Accompanying code for “Estimating causal effects when treatments are entangled by a network of units” — Yahoo! Data Example

This document describes the code to reproduce the results of the Yahoo! Go application in Section 6 of the paper. Two R script files are necessary: `yahoo_preprocess.R` and `yahoo.R`. These can reproduce Table 3 in the paper. They can also be used to reproduce the power simulated study in the Supplementary Material

1 Obtain the data

To obtain the data, follow these steps:

1. Make sure you have an AWS account. You will need a ’Canonical User ID’ and, later, the public/secret keys to download the data.
2. Open up a simple Yahoo account if you haven’t one already.
3. Go here `https://webscope.sandbox.yahoo.com/catalog.php?datatype=i&did=67`.
4. Open a request for dataset ”G7 - Yahoo! Property and Instant Messenger Data use for a Sample of Users, v.1.0 (4.3 Gb) (Hosted on AWS).” You will need a short description for the project you need it for. You will also need to use an @edu email account.
5. Wait for instructions on your @edu email account on how to download the data on AWS S3. It shouldn’t take more than 2-3 days. Make sure to check your spam.

After you download the data and decompress the files, you should be able to see the “Yahoo” folder. In this folder, you should see several README files (in .docx and .pdf format) and also folder named “dataset” containing all dataset files in .dat format.

2 Preprocess the data

To correctly pre-process the data, execute the following steps:

1. Place `yahoo_preprocess.R` inside the “Yahoo” folder created in Step 1 described before.
2. Source the file.
3. This should create a sub-folder named “simple” inside the “dataset” folder. This should contain the files `user.rda, go.rda` and several `im_*.rda` files. If the script has failed due to access issues, then create an empty subfolder “simple” manually, and re-source the script.

3 Reproduce Table 3

Make sure `dplyr` is installed. Then, to reproduce Table 3 follows these steps:

1. Open the file `yahoo.R`, which may located anywhere on your computer.

2. Make sure to set the “DATA_FOLDER” variable to point to the root “Yahoo” directory on your machine.

3. Source the script.

4. Run `Process()` on the command line. This should take about a minute to run, and also prints some messages summarizing the communication network being generated.

5. Next,

   - `Table3_standard()` can be used to reproduce the point estimators and randomization-based confidence intervals for the naive method reported in Table 3 of the paper.
   - `Table3_entanglement()` can be used to reproduce the point estimators and randomization-based confidence intervals for the entanglement-aware method reported in Table 3 of the paper.

   Note: The functions have been preset to calculate the p-values only for the endpoints of the confidence intervals. All of these p-values should evaluate to 0.05 showing that the corresponding parameter values are indeed approximate endpoints of the corresponding confidence interval.

6. For the placebo study in the Supplementary Material, please run `Placebo_study()`.

4 Contact

For any questions please email Panos Toulis (panos.toulis@chicagobooth.edu).