Thesis Portfolio

Health Modeling Using Smart Device Data
(Technical Report)

Physiological Data Privacy in the Digital Age
(STS Research Paper)

An Undergraduate Thesis

Presented to the Faculty of the School of Engineering and Applied Science
University of Virginia • Charlottesville, Virginia

In Fulfillment of the Requirements for the Degree
Bachelor of Science, School of Engineering

Aldrick Johan
Spring, 2021

Department of Computer Science

Health Modeling Using Smart Watch Data

Physiological Data Privacy in the Digital Age

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

Technical Project Team Members
Aldrick Johan

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

Signature  Date: 12/04/2020

Approved _____ Date _____
Department of

Approved _____ Date _____
Department of Engineering and Society

Physiological Data Privacy in the Digital Age

Introduction

Due to the severity of the COVID-19 crisis, governments have been scrambling to find any means to slow down or prevent any spreading of the disease. They have used various methods to hold back the disease such as quarantining, limiting travel, and even using apps for contact tracing and detecting the illness. This contact tracing is accomplished by collecting physiological data from mobile phones and wearable devices. Anybody who uses their smartphone or wears a "smart" device is subject to this collection of data. This data is collected by large companies employed by governments to supposedly detect signs of the illness. However, the collection of this data can violate a person's privacy or be used malevolently.

For my STS research I will be exploring the concept of privacy in relation to physiological data collection. As wearable devices become more common, the topic of physiological data collection will become more crucial. I will be exploring the topic using a series of targeted questions and by reviewing various sources as outlined in my STS Prospectus. My technical research is connected to this STS topic of physiological data collection. For my research, I will be exploring the use cases of the personal health data that is collected. One of the supposed benefits of collecting this information is that it can detect illnesses in users. I will be verifying whether this is true by using machine learning on data collected by wearables to see what useful information can be determined about a user. I will also be exploring the bounds of what data is available and how personal that data is.

Technical Topic

(Project assigned by your home department)

For my technical research I will be working on extracting useful information, such as a user's current mental state or detecting illnesses, from data that is collected from the user's wearable devices. To perform the research, I will primarily be using machine learning techniques to identify patterns that may be indicative of certain physical or mental states. Multiple different machine learning models will be put together and trained using the user data. I will be using different types of models including regression models, classification models, clustering models, and potentially even neural network models. Machine learning models are intrinsically designed to be more favorable for some tasks over other tasks. Because of this, I will be exploring further into each model type based on how preliminary experiments go. If one group of models does not perform well with the data, I will move away from those models and use better suited models.

The required data to perform the research is user health data that is collected from wearable devices and data pertaining to the user's current mental and physical health. The health data will be sourced from participating users' devices. The data pertaining to the mental and physical health of the user will be acquired by having users fill out a short survey or questionnaire at regular intervals. Another option to collect this data will be to conduct interviews, however this may be too time consuming and may be reserved for special cases. Using the surveys, the health data collected from the wearable devices can be labelled. Labels are used in machine learning to run supervised learning models. Many of the models that I will be exploring are considered supervised learning models so this labelling is necessary. Some labels that may be used are mentally healthy, mentally unhealthy, physically healthy, and physically unhealthy. These categories can be further broken down into specific illnesses, but

this option will only be pursued if there is enough data to enable it. There is one more option for sourcing the required data. Due to time and resource constraints, I may not be able to collect enough data to conduct the research. If this happens, I will look to already published datasets. These datasets are found on public databases and are made for free use. If needed, these databases will be searched for data that can be used to further this project.

I am currently planning to complete this research next semester. I will be outlining my timeline based on a 15-week college semester. For the first 5 weeks I expect to be collecting the data. This will involve collecting health data from users' devices and conducting the survey. I will evaluate the amount of data collected around 3 weeks into the process and if there is not enough information, I will check public datasets. For the next 5 weeks I will be building and testing various machine learning models. Based on initial performance, I will be further exploring some of the models. During this time period I will be working on tuning these models to be as accurate as possible. For the last 5 weeks, I will be aggregating the results and writing a formal report about my findings.

STS Prospectus

Introduction

The COVID-19 virus has surged throughout the world, forcing governments to employ various methods to slow its spread. One of these methods is using contact tracing apps to track which people have been in contact with each other. To achieve this, data must be collected from people's smart devices. The collection of this data has many social ramifications. It normalizes the monitoring of the general public, while also desensitizing them from losing their privacy. The collection of the physiological data would serve to contribute to this and could lead to a culture that believes it is normal to be monitored.

While this idea is alarming from the perspective of a Western society, other societies view this data collection as a necessity. As the world gets further into this digital age, the concept of privacy is evolving. Almost every person owns smart devices that collect information on themselves. Some societies believe this is a positive change because the data can be used for the public good, especially during a health crisis such as the COVID-19 pandemic. It can be used to detect the illness in individuals and can identify at-risk individuals. As the number of cases across the world increases, this could be one of the many possible solutions to slowing these cases down.

Research Question

The question of data privacy in regards to physiological data collection does not have a simple answer. Many aspects of the problem will have to be explored. To what extent should this data be collected? Who should have access to this data? What exactly should the data be used for? What defines privacy, especially during this digital age? I will study these questions by

exploring what data is already being collected and how that data is being used. I will do research on the prevalent attitudes towards privacy and data collection from the perspective of different cultures. This will allow me to explore how the concept of privacy is constructed and how it is changing.

Literature Review

I examined prior works of literature to provide context regarding the topic of physiological data collection. I began by reading articles that discussed how the COVID-19 pandemic impacted physiological data privacy. Then I explored the different viewpoints on the topic and why they exist. Finally, I outlined my data collection plan.

