Spring 2020

Ethics, Privacy and Data Collection: A Complex Intersection

Matthew S. Brown
Bucknell University, msb027@bucknell.edu

Ethics, Privacy and Data Collection: A Complex Intersection
by

Matthew S. Brown

A Thesis

Presented to the Faculty of

Bucknell University

in Partial Fulfillment of the Requirements for the Degree of

Bachelor of Science with Honors in Computer Science

May 11, 2020

Approved: _____

Professor L. Felipe Perrone
Thesis Advisor
Chair, Department of Computer Science

_____

Professor Darakhshan Mir
Second Reader

Acknowledgements

I would be remiss in not thanking the many people who, without whom, I would not have been able to complete this work.

First, my advisor throughout this process, Professor L. Felipe Perrone. Thank you for recognizing my potential and encouraging me to take on the challenge of writing an undergraduate thesis. I enjoyed every minute of our weekly meetings and the many conversations we had regarding my topic. I learn something from you every time we speak. Without your support, I would not have been able to complete this thesis.

I would also like to acknowledge my Thesis Committee, Professor Jim Maneval, and Professor Darakhshan J. Mir. Thank you for the time you dedicated to reading and supporting my work. I appreciate your commitment to undergraduate research and scholarship.

Thank you to the Bucknell community for participating in my research and for inspiring me to explore the issues around internet privacy.

To my friends and family, thank you for supporting me throughout my education. Special thanks to my parents, for encouraging me to take on new challenges and explore the issues that interest me. Thank you to my brother and sister, Adam and Abigail, for being constant examples of hard work and for inspiring me to push myself to the same standard.

Finally, I want to thank Kat Swank. Thank you for being a constant source of support, and always making time to give me feedback on an idea or my writing when I needed help.

Contents

Chapter 1

Introduction

The technology around us enables incredible abilities such as 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 are constantly receiving and, more importantly, sending data to many more places than is apparent (Valentino-DeVries et al. 2018). In order to make their services free, many online companies sell data or access to targeted demographics in the form of advertising 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 with peers, professors, friends, and family. 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 does not understand how their data is used.

This shift has made me curious about people's relationship with their data. In this work, I explore 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.

1.1 Motivation

The potential risk to people's privacy and security due to data collection should be concerning for everyone. My thesis has two primary goals. The first is to bring attention to this issue in the computing community to encourage software developers to 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 web service is asking for certain information instead of blindly giving permission.

1.2 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 intensely 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 Winkler and Zeadally (2016), who studied the readability of terms of service agreements for popular social media platforms. Ultimately, Winkler and Zeadally (2016) discovered 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 my further interest in who collects and controls user data and also how users interact with computers when it comes to their data.

My internships in the last two summers provided me with learning experiences and information on the topic of data collection and personal attitudes toward privacy that are 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. A year later in 2019, at ASML, a world leader in optical lithography, 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.) Considering that a large portion of this data will be personally identifiable information, 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 use two-factor authenticatoin to confirm a login attempt. To aid in the second goal of this thesis, I aim to describe the benefits and importance of certain practices with regard to security and privacy to encourage better habits.

1.3 Background

This background section serves to provide an intuition for how common privacy issues are. Additionally, I will comment on and provide a general explanation of Bayesian Inference as a method of extracting meaningful information from user data. This section also includes a primer on ethical theories used later in this thesis to examine the moral dilemmas presented by social media data collection. I argue that knowledge and consideration of these theories can benefit both the creators and users of online technologies.

1.3.1 Bayesian Inference

As seen later in this thesis, most users do not understand how their seemingly inconsequential actions each time they use a social media platform are recorded and used to inform advertising decisions made by the platform. To address this lack of un-

derstanding, this section describes one mechanism by which user behavior and data can be used to inform targeted advertising. Tanaka et al. (2016) demonstrate the power of monitoring user actions by creating a model that considers the purchase history and media advertisement views for a user and can produce valuable information for advertisers. Tanaka et al. (2016) use an efficient Bayesian inference approach to model likely purchase behavior of the users. In this section, I introduce a common and powerful statistical tool that indicates the likelihood that a certain prediction is true, given that another event or series of events have occurred.

Let's start with an intuitive example. Say we are a social media company trying to serve the most relevant ads to consumers to make them more likely to click them. We have a user who we know recently booked a hotel in a tropical location, bought plane tickets to that location, and bought sunscreen online. If we take all of these events into account, we might conclude that the user can be convinced to purchase a new swim suit. Similarly, if we have a user who is visiting extremist websites, looking for instructions on making explosives, and following extremist accounts on our website, then it is reasonable to expect the person has the potential to commit an act of terror.

I introduce here the foundational framework for Bayesian Inference. This is a simpler version than what was used by Tanaka et al. (2016), but this form is enough to understand the work. The formula below describes the general form of Bayesian Inference.

$$P(H|E) = \frac{P(E|H) \cdot P(H)}{P(E)} \tag{1.1}$$

In (1.1), the variable H takes the place of the hypothesis we are interested in while E is the event we have or are expecting to observe. The value we are interested in, $P(H|E)$ is the probability that H will occur, given we observe E. $P(H)$ is simply the

likelihood of H generally being true. $P(E)$ is the probability that we will observe the specific event E. For instance, if $P(H)$ is the probability that there will be a snow day, and $P(E)$ is the probability that it will snow. We assume $P(E|H)$, which is the probability that it is snowing given we have a snow day. From these values, we can calculate $P(H|E)$ which indicates the probability of having a snow day given it is snowing.

1.3.2 Primer on Ethical Philosophy

This section draws heavily on reading from *The Elements of Moral Philosophy* by James Rachels (7th ed.). The ethical theories presented here are used later to examine dilemmas at the intersection of data collection and privacy. These descriptions of several philosophies are not meant to be exhaustive, rather simply to provide a basic understanding which will serve the reader well later in this thesis. Further reading in Rachels (2012) or other sources is necessary to fully understand these ideas.

Kant's Moral Philosophy

Many of Kant's ideas focus on resolving moral dilemmas through the laws of "duty", respect for people, and what he calls the Categorical Imperative. According to Rachels (2012), Kant presents the Categorical Imperative in a few different ways; one of the formulations is as follows:

> "Act so that you treat humanity, whether in your own person or in that of another, always as an end and never as a means only."

Rachels (2012) explains that this idea of treating people "as an end and never as

a means" is essentially saying that we need to treat people with respect and never manipulate other people in order to achieve our goals. Rachels (2012) continues that treating people as an end extends to acting in ways to promote their well-being and avoid harm.

