Building a New Grading System for CS 2110: Software Development Methods

(Technical Paper)

The Rise of the Problem in Data and Privacy

(STS Paper)

A Thesis Prospectus Submitted to the

Faculty of the School of Engineering and Applied Science

University of Virginia • Charlottesville, Virginia

In Partial Fulfillment of the Requirements of the Degree

Bachelor of Science, School of Engineering

Stephen Shiao

Spring, 2020

Technical Project Team Members

Stephen Shiao

Kenneth Chen

On my honor as a University Student, I have neither given nor received unauthorized aid on this assignment as defined by the Honor Guidelines for Thesis-Related Assignments

Stephen Shiao, Spring 2020, Department of Computer Science

Technical Advisor

Dr. Nada Basit, Department of Computer Science

STS Advisor

Michael Gorman, Department of Engineering and Society

Peer Reviews, Comments, and Additional Help and Advice

I got great feedback from my peers and Professor Gorman while writing this thesis prospectus. This section hopes to show my appreciation as well as details their contributions.

My classmates Kenneth Chen and Kelsi Loudenslager encouraged me to move forward with the topic of data and privacy. Kenneth Chen had helped me to focus on a more specific topic within data and privacy: how data breaches affect society and trust in corporate control of data. After talking to Kenneth more, I was able to narrow my topic down, as data and privacy as a topic is very broad.

Another classmate, Julianna Chaput, helped me to develop an STS framework to work towards the causation of the issue of data and privacy. Namely, using Actor Network Theory and normalized deviance to analyze the issue since many companies are young and prone to mistakes and new research.

Professor Gorman helped me develop the STS framework trading zones to come up with ways our system could alleviate the issue in the future. In a time where many people are attempting to come up with solutions, this became a new topic for me to research for my prospectus and thesis.

I am thankful for James Dellaripa, Sudhir Shenoy, and Professor Gorman for teaching this class in order for me to research a topic and look at it with new STS frameworks.

Introduction

In a world of growing information and technology, companies have more and more data on customers, which they utilize to grow and run their businesses. However, with this data, there is a growing concern for privacy. To what extent do companies collect data on customers? What are the risks involved with putting this information on their servers? What potential effects do these have on society?

To investigate this phenomenon, I will look into different companies and their products to see what data they collect, if they are limited by laws in, or if they choose not to collect certain types of data. I will also look into risks involved with having different types of data in a database, like potential exploits that people could abuse, as well as examples of when a company had their public image lowered because of these possible exploitations. To analyze the problem, I intend to use Actor Network Theory (ANT) on the different examples, as well as discuss how multiple stakeholders could form trading zones in order to protect data and their users. I will also look into normalized deviance – could the creators of the product have thought about the consequences, but figured it would be fine?

Background

Data in its simplest form is just information, but it can be split into many categories. In particular, with respect to people, it can be classified as: person-specific data, anonymous data, explicit identifying data, and de-identified data (Sweeney, 2000). Person-specific data refers to information that is specific to an individual, but not necessarily unique to any one individual, like age. Anonymous data are information that cannot be manipulated or linked to uniquely identify the subject of the data. Explicit identifying data refers to data that provides a direct medium of communication with an individual, so that they can be directly and uniquely contacted, like a

phone number. De-identified data refers to a set of data where all explicit identifying data is removed, replaced, or generalized.

In our world of ever-changing technology and products, businesses have to collect data on customers in order to be successful (M, Atif, 2019). They analyze the data they collect to gain insight on consumers, to attract and engage potential customers, and to better retain customers. With the amount of data that companies collect, the term is now coined "big data." Companies now utilize several big data strategies to accomplish their goals (Parise, Iyer, & Vesset, 2012). First is performance management – they try to understand the data they collect to create better short-term business decisions and longer-term plans. Next, data exploration uses statistics to experiment and find answers to questions they might not have been thinking of originally. Social analytics aims to understand the social aspect of the internet, examining exposure of social content and interconnectivity between platforms and their members. Finally, decision science aims to improve the general decision-making process based on past data.

Businesses build their products by catering them to the customers' needs or wants, but they may not consider their safety or privacy. The business themselves could use it with alternative intent than just what they state they use it for. Typically, this involves selling that information to another company (Matsakis, 2019). Hackers with malicious intent could obtain unauthorized access to the data, and profit off that. Finally, products that connect with third party applications may have vulnerabilities that the third parties could take advantage of. Furthermore, much of data that doesn't seem like it could uniquely identify someone actually could. In a publication in 2000, 87% of the US population could be identified with a 5-digit ZIP code, their gender, and date of birth. Additionally, 53% could be identified by place, gender, and date of

birth, and 18% could be identified by county, gender, and date of birth.4 Thus, even giving basic person-specific data to companies is valuable.

