data via Facebook ads. Summary statistics are provided in Online Appendix B. A codebooks is not provided.

> [REQUIRED] Location of survey instruments is mentioned in the title page footnote. Please deposit the survey instruments at a repository, which can be the AEA Data and Code Repository, or other reputable repositories. Private websites are not acceptable archival locations.

> [NOTE] It seems like Online Appendix A are the hard-copies of the survey instruments. If that is the complete set, then a correction of the footnote is in order, and sufficient.

> [REQUIRED] There does not seem to be any mention of access to the confidential data. The authors should specify (a) if the collected confidential data has been preserved (b) whether access by third-

> parties for the purpose of replication or further analysis is permissible or, in the absence of (b), if (c) support by the authors for replication is provided. An example of agreements which allow third-parties to access confidential data for the purpose of replication is provided at https://social-science-data-editors.github.io/guidance/Requested_information_data.html.

L2 "voting data"

- L2 database: No information on "L2" is provided other than "L2, a voting data provider". No codebook is provided, no summary statistics are provided.

> [REQUIRED] Please cite the L2 data, and provide information on how data could be obtained by replicators.

Other data

- The file `analysis/raw/readme.txt` cites Brynjolfsson et al. (2018) as source for data, which seems to be used only in Figure A37.

> [REQUIRED] Describe the use of this data by programs more clearly in the README.

- The files in `analysis/GeoData/` seem to be county shape files. These files are references to Git LFS which are non-functional in the openICPSR archive.

> [REQUIRED] If used, the data should be cited, and the proper source identified. If not used, then it should be removed.

- The file `data/raw/us_timezones.xlsx` is not mentioned in the manuscript or online appendix. If actually used, it should be cited. The source noted seems to be under CC-BY-SA/

> [REQUIRED] If used, please cite this data. If not, remove from replication archive.

> [NOTE] The disambiguation rule "When there is more than 1 time zone per state, we kept the first-listed time-zone" seems problematic, depending on how the data is used.

data deposit

- [REQUIRED] ZIP files should be uploaded via "Import from ZIP" instead of "Upload Files" (there should be no ZIP files visible, except in certain circumstances, like when there are too many files). Please delete the ZIP files, and re-upload using the "Import from ZIP" function.
- [REQUIRED] Please delete the `__MACOS` directory that may be generated by the above import
- [REQUIRED] Please delete empty directories
- [REQUIRED] Please delete any redundant (obsolete) files
- Further guidance can be found at https://aeadataeditor.github.io/aea-de-guidance/data-deposit-aea-guidance.html

Code description

> [STRONGLY SUGGESTED] The README does not identify which programs produce what tables or figures. This should be remediated. [STRONGLY SUGGESTED] The code was manually inspected to identify which of the 6 tables and 12 figures from the main manuscript could be identified in the code.

- The word "Table" or "Figure" does not show up in any of the Stata files under `analysis/code` in the context of numbered tables from the manuscript.
- Programs contain hard-coded parameters (code/descriptive/MakeTablesAndFigures.do)

Data checks

> INSTRUCTIONS: When data are present, run checks:
>
> - can data be read (using software indicated by author)?
> - Is data in archive-ready formats (CSV, TXT) or in custom formats (DTA, SAS7BDAT, Rdata)?
> - Does the data have variable labels (Stata: run `describe using (name of DTA)` and check that there is content in the column "variable label")?
> - Run check for PII (PII_stata_scan.do, sourced from here if using Stata) and report results. Note: this check will have lots of false positives - fields it thinks might be sensitive that are not, in fact, sensitive. Apply judgement.

> [REQUIRED] The README provides no information whatsoever about structure or content of provided data files. Please describe the data that is part of the replication archive. As per AEA Data and Code policy, "enough information should be provided (a) to accurately describe the data so that somebody who doesn't have knowledge of the data can understand its principal (and salient) characteristics (INFORMATION)".

- Data seem to be present in `analysis/raw` and `confidential/main_experiment/output`.
- The Online appendix B provides Variable Definitions and Descriptive Statistics. It is not clear which datasets (confidential or otherwise) this corresponds to. The "Variable definitions" do not identify values or encoding (are these encoded as "yes/no" variables, fully expanded, or are they mapped to values "A/B/C/D" with value explanations?) For instance, "Facebook minutes" seems to be in categories (pg. 95 of 144 of PDF_PROOF), but Table A3 (pg 103 of 144) identifies only a mean/std, and does not specify how categories are converted to scalars.

> [SUGGESTED] The README should map the Variable definitions from Appendix B to actual data files, both confidential and provided.

Replication steps

- Downloaded ZIP file from openICPSR
- Created a Ubuntu VM with Anaconda3 (installer as 2019-09-02) and Stata 14
- Ran setup as per README
    - Installed git-lfs as per Ubuntu instructions

```
sudo apt install curl
curl -s https://packagecloud.io/install/repositories/github/git-lfs/script.deb.sh | sudo bash
sudo apt-get install git-lfs
git lfs install
```

- Further steps per README
    - Ran `python -m pip install --user -r requirements.txt` successfully

- Ran `python check_setup.py` - failed, no module named "gslab_make", and no instructions on how to install.
- Ran `stata-mp -e setup_stata.do` which installed 21 packages
- Created a "config_user.yaml" file and adjusted to my system

> [SUGGESTED] The setup process is intrusive on replicator's system. It would be useful to have a Docker or VM image provided as part of the process.

> [NOTE] The instructions on lab website do not provide instructions for Linux, which might be useful for the above [SUGGESTION]

> [SUGGESTED] Redundant or useless steps for replicators should be eliminated.

- Examples of redundant setup:
  - Replicators have no access to files by the authors stored on Git-LFS
  - It is not clear that Lyx is necessary
  - Are all 21 Stata packages used by the code in this repository?

> [SUGGESTED] We applaud the authors of having specified a LICENSE. To increase machine-readable visibility, we strongly suggest this license be saved as a separate file `LICENSE` or `LICENSE.txt`.

Findings

The README specifies 5 steps, connected to 5 directories.

Issues:

- The code in `confidential` will not run due to the absence of confidential information. This is not mentioned in the main README, only in the readme of the subdirectory.

- As an alternative to running the code in `confidential`, anonymized output data is promised in the `output` directories (presumably `./main_experiment/output` and `./L2/output`). HOwever, the files in those directory are Git-LFS references, and are not accessible to replicators.

- The code in `data` will not run due to the above absence of output files.

- The Python make file will not run due to the absence of the "gslab_make" module.

- Manually configuring based on inspection of `inputs.txt` leads to the expected failure due to the absence of `confidential_main/baseline_anonymous.dta`

- The directory `paper` (step 5) does not exist. This is presumably `paper_slides`.

- The code in `analysis` does not run, in the absence of confidential data. This is not mentioned in the main README, only in the readme of the subdirectory.

STOPPED HERE

> [REQUIRED] Please update the README so that it accurately reflects what a replicator should and can run, and what workarounds are in place, if any, to palliate the absence of confidential data.

> Please provide all files that seem to be necessary, and provide a manifest of the necessary and sufficient files to run the available code, as part of the README.

Classification

- [ ] full replication
- [ ] partial replication (see above)
- [x] not able to replicate (reasons see above)