Sept 15:

I added two of the three things Christo mentioned into the /net/data/brand-msgs/ folder:

1.       survey.csv: a file containing survey responses, the main column you want is `psid` which links to `user\_id` in the browser history data. There will be many more survey respondents than extension users (low uptake), and many duplicates since the survey was conducted over time. The main use of this file is to link user demographics to the browser history.

2.       bh-clean.tsv: a file containing users’ visits to websites, is equivalent to the browser history you see in your own browser if you use Chrome or Firefox

I added a README.md file to that folder as well, which has descriptions for all the columns in bh-clean.tsv. I don’t have documentation for the survey file, but most of the columns should be self-explanatory (e.g. “age”), and I’ll forward the documentation for the rest as soon as I get it from the folks who were running it.

The third item – the ad preference profiles – we never pulled from the raw SQL database, so it’ll take a little longer to get those. I briefly looked into it though, and we were periodically collecting data from Google ([https://adssettings.google.com/authenticated](https://nam12.safelinks.protection.outlook.com/?url=https%3A%2F%2Fadssettings.google.com%2Fauthenticated&data=04%7C01%7Cy.bart%40northeastern.edu%7Cea92d95a4956492e1b4b08d99813784e%7Ca8eec281aaa34daeac9b9a398b9215e7%7C0%7C0%7C637708031662579584%7CUnknown%7CTWFpbGZsb3d8eyJWIjoiMC4wLjAwMDAiLCJQIjoiV2luMzIiLCJBTiI6Ik1haWwiLCJXVCI6Mn0%3D%7C1000&sdata=bLVUx%2BKHejGo9SOck8sTm89fT6lbIeW6EiKmI77DHu8%3D&reserved=0)) and BlueKai ([https://o.bluekai.com/registry](https://nam12.safelinks.protection.outlook.com/?url=https%3A%2F%2Fo.bluekai.com%2Fregistry&data=04%7C01%7Cy.bart%40northeastern.edu%7Cea92d95a4956492e1b4b08d99813784e%7Ca8eec281aaa34daeac9b9a398b9215e7%7C0%7C0%7C637708031662589541%7CUnknown%7CTWFpbGZsb3d8eyJWIjoiMC4wLjAwMDAiLCJQIjoiV2luMzIiLCJBTiI6Ik1haWwiLCJXVCI6Mn0%3D%7C1000&sdata=RyLja%2Bep69bULBycobeUWTo0FnoeSuqVXRUgf9V5Dns%3D&reserved=0)) for all users. Will ping you when I get those moved over to the folder, but I imagine there will be substantial data munging needed on that, whereas the browser history data should be good to go as is.

Oct 25:

I got the rest of the data you requested and put it in two new folders in /net/data/brand-msgs/: bluekai and google ads. Each folder has two files in it. One has a “.json.bz2” file ending, and contains bzip2 zipped json where each line is a dictionary containing two keys: id and html. The value stored in the html key should be the raw HTML obtained by sending requests to the urls I listed in my previous email from each participant’s browser. We sent those requests once a week for the duration of the study. The other key (“id”) allows you to link the HTML to the data in the other file in each folder “subset.tsv” (e.g. bluekai/subset.tsv). Besides containing an “id” column to link the HTML data to, the subset.tsv file also contains other metdata, like when the request was made and the user ID of the participant we collected it through. It’ll take some exploring to extract structured data from the HTML, hope that turns out to be useful. Let me know if you have any questions.