Measuring the Digital Footprint

A Tool for Understanding the Data Collected from Internet Users

Zachary M. Shaffer
Department of Computer Science
Allegheny College
shafferz@allegheny.edu
http://shafferz.github.io

May 4, 2018

Abstract

Every week, hundreds of millions of Americans access the internet for billions of hours. In this time, the user generates a distinct digital footprint that can be used to identify the user's personal information and browsing habits, such as legal names, addresses, phone numbers, entertainment viewing habits, social media accounts, email addresses, and much more. These deeply personal dossiers of information can be linked to a given user with a plurality of tools, cookies, and other unique pieces of identifying information. Since most Americans have concerns about their digital privacy, this thesis proposes a tool that will identify, measure, and present a user with their own digital footprint, and give the user advice on how to reduce the transparency of this information. This tool is being developed on the forefront of an age when digital privacy is a major concern throughout the United States as a topical and poignant subject.

1 Introduction

In 2016, the average American internet user spent about 23.6 hours online each week [1]. As of June 30, 2017, there are over 286 million American internet users. That is a collective 6.7 billion hours spent online in America alone, and America only accounts for 9.8% of all internet usage worldwide [2]. Keeping track of what any one given individual does online, then, should seem to be fairly difficult. That is not the case, however, as websites are using information given by users who access the internet to build and sell profiles of any given individual user.

This collection of information on a user, known as the user's digital footprint, can tell a website anything ranging from their location and IPv6 address, to what kinds of videos the user likes to watch online, or even what kind of products they like to purchase. The average user, however, is typically unaware of the kind of information they are providing to these websites. The motivation for the creation of a digital footprint measuring tool comes from the simple fact that millions of users who spend billions of hours online do not typically understand what is revealed when interacting with the internet. Furthermore, the issue of privacy online is important to most Americans, as 93% of American adults want to control who can access their information, and 90% of American adults want to control what information can be accessed about them [3].

The personal information of any given individual user that is recorded digitally is what we will define as a **digital footprint**. In some articles, this is also referred to as a *cyber shadow*, *electronic footprint*, or *digital shadow*, but I will synonymize these terms with digital footprint. This

thesis will **not** focus on tracking the digital footprint of corporations, companies, or groups, as multiple commercial tools already exist to help a company measure their digital footprint. However, these tools are largely unhelpful to an individual user for tracking, measuring, or understanding their personal digital footprint. Statistics and information will be largely limited to Americans or North Americans where possible for this thesis.

2 Related Work

Numerous resources exist for directing a user to websites that aid in the process of investigating and measuring your public digital footprint manually. The Centre for the Protection of National Infrastructure in the United Kingdom, for example, produced a simple guide for the common user for discovering and managing their digital footprint [4]. In the guide, they reference multiple websites that are free to use for discovering the scope of the user's digital footprint, such as *Pipl* for personal information, *Whois.com* for information on website owners, and *TinEye* for image searching. Pipl is a search engine that uses deep-diving algorithms to find information that is publicly available and associated with specific names. The searches on Pipl can be refined with email addresses, phone numbers, or locations. Whois.com is a website for searching domains and finding information regarding the websites' owners. TinEye is a reverse-image search engine that takes a digital photo as input and searches the web for it using computer vision, pattern recognition, neural networks, and machine learning [5]. None of the aforementioned resources, however, automatically track data of an individual user of the internet. They are strictly manual searches, requiring names or images to be searched manually.

A user's internet browser plays a large role in indentifying any given individual user from the millions of other internet users, which consequently can influence the amount of information associated with a user's digital footprint. The technique of analyzing a user's internet browser to identify them is known as **browser fingerprinting**, and can play a substantial role in identifying a user. *Panopticlick*, a research project by the Electrionic Frontier Foundation, is one tool available for free online to help a user determine the uniqueness of their browser fingerprint [6]. This technology can help the tool proposed in this thesis further generate an in-depth and personal dossier of a user's browsing habits.

To discuss or investigate measuring a user's digital footprint without discussing **trackware** or **tracking cookies** would be like discussing identifying cars without mentioning license plates. Trackware is, in broad terms, any piece of software that tracks system or user activity. Tracking cookies are a type of trackware deployed by websites to uniquely identify a visitor. One open-source tracking cookie is *evercookie*, an aggressively persistent cookie that restores deleted cookie data everywhere if any piece of the cookie remains on the user's system [7]. Paired with previously mentioned tools, tracking cookies can be created and implemented to further increase the effectiveness of the proposed tool.

3 Method of Approach

The first step of the proposed thesis is to identify free and open source software used in identifying an individual user. This would include software such as the evercookie, a cookie that is excessively difficult to remove from your system [7]. Evercookie, paired with a service like

Panopticlick [6], can go a long way towards identifying the unique user. Once a method for identifying any given user is established, the proposed tool will be able to consistently pair a user to the user's browsing session. These technologies, when used together, can theoretically create an extensive profile of a user, tracking things such as browsing history, video viewing habits, physical location within the range of a city and state, and other personal information. The only thing lacking in this regard, then, is associating this information with a legal name, face, email address, or home address.

