ETHICS, PRIVACY, AND DATA COLLECTION: A COMPLEX INTERSECTION

by

Matthew S. Brown

A Proposal Submitted to the Honors Council

For Honors in Computer Science

September 20, 2019

Approved by:

Adviser and Department Chair:

_____

L. Felipe Perrone

Second Evaluator in major

_____

Darakhshan J. Mir

Introduction

The technology around us enables amazing things like high-resolution video calls and the ability to stay connected with everyone we care about through social media. Yet, it typically comes with an unseen cost. The computers we carry with us in our pockets, backpacks, and on our wrists is constantly receiving and, more importantly, sending data to many more places than most people know. In order to make their services free, many online companies sell data or access to targeted demographics as a means of making their profit.

On its surface, this may not seem so bad. Who cares if Facebook knows what websites I go to? Why does it matter if Uber knows where I am all the time? These questions are not too far from the ones I have heard in everyday conversation. The issue comes down to what these companies do with the data they collect and who they share it with. When you take time to fully consider the consequences of one app or website having your name, birthday, a record of your location, and your payment information, it becomes easier to recognize the hazards of this broad data collection.

A trend in computer science and examples could be drawn for any field, in recent history has been the invention of new technology and processes with limited consideration for the consequences or long term negative effects. This is especially relevant to user privacy and data collection. Many current college students can probably recall growing up hearing adults say ”Don’t use your real name online” and ”Don’t tell anyone online where you live or any personal information”. Yet, we now live in a world where people’s entire lives exist in their Facebook and Instagram profiles and massive amounts of personal information are available publicly. The issue

is that the average user either does not know enough to care or knows but still does not see the issue as an immediate problem.

This shift has made me curious about people's relationship with their data. Specifically, I am exploring what data is collected online by different websites and applications, what do users know is being collected, how do they feel about the data being collected, and what ethical considerations should be made by the companies collecting the data.

Motivation

The potential for a significant impact on people's lives and the world as a whole through the amount of data created by and about every individual should concern everyone. For example, there is still controversy around whether the work of Cambridge Analytica on President Trump's campaign actually impacted the outcome of the election. However, the fact that this issue is being discussed seriously and it is conceivable that a company could produce extremely effective, targeted messages that have the potential to alter an individual's perception of a person or event should concern everyone.

My thesis has two primary goals. The first is to bring attention to this issue in the computing community to encourage engineers to consider the implication of what they are making and think critically about the positive and negative outcomes of their work. Secondly, I hope to motivate people to be more conscious of their online privacy and question why a website is asking for certain information instead of blindly giving permission.

Previous Work

The concept of this thesis came about from a collection of sources, including readings of scholarly literature, previous coursework, internship experience, and general observations of my peers.

In my final paper for CSCI 245 Life, Computers, and Everything, a course on computer ethics, I explored the ethics of data collection by social media companies in connection with the profit made on collected user data. These companies treat information on users like commodities and package it as a product valued for laser beam targeted marketing. Through my research for this paper, it became evident that most users are unaware of the extent and the amount of data that is collected about them. This concept is the focus of Zeadally and Winkler (2016), who studied the readability of terms of service agreements for popular websites, ultimately discovering that the terms of service agreements that they studied were written at a high school level, but half of all Americans do not read above an eighth grade level. This class inspired further interest in who collects and controls user data and also how users interact with computers when it comes to their data.

My internships the last two summers provided me with learning experiences and information on the topic of data collection and personal attitudes toward privacy that are very relevant to this thesis. At the Federal Housing Finance Agency in 2018, all employees were brought to a lecture on security best practices where I learned key points of information security and again saw cases where highly educated people were previously unaware of how important strong security practices were to protect the information to which they had access. At ASML in

2019, I attended a company-wide meeting where it was projected that, by the end of 2020, there would be 44 zettabytes of data in the world. (One zettabyte is one trillion gigabytes of data.) Granting that not all of this data will be personal data from individual users, there is still potential that this data can be used in hazardous ways.

Finally, interacting with my peers has led me to conclude that many students do not consider the risks of how much data is collected or are ambivalent to issues regarding online security and privacy. I frequently hear other students complain about the password requirements to log into myBucknell or having to grab their phone to physically authenticate a login attempt. To aid in the second goal of this thesis, finding a way to describe the benefits and importance of certain practices with regard to security and privacy can lead to broader acceptance or, at the very least, adoption of better habits.

Project Description

To achieve these goals I intend to approach this problem three different ways. Initially, I will start with a review of current literature and work. I will include both a collection of scholarly articles and journalistic narratives from reputable news sources. The scholarly articles will provide greater insight and thought into the issues, but the news articles are frequently more up to date on recent events. Secondly, I will conduct a survey of Bucknell students and faculty to gain a clear perspective on the attitudes and approaches that are taken when it comes to data privacy.

Lastly, I will conduct conversations with experts in the field to gain greater insight, similar to scholarly articles. Hopefully the discussion in these conversations can prompt new

insight that is not contained in popular scholarly works. Additionally, I will expand on previous learning in the field of applied ethics and consider some of the real world situations that many people experience in relation to their privacy through philosophical lenses such as Kantianism and Utilitarianism.