Another formulation of Kant's Categorical Imperative is: (Rachels 2012)

> "Act only according to that maxim by which you can at the same time will that it should become a universal law."

This establishes a tool to guide one's moral compass in the resolution of moral dilemmas. (Rachels 2012) When deciding how to act, one can use this to figure out what the "maxim", or rule, they would be following by taking an action and then considering whether they would be comfortable with that rule being followed universally.

In one example Kant provides on applying this tool, he suggests that, perhaps,"I refuse to give help to others thinking that I do not care and that each person can fend for their self." We cannot will this to be a universal rule because in times where we need help from others, we would not want them to turn us away.

Utilitarianism

Utilitarianism is a consequentionalist ethical theory which argues that we must act in a way to increase the net "happiness" in the world. Rachels (2012) describes the three main propositions of Classical Utilitarianism as follows:

> a) The morality of an action depends solely on the **consequences** of the action; nothing else matters.

b) An action's **consequences** matter only insofar as they involve the greater or lesser happiness of individuals.

c) In the assessment of **consequences**, each individual's happiness gets "equal consideration." This means that equal amounts of happiness always count equally; nobody's well-being matters more just because, for instance, one is rich, or powerful, or good-looking. Morally, everyone counts the same.

This description of Utilitarianism leads people to think that each particular action needs to be judged for the effect it has on total happiness. Some argue, however, that instead we should judge actions based on whether they abide by a set of rules designed to optimize happiness. It is up to us to determine what this optimal set of rules is. This new form of Utilitarianism is called Rule Utilitarianism while the original theory is generally referred to as Act Utilitarianism.

1.3.3 Recent Major Privacy Blunders

This section highlights two major privacy blunders from the last few years that add legitimacy to why privacy needs to be continuously studied and audited in order to best protect users.

Facebook and Cambridge Analytica

The controversy around Cambridge Analytica from a privacy perspective is less discussed than the debate around whether the firm managed to sway the tide of the 2016 United States Presidential Election. When collecting the data from Facebook,

Cambridge Analytica worked with an academic, Aleksandr Kogan, running a survey that disguised as a personality test (Cadwalladr and Graham-Harrison 2018). The users who took the personality test agreed to participate in academic research, however, what was not made clear is that the information of these users' friends was also collected (Cadwalladr and Graham-Harrison 2018). This meant that users did not have to personally give consent if one of their friends did. While Facebook did not actually collect the data, the company should still be responsible for how external companies can access user information and what kinds of consent are necessary on the user level. In this case especially, where the political direction of a country may have been influenced the burden is significantly higher.

Zoom

When nearly every private and public school from kindergarten through 12^{th} grade and higher education switched to a remote learning method as well as many people working remotely in Spring 2020 in response to the COVID-19 pandemic, many chose Zoom as their solution to host video lectures and online classrooms. This led to Zoom seeing 200 million daily users in March 2020 compared to just 10 million in December 2019 (Bond 2020). While the article from NPR spends a fair amount of time discussing the issue of 'Zoombombing', what is more interesting for my work is a sentence at the end pointing to an article. The article references a report where Zoom was sharing data with Facebook, even on Zoom users who did not have Facebook accounts (Bond 2020). Zoom acknowledged this report claiming that this functionality was a mistake, but it begs the question of how this even happened and makes the oversight on Zoom's part extremely evident.

Chapter 2

Literature Review

This chapter highlights existing work on this topic, comparing opposing views, both in support of data collection and in support of user privacy. In general, the literature I found that supports data collection demonstrates that data collected from online users can be used for justifiably beneficial purposes. Academics work toward improving the world and advancing technology, and so are less concerned with the profit earning side of data collection that is a concern for social media companies. I have chosen to explore current and recent events for content of this literature review as well to broaden the range of available resources and discussion. To that end, I am including references from news articles, doing my best to share only facts to accurately represent the situations at hand.

2.1 Understanding Privacy

It is important to ground the meaning of privacy in literature. Nissenbaum (2015) argues that concern over privacy should not be focused on the sheer amount of data used

but how appropriate the use is, calling this *contextual integrity.* citeMartin2016MeasuringVariables elaborate that the advantage to defining privacy as contextual integrity creates a distinction between loss of privacy and loss of information. Thus, the flow of information itself is not seen as harmful as long as it fits the context. For example, the sharing of medical information with a physician, or financial information with the Internal Revenue Service (Martin and Nissenbaum 2016). However, when the flow of information is no longer appropriate for the context, then this is a loss of privacy (Martin and Nissenbaum 2016). Referring to the previous example, if a healthcare provider is then selling the shared information without the individual's permission then we see a privacy violation.

While this thesis focuses mostly on how the individual interacts with privacy, it is important to note the work of Boyd (2012) on the idea of "networked privacy". Networked privacy addresses how the actions of an individual or group affect people beyond that individual or group (Boyd 2012). Boyd (2012) gives the example of the popular DNA testing company *23andMe*. When an individual submits their DNA sample to be sequenced and studied, they are also providing information on anyone related to them and any descendants they may have. This can all be done without obtaining permission from all of these other people. Networked privacy is one key aspect of the privacy failure discussed earlier in Section 1.3.3. Boyd (2012) argues that privacy models need to be developed that put groups and communities at the center of the discussion instead of the individual.

2.2 In Support of Data Collection

The work of Sinha et al. (2016) demonstrates the value of social media companies making user data accessible. Sinha et al. (2016) developed a natural language pro-

cessing (NLP) system to process posts on Facebook and other social media platforms. They used this information to identify user posts regarding issues on the public transit system in Bangalore. The expected application of this research would be to more rapidly identify and correct issues with the public transit system. While this work was in very early stages at the time of publishing, the potential to improve a city's awareness of issues with the transit system and accelerate the response is incredible. This work is a strong advocate for continued access to user data on social media platforms.