To acquire more personal information, the strategy the tool will employ is to create cookies for websites that will try to acquire the user's email address. The tool will continue to monitor and track the user's activity over the course of 30 days, or until it has acquired an email address of the user. This can be done by reading the headers of a website, but may be difficult to do. Another option is to try to capture packets of information containing other websites' cookies and decrypt/dehash them, but there are two issues with this. First, cookies are time-sensitive so this may not be feasible. Second, this is not legal and therefore not a valid solution. If an email address is successfully acquired by the tool, though, it can use an open-source API of the Pipl engine to search usernames and real names associated with the email address [**?**]. After using the Pipl search engine on my own email addresses, I was able to yield my full legal name, current home address, previous home addresses, former telephone numbers, and even links to my social media accounts.

Once the information has been gathered by the tool, it will need to parse the data. Using an XML file, the tool will organize all of the data it has collected. Then, using a visualization tool created in Java or Python, the data will be displayed in a format that is easy to comprehend for the average user. It will use response feedback from the user to verify the pieces of information that were correctly gathered by the tool, and discard the pieces of information that are inaccurate. This process can be repeated, and it will be the responsibility of the user to see if this information is accurate. Whenever the user decides they have used the tool for long enough, the tool will then try to help the user identify ways to improve their own personal security. For this, I plan on researching already existing tools that increase a user's security online, such as browser plugins, alternative browsers, virtual proxy networks (VPNs), and settings recommendations for existing popular browsers.

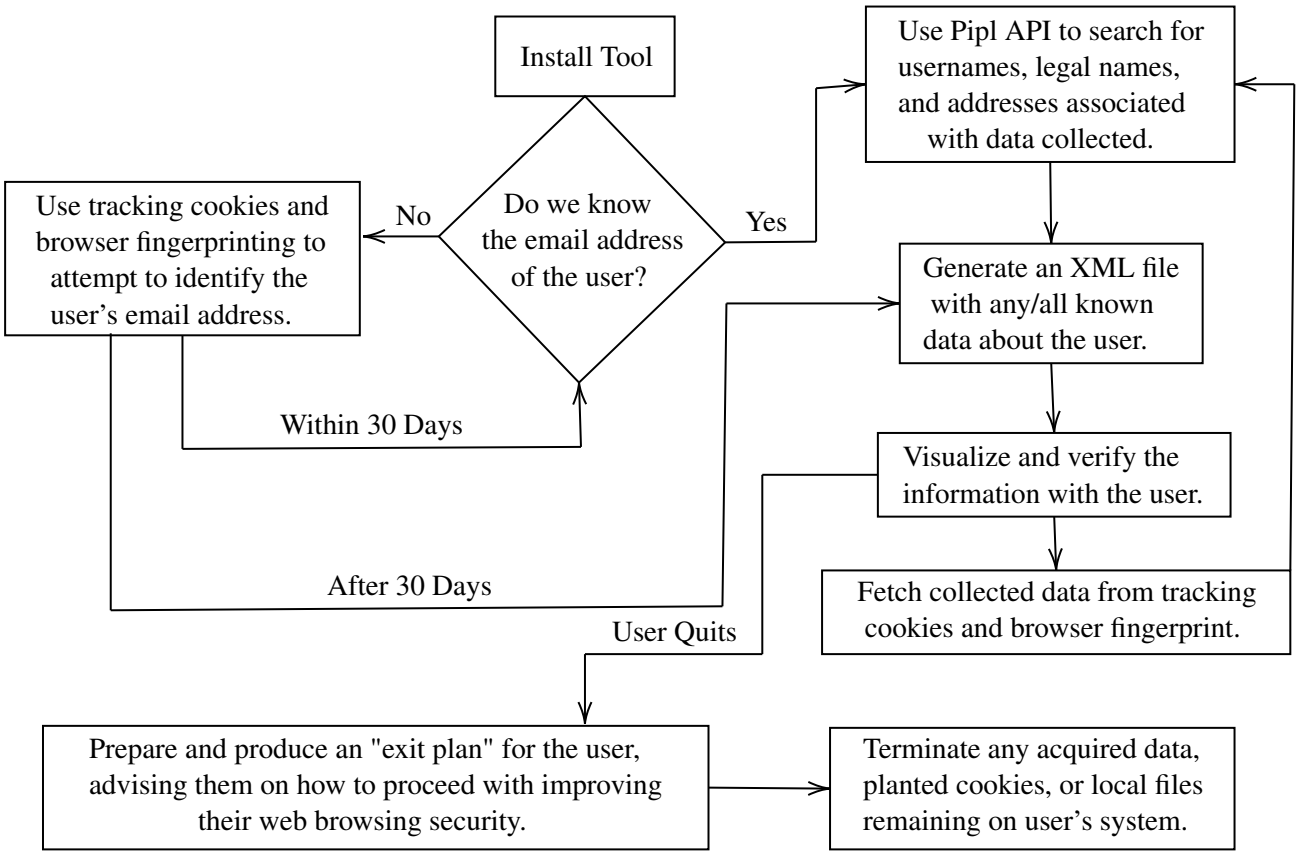

Figure 1. The Tool Model.

