Project Proposal: Creating Comparison Groups for Event Studies on Reddit

Dan Connolly & Shilin Ma

March 23, 2022

1 Original Proposal

Reddit is a link aggregation and user discussion-based website. The site features many different subcommunities ("subreddits") focused on different topics. Last year, I started a research project on how the behavior of regular Reddit users changes after they lose their job. In order to answer this question, I scraped user posts from an unemployment-focused subreddit ("r/Unemployment") on which users post questions about how to apply for unemployment insurance in the United States. By restricting my attention to users who had been regular Reddit posters before they posted to r/Unemployment, I was able to analyze how their posting activity changed after losing their job (proxied by their first post to r/Unemployment). To further validate my findings, I constructed a comparison group of users by examining the subreddits in which the original group of users posted most frequently prior to unemployment, and scraped those subreddits for other users with similar profiles. This matching was done very simplistically [1], and I think there is a significant opportunity to both improve my research project and/or develop a package available to anyone who wants to match users on Reddit.

In the social sciences (and in economics in particular), the "event study" is a paradigm in which researchers examine an outcome of interest before and after a plausibly exogenous event and compare differences in the outcome to individuals who do not experience the event. Reddit data is extremely rich, but the ability to make inferences about users in an event study paradigm is limited by how difficult it is to construct an appropriate matched control. I propose to revisit my initial implementation of user matching on Reddit and design and implement at least one new matching strategy. The current implementation uses a squared-loss function based on users' account age and "karma" (a measure of how well or poorly posts by the user have been received by other users). Calculating this loss is both computationally expensive and probably not very accurate.

[1] Inspired by Chandrasekhan et. al. 2017, available here: https://comp.social.gatech.edu/papers/cscw18-chand-hate.pdf

2 Proposal Updates

After presenting the idea to the class, Shilin expressed interest in working on the project. We are excited to work together!

We'd like to amend the proposal to include the following goals (in order of feasibility):

1. "Clean up" the original code. The original code was written for a very specific use, and we'd like to genericize it to accept (at minimum) arguments for the "event" subreddit of interest and a desired number of matched users.

2. Write a set of test cases for each of the component functions. I wrote the original code quickly, and I am sure there are hidden errors that might be of importance for someone else trying to use the software. For example, we should write a series of tests for the matching function that ensure that the matched user is returned as a "Redditor" object, is not equivalent to the original user, and has a nonzero number of comments or posts.

3. Write a Jupyter notebook that summarizes and visualizes match quality. For example, we could plot overlapping histograms showing post activity over time for treatment and control users to assess how similar the activity is on average.

4. Implement a more sophisticated matching function. As discussed above, the current matching function is highly approximated. We would decide on a set of implementable criteria for a match and rewrite the matching function to assess these criteria.

5. Post the software as a public package, with separate files for scraping a set of treatment users, matching them with control users, scraping control user activity, and then running a basic set of diagnostics (per #3 above) on match quality.