Cybersecurity: Protecting Genomic Data by Improving the Security Hygiene of DNA Processing Programs and Databases
(Technical Paper)

A Sociotechnical Analysis of Consumer Genetic Testing on The Understanding of Privacy
(STS Paper)

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
Stephanie M Skahen

October 27, 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

Briana Morrison, PhD, Department of Computer Science

Bryn E. Seabrook, PhD, Department of Engineering and Society

Introduction

Utopias, or often 'Dystopias' have been a common element of Science Fiction media for decades. Many of these Dystopian stories such as: *Gattaca, Unwind,* and *The Giver*, base their societies on the superiority of certain people's genetics. *Gattaca*, for instance, bases social standing, job prospects, and dating options solely on genetic makeup. Much of American popular culture portrays the use and knowledge of genetic information in negative lights and is used for the discrimination of the genetically inferior in the societies.

Currently, the consumer genetic testing market, which includes companies such as 23AndMe, is expected to grow by 12.5% every year according to a Market Research Report from Souring Intelligence, LLP (2021). The growing use and storage of genetic data combined with the connotations presented in literature and films, prompted me to begin thinking about privacy. Privacy is difficult to define. NIST defines the term in a cybersecurity sense as the "assurance that the confidentiality of, and access to, certain information about an entity is protected" (NIST, n.d.). Determining the general perception surrounding whether genetic information in databases is considered private is the inspiration for the STS research paper portion of this prospectus. Determining the current level of security that the databases containing this information have and how to improve that security is the topic for the Technical Report.

Technical Report

With the rise of DNA testing sites such as 23andMe, a high volume of genetic information is being processed, stored, accessed, and shared on the internet. Without proper protections and security practices, this increases the risk that genetic information containing an individual's health conditions, medical risk factors, and other potentially damaging information can be taken and used with ill intent. As previously mentioned, movies like *Gattaca,* show many

of the possible negative uses of genetic information in healthcare, jobs, and overall societal standing. While *Gattaca* is one, extreme negative use of genetic data, it prompts concern over other potential consequences to individuals if their genetic data is exposed.

Databases have been hacked in the past, in 2020, the online DNA database GEDmatch was hacked altering the privacy settings of users (Mullin, 2020). Hacks like this mean that the GEDmatch database is vulnerable to attacks, and likely, so are others. One way to decrease the likelihood of another leak, is to improve database security hygiene by creating policies and mandates in line with common security hygiene practices. This report will synthesize the knowledge learned in CS 3710: "Introduction to Cybersecurity" and CS 4640: "Database Systems" to produce an initial assessment of the hygiene status of 23andMe, Ancestry.com, and GEDmatch; and propose solutions to the vulnerabilities found. Introduction to Cybersecurity will be used to aid in the terminology, types of attacks, and identifying the different types vulnerabilities that exist within these databases. Database Systems will be used to associate those vulnerabilities with database security best practices in order to synthesize a solution.

The research will be conducted using the methods Discourse Analysis and Policy Analysis. The policy analysis will be used to analyze existing restrictions and requirements for genetic database security as well as individual companies' privacy policies. The discourse analysis will aid with the synthesis of information about previous breaches, penetration tests, and database security sources.

STS Research Paper

In recent years, there has been the development of DNA sequencing for the everyday person. While this sequencing is primarily used to inform the individual of their ancestral heritage, there have been instances of that data being used to solve crimes. GEDmatch, FamilyTreeDNA, and MyHeritage were used by law enforcement and private companies to find and apprehend the Golden State Killer, Joseph DeAngelo, in 2018 (St. John, 2020). This event prompted the question: How could the widespread documentation and storage of genomic data and the sociotechnical systems associated with it, impact public understanding and perception of privacy in the United States?

