Masthead Logo

FDLA Journal

Volume 4 *Winter 2019*

Article 2

2017

Are you really anonymous online? Your friends on Twitter may give you away

Jessica T. Su

Stanford University, jtysu@stanford.edu

Are you really anonymous online? Your friends on Twitter may give you away

Cover Page Footnote
Republished from: The Conversation: Academic rigor, journalistic flair (February 7, 2017 12.07pm EST)

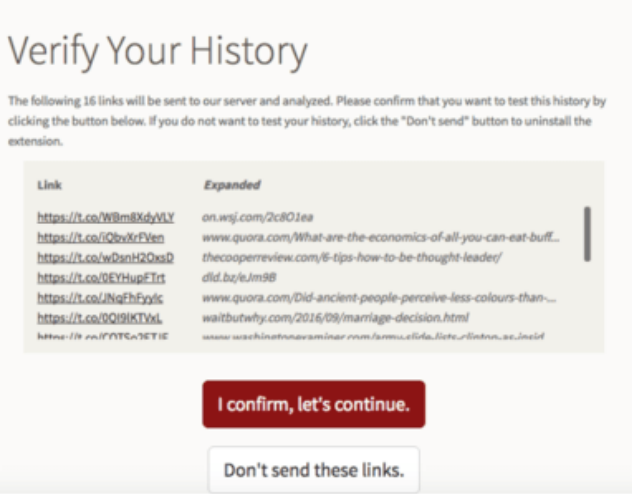
Are you really anonymous online? Your friends on Twitter may give you away

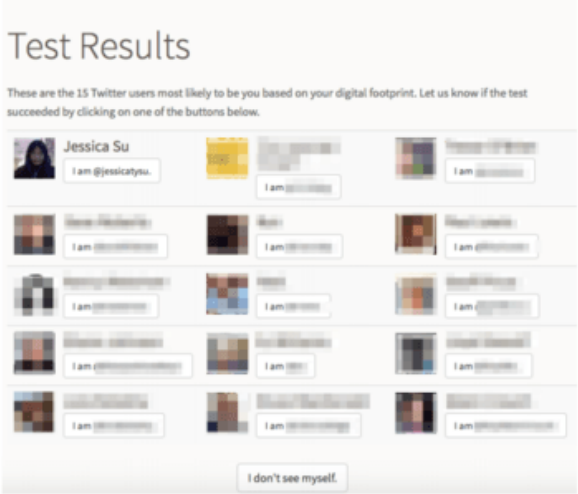
As you browse the internet, online advertisers track nearly every site you visit, amassing a trove of information on your habits and preferences. When you visit a news site, they might see you're a fan of basketball, opera and mystery novels, and accordingly select ads tailored to your tastes.

Advertisers use this information to create highly personalized experiences, but they typically don't know exactly who you are. They observe only your digital trail, not your identity itself, and so you might feel that you've retained a degree of anonymity.

But, in a paper I coauthored with Ansh Shukla, Sharad Goel and Arvind Narayanan, we show that these anonymous web browsing records can in fact often be tied back to real-world identities.</p>

To test our approach, we built a website where people could donate their browsing history for the purposes of this study. We then tried to see if we could link their histories back to their Twitter profiles using only publicly available data. Seventy-two percent of people who we tried to deanonymize were correctly identified as the top candidate in the search results, and 81 percent were among the top 15 candidates.





This is, to our knowledge, the largest-scale demonstration of deanonymization to date, since it picks the correct user out of hundreds of millions of possible Twitter users. In addition, our method requires only that a person clicks on the links appearing in their social media feeds, not that they post any content – so even people who are careful about what they share on the internet are still vulnerable to this attack.

How it works

At a high level, our approach is based on a simple observation. Each person has a highly distinctive social network, comprising family and friends from school, work and various stages of their life. As a consequence, the set of links in your Facebook and Twitter feeds is highly distinctive. Clicking on these links leaves a tell-tale mark in your browsing history.

By looking at the set of web pages an individual has visited, we were able to pick out similar social media feeds, yielding a list of candidates who likely generated that web browsing history. In this manner, we can tie a person's real-world identity to the nearly complete set of links they have visited, including links that were never posted on any social media site.

Carrying out this strategy involves two key challenges. The first is theoretical: How do you quantify how similar a specific social media feed is to a given web browsing history? One simple way is to measure the fraction of links in the browsing history that also appear in the feed. This works reasonably well in practice, but it overstates similarity for large feeds, since those simply contain more links. We instead take an alternative approach. We posit a stylized, probabilistic model of web browsing behavior, and then compute the likelihood a user with that social media feed generated the observed browsing history. Then we choose the social media feed that is most likely.