Similar to public transit, public health efforts can be supported by data collection. With the COVID-19 pandemic currently wreaking havoc around the world, data is being collected in a number of ways to try and monitor the spread of the disease. One extremely surprising development is a partnership announced between Apple and Google to develop a set of application programming interfaces (APIs) that developers can include in their applications for either platform Newsroom (2020). At its core, this partnership is creating technologies that are meant to help track the spread of COVID-19. This includes using the Bluetooth radios in smartphones to monitor contact between individuals and alert users if they have been in contact with someone who is later confirmed to have the disease after the user reports the diagnoses through an application linked to their medical provider (Greenberg 2020). Apple's press release emphasizes that "Privacy, transparency, and consent are of utmost importance in this effort..." indicating that this would be purely opt in and users would be fully aware of their participation (Newsroom 2020). The advantage of the design of this system is that it uses Bluetooth to connect two phones and leave an anonymous token behind that would then be compared to a database of tokens of people who have reported receiving a positive diagnosis of viral infection. The people who have been near the person with the virus would then receive a notification informing them that they may

have been exposed. The advantage of Bluetooth is that it eliminates the need for GPS tracking and broadcasting the location of every user, instead just relying on the individual device to communicate with the nearby devices and swap tokens, reducing privacy concerns. It should also be noted that there are legitimate situations where data can be collected ethically and the context determines the appropriate flow of information, as discussed in Section 2.1

Based on contextual integrity as discussed by Nissenbaum (2015), contact tracing seems appropriate assuming it is done safely. However, Selinger and Leong (2020) raise an important question as to whether contact tracing is medically necessary to protecting public health during the pandemic. The point being that if it is not medically necessary, then the risks are too great (Selinger and Leong 2020). Additionally, if it is not necessary then the context is no longer appropriate from a contextual integrity persepctive.

Another example of data collection to monitor the pandemic came from Ghost Data, a research group in Italy and the United States that collected location and post information from over 500,000 public Instagram accounts in March 2020 (Alfred Ng 2020). The purpose of the data collection was to provide data to the Italian government as to whether citizens were obeying the quarantine order issued by the Prime Minister on March 9, 2020. Facebook, Instagram's parent company, responded that the act of scraping user profiles violates Facebook's policies and that they are investigating (Alfred Ng 2020).

2.3 In Support of User Privacy

The literature that acts in support of user privacy typically tries to explain why users might feel their privacy is being limited, or what problems exist that allow so much data to be collected. Herder and Zhang (2019) discuss how advertising on social media, specifically Facebook, frequently appears creepy to users. They identify that "the key to creepiness appears to be the uncertainty about possible threats and the uneasiness due to a lack of social norms" (Herder and Zhang 2019). Following their study and discussion with users, they identify the key point that "explanations and transparency seem to reduce users' anxiety and increase trust only to a certain extent...". In their study, Herder and Zhang (2019) found that users were dissatisfied with explanations for the advertisements that the user did not deem specific enough. What causes users surprise and discomfort is that the ads seem to indicate that the system has access to information that it should not.

While explanations for advertising seem to placate user fears slightly, the work of Jordan and Rand (2019) says that it should not be surprising that users do not know what is being collected. In this study, the researchers created a fake social media page and monitored how much time the users actually spent reading the Terms of Service agreement and the Privacy Policy. They found that users frequently indicated that they do care about privacy. When they asked the users to self report how much time they spent reading these documents while signing up for a new service, the responses indicated that the average should be about five minutes. In reality, based on the data collected by the fake service, average engagement was about one minute and the median was a mere fourteen seconds. This seems to indicate that while users claim to be concerned about privacy, they value their time spent both reading the policies and understanding to what they are agreeing more than the privacy itself.

One argument is that users value their time more than understanding the ToS, but they may not be offered a meaningful choice. It can be argued that the social and, in some cases, financial cost of **not** using web services is so great that users have no choice but to accept the terms presented to them (Nissenbaum 2015). Referencing Nissenbaum (2009), Susser (2019) argues that privacy is a set of social norms not individual decisions. Individuals can not be made to decide the fate of a "social good" and this decision should be made before the service is presented to the user (Susser 2019).

Focusing on application level security, an experiment on the campus of the University of California Santa Cruz demonstrated the need for companies to consciously develop their software with user privacy in mind. Xue et al. (2016) used machine learning to identify the original location of posts to Yik Yak (a, now defunct, social media where users could post anonymous messages only visible to those within a certain geographic range) with an average error of just 106 meters. Yik Yak was advertised as a fully anonymous platform, offering users a sense of security and privacy, however, Xue et al. (2016) managed to strip away a level of that privacy in pinpointing the location of the posts. While this research was reported to Yik Yak so they could pursue a fix, it still points to an oversight made by the company in the first place.

2.4 Conclusion

Ultimately, either side of this debate make strong points. The examples in support of data collection present situations where having more data is indisputably valuable. This is especially obvious in the case of the COVID-19 pandemic. However, one might argue that the emergency situation might make people more willing to give up their

privacy which may be difficult to recover once the disease is contained. Similarly, the literature supporting user privacy places the burden on companies to provide greater transparency and choice to the user. Additionally, as seen in the work of Xue et al. (2016), the burden is on the developers to provide a high standard of security and privacy.

Chapter 3

A Survey Experiment

My reading of the literature made me question how the people at Bucknell interact with privacy ideas leading me to create a survey. The survey was motivated by attempting to understand the source of the ambivalence regarding online privacy and data collection informally observed within the student population at Bucknell. The survey seeks to put data to the work of Winkler and Zeadally (2016), either confirming the assertion that users do not understand the privacy policies to which they are agreeing, or indicating that the ambivalence comes from another source.

3.1 Survey Method

The survey was conducted with approval by the Institutional Review Board (IRB) via Qualtrics over a period of about two and a half weeks. There were 130 participants, consisting of faculty, staff, and students at Bucknell University. The questions gauge user experience with topics of online privacy and security, then inquire as to the user's typical behavior when interacting with social media websites. The survey consisted

of mostly multiple choice questions where participants rated their agreement with statements. Two questions asked participants to enter, into a text box, all the types of information they thought were collected by Facebook and Google. Participants were given the opportunity to enter a drawing to win one of four gift certificates, each worth $25, to encourage a greater number of participants. The raffle was conducted through Qualtrics as well, and the entries were kept separate from the responses to the actual study. The full survey instrument can be found in the Appendix.

3.2 Expected Results Prior to Survey

Based on the work of Winkler and Zeadally (2016) as well as the work of Kaiser (2016), my assumption is that most users, especially if they do not routinely study privacy and security, are unlikely to have a significant understanding of the amount of information that is collected and the various types. The work of Winkler and Zeadally (2016) indicates that around half of all users likely cannot read at a high enough grade level (based on the Fleisch-Kincaid metric) to fully understand the terms of service or privacy policies of most popular social networks. Kaiser (2016) conducted a similar survey, concluding that younger demographics are more aware of privacy issues and perceive less privacy when searching online than older demographics. This conclusion will be interesting to compare to the results of my survey. I should be able to compare the general attitudes of students to those of faculty and staff. My expectation is that that vast majority of respondents do not read the terms of service or privacy policy in its entirety when signing up for new services.

