The Illusion of Privacy on Reddit

Renzo Lucioni ‘14, R.J. Aquino ‘14

rlucioni@college.harvard.edu, rjaquino@college.harvard.edu

Harvard University

The Politics of Personal Data

Gov 1430

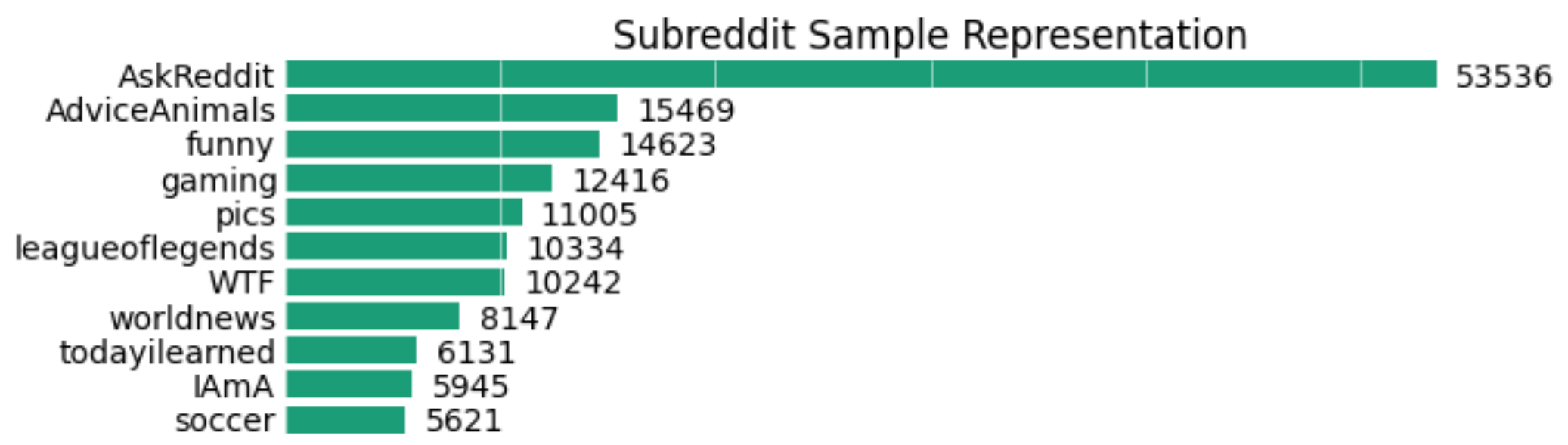
ABSTRACT

The goal of this project was to explore privacy on Reddit, the popular social news and entertainment website. On Reddit, users register under a pseudonym, submit content primarily in the form of links, and make text comments on content submitted by others. We set out to harvest personally identifying information from a collection of 15 million comments posted to Reddit between August 28 and September 12. We succeeded in finding many Reddit users’ Facebook profiles, Twitter handles, email addresses, phone numbers, and ZIP codes. This information allowed us to identify many Redditors with a high degree of confidence. These results demonstrate that the notion of privacy on Reddit is an illusion for many users. Creating a truly anonymous environment on Reddit would require measures preventing users from posting this kind of information. However, such changes would be in direct opposition to the way the community on Reddit behaves.

INTRODUCTION

For many, reddit.com is a website that users browse anonymously, or under pseudonyms/usernames. In their Privacy Policy, Reddit states that “we will only share your personal data with your consent”. For these reasons, it is reasonable for Reddit users to expect that their personal information is indeed private, and that their “real world” data (name, date of birth, phone number, etc) is not connected to their “digital” data (i.e. their Reddit username). We intend to demonstrate that a number of reddit users have, either on purpose or accidentally, revealed identifying information, and that the information they’ve revealed can be tied back to a real identity.

Many Reddit users publish phone numbers, email addresses, Twitter handles, and Facebook profiles. This information can sometimes be used automatically to generate an identity, and often times can be combined with some extra sleuthing to reveal personal information. We combed 15 million comments looking for personally identifying information.



Above: “Subreddit” counts over a sample of 500,000 comments.

MATERIALS AND METHODS

Our project consists of two parts: (1) surveying Reddit as a whole for personal information, and (2) attempting to identify specific users based on the personal information we found on Reddit.

Here is a representative comment (with only interesting fields shown):

```
{
    u'author': u'olafthebent',
    u'body': u"The island airport was one of
    my all-time pleasant flying
    experiences... 100 times better than
    Pearson.\n\nAs long as the jest don't
    produce more sound (and we're told they
    don't) then it shouldn't be a problem",
    u'id': u'cbwxttd',
    u'subreddit': u'toronto',
    u'ups': 1
}
```

We decided to search for personally identifying information at various levels of specificity. First, we searched for keywords, like “birthday” and “email”.

