REPLICATION INSTRUCTIONS: Birds of a Feather Tweet Together. Bayesian Ideal Point Estimation Using Twitter Data.

Pablo Barberá
pablo.barbera@nyu.edu

Version 1.0, July 6th 2014

This document provides details about the replication materials for the article. The **Data** section describes the datasets of political elites collected for the estimation in all six countries included in the paper, as well as the tweet collections. The **Script** section summarizes the purpose of each R script, the data required for each of them to be run, and the output. The **Output** section provides additional information about the files generated after running the scripts, also included in the replication materials.

All files containing personal information about individual users have been anonymized for privacy reasons, and the original User IDs have been replaced by randomly-generated ID numbers. The complete version of all datasets is available upon request.

Data

The following data files are necessary in order to replicate the analysis in the article:

- *data/elites-data.rdata* contains the names of the political elites included in the estimation for all six countries, as well as their Twitter profile information as of November 2012 for the US, Spain, the Netherlands, and the UK, February 2013 for Italy, and August 2013 for Germany. The data set also indicates the party to which each politician belongs, and their DW-NOMINATE scores in the case of Members of the U.S. Congress.

- *data/state-data.rdata* contains information about the mean liberal opinion in each state (Lax and Phillips, 2012), and measures of "Republican advantage" by Gallup.

- *data/contributor-data.rdata* contains information about campaign contributions from the dataset compiled by Bonica (2013), such as the amount donated to Democratic and Republican candidates and each contributor's CFscore. Personal information such as their unique ID numbers has been anonymized.

- *data/ohio-data.rdata* contains information about a sample of Twitter users from Ohio that were matched with the voter registration file, as of November 2012. Their voter ID and Twitter ID have been anonymized.

- *data/obama-tweets.txt* and *data/romney-tweets.txt* contain the tweet IDs for all tweets mentioning "Obama" or "Romney" during the 2012 presidential election campaign. In compliance with Twitter's Terms and Conditions, the entire dataset of tweets used in the article cannot be distributed, only the IDs of the tweets. The script *recover-tweets.R*, which provides functions to re-generate the entire dataset of tweets by querying the API, is also included in the replication materials.

- *data/retweets-data.rdata* contains the list of retweets used in the analysis in Section 6, extracted from the "obama" and "romney" collections (IDs of "retweeter" and "retweeted"). Note that the original user IDs have been replaced by randomly-generated ID numbers for privacy reasons.

Scripts

The following R scripts replicate each step in the analysis

- *00-install-packages.R* installs all the R packages used in the analysis, and provides instructions to create an authentication token to query the Twitter API.

- *01-get-twitter-data.R* downloads the lists of followers for each of the elite accounts in each country.

- *02-get-users-data.R* creates the full list of users to be included in the analysis, downloads user information from Twitter's REST API and applies the basic spam and geography classifier described in Section 3.

- *03-create-adjacency-matrix.R* generates the adjacency matrix (**Y**) indicating what users follow each political elite, which is then used to estimate the ideal points. The output of this script are the files *adj-matrix-XX.rdata*, where XX is the name of each country.

- *04-model-first-stage.R* fits the first stage of the "spatial following model" (see Section 2.4 for more details). It generates the files *samples-XX.rdata* and *resultes-elites-XX.rdata*.

- *05-model-second-stage.R* fits the second stage of the "spatial following model" (see Section 2.4 for more details). The exact code used in the article to estimate all 300,00 users' ideal points in the US is in *05b-model-second-stage-complete.R*.

- *06-analysis-section-4-1.R* replicates the analysis and figures in Section 4.1 of the paper, comparing Twitter-based ideology estimates with DW-NOMINATE for Members of the US Congress, and with expert surveys in five European countries.

- *07-analysis-section-4-2.R* replicates the analysis and figures in Section 4.2 of the paper, examining the ideological distribution of ordinary users and political elites in the US, and users who self-identify as "conservative", "moderate", and "liberals"; and comparing state-level aggregates of political ideology with other existing measures of ideology at the state level.

- *08-analysis-section-4-3.R* replicates the analysis and figures in Section 4.3 of the paper. It compares ideal point estimates for individual Twitter users with their campaign contributions and voter registration history.

- *09-analysis-section-5.R* replicates the analysis and figures in Section 5 of the paper. It examines the distribution of activity on Twitter by ideology, as well as the level of ideological polarization on retweet activity.

- *functions.R* is an additional script that contains some of the functions used in the previous scripts.

- *recover-tweets.R* shows how to re-generate the entire collection of tweets used in Section 5 of the paper.

Output

The following results data files, which provide the output of the intermediate and final steps of the analysis, are also provided as part of the replication materials, in order to facilitate the replication of the paper.

- *output/adj-matrix-XX.rdata* contains the adjacency matrix (**Y**) used in the estimation of the ideal points.

- *output/samples-XX.rdata* contains the MCMC samples resulting from running the first stage of the model.

- *output/results-elites-XX.rdata* is a summary dataset of the ideal point estimates for the political actors in each country.

- *output/users-data-US.rdata* combines anonymized user information about each Twitter user in the US sample with their ideal point estimates.