Currently, many consumers are concerned with privacy, data breaches, and identity theft, even if they haven't actually experience financial harm (Sweeney, 2000). Specifically, they're anxious about the potential for data misuse. When a data breach occurs, the focal firm's stock always drops. However, the effect on the focal firm's rivals varies – it may lower due to concerns about similar data breaches at a rival firm, or it may increase since customers are switching firms within a line of business (Martin, Borah, & Palmatier, 2016). However, when a data breach occurs, there are ways companies can mitigate the loss of trust with customers. The more transparent a company is with data policies, the less negative effects they'll have. Additionally, the more control that customers have over their own data, the more trust they'll get from customers (Martin, Borah, & Palmatier, 2016).

Customers also care about privacy, which refers to a customer's perception of the potential for harm as a result of using their personal data (Martin, Borah, & Palmatier, 2016). As a firm collects more personal information on a customer, it increases susceptibility to harm, thus increasing customers' feelings of vulnerability (Martin, Borah, & Palmatier, 2016). Once again, negative customer feelings stem from anxiety for potential damage. Furthermore, even if increased control is given to customers, there is not much they can do. Data put onto the internet is difficult or impossible to remove, since an account could be deleted on a product, but the data may have already been given to another firm (Martin, Borah, & Palmatier, 2016).

Consumers also judge companies at a certain level, while not participating in secure practices on their own, so not all blame can be put on companies if a customer's data is leaked. They figure out how much to trust a company based on whether or not they have had a data

breach before and how much data they are collecting from a consumer, despite not always participating in secure online practices. Companies should encourage personal security behaviors, while maintaining their own brand by engaging in transparency with the consumers (Curtis & Jones, 2018).

Policies are being put into place to regulate the exchange of data, but many misuses of data have already occurred that brings this issue of privacy and data into the spotlight. One such example is the Facebook-Cambridge Analytica data scandal.

Facebook – Cambridge Analytica Data Scandal

Cambridge Analytica acquired Facebook data in multiple ways. First, they had purchased the data on tens of millions of Americans, who did not know about this acquisition (Matsakis, 2019). Next, a professor at the University of Cambridge had created a third-party quiz application to share on Facebook (Lapowsky, 2019). Once users got the app, the app would collect their personal information. Not only that, but it would continue onto users' friends, if their privacy settings allowed it, getting data on about 87 million users. This data was then given to Strategic Communication Laboratories, who owns Cambridge Analytica.

Previously, Cambridge Analytica had published a paper, describing how they could predict people's personalities and other sensitive details based on likes of Facebook posts (Kozlowska, 2019). Using this method and the data they had acquired, they were able to create personality profiles and influence political ads during the 2016 presidential election (Lapowsky, 2019).

This data scandal led to public outrage and criticism on policy. Facebook lost many users through the #DeleteFacebook boycott campaign and lost about $50 billion in market capitalization within 3 days (Kozlowska, 2019). They also faced many lawsuits and were called

into Congress to answer questions about actions and privacy. The CEO of Cambridge Analytica, Alexander Nix, was suspended after the scandal and later stepped down. The scandal was not a data breach – all data obtained was technically consensual, however data was misused. As such, Facebook and Cambridge Analytica both breached customers' trust. However, Facebook was transparent about the misuse, and apologized for breaching customer trust. As a result, negative effects on Facebook were mitigated.

This fiasco led to a spark in policymaking for privacy and data, and the questioning of some companies' behaviors. Facebook tightened their API's so that the information third-party apps can collect is limited. Furthermore, if a user hasn't used their app in three months, they would not be able to collect user data from them. Finally, Facebook now authorizes those who want to put political or issues advertisements, including information on who posted it (Lapowsky, 2018). They also plan to implement end-to-end encryption for their Messenger and Instagram products (Kaslowska, 2019). Furthermore, state governments have started pushing new laws: Vermont implemented a new law that requires data brokers to register with the state, and California implemented a law that would give residents the ability to opt out of having their data sold. Additionally, Congress is in talks of federal data protection law (Lapowsky, 2019).

The scandal also leads to other lessons learned. Consumers should recognize that their personal data is valuable – companies use it to drive their businesses. Companies should also better balance privacy risks and privacy controls – the more sensitive the data, the more protections should be put in place. Finally, governments should continue to reform legislation to address the modern problems that data has given rise to (Lapowsky, 2019).

What allowed the data scandal to occur should also be examined. According to a report, in the early stages of Facebook, they had made the decision to allow data access from third-apps

on users' friends. Zuckerberg had wondered whether or not allowing this was risky, but ultimately allowed it (Lapowsky, 2019). Furthermore, Cambridge Analytica had received warnings from its lawyer when they employed European and Canadian data scientists whilst working on the American presidential campaign. The CEO, Alexander Nix, had also suggested that the company used other unethical practices to influence elections around the world (Confessore, 2018).