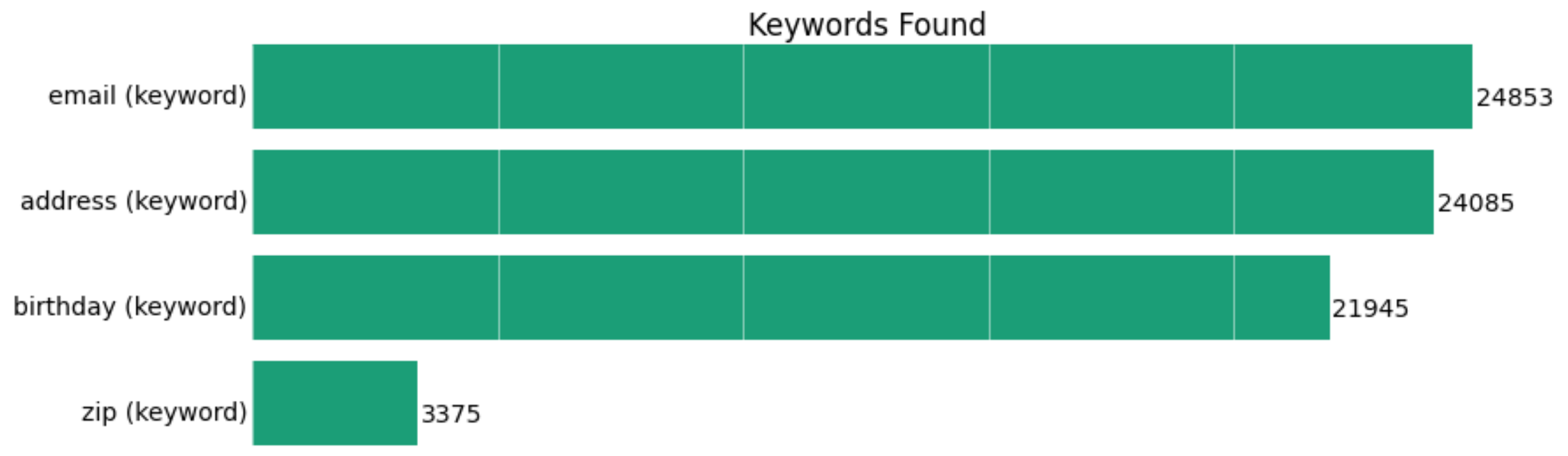
Next, we searched for actual information, in particular email addresses and phone numbers, using regular expressions for pattern matching.

Finally, we looked for social networking information, both facebook.com URLs and Facebook profile URLs (of the form facebook.com/profile.php?id=XXXX)

For the second goal of our project, we took just the comments output by the above searches, and refined our methods, searching for personally identifying words like “my”, “personal”, “work”, “private”, “brother”, “sister”, among others. Using this filtered list of comments, we tried to identify particular users, aggregating the information we could accrue from the comments themselves. The Facebook profile URLs alone would likely identify a user, but we also looked at phone numbers and email addresses.

