THE USE OF DECISION TREES TO DEFEND AGAINST SPECIFIC CYBERATTACKS

HOW UNDERSTANDING LEGISLATION IN CALIFORNIA CAN CLARIFY DATA PRIVACY RESPONSIBILITIES

A Thesis Prospectus
In STS 4500
Presented to
The Faculty of the
School of Engineering and Applied Science
University of Virginia
In Partial Fulfillment of the Requirements for the Degree
Bachelor of Science in Computer Science

By
Jordan Crawley

August 5, 2022

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.

ADVISORS

Catherine D. Baritaud and Bryn Seabrook, Department of Engineering and Society

Sebastian Elbaum, Computer Science Department

The Internet has come a very long way since its inception in the late 20th century. Many would likely agree that overall, the Internet has been a boon to society in many ways. To name just a few, the Internet allows us to experience generally increased ease of living, increased and speedier access to information, and even increased social connection over distances and with people we might not otherwise interact. This makes the Internet one of the most important technological innovations in the last century. As predicted by Stansberry, Anderson, and Raine in their 2020 article *The Internet will continue to make life better*, the Internet's relevance will only increase as time goes on (Stansberry et al., 2020, para. 3).

However, just as the Internet continues to expand, so too do the concerns that arise around it. For example, cybercriminals and cyberattacks have become a big concern for both individual users of the Internet and large companies, as they often threaten to steal important information. As a society, we've seen an increase in large-scale data breaches in recent years, which demonstrates the importance of the cybersecurity field and the necessity of thorough practices within cybersecurity.

This risk of data breaches, in addition to general concern of personal data misuse by companies, the government, etc., also give rise to another important issue: the risk makes it important for people in our society to understand what responsibilities exist between users, companies, the government, cybercriminals, etc., when an incident regarding user personal data does occur. However, as it is now, there is a distinct lack of understanding of these responsibilities and how they may impact consequences of a data breach or similar incident. This is especially true in the United States, as legislation regarding these areas is somewhat unclear. One state in particular, California, is the exception to this statement, as their legislation has been made much clearer over time.

The focus of this thesis portfolio, including both a technical report and an STS research project, will be on these concerns regarding cybersecurity attacks and data privacy. In particular, the technical report will blend relevant learnings from the University of Virginia's Algorithms class with learnings from the University's Introduction to Cybersecurity and Defense against the Dark Arts classes in order to detail how specific algorithms can aid in defending a network against specific cyberattacks. The STS research paper will instead focus on exploring how to go about determining what responsibilities should exist regarding data privacy and how to make these responsibilities clearer to the public. In particular, it will do this by making the case that exploring the history of California's data privacy legislation can help to provide an answer to these concerns. The entirety of the work will be carried out during the Fall 2022 and Spring 2023 semesters, as depicted in the Gantt Chart shown in Figure 1.

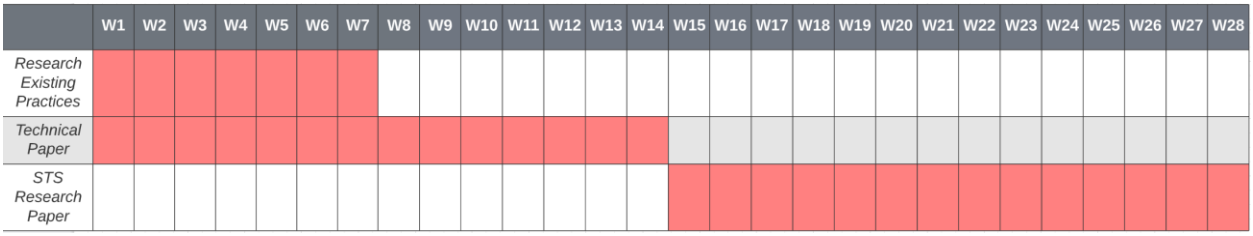

Figure 1: Gantt chart UVA CS Thesis Portfolio. This figure visualizes the expected timeline for the most important milestones of the portfolio

THE USE OF DECISION TREES TO DEFEND AGAINST SPECIFIC CYBERATTACKS

As detailed in the introduction, modern cybersecurity has become a mainstream concern for users of the Internet everywhere. Corporations and even governments are faced with the possibility of data breaches, theft of classified information, or loss of internal system control. Individual users are faced with similar concerns, with the most prevalent generally being the theft of personal data and loss of system control. In order to dissuade these concerns and prevent

cyberattacks, cybersecurity has become a very relevant discipline, with companies showing increased interest in having able cybersecurity personnel in place and a surge in cybersecurity studies in recent years. We've even seen the rise of mainstream public antivirus and security software such as Norton, which aims to provide the availability of automated cybersecurity practices to the general public, even including individual users who may not have a proper understanding of cybersecurity concepts.