Conclusion

Companies collect personal data on customers in order to get a competitive edge over rivals and to maximize their profits. To what extent depends on what they can get from customers – or rather what customers are willing to give them. Consumers risk having their data misused when they give it to companies. The risk comes from potential hackers attempting to breach data, or even just the company itself finding a loophole to profit off of the data. These may affect society on different scales, but as Cambridge Analytica has shown us, the effects could be drastic.

Using the framework of Actor Network Theory (ANT), several actors can be identified. The first major actors of products requiring data and privacy include the companies that make the products, the customers that use the products, and the actual product. Within product creation, companies are concerned with making products that customers want to use. Customers will only pay money or use a product if it fits their tastes. This network is driven by consumer needs, but doesn't consider safety or privacy as an actor. This can be seen in the Facebook-Cambridge Analytica scandal: Zuckerberg had decided that there posed no risk to allow third-party applications to access users' friends' data, and that policy became the norm until another company had abused that fact, a result of normalized deviance. The introduction of risk as an

actor breaks down the first model constructed by ANT, as new actors that are introduced will require a new network to be constructed. However, to enforce this new actor to be recognized by companies, the government must also be introduced, which comes with lawmakers and voters.

This new network will consist of the products, their companies, their users, voters, lawmakers, and the laws they create (involving data and privacy). The first three are connected in the same way. Lawmakers will take into consideration the best interest of users and voters, and create laws based on their opinions. These laws will enforce certain practices for companies, making them keep users' safety and privacy in mind.

In order to actually create good laws to handle data and privacy, we will need to look into trading zones. Voters and users must tell policymakers to know what to base their policies around. Experts in data and privacy must also be involved in order for them to know how to implement these policies. Once these stakeholders figure out sufficient laws, government agencies and the legal system must work to enforce these. After, we may limit the growing problem within data and privacy.

Future Steps

I would like to continue to analyze cases of data breaches in order to keep learning about the effects of data misuse on society. This could include past cases like Yahoo's, Uber's, or Equifax's data breaches, or very recent problems like the FBI collecting facial recognition data. I also would like to apply the ANT framework, in detail, on many of these cases, to get a better understanding of causation – how we have approached this point in time about privacy and data, and why it may fall apart or give rise to more issues. Additionally, in recent news many presidential candidates are proposing more and more ideas to combat the issue of data and

privacy – I would like to take a look at these to find examples of trading zones that are occurring within our legal system now.

References

Sweeney, L. (2000). Simple Demographics Often Identify People Uniquely. Retrieved from https://dataprivacylab.org/projects/identifiability/paper1.pdf.

M., Atif. (2019, May 18). Predictive Analytics: Retrieved from https://towardsdatascience.com/predictive-analytics-predicting-consumer-behavior-withdata-analytics-8ca51abb8dc2.

Parise, S., Iyer, B., & Vesset, D. (2012, July). Four Strategies to Capture and Create Value from Big Data: Retrieved from https://iveybusinessjournal.com/publication/four-strategies-tocapture-and-create-value-from-big-data/.

Martin, Kelly & Borah, Abhishek & Palmatier, Robert. (2016). Data Privacy: Effects on Customer and Firm Performance. Journal of Marketing. 81. 10.1509/jm.15.0497.

Matsakis, L. (2019, February 19). The WIRED Guide to Your Personal Data (and Who Is Using It). Retrieved from https://www.wired.com/story/wired-guide-personal-data-collection/.

Kozlowska, Iga (2019, July 11) Facebook and Data Privacy in the Age of Cambridge Analytica. Retrieved from https://jsis.washington.edu/news/facebook-data-privacy-age-cambridgeanalytica/.

Confessore, N. (2018, April 4). Cambridge Analytica and Facebook: The Scandal and the Fallout So Far. Retrieved from https://www.nytimes.com/2018/04/04/us/politics/cambridgeanalytica-scandal-fallout.html.

Lapowsky, I. (2019, March 18). How Cambridge Analytica Sparked the Great Privacy Awakening. Retrieved from https://www.wired.com/story/cambridge-analytica-facebookprivacy-awakening/.

Sources that were not cited in line

Armerding, T. (2018, December 20). The 18 biggest data breaches of the 21st century. Retrieved from https://www.csoonline.com/article/2130877/the-biggest-data-breaches-of-the-21stcentury.html.

Crockford, K. (2019, October 31). ACLU News & Commentary. Retrieved from https://www.aclu.org/news/privacy-technology/the-fbi-is-tracking-our-faces-in-secretwere-suing/.

Edwards, Benjamin, Hofmeyr, Steven, Forrest, & Stephanie. (2016, December 30). Hype and heavy tails: A closer look at data breaches. Retrieved from https://academic.oup.com/cybersecurity/article/2/1/3/2736315.

Pros and Cons of Predictive Analysis. (2018, September 28). Retrieved from https://scsonline.georgetown.edu/programs/masters-technologymanagement/resources/pros-and-cons-predictive-analysis.