RESULTS

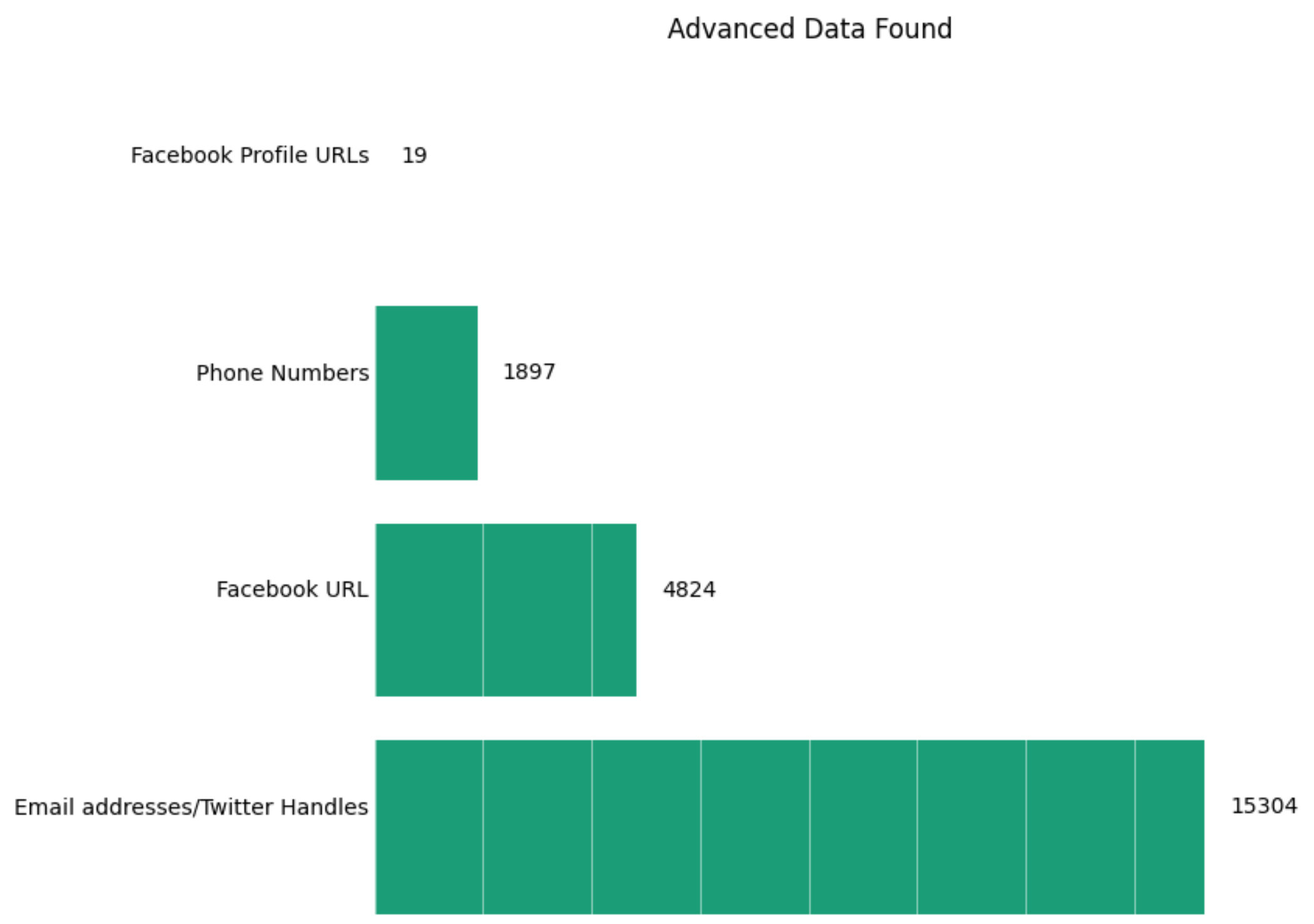
The keyword search revealed that Reddit users are at least discussing personal information. In our data set of 15 million comments, the keywords email, address, and birthday occurred over 20,000 times each.



RESULTS

Looking at the comments returned by this search revealed that the signal to noise ratio is pretty low– most of the comments do not contain identifying information, just the keyword in question. For instance, discussion of the NSA surveillance often contained the word “email”, without the user actually posting a real email address. We determined, then, that this data would not easily automatically give us results. We saved it for part (2) of our project, however.