Methods

There is a significant body of work on data privacy from academic, industry, and governmental sources. I plan to study the literature from these different kinds of sources and consolidate perspectives into a document that presents the multiple facets of the discussion on how private data is protected and exploited. There is an abundance of literature condemning the broad collection of user data, but I am also interested in the arguments that support data collection and monetization. While my personal views would be to limit the collection of user data, it would be irresponsible to ignore the arguments against my position. Zeng et al. (2010) discuss many positive uses for data collected from social media sites, such as gauging public opinion of political issues and monitoring public health issues. Cecchinel et al. (2014) are another example, proposing a structure for the storage and manipulation of large amounts of data created by various devices. Kaiser (2016) conducted a study similar to the one I am proposing, making my proposal even more significant because being able to compare the two data sets will add more to the discourse on the topic. Furini and Tamanini (2015) similarly study the opinions and behaviors of users specifically around location data and its use. Xu et al. (2014), Lucas and Borisov (2008), and Rosenblum (2007) all focus on the implications to the individual privacy of

users due to online data collection. Lucas and Borisov go as far as implementing a tool to protect some data and secure it before it is shared.

I will conduct a survey of Bucknell students and faculty, based on IRB approval, in order to gain a better understanding of the awareness of how their data is collected and used, in addition to gauging opinion and adoption of various security practices, such as the Duo authenticator. I recognize the irony of wanting to collect data in a thesis criticizing data collection. However, the issue is not so much the amount of data collected, but the transparency of what data is collected and how it is used. With this in mind, the survey responses will be completely anonymous and a clear, non-technical description of the purpose would accompany the request. With the goal of removing my bias from the survey, I will consult with Bucknell faculty, such as Prof. Darakhshan Mir in Computer Science, and specifically members of the IRB, to design a survey instrument that does not skew responses in any particular direction.

Similarly, I will reach out to faculty and researchers, at Bucknell and beyond, to conduct interviews and gain expert perspectives on the issue. In particular, I will reach out to Prof. James Mickens at Harvard. Prof. Mickens primarily researches security issues. In 2018, he gave the keynote speech at the USENIX Security Conference primarily emphasizing the importance of computer scientists questioning whether or not they should make something even if they can. Another significant contributor in the field is Seth Stephens-Davidowitz. He is the author of *Everybody Lies*, a book that focuses on trends he has been able to find from scraping large amounts of data from different web sources. Additionally, Stephens-Davidowitz was formerly a data scientist at Google and he has written several op-ed pieces for the New York Times. I will leverage the work he has done to bolster my argument with the amount of data he has accessed

and reach out to him with any questions not answered by his book or other writings. Interestingly, Stephens-Davidowitz makes all of his data sets public on his website. I will download these data sets and investigate the potential hazards in the data. An immediate question arises though, whether he receives the data stripped of any identifying information, or if he santizes it before publishing it on his website.

With regard to timing, my plan is to have all of my data collected and conversations with individuals outside of Bucknell completed leading into Thanksgiving Break. Simultaneously, through the course of this semester I plan to complete my review of relevant literature. By the end of the semester, I plan to have completed all of the literature review and statistical analysis. Over the winter break I plan to outline a focused structure of the final document. Going into the spring semester, I plan to take CSCI 376 which would dedicate time in my schedule to work on writing my thesis. This would aid in my goal of completing about eighty percent of the writing by March 1, with the full draft completed going into spring break. This leaves the rest of March to revise with some flex time.

Conclusion

This research will contribute to the ongoing study of online privacy and provide greater data to the academic community exploring the relationship that users have with their data online. Additionally, in response to the conversations I have had with my peers, this research will better inform more productive and safer attitudes toward the online exposure of personal data that should be respected as private.

References

Cecchinel, Cyril, Matthieu Jimenez, Sebastien Mosser, and Michel Riveill. 2014. "An Architecture to Support the Collection of Big Data in the Internet of Things." *2014 IEEE World Congress on Services:* 442–49. https://doi.org/10.1109/services.2014.83.

Furini, Marco, and Valentina Tamanini. 2015. "Location Privacy and Public Metadata in Social Media Platforms: Attitudes, Behaviors and Opinions." *Multimedia Tools and Applications* 74 (21): 9795–9825. https://doi.org/10.1007/s11042-014-2151-7.

Kaiser, Alena Fiona. 2016. "Privacy and Security Perceptions between Different Age Groups While Searching Online."

Lucas, Matthew M., and Nikita Borisov. 2008. "FlyByNight: Mitigating the Privacy Risks of Social Networking," 122.

Rosenblum, David. 2007. "What Anyone Can Know: The Privacy Risks of Social Networking Sites." https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=4218550.

Winkler, Stephanie, and Sherali Zeadally. 2016. "Privacy Policy Analysis of Popular Web Platforms." *IEEE Technology and Society Magazine* 35 (2): 75–85. https://doi.org/10.1109/MTS.2016.2554419.

Xu, Lei, Chunxiao Jiang, Jian Wang, Jian Yuan, and Yong Ren. 2014. "Information Security in Big Data: Privacy and Data Mining." *IEEE Access* 2: 1151–78. https://doi.org/10.1109/ACCESS.2014.2362522.

Zeng, Daniel, Hsinchun Chen, Robert Lusch, and Li Shu-Hsing. 2010. “Social Media Analytics and Intelligence.” *IEEE Intelligent Systems*, 2010. https://ieeexplore.ieee.org/abstract/document/5678581.