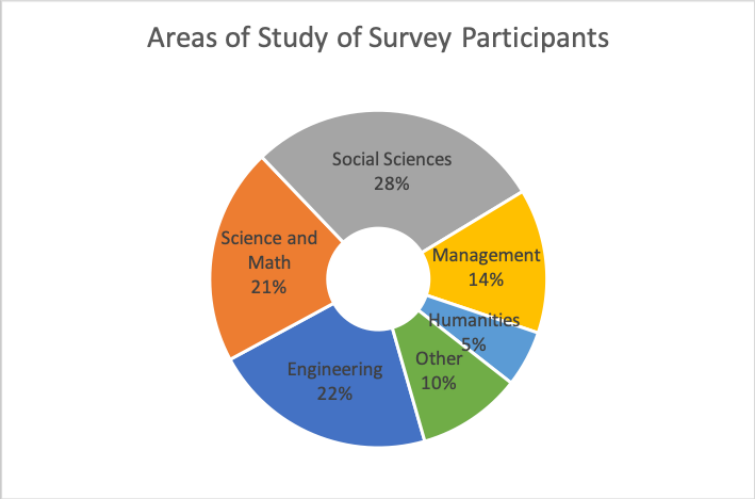
Ideally the responses will be diverse in terms of departmental affiliation and research experience on the topic. This will provide the most useful result especially because computer science and technology focused individuals are probably more likely

to be more educated on the topic.

3.3 Survey Results

The survey gathered 130 participants, primarily coming from Science and Math, Engineering, and Social Sciences. The other category in the chart below includes participants from various non-academic departments, and students who indicated they were undecided for their major. This spread of backgrounds for the participants ensures that the data provides a reasonable image of behaviors and opinions at Bucknell because the results are not dominated by any particular field or discipline. This distribution was not specifically selected, although the survey was advertised through my various social networks on campus which I knew would be diverse in terms of departmental affiliation. I also asked faculty I knew well to distribute the survey within their department.

Figure 3.1



The figure below illustrates the university affiliation of all of the survey participants. The vast majority of survey participants were current students. Six and seven

percent of the participants were staff and faculty respectively.

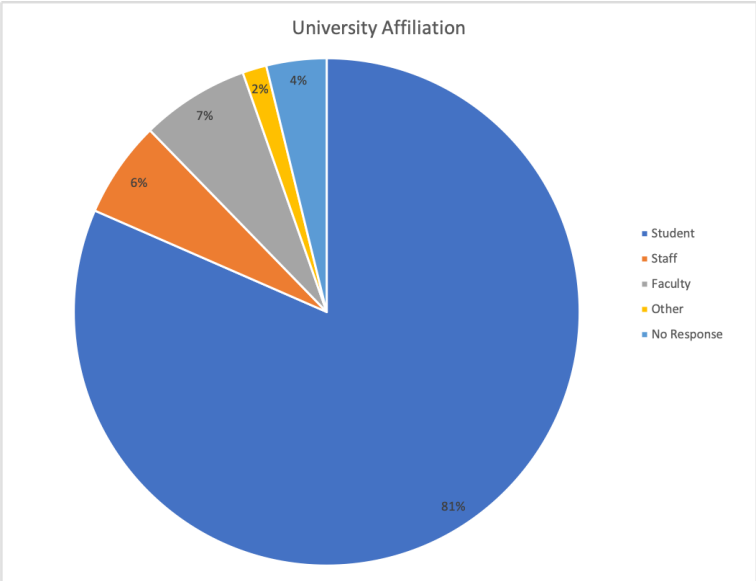


Figure 3.2

3.3.1 Qualitative Results

The survey asked participants to enter in a text box what information they thought Facebook and Google collected on their users. The responses are organized according to the categories identified in Winkler and Zeadally (2016)(see Figure 3.3) In general, most users recognized that these companies collect and keep records of Personal Identifiable Information (PII). Similarly, many users recognized that these platforms collect both location information and on-platform-activity. Some participants correctly identified that Facebook also monitors off-platform-activity as well. Many of the same participants also believed that Google monitors off platform activity, however, according to the work of Winkler and Zeadally (2016), Google's privacy policy does not indicate that off platform activity is monitored. No participants explicitly mentioned the collection of device information with regard to either platform. Three specific participants may have been trying express this idea when they entered that

the platforms collect IP addresses and photo metadata. I reached this conclusion based on the fact that photo metadata typically includes what device was used to take the photo. Device information is more specific with regard to the type of device (laptop/desktop, tablet, phone, etc.) and even browser type, as this information is used by developers to maintain support for different devices. It is interesting to note that while both Facebook and Google collect and store payment information, no participants entered this when asked about Facebook, and very few identified that Google stores payment information.

TABLE 1. Comparing Data Collected According to Privacy Policies

Type of Information	Facebook	Google
Personally identifiable information	Yes	Yes
User-generated content	Yes	No
Device information	Yes	Yes
Location information	Yes	Yes
Payment information	Yes	No
Off-platform activities (other platforms visited and user engagement)	Yes	No
How the user interacts and uses the web platform	Yes	Yes

Figure 3.3: From Winkler and Zeadally (2016), cropped to include only platforms of interest

While many participants failed to correctly identify one or more of the categories used by Winkler and Zeadally (2016), others suggested some types of information are collected that Winkler and Zeadally (2016) do not identify. Some examples include, medical information, or that their phones are constantly recording everything they say to inform ad targeting. Several users also simply said "all of it" while others listed a few specific categories then included "everything" or "all of it" as a catch-all. While these answers are technically incorrect, they are indicative of the perception that some users have regarding the sheer quantity of data collected by these platforms.

This idea is evident in that when asked "How frequently do you feel that ads you see online or on social media are unnervingly relevant? (For example, you see an advertisement for something you were just discussing with another person.),"

with little exception, the responses were overwhelmingly towards the affirmative. This phenomenon is likely why people frequently share anecdotes about how they think their phones are listening to them because they talk about something and then immediately see an advertisement for it.

3.3.2 Quantitative Results

Figure 3.4 contains a significant amount of insight into the behaviors and attitudes of the participants. The survey presented the participants with a series of statements (seen below) and asked them to rank their level of agreement. Of particular interest, are the first three statements on the left side of Figure 3.4. The general opinion is that internet companies collect too much information, but also users generally do not know what information is being collected or how to access the records of their data. While the second and third statements on the right half of Figure 3.4 do not have a significant skew towards agree or disagree, this result still tells an interesting story. Because in both cases, about half of the respondents indicate that they either do not know how to use the privacy features available to them or they do not use these features. It is probably safe to assume (and should be confirmed with further study) that those who do not know what features exist, are also the users who are not using them. There should be a reasonable amount of responsibility put on internet companies to find and fill the gap in user understanding of how to protect their privacy while using each particular platform.[1]

[1]The last statement on the right side of the figure was included to confirm the results of the first statement.