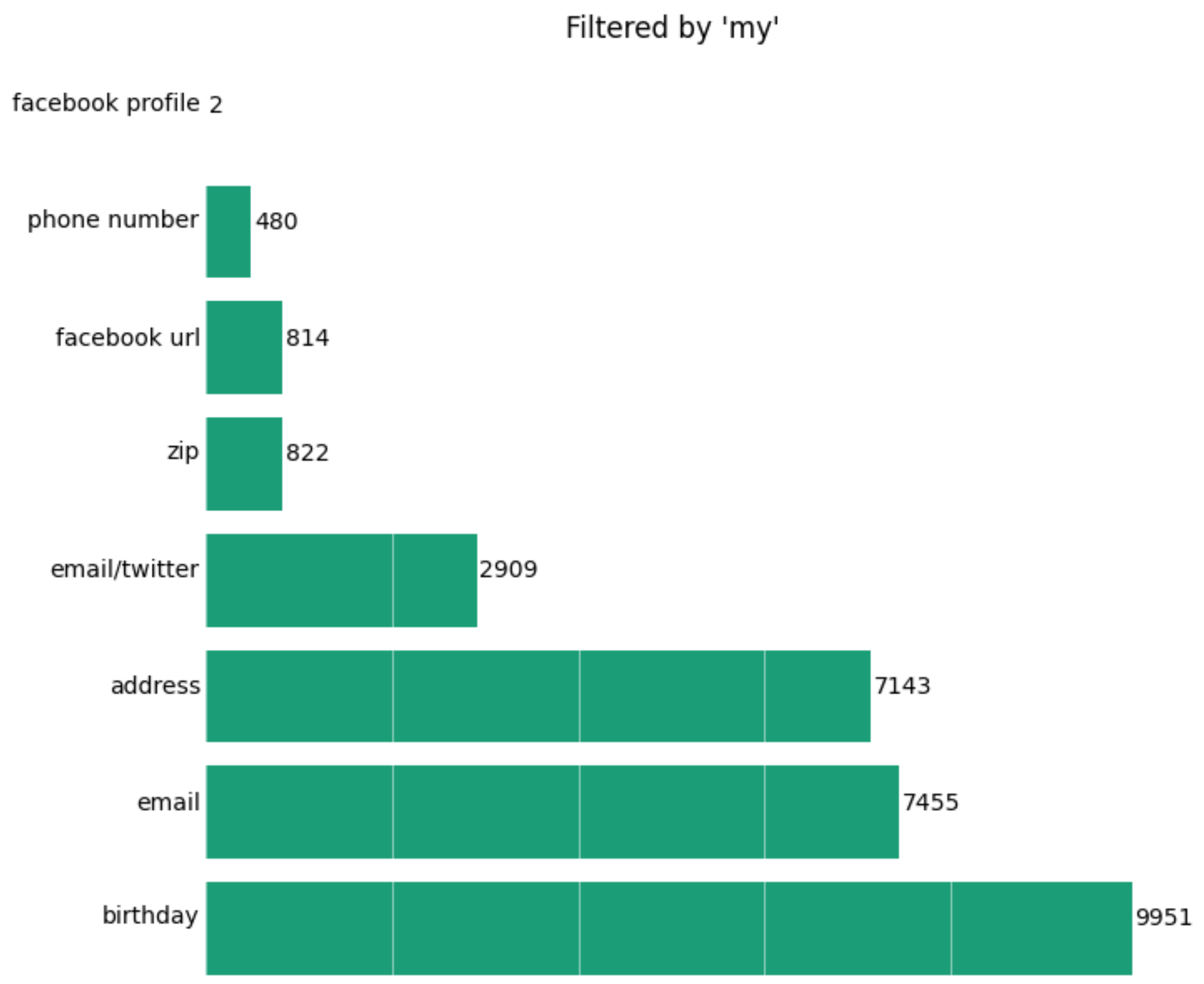
We decided to search for more specific information – actual email addresses, phone numbers, and Facebook URLs.



We found almost 5,000 Facebook URLs, and even 20 Facebook profile URLs. Looking at these, they were much more identifying, often tying Reddit accounts to a user’s band’s Facebook page, or even their personal Facebook account! In one instance, the user’s Facebook ID was embedded in the URL they posted, almost definitely accidentally.

For part (2), we decided to invest a bit more manual time. We went back to our first results, and filtered on keywords like “my” and “today”. We found that this resulted in a much smaller dataset that could be surveyed manually.

Specifically, 929 of the comments containing “birthday” also contained “today”, and 814 of the comments containing a facebook.com url contained the word “my”. Scanning these, we saw that most of them seemed personal – comments like “today is my birthday” and “like my page” were common, leading us to believe that this dataset would be much more useful for identifying users.



CONCLUSIONS

As we expected, there is a lot of personally identifying information on Reddit. It was not as easy as we hoped, however, to separate the signal from the noise. By searching for specific patterns of information and by filtering broad searches with “personal” words, however, we were able to connect personal information, like birthdays, to Reddit users.

Furthermore, we were able to identify a handful of Reddit users, specifically through their Facebook profiles. Using the information we connected to their accounts did not lead to identifications directly, but would aid in verifying connections made through other means.

NEXT STEPS

Moving forward, we would hope to run more complex analyses on the data, looking for a better way to identify “personal” comments, possibly through a combination of natural language processing and machine learning techniques. In particular, if our methods can reliably identify posts with personal information, we would like to set up a scraper that continuously polls the Reddit data feed, and alerts us to new personal posts (without the need to download and save every Reddit comment.

REFERENCES

The Reddit comment data set was collected and cleaned by Reddit user w2m3d and posted here:
http://www.reddit.com/r/datasets/comments/1mbsa2/155m_reddit_comments_over_15_days/

The Reddit logo (“Snoo”) was taken from
http://www.redditstatic.com/about/assets/reddit-alien.png

Harvard University
Department of Government

The Institute
for Quantitative Social Science
at Harvard University

DATA PRIVACY LAB