An Analysis of Data Harvesting in the Facebook-Cambridge Analytica Scandal

A Prospectus submitted to the Department of Engineering and Society

Presented to the Faculty of the School of Engineering and Applied Science

University of Virginia • Charlottesville, Virginia

In Partial Fulfillment of the Requirements for the Degree

Bachelor of Science, School of Engineering

Micah William Cho

Spring 2023

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

Advisor

Alice Fox, Department of Engineering and Society

Overview:

In 2006, data scientist Clive Humby coined the phrase "Data is the new oil," alluding to the value that data possesses in the 21st century. After two decades of rapid growth in technology and the emergence of big data, Humby's statement is spot on. Data is more valuable than ever and fuels some of the most valuable companies in the world, including Apple, Microsoft, Google, and Meta. Consequently, data harvesting, or the collection of data, has become a popular practice. Although data harvesting methods are rapidly improving and advancing, a lack of regulation has caused data harvesting to look more like data exploitation. My STS research will explore data harvesting in the Facebook-Cambridge Analytica scandal and determine ways we can best protect private data from exploitation by big data companies. I will perform this research from an Actor Network Theory perspective in order to inspect the major players in the scandal and how their relationships and interactions resulted in the scandal. My technical report will analyze the intersect between banking, technology, and cybersecurity through my experiences as an intern at Capital One and analyze how university computer science departments can better prepare students for careers at companies that focus on the intersect between technology and other fields.

Problematization:

In 2018, Christopher Wylie, a data scientist for political consulting firm Cambridge Analytica, revealed that the company had obtained harvested data on over 87 million Facebook profiles and had used the information to attempt to influence the 2016 U.S. presidential election (Rehman, 2019). This event, which later became known as the Facebook-Cambridge Analytica scandal, sparked massive concerns over data privacy policy in the United States and across the

world. In the years since the scandal, it has become evident that there are issues with data collection, especially when it comes to personal and private information.

Guiding Question:

How can we prevent or discourage big data companies from exploiting personal and private data?

Projected Outcomes:

I am performing my research with two goals: to empower the public and to impact legislation. In my research, I will address the problem of unethical data collection by attempting to define what is “public” data and what is “private” data. With these definitions, I hope to empower individuals to protect their own clearly defined “private” data. Additionally, these definitions can be worked into legislation on data harvesting, which can prevent data harvesters from collecting “private” information.

Technical Project Description:

My technical report will focus on my experiences as an intern at Capital One. Software engineering projects at Capital One focus on the intersect between banking, technology, and security. However, it is rare to find a software engineer that is well educated in all three of these fields. Thus, my goal with my technical project is to analyze how computer science departments at universities can better prepare students for jobs like these, where there is a focus on technology and its application in other fields. I will have completed this internship by the end of this summer and hope to use my experiences at Capital One to construct meaningful solutions for better education practices.

Preliminary Literature Review & Findings:

In the early aftermath of the Facebook-Cambridge Analytica scandal, academia, researchers, and professionals have published a number of analysis papers, potential solutions, and commentary on data privacy policy. Furthermore, policy reform has been analyzed in a number of other countries as concerns over data privacy have risen across the world since the event. In my research of national and international commentary, I have come across several key findings and potential solutions which I will attempt to build upon in my STS research.

Current scholars have proposed solutions from a variety of perspectives for the problems from the scandal. (Mancuso and Vegetti, 2020) proposed a tool for collecting Facebook data that is considered "public." (Sun, 2020) suggested legislative changes that would recognize data as property. (Fuller, 2019) composed an article promoting theological responses to the underlying ethical issues of the scandal. Additionally, (Isaak and Hanna, 2018) discussed the need for engineers and technologists to contribute to policy change. In my continued research, my goal is to identify which strategies or combinations of strategies will be most effective in protecting data privacy long-term.

STS Project Proposal:

The Facebook-Cambridge Analytica scandal was a wake-up call to the world about the dangers of unregulated data harvesting. My STS project will examine data harvesting in the Facebook-Cambridge Analytica scandal and will specifically investigate the ethics behind how the data was harvested. The goal of my research is to identify the problems in the ways that data was harvested and to look for potential solutions, through both societal change and legislation. I hope to discover the impacts of both data and data harvesting on the digital world and ways we can create a more secure digital world. This examination of how our modern society is affected by technologies involved in data harvesting is what makes this project an STS project.

Through my STS research, I will take two main approaches to thoroughly analyze the Facebook-Cambridge Analytica scandal. Firstly, I intend to review the actions of Facebook, Cambridge Analytica, and other main actors, from an ethics and value standpoint. It is evident that the data harvesting was unethical, but in what ways? Additionally, can we place the blame on any specific actor in the scandal? In analyzing questions in ethics like these we can begin to understand what went wrong. Secondly, I intend to analyze the scandal from a policy standpoint. Clearly, data privacy policy at the time of the scandal was insufficient, but how can we learn from these insufficiencies? Is policy reform adequate enough as a solution, or will we require different types of social change? I intend to review policy surrounding data privacy and data harvesting in efforts to answer these questions.

I choose to frame this issue from an Actor Network Theory perspective. I believe this will be most effective since I can investigate how each actor in the Facebook-Cambridge Analytica scandal contributed to the data harvesting issue. Furthermore, I can examine the relationships between actors in order to devise solutions that will keep similar actors accountable in the future. Using Actor Network Theory, I will perform my analysis of the scandal mainly through review of literature written by academia and researchers along with legislation written both before and in response to the scandal.

Barriers & Boons:

My approach, which focuses on ethics and policy, is limited by the fact that I have very little formal education in public policy as an engineering student. To accommodate for this, I will work to validate my approach by extensively reading, studying, and performing research on legislation and commentary regarding ethics in data privacy. One significant barrier to solutions I will propose in my STS project is the fact that policy implementations can take years in the U.S.

legislative system. Thus, my solutions will have to be able to endure the test of time, especially as technology and data harvesting methods advance more rapidly than policy can keep up with. Additionally, my choice of using Actor Network Theory can have limitations, as Actor Network Theory only analyzes actors and their relationships but does not expand much beyond that. I will accommodate this in my analysis by critically analyzing each player in the scandal and by using relationships between players as major points of analysis.

References

Fuller, M. (2019). Big Data and the Facebook Scandal: Issues and Responses. *Theology*, 122(1), 14-21.

Isaak, J., and Hanna, M. J. (2018). User Data Privacy: Facebook, Cambridge Analytica, and Privacy Protection. *Computer*, 51(8), 56-59.

Mancosu, M., & Vegetti, F. (2020). What You Can Scrape and What Is Right To Scrape: a Proposal for a Tool to Collect Public Facebook Data. *Social Media+ Society*, 6(3), 2056305120940703.

Rehman, I. u. (2019). Facebook-Cambridge Analytica Data Harvesting: What You Need to Know. *Library Philosophy and Practice*, 1-11.

Sun, S. C. J. (2020, May). Cambridge Analytica: A Property-based Solution. In *Cognito-studentisches Forum fur Recht und Gesellschaft*.