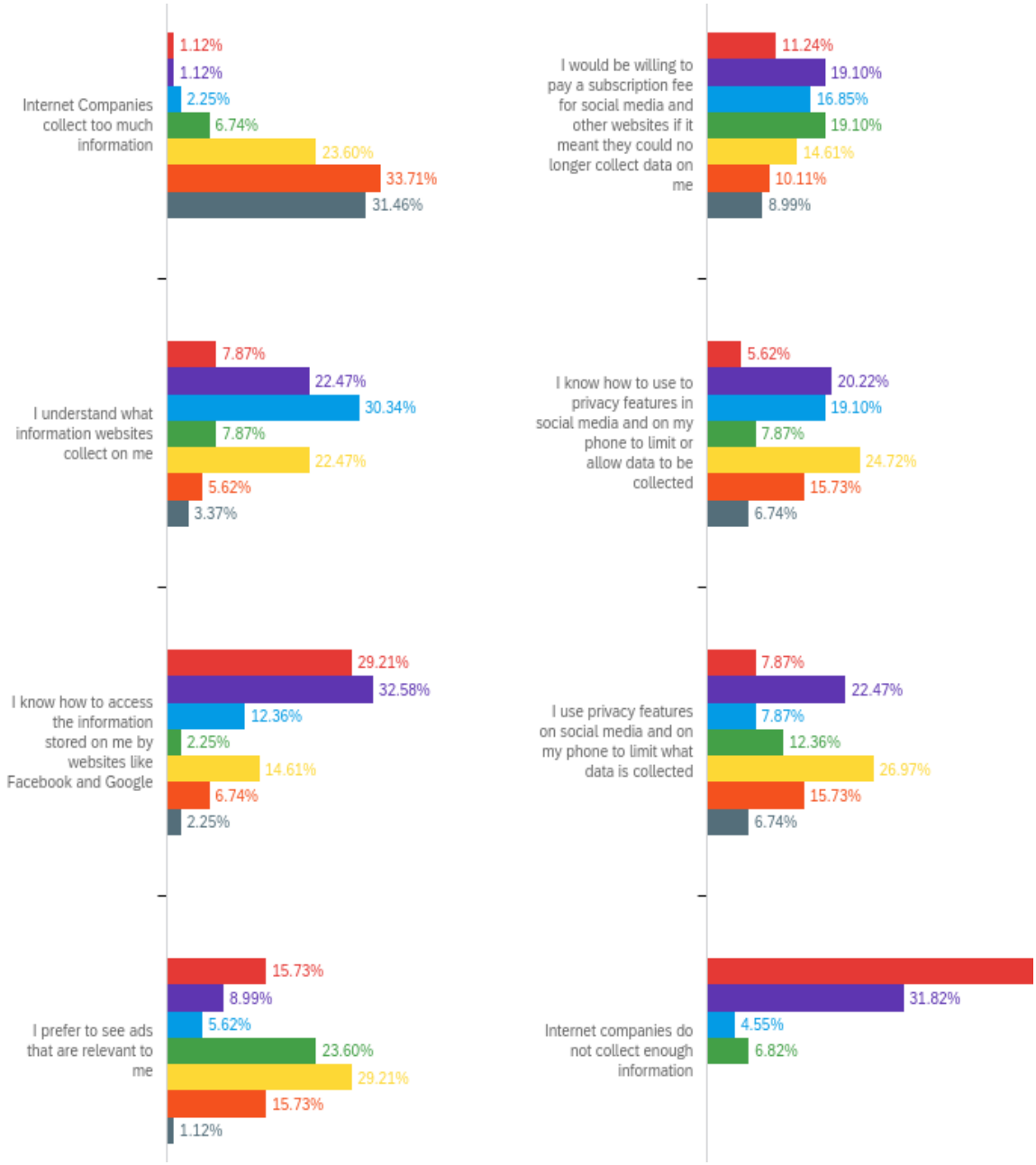


Figure 3.4: Levels of agreement to different statements regarding online behavior and opinions. See Figure 3.5 for legend.

Figure 3.5: Legend for Figure 3.4

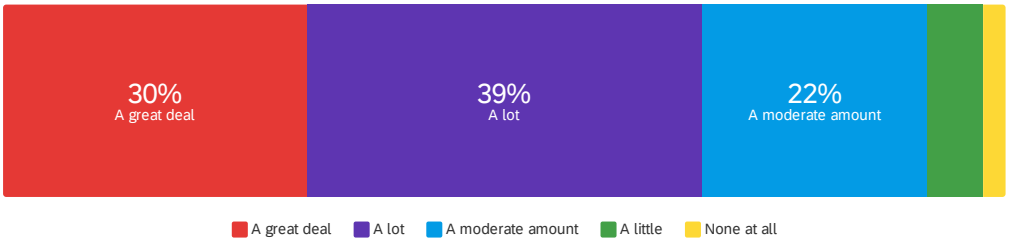


Figure 3.6: Participant reactions to how frequently they feel the ads they see are unnervingly relevant.

Figure 3.6 shows that participants overwhelmingly said that they frequently see advertisements that appear to be unnervingly relevant. With 91% of the participants saying that at least a moderate amount of advertisements they see online are unnervingly relevant. This connects to anecdotal conversations I have had with peers, where they report believing that their phones and other devices are listening to their conversations and that this informs advertising. This survey data adds to the evidence that users do not understand the mechanisms behind targeted advertising and what they can do to protect their information.

Figure 3.7 indicates that most users typically do not read any part of the Terms of Service (ToS) or Privacy policy when registering for a new service. It is interesting

to note that in Figure 3.8, 7% of respondents said that they do not read any of the ToS or Privacy Policy. I find this interesting because participants were only asked how much of the ToS or Privacy Policy they read if they indicated that they read these documents with any frequency in the previous question.

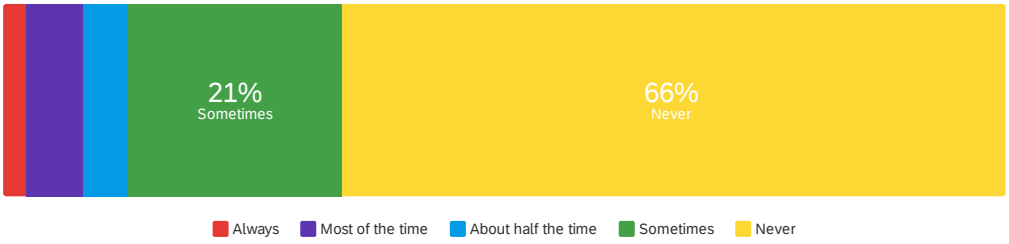


Figure 3.7: Participant rating of how frequently they read the Terms of Service or Privacy Policy for any service for which they register.