4 Evaluation Strategy

The tool will be largely evaluated by the user of the software, and will ideally use feedback from users to improve the accuracy and usefulness of results of the scale up to and through final deployment. User curation and feedback will help me to know what kind of information is useful in the future. If it can be done within the scope of the project, a future goal for the project would be to implement some level of machine learning to the tool for learning how a user will behave online. This can be done with a neural network, linear regression, or some other machine learning technique. In the meanwhile, however, simple aggregation and organization algorithms will suffice for this thesis.

Testing and evaluation performed during the development stages of the software will be difficult. As indicated in Figure 1, the tool will ideally use a 30-day buffer period to attempt to acquire the user's email address. If the tool cannot acquire the email address, I speculate that it will struggle to acquire the bulk of the information. This is because things like a user's real name and address are inherently difficult to obtain, as that information is only either freely given by the user (if we are lucky), or withheld by the user's Internet Service Provider (ISP). Tracking cookies and browser fingerprinting can still colelct an impressive amount of data on an individual user, but without the email address, particularly personal or shocking information may be hard to come by.

For this reason, a large period of time should be given to the tool to have the opportunity to find this information, with regular refreshes to check whether or not the email has been found. This creates an interesting difficulty within testing the software, as trial runs of the program would take literal months to complete on any given machine. To circumvent this, I would attempt to deploy the software for testing to a research or trial group of volunteers. Many previous theses have struggled when it came to the aspect of getting and relying on volunteers, and for a comprehensive project that deals with data as sensitive as this, it may be more difficult.

There is hope, though, in that the email field is not necessarily the most important part of the tool. Therefore, I can circumvent the need for months to execute the project altogether (at least for the majority of the development process) by simply providing the tool with an email address from the start. Users who volunteer to use my tool can also do the same, truncating the time required to complete the tests by magnitudes of weeks.

5 Conclusion

I believe that the impact of the proposed research has ramifications that supercede the field of Computer Science. In the contemporary digital age, users of the internet worldwide are more connected than ever. We, as a society, largely prioritize building a digital life online for others to see. Our day to day interactions with the internet drive is together, but also can drive us apart. The average user leaks data from their fingertips onto their internet selves, and it is this data hemorrhaging that the average user does that has enabled the rise of web brigades, popularly referred to as Russian "troll farms," designed to use easily accessible online data on users to spread misinformation and propaganda. The collection and sale of a user's data is also in the American public eye with regards to Facebook CEO Mark Zuckerberg's Congressional hearings with regards to a massive data scrape performed by a third party company that violated the privacy of millions of users. In the coming years, both the private sector and governments will be making huge changes to policies and regulations with regards to personal information privacy. This tool has the potential to both inform and frighten average users into understanding the importance of carefully curating their online exposure and how much data they reveal about themselves online.

The proposed tool will be on the forefront of this monumental wave of changes to user data usage, privacy, and online profiling of users. It will exist in a time when almost every American is aware of digital security concerns associated with using the internet and social media.

References

[1] H. Lebo, "The Digital Future Report," University of Southern California, Tech. Rep., 2017. [Online]. Available: http://www.digitalcenter.org/wp-content/uploads/2013/10/2017-Digital-Future-Report.pdf

[2] "Internet World Stats," Copyright © 2017, Miniwatts Marketing Group. All rights reserved worldwide. [Online]. Available: http://www.internetworldstats.com/

[3] M. Madden and L. Rainey, "Americans' Attitudes About Privacy, Security and Surveillance," Pew Research Center, 2015. [Online]. Available: http://www.pewinternet.org/2015/05/20/americans-attitudes-about-privacy-security-and-surveillance/

[4] "Tracking my Digital Footprint: A guide to digital footprint discovery and management," © Crown Copyright 2015, 2015. [Online]. Available: https://www.cpni.gov.uk/system/files/documents/59/06/10_Tracking%20my%20digital%20footprint_FINAL.pdf

[5] "Tineye." [Online]. Available: https://www.tineye.com/

[6] "Panopticlick 3.0." [Online]. Available: https://panopticlick.eff.org/about

[7] S. Kamkar. (2010, September) evercookie. [Online]. Available: https://samy.pl/evercookie/