To respond to this issue, the technical project of the portfolio will be a State-of-the-Art Technical Report which will focus on how learnings from the University of Virginia's Algorithms class can be blended with learning from the University's Introduction to Cybersecurity and Defense Against the Dark Arts classes in order to defend a network against specific cyberattacks. In particular, the technical project will focus on the use of decision trees as a way to respond to cyberattacks. Trees are one of the most important concepts taught in the University's Algorithms class, and their many different varieties are used in a number of specific algorithms for various purposes. 2U, Inc defines a decision tree as "a type of supervised machine learning used to categorize or make predictions based on how a previous set of questions were answered" (2U, Inc, 2022, para. 1). In other words, a decision tree is a specific type of tree that uses some kind of algorithm in order to decide on a specific outcome or response to a situation. In cybersecurity, decision trees have many uses. In particular, cybersecurity professionals note that some form of decision trees are commonly used in the detection of cybersecurity attacks. For example, Kaur, Vashisht, and Saurabh note that decision trees are an essential part of their adaptive cybercrime detection algorithm (Kaur et al., 2012, p. 2). Attackers use algorithms and machine learning themselves in order to determine how to attack a target, so identifying these types of attacks through a decision tree algorithm makes the next steps of responding to an attack

easier. Because of this, it is worthwhile for the technical report to give information on how these decision trees can be used to accomplish this in detail.

The technical report will begin by giving background information on decision trees themselves. Next, it will give context to the most common types of attacks which a decision tree can be used to defend against. For each of these attacks, the report will detail exactly how the use of the decision tree increases the security of the network and appropriate responses to said attack. Finally, it will conclude by offering a conclusion which summarizes the most important points of the report. Hopefully, the completion of this technical report will increase the reader's understanding of decision trees and aid the general state of cybersecurity by drawing attention to the issue.

It should also be noted that as a State-of-the-Art report, the project should not require the use of any physical resources such as labs, large equipment, etc., and neither should it require any amount of funding by the University. All research should be able to be carried out through the use of online resources and physical resources found in the University library alone. The paper will be written in the form of a scholarly article, with the goal of being able to be decently understood by anyone with a general education, but with an intended audience of college-level students or those studying cybersecurity or algorithms.

HOW UNDERSTANDING LEGISLATION IN CALIFORNIA CAN CLARIFY DATA PRIVACY RESPONSIBILITIES

As important as cybersecurity is, it is inevitable that sometimes it does fail. When this happens, it can sometimes be a struggle to determine who exactly should face the consequences for the repercussions that may occur as a result of the incident. As stated in the introduction, a growing concern among members of society in the 21st century has been determining the

responsibilities that exist between different entities with regards to data privacy concerns. While this is less of an issue in certain countries like the European Union, particularly in the United States, there is lack of true clarity in legislation and precedent regarding who holds what responsibilities in data privacy incidents. This lack of understanding has been evident in a number of recent incidents, for example the 2019 trials involving Facebook CEO Mark Zuckerberg after a largescale data breach on the platform. The graph in Figure 2 depicts the results of a Cisco Consumer Privacy Survey conducted in 2019, showing the disparity between different ideas of responsibilities for data privacy.

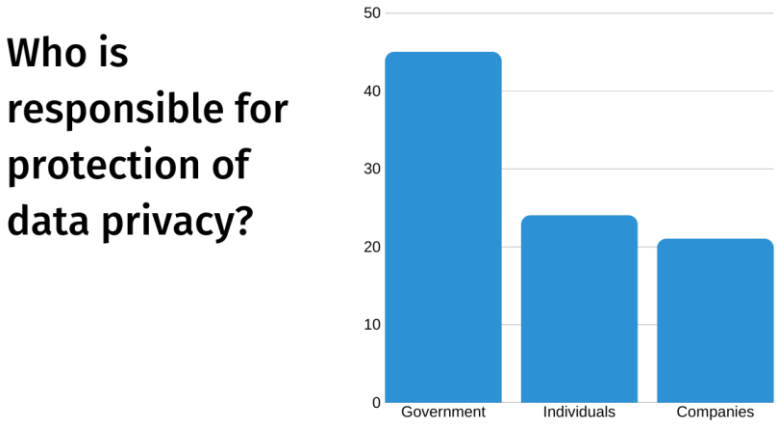

Figure 2: Differing public ideas of data privacy responsibilities (Cisco Consumer Privacy Survey, 2019).

However, California is the exception to this. Some regard California as having the clearest and most well-developed legislation regarding data privacy in the world. This is because California's legislation has gone through several iterations over time and puts focus on making sure that the law is understandable for people with minimal technical knowledge, and also likely because California's technology industry is so large. The STS Research paper will make the case that studying and understanding this legislation can make it easier to determine and clarify the

responsibilities that should exist between different entities with regards to data privacy. It will accomplish this through the use of Actor Network Theory.