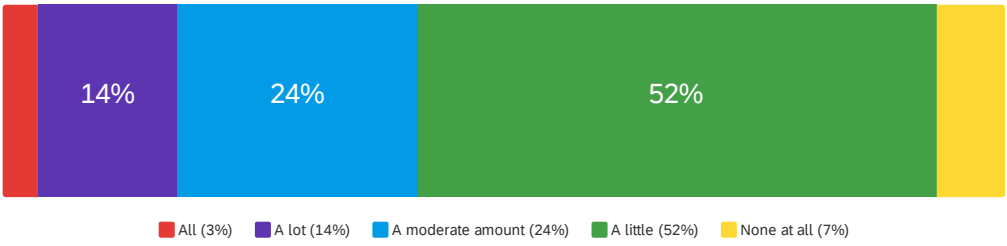


Figure 3.8: Participant rating of how much of the the Terms of Service or Privacy Policy they read for any service for which they register.

Gauging Expertise on Privacy Issues of Survey Participants

Figure 3.9, Figure 3.10, and Figure 3.11 all indicate that the majority of participants are not generally familiar with privacy issues. Most participants do not frequently take

an interest in discussing or researching the recent events surrounding online privacy. For these questions it is especially important to note that there was at least one outlier who identified in the survey that they have made these issues part of their primary research. Removing this individual makes the claim even stronger that the typical user in the study rarely discusses or studies issues and ideas around online privacy. While it is interesting to have data to support this idea, it comes as no surprise that most participants do not discuss or research these issues regularly. However, specific interest in this niche field should not be a prerequisite for understanding what information social media companies are collecting.

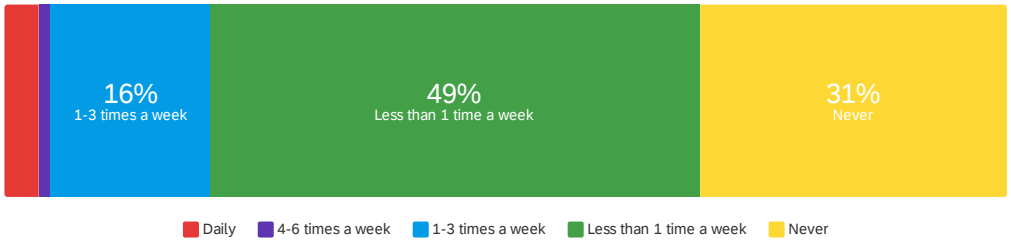


Figure 3.9: Frequency of which participants read articles related to online privacy issues.

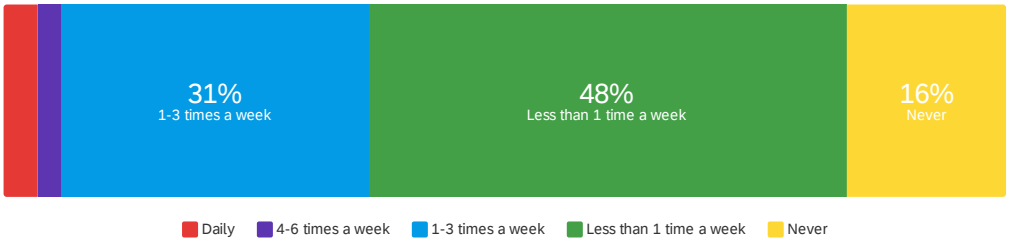


Figure 3.10: Frequency of which participants discuss online privacy issues with their peers.

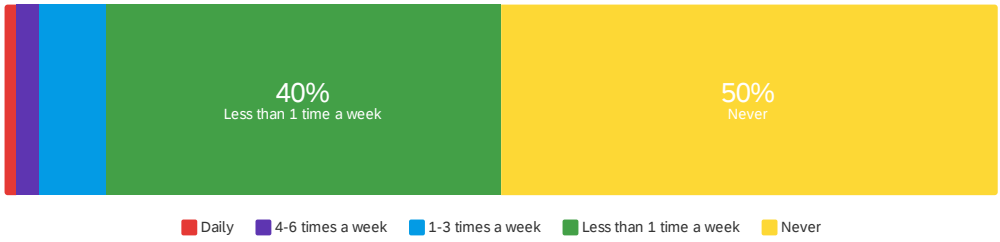


Figure 3.11: Frequency of which participants research online privacy issues.

3.4 Conclusion

The survey confirmed my expectations and generally agreed with the work of Winkler and Zeadally (2016) and Kaiser (2016) which motivated my study. From this data, it is important to note that the average Bucknell student is not very familiar with privacy issues, most users do not have a complete understanding of what information is collected, and most participants do not read any significant portion of the Terms of Service or Privacy Policy when signing up for a new service or website. Additionally, there was not a significant difference between the understanding demonstrated by science, technology, engineering, or math affiliated participants and the understanding of participants affiliated with humanities and social sciences.

Chapter 4

A Better Business Model For Social Networks

Based on my survey, it is clear that many people at Bucknell feel too much data is collected, but I did not address how highly they value their privacy. This is addressed in a 2015 Pew Research Center study where participants were asked generally how important privacy is to them (Madden and Rainie 2015). In this study, 93% of participants said that controlling who can get information about them is important and 90% said it is important for them to be able to control what information is collected about them (Madden and Rainie 2015). A later study also from the Pew Research Center found that 81% of U.S. adults feel they lack control over what information companies collect on them and that the risks outweigh the benefits (Auxier et al. 2019). Additionally, 79% also said they are concerned over who uses the data that is collected (Auxier et al. 2019).

I have proposed an alternative business model for social media companies that avoids the ethical issue of selling user data (Brown 2019). The idea I propose is a

tiered structure to social media user types (See Figure 4.1). A free tier where the user agrees to let the company collect and sell their information, and a paid tier where the user still uses the service, but the company cannot collect or sell their information or show them advertisements. While this does not fully satisfy Kant because technically the users are still a means to an end, the choice and greater level of informed consent are a step in the right direction.

Price	Privacy Control
$$	👍👍👍
$	👍👍
Free	👍

Figure 4.1: Tiered account type for proposed business model.

One problem with this solution is that it disadvantages those who cannot afford to pay for their privacy. An alteration to the solution above might also include a paid, but much less expensive, middle tier that still serves ads to the user but the advertising is only informed by an even more limited scope of their information beyond restrictive privacy settings. These solutions also put the social media companies on firmer ground with the Utilitarian philosophy because restoring some choice to the users should, theoretically, be a step towards maximizing happiness.