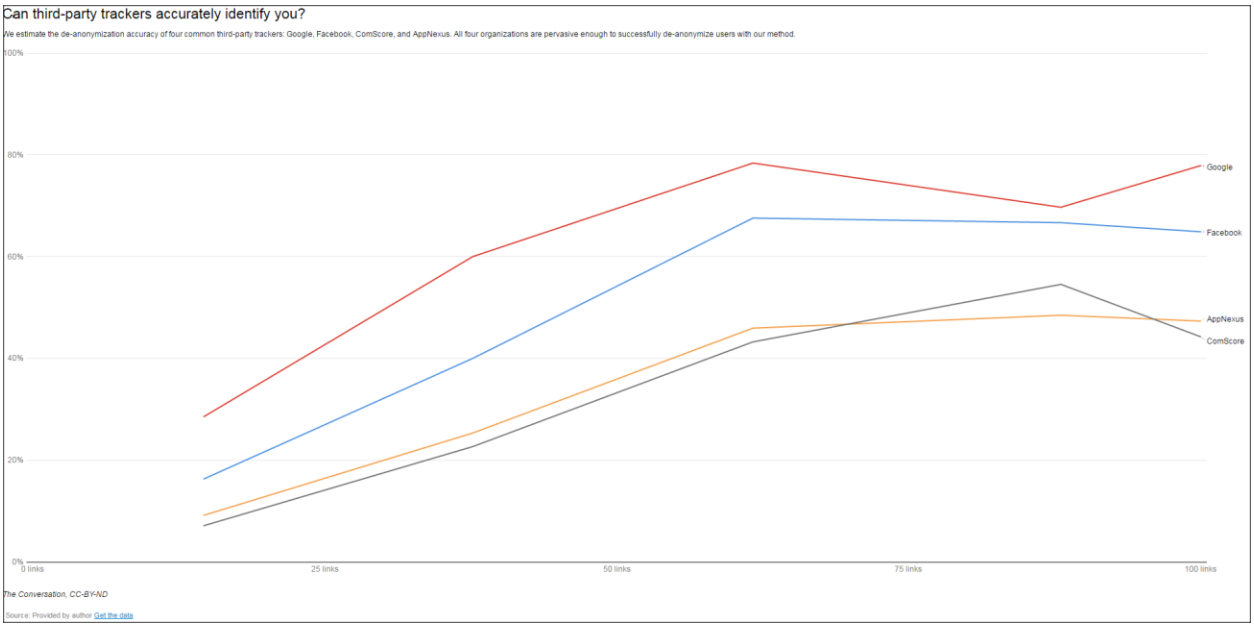
The second challenge involves identifying the most similar feeds in real time. Here we turn to Twitter, since Twitter feeds (in contrast to Facebook) are largely public. However, even though the feeds are public, we cannot simply create a local copy of Twitter against which we can run our queries. Instead we apply a series of techniques to dramatically reduce the search space. We then combine caching techniques with on-demand network crawls to construct the feeds of the most promising candidates. On this reduced candidate set, we apply our similarity measure to produce the final results. Given a browsing history, we can typically carry out this entire process in under 60 seconds.

Our method is more accurate for people who browse Twitter more actively. Ninety percent of participants who had clicked on 100 or more links on Twitter could be matched to their identity.

Many companies have the tracking resources to carry out an attack like this one, even without the consent of the participant. We attempted to deanonymize each of our experiment participants using only the parts of their browsing histories that were visible to specific tracking companies (because the companies have trackers on those pages). We found that several companies had the resources to accurately identify the participants.

Can third-party trackers accurately identify you?

We estimate the de-anonymization accuracy of four common third-party trackers: Google, Facebook, ComScore, and AppNexus. All four organizations are pervasive enough to successfully de-anonymize users with our method.



Other deanonymization studies

Several other studies have used publicly available footprints to deanonymize sensitive data.

Perhaps the most famous study along these lines was performed by Latanya Sweeney at Harvard University in 2002. She discovered that 87 percent of Americans were uniquely identifiable based on a combination of their ZIP code, gender and date of birth. Those three attributes were available in both public voter registration data (which she bought for US$20) and anonymous medical data (which were widely distributed, because people thought the data were anonymous). By connecting these data sources, she found the medical records of the governor of Massachusetts.

In 2006, Netflix ran a contest to improve the quality of its movie recommendations. They released an anonymized dataset of people's movie ratings, and offered $1 million to the team that could improve their recommendation algorithm by 10 percent. Computer scientists Arvind Narayanan and Vitaly Shmatikov noticed that the movies people watched were very distinctive, and most people in the dataset were uniquely identifiable based on a small subset of their movies. In other words, based on Netflix movie choices and IMDB reviews, the researchers were able to determine who those Netflix users actually were.

With the rise of social media, more and more people are sharing information that seems innocuous, but actually reveals a lot of personal information. A study led by Michal Kosinski at the University of Cambridge used Facebook likes to predict people's sexual orientation, political views and personality traits.

Another team, led by Gilbert Wondracek at Vienna University of Technology, built a "deanonymization machine" that figured out which groups people were part of on the social network Xing, and used that to figure out who they were – since the groups you are part of are often enough to uniquely identify you.

What you can do

Most of these attacks are tricky to defend against, unless you stop using the internet or participating in public life.

Even if you stop using the internet, companies can still collect data on you. If several of your friends upload their phone contacts to Facebook, and your number is in all of their contact lists, then Facebook can make predictions about you, even if you don't use their service.

The best way to defend against deanonymizing algorithms like ours is to limit the set of people who have access to your anonymous browsing data. Browser extensions like Ghostery block third-party trackers. That means that, even though the company whose website you're visiting will know that you're visiting them, the advertising companies that show ads on their page won't be able to gather your browsing data and aggregate it across multiple sites.

If you are a webmaster, you can help protect your users by letting them browse your site using HTTPS. Browsing using HTTP allows attackers to get your browsing history by sniffing network traffic, which lets them carry out this attack. Many websites have already switched to HTTPS; when we repeated our deanonymization experiment from the perspective of a network traffic sniffer, only 31 percent of participants could be deanonymized.

However, there is very little you can do to protect yourself against deanonymization attacks in general, and perhaps the best course of action is to adjust one's expectations. Nothing is private in this digital age.

Jessica Su, Ph.D. Student at Stanford, Stanford University