In her online article on the subject, Charlotte Nickerson loosely defines Actor Network Theory as an "approach to understanding humans and their interactions with inanimate objects" (Nickerson, 2022, para. 5). In Actor Network Theory, a particular system is analyzed as a network, focusing on the relationships between different entities in the network and how they interact. Individual entities or points in this network, both human and nonhuman, are referred to as actors or actants (Nickerson, 2022, para. 14). In the case of data privacy legislation, actors might include human components like companies and nonhuman components like online agreements. By employing the use of Actor Network Theory, the STS research paper will examine each of the important actors that exist in the history of California's data privacy legislation, including but not limited to the government, individuals, cybercriminals, and malware, in addition to the aforementioned actors. By exploring the relationships between these actors, the research paper will detail how these relationships and actors within the network helped to get legislation in California to where it is now and how this could be reflected elsewhere. Hopefully, after the paper is complete, it will give readers a better understanding of how the relevant responsibilities might be assigned elsewhere across the United States based on how California's has handled data privacy policies. This paper will also take the form of a scholarly article, presenting information on modern interpretations of data privacy with graphics, as well as presenting information on and analyzing data privacy policies in California.

THE IMPORTANCE OF DATA PRIVACY

As a whole, the portfolio is intended to shed light on some of today's most relevant concerns regarding data privacy. It will do this mainly from the perspective of companies in the

technical report by examining cybersecurity practices with decision trees. It will also take an issue more relevant to individual users with the STS research paper by examining data privacy responsibilities, a subject relevant to all users of the Internet.

REFERENCES

2U, Inc. (2022, July 11). Decision tree. Master's in Data Science. Retrieved August 4, 2022, from https://www.mastersindatascience.org/learning/machine-learning-algorithms/decision-tree/

Cheng, L., Liu, F., & Yao, D. D. (2017). Enterprise data breach: causes, challenges, prevention, and future directions. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 7(5). https://doi.org/10.1002/widm.1211

Cisco Consumer Privacy Survey (2019). Figure 2: Differing public ideas of data privacy responsibilities

Duan, Y., Ge, Y., & Feng, Y. (2020). Pricing and personal data collection strategies of online platforms in the face of privacy concerns. *Electronic Commerce Research*, 22(2), 539–559. https://doi.org/10.1007/s10660-020-09439-8

Giles, J. (2013). The truth behind the big data hype. *New Scientist*, *217*(2905), 48–49. https://doi.org/10.1016/s0262-4079(13)60507-2

Hall, A. A., & Wright, C. S. (2018). Data security: A review of major security breaches between 2014 and 2018. Retrieved July 8, 2022, from https://fbdonline.org/wp-content/uploads/2021/02/2018-V6.P50-63-Data-Security-Breaches.pdf

Hochheiser, M. (n.d.). The truth behind data collection and analysis, 32 J. Marshall J. Info. tech and privacy L. 33 (2015). UIC Law Open Access Repository. Retrieved July 8, 2022, from https://repository.law.uic.edu/jitpl/vol32/iss1/3/

Jung, K. (2021). Extreme data breach losses: An alternative approach to estimating probable maximum loss for data breach risk. *North American Actuarial Journal*, *25*(4), 580–603. https://doi.org/10.1080/10920277.2021.1919145

Kaur, M., Vashisht, S., & Saurabh, K. (2012). Adaptive algorithm for cyber crime detection. *International Journal of Computer Science and Information Technologies*, 3(2012), 1–3. https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.438.1130&rep=rep1&type=pdf

Kröger, J. L., Miceli, M., & Müller, F. (2021). How data can be used against people: A classification of personal data misuses. SSRN Electronic Journal. https://doi.org/10.2139/ssrn.3887097

Nickerson, C. (2022, March 3). Latour's actor network theory. SimplySociology. Retrieved August 4, 2022, from https://simplysociology.com/Actor-Network-Theory.html

Norian, P. (n.d.). The struggle to keep personal data personal: Attempts to reform online privacy and how Congress should respond. *CUA Law Scholarship Repository*. Retrieved July 8, 2022, from https://scholarship.law.edu/lawreview/vol52/iss3/8/

Rusted, M. L., & Koenig, T. H. (2019). Towards a global data privacy standard - University of Florida. Retrieved July 8, 2022, from https://scholarship.law.ufl.edu/cgi/viewcontent.cgi?article=1446&context=flr

Sen, R., & Borle, S. (2015). Estimating the contextual risk of data breach: An empirical approach. *Journal of Management Information Systems*, *32*(2), 314–341. https://doi.org/10.1080/07421222.2015.1063315

Shackelford, S. (2017). Exploring the shared responsibility of cyber peace: Should cybersecurity be a human right? *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.3005062

Stansberry, K., Anderson, J., & Rainie, L. (2020, July 9). The internet will continue to make life better. *Pew Research Center: Internet, Science and Tech.* Retrieved August 1, 2022, from https://www.pewresearch.org/internet/2019/10/28/4-the-internet-will-continue-to-make-life-better/

Strawser, B. J., & Joy, D. J. (2015). Cyber security and user responsibility: Surprising normative differences. *Procedia Manufacturing*, 3, 1101–1108. https://doi.org/10.1016/j.promfg.2015.07.183