This presents the question of why should anyone have to pay for their privacy? While it would be nice if everyone could retain full control of their information for free, selling targeted advertisements is how many internet companies make money and are able to pay their employees. If a significant portion of a platform's users wish to keep their data private, it lessens the value of the advertising products the company offers to its customers. Thus, the lost revenue must be made up in order for the company to remain in business. This is why users who choose to opt out of

data collection would need to pay for the service.

I tested this idea in my survey, asking participants if they would be willing to pay for a service that did not collect their information. It did not gain a lot of favor with about 66% of respondents saying either that they would not want to pay or that they are neutral towards the idea. This is contrary to the result of a Pew Research Center study where 51% of U.S. adults in the study deemed the trade-off of a free social media platform being able to use their data being unacceptable while 33% said it was acceptable and 15% commented that it would depend on context (Rainie and Duggan 2016). This response indicates that there would likely be support for a paid service that could not collect information on its users.

Regardless of whatever pay structure a company decides to implement, meaningful choice is vastly important. As seen in the work of Winkler and Zeadally (2016), Jordan and Rand (2019), and Herder and Zhang (2019) and the survey that I ran, users often simply do not understand to what they are agreeing. This can be due to their unwillingness to read a long legal document, their own lack of understanding of the technology, or the obfuscation of what the user needs to know by unnecessary legalese. Or, as argued by Nissenbaum (2009), users lack meaningful choice and simply have to agree because the cost of not using the service is too great. But it should not require a college degree to large amounts of outside research to understand how social media companies know so much about us. As stated in the Institute of Electrical and Electronics Engineers (IEEE) guidelines for Ethically Aligned Design:

> A fundamental need is that people have the right to define access and provide informed consent with respect to the use of their personal digital data. Individuals require mechanisms to help curate their unique identity and personal data in conjunction with policies and practices that make

> them explicitly aware of consequences resulting from the bundling or resale of their personal information (IEEE SA 2017).

From what I have presented in this work, it's clear many companies are failing to meet this standard. To better follow the IEEE standard above, companies should provide a brief page, in language no more advanced than a seventh grade level (when many children turn 13 and can register for a Facebook account according to Facebook's ToS) that outlines what information is collected and what is done with it. Social media companies could also implement privacy features akin to smartphone operating systems that allow third parties temporary access certain information. Similar to how phones provide an option for an application to access your location "just this time" or for an hour, a similar feature on social media would restore a much higher level of control to the user.

A common question with this recommendation is whether the legalese is just a necessary evil to create a binding document. Simply put, it is unnecessary. GE Aviation's digital services business began an initiative in 2014 to write plain-language contracts with an overwhelmingly positive result Burton (2018). So if this went so well, why are other companies not following suit? They seem keen to hide behind confusing language and should be compelled to provide the ToS in plain language.

4.1 Conclusion

Ultimately, in order to maintain their revenue, social media companies could choose to offer various paid options that limit or eliminate the data the company is allowed to collect and share with advertisers. However regardless of a paid offering, social media companies should absolutely be required to gain informed consent from their

users, clearly indicating what data is collected and how it is used.

Chapter 5

Conclusion

In this thesis, I have shown my motivation for exploring this topic, provided a synopsis of some of the literature and current events around data collection and user privacy issues. Additionally, I reported the results of a survey that I created and ran on Bucknell University's campus that explored the opinions and behaviors of Bucknell students, faculty, and staff. Finally, I demonstrated that users feel their privacy is important and proposed a potential solution to the issues I identified in how social media platforms collect data and allow users to control their data.

In Chapter 2, the work Herder and Zhang (2019) and Jordan and Rand (2019) makes it clear that users are uncertain about what is collected and do not take the time to read the Terms of Service. These findings were confirmed by the results reported in Chapter 3, strengthening the argument made by Winkler and Zeadally (2016) that Terms of Service and Privacy Policies are not written in language accessible to many users or presented in a way that encourages the user to actually pay attention to the information. Considering these observations and arguments, along with the evidence from Madden and Rainie (2015) and Auxier et al. (2019) that people value their

privacy, in Chapter 4, I propose an alternative to the free, advertisement supported social media platform. I also emphasize that users give informed consent to their data being collected meaning that the ToS and Privacy Policy need to be presented in a quickly digestible and understandable way.

5.1 Future Work

Over the course of my research I ran into questions that were out of scope for this thesis, but might provide interesting extensions or continuations of my work. Regarding more concise ToS and privacy policies, it would be interesting to know where the middle ground between "too short to be informative" and "too long to read to hold the user's attention" lies. Another topic to explore would be whether an equivalent to the EU's GDPR would be helpful in anyway to American users. It would also be relevant to explore the economics of a paid social media platform using the tiered structure I suggest to determine its feasibility.

Additionally, another topic to explore would be an annual privacy audit that the platform directs the user through where they learn about the privacy features available and what data is being collected. Studying this for user opinion and willingness to actually do it along with the proper format and allotted could potentially completely change the priority that many people assign to their privacy

Finally, while I focused on how individuals interact with privacy, the ideas brought up in Section 2.1 by the work of Boyd (2012) provide an interesting counterpoint to my solution. Modelling privacy with the public as the focus rather the individual might provide generally cleaner solutions.

References

Alfred Ng (2020, 5). Governments could track COVID-19 lockdowns through social media posts.

Auxier, B. Y. B., L. Rainie, M. Anderson, A. Perrin, M. Kumar, and E. Turner (2019). Americans and Privacy: Concerned, Confused and Feeling Lack of Control Over Their Personal Information. Technical report, Pew Research Center.

Bond, S. (2020). The Coronavirus Crisis A Must For Millions , Zoom Has A Dark Side — And An FBI Warning.

Boyd, D. (2012). Networked privacy. *Surveillance and Society 10*(3-4), 348–350.

Brown, M. (2019). Internet Companies : Balancing Privacy and Profit. Term Paper for CSCI 245 Life, Computers, and Everything. Spring 2019.

Burton, S. (2018). The Case for Plain Language Legislation.

Cadwalladr, C. and E. Graham-Harrison (2018, 3). Revealed : 50 million Facebook profiles harvested for Cambridge Analytica in major data breach.

Greenberg, A. (2020, 4). How Apple and Google Are Enabling Covid-19 Contact-Tracing.

Herder, E. and B. Zhang (2019). Unexpected and unpredictable: Factors that make personalized advertisements creepy. *ABIS 2019 - Proceedings of the 23rd Interna-*

tional Workshop on Personalization and Recommendation on the Web and Beyond, 1–6.