To answer the research question above, Sheila Jasanoff's framework of Co-Production will be used. Jasanoff defines the framework on her website as " scientific ideas and beliefs, and (often) associated technological artifacts, evolve together with *representations*, *identities, discourses,* and *institutions* that give practical effect and meaning to ideas and objects" (Jasanoff, n.d.). Co-production is divided into *constitutional* analysis and *interactional* analysis. Constitutional analysis deals with the "emergence of new socio-technical formations" and "seeks to account for how people perceive elements of nature and society". Interactional analysis pertains to the "conflicts within existing formations" and "how we know" not what we know (Jasanoff, 2004, p.15-19). These two aspects will be used to discuss the current understanding of privacy under the current sociotechnical system, as well as the emerging understanding and discourse with the emerging sociotechnical system associated with genetic data. The analysis will be constructed as a discussion privacy in terms of the four themes mentioned in the definition. One of the limitations of this framework proposed by a collection of STS, politics, and environmentalism professors, is that Co-production often "fails to adequately account for power within science-society relationships" and can inadvertently reinforce the power of policy

elites and marginalize those with alternate perspectives (Wyborn et al., 2019, p. 323). In order to mitigate the effect of this limitation, a large range of stakeholders will be considered in the analysis of the effect of the sociotechnical systems involved with genomic data.

A discourse analysis method will be used to answer: How could the widespread documentation and storage of genomic data and the sociotechnical systems associated with it, impact public understanding and perception of privacy in the United States? An initial search for background sources like news articles and blog posts will be conducted using Google and keywords such as "genetic data", "privacy", and "genomic databases". Many news articles reference scholarly articles pertaining to the topic, which will be analyzed as other potential sources. News articles and blog posts will also provide insight into the societal aspect of the sociotechnical systems. An analysis of scholarly articles will supplement the non-traditional sources. These articles will be found using the database "Web of Knowledge, Web of Science" and other databases offered by the University of Virginia Library consisting of studies conducting public surveys, papers synthesizing perspectives, and articles. The research will be ordered thematically as well as between the constitutional and interactional components of the framework. The themes explored will be: anonymized genomic data use for research, database hacks/breaches, and genomic data use in forensics.

Conclusion

This prospectus outlines the research that will be conducted for two projects: a technical report and an STS research paper. The Theme connecting the two projects is genomic data privacy, and each project will explore a different aspect of privacy. The technical report will synthesize two courses I have taken at the University of Virginia, Introduction to Cybersecurity

and Database Systems, in order to determine consumer genetic testing companies' ability to keep information private and prevent leaks, as well as propose solutions to vulnerabilities. The STS research paper will use Sheila Jasanoff's theory of Co-Production in order to analyze the sociotechnical systems containing consumer genetic testing and how those systems impact public understanding of privacy. The goal of this project is to inform readers of the different perspectives on the privacy of genetic data as well as propose solutions to increase the security hygiene of consumer genetic testing websites and databases to protect people genetic data from unanticipated use.

References

ltd, R. and M. (2021, May). *Consumer DNA (genetic) testing market - forecasts from 2021 to 2026*. Research and Markets - Market Research Reports - Welcome. Retrieved October 24, 2022, from https://www.researchandmarkets.com/reports/5351043/consumer-dna-genetic-testing-market-forecasts?utm_source=GNOM&utm_medium=PressRelease&utm_code=9w5mxv&utm_campaign=1555375%2B-%2BWorld%2BConsumer%2BDNA%2B%28Genetic%29%2BTesting%2BMarket%2BReport%2B2021&utm_exec=chdo54prd

Jasanoff, S. (n.d.). *Co-production* [Sheila Jasanoff]. Retrieved October 31, 2022, from https://sheilajasanoff.org/research/co-production/

Jasanoff, S. (Ed.). (2004). *States of Knowledge* (0 ed.). Routledge. https://doi.org/10.4324/9780203413845

Mullin, E. (2020, July 30). *The Era of DNA Database Hacks Is Here*. OneZero. https://onezero.medium.com/the-era-of-dna-database-hacks-is-here-85a860190622

NIST. (n.d.). *CSRC - Glossery*. Information Technology Laboratory Computer Security Resource Center. Retrieved October 27, 2022, from https://csrc.nist.gov/glossary/term/privacy

St. John, P. (2020, December 8). *The untold story of how the Golden State Killer was found: A covert operation and private DNA*. Los Angeles Times. https://www.latimes.com/california/story/2020-12-08/man-in-the-window

Wyborn, C., Datta, A., Montana, J., Ryan, M., Leith, P., Chaffin, B., Miller, C., & van Kerkhoff, L. (2019). Co-Producing Sustainability: Reordering the Governance of Science, Policy, and Practice. *Annual Review of Environment and Resources*, *44*(1), 319–346. https://doi.org/10.1146/annurev-environ-101718-033103