IEEE SA (2017). Ethically Aligned Design: A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems - Version 2. Technical report.

Jordan, J. J. and D. G. Rand (2019). The Biggest Lie on the Internet: Ignoring the Privacy Policies and Terms of Service Policies of Social Networking Services. *1*(June), 1–18.

Kaiser, A. F. (2016). Privacy and security perceptions between different age groups while searching online. Technical report, University of Twente.

Madden, M. and L. Rainie (2015). Americans' Views About Data Collection and Security. Technical report, Pew Research Center.

Martin, K. E. and H. Nissenbaum (2016). Measuring Privacy: Using Context to Expose Confounding Variables. *SSRN Electronic Journal*, 1–40.

Newsroom, A. (2020). Apple and Google partner on COVID-19 contact tracing technology.

Nissenbaum, H. (2009). *Privacy in Context*. Stanford University Press.

Nissenbaum, H. (2015). Respecting Context to Protect Privacy: Why Meaning Matters. *Science and Engineering Ethics 24*(3), 831–852.

Rachels, J. (2012, 7). *The Elements of Moral Philosophy* (7th ed.). New York: McGraw-Hill.

Rainie, L. and M. Duggan (2016). Privacy and Information Sharing. Technical Report December 2015, Pew Research Center.

Selinger, E. and B. Leong (2020). The Lasting Privacy and Civil Liberties Impacts of Responses to COVID-19.

Sinha, M., P. Varma, G. Sivakumar, M. Singh, T. Mukherjee, D. Chander, and K. Dasgupta (2016). Improving urban transportation through social media analytics. *Proceedings of the 3rd ACM IKDD Conference on Data Sciences, CODS 2016*, 1–2.

Susser, D. (2019). Notice after notice-and-consent. *Journal of Information Policy 9*(May), 132–157.

Tanaka, Y., T. Kurashima, Y. Fujiwara, T. Iwata, and H. Sawada (2016). Inferring latent triggers of purchases with consideration of social effects and media advertisements. *WSDM 2016 - Proceedings of the 9th ACM International Conference on Web Search and Data Mining*, 543–552.

Valentino-DeVries, J., N. Singer, M. H. Keller, and A. Krolik (2018). Your Apps Know Where You Were Last Night, and They're Not Keeping It Secret.

Winkler, S. and S. Zeadally (2016, 6). Privacy Policy Analysis of Popular Web Platforms. *IEEE Technology and Society Magazine 35*(2), 75–85.

Xue, M., C. Ballard, K. Liu, C. Nemelka, Y. Wu, K. Ross, and H. Qian (2016). You can yak but you can't hide: Localizing anonymous social network users. *Proceedings of the ACM SIGCOMM Internet Measurement Conference, IMC 14-16-Nove*(i), 25–31.

Appendices

Block 1

Online Privacy and Social Media
Informed Consent Bucknell University

You are invited to participate in a research study about attitudes and behaviors related to online privacy and data collected by social media. If you agree, you will be presented with a brief, anonymous survey (expected to take no more than 10 minutes) that will ask about your perception of social media, your engagement with privacy issues, and your online behavior as it pertains to privacy. We believe that this will help our research team understand how the majority of users engage with online privacy and inform analysis of the practices employed by social media companies. Your participation is fully voluntary and you may discontinue your participation at any time. There are no risks anticipated for this study. We will not be asking you for any personally identifying information. The research team hopes to recruit around 400 participants.

At the end of the survey, you will be asked if you would like to be entered in a drawing for one of four $25 gift certificates to Downtown Lewisburg. The drawing will take place on February 7th, 2020 with notifications to winners coming shortly after.

If you have any questions or concerns about this study, you may contact the Principal Investigator, Matt Brown, by phone at (203) 501-9875 or email at msb027@bucknell.edu. General questions or concerns about the rights of human subjects of research may be directed to the chair of the Institutional Review Board at Bucknell: Matthew Slater at matthew.slater@bucknell.edu or x72767

By clicking 'Agree' below, I affirm that I am 18 years of age or older:

Agree ○ Disagree ○

Default Question Block

What is your primary area of study? (e.g. major, department, or specialization)

What is your affiliation with Bucknell?

- ○ Student
- ○ Faculty
- ○ Staff
- ○ Other (please specify)

Block 2

What information do you think Facebook collects on its users? Please enter as many as you can think of separated by commas.

What information do you think Google collects on its users? Please enter as many as you can think of separated by commas.

Block 3

To what degree to you agree with the following statements?

	Strongly disagree	Disagree	Somewhat disagree	Neither agree nor disagree	Somewhat agree	Agree	Strongly agree
Internet Companies collect too much information	○	○	○	○	○	○	○
I understand what information websites collect on me	○	○	○	○	○	○	○
I know how to access the information stored on me by websites like Facebook and Google	○	○	○	○	○	○	○
I prefer to see ads that are relevant to me	○	○	○	○	○	○	○
I would be willing to pay a subscription fee for social media and other websites if it meant they could no longer collect data on me	○	○	○	○	○	○	○
I know how to use to privacy features in social media and on my phone to limit or allow data to be collected	○	○	○	○	○	○	○
I use privacy features on social media and on my phone to limit what data is collected	○	○	○	○	○	○	○
Internet companies do not collect enough information	○	○	○	○	○	○	○

Block 4

How frequently do you feel that ads you see online or on social media are unnervingly relevant? (For example, you see an advertisement for something you were just discussing with another person.)

A great deal	A lot	A moderate amount	A little	None at all

How frequently do you read the Terms of Service/User Agreement/Privacy Policy when signing up for a new website or service?

Always	Most of the time	About half the time	Sometimes	Never
○	○	○	○	○

How much of the Terms of Service/User Agreement/Privacy Policy do you read?

All	A lot	A moderate amount	A little	None at all
○	○	○	○	○

How frequently do you read news articles or academic papers related to online privacy or data privacy?

- ○ Daily
- ○ 4-6 times a week
- ○ 1-3 times a week
- ○ Less than 1 time a week
- ○ Never

How frequently do you discuss online privacy or data privacy with peers?

- ○ Daily
- ○ 4-6 times a week
- ○ 1-3 times a week
- ○ Less than 1 time a week
- ○ Never

How frequently do you research online privacy, data privacy, or related issues?

- ○ Daily
- ○ 4-6 times a week
- ○ 1-3 times a week
- ○ Less than 1 time a week
- ○ Never

Block 5

Thank you for your participation! Would you like to be entered in the drawing for one of four Lewisburg Dollars gift certificates worth $25?

Yes ○

No ○