One of the major topics in the literature was how data collection should be handled during a global health crisis. The most prevalent school of thought on this problem is to accept the tradeoff between personal privacy and information, in the name of public health. This ideology posits that the cost of not collecting this data is far greater than the loss of some personal privacy (Cho et al., 2020). They argue that although some privacy is lost, the amount of personal information that is collected is minimal and is only revealed to authorized personnel. This ideology acknowledges that the current methods are not perfect, but they will improve as the pandemic progresses and that the information gained will be very valuable in terms of public health. The flow of the data would be as secure as possible using modern data transfer techniques and would anonymize as many parties as possible (Beskorovajnov et al., 2020). Using these methodologies would make it harder for harmful entities to attack them and serves to minimize the risk of potential attacks. These methods would also ensure that not much personal information would be revealed even if an attack was successful (Liu & Sun, 2016). The data will be collected from users' devices then would be sent to secure

databases using the described technologies. Once it reaches these databases, it can be processed and used to trace the spread of the virus.

Proponents of this ideology also argue that collecting this data would not cause significant changes to personal privacy, because it has already been impacted by the rise of big data (Perera et al, 2015). The collection of large amounts of data on the public, also known as big data, has been occurring since Internet of Things (IoT) devices have become more commonplace. These devices include cellphones, smart watches, medical sensors, and other devices that can connect to the internet. These devices collect metrics from their users and it is argued that true personal privacy is no longer possible due to them. When big data first became popularized many people took issue with it because of weak infrastructure and security (Terzi et al., 2015). However, many of these concerns have been addressed and infrastructures are now secure enough to prevent leaking of the data. Furthermore, the majority of the data will be useless to attackers as the data will be encrypted within the database.

Another recurring theme in the literature was the difference in global viewpoints when it came to data sharing and privacy. Privacy seemed to have a different definition depending on where people were from. Although many governments across the world already collect information on their citizens, the lack of global data sharing standards prevented mass cooperation amongst multiple governments (Allam & Jones, 2020). Many countries have “smart cities” where there are copious IoT networks. This allows the government to collect detailed information on their citizens. During a global crisis, it would be mutually beneficial for countries to share relevant data with each other. Unfortunately, there is no global consensus on how to achieve this. The World Health Organization (WHO) implemented some policies to solve this problem, however there are many doubts whether the data shared with WHO is accurate. Another aspect that is holding back global

participation with these programs is the difference in viewpoints concerning personal privacy. In the United States, personal privacy is considered to be important to the people (Bellman et al, 2010). This is due to the individualistic culture that has developed within the country. People from the United States have always been more focused on their own personal rights and have been wary of the government. This culture has led to citizens wanting the most secure data regulation while also collecting as little as data as possible (Ballman et al, 2004). Since the United States is an influential nation, other countries have also followed suit with restrictive data regulations including the United Kingdom. These policies make it difficult for global collaboration to grow especially during a pandemic and prevent the implementation of systems that would help prevent future outbreaks.

An additional trend I noticed while reading the literature was that many people still bought and used smart devices, knowing that it would affect their personal privacy. Consumers know that these devices have sensors and can connect to the internet without their permission (Perez & Zeadally, 2018). The sensors allow the producers of these devices to collect personal information on the users. The internet functionality of the devices allows them to transmit user data to other parties. While this may be alarming to some, others said that they had "nothing to hide" and were not concerned with their data being collected (Udoh & Alkharashi, 2016). This mentality was surprising as the data could be used to track a person's location in real time and to get a detailed look into their lives. This carefree attitude towards data collection may stem from these people growing up with devices and perceiving the data collection as inevitable. Fortunately, there are new data collection methods that manage to preserve privacy while getting the necessary data (Kim et al, 2019). These methods can help assuage fears about loss of privacy and can lead to the normalization of data collection.

STS Framework

There are various factors that both affect and are affected by the technologies that are relevant to physiological data privacy. These factors, known as actors, have varying levels of influence on the other actors. This influence is known as agency in the Actor Network Theory (ANT) and any actor with agency is known as an actant. By using the ANT framework, the various actants and their agency can be analyzed to see how wearable smart devices reached their current state.

The major actors in this network are governments, public policies, wearable device producers, the wearable devices themselves, the collected data, and the users of the devices. In this network the governments are connected to public policies because they create the laws. The policies are then applied to the remaining actants. This provides the governments with agency over the public policies and the public policies with agency over the producers, the devices, the data, and the users. The producers are connected to the wearable devices and the data collected by them. The producers have significant agency over both of these actants as they design the devices and how the data will be collected and stored. The devices have agency over the data and the users of the device because they interact directly with them. The data has ties with both the users and the government. The data has some agency over governments because if this data leaked it could be potentially be used as intelligence against the government. It has agency over users because users want their data to be private and would be upset if this data was leaked. The users have agency over the producers of the devices and the government. The designers utilize user feedback to improve their product going forward. Users have agency over the government as they can influence what topics are addressed in new laws. Users can also vote for new leaders in many countries so the government officials want to keep those users happy.

The actants in this complex web interact with each other, indirectly causing many different outcomes that eventually lead to major changes within the network.

The ANT framework also includes a concept known as translation. This occurs when actants are displaced and transformed to fit into the network. A successful translation in this network occurs between the wearable devices and the users. Before the translation the actors are two independent parties: a wearable smart device and potential consumers in the market. The wearable smart device then recruits the potential consumers and makes them users of the device. When this occurs, the consumers become a data collector who impact the network. As a data collector they have agency over the collected data because they are enabling the collection of the data. This alters the overall dynamic of the system by pushing the focus from the means of collecting the data to how the data should be used. This will eventually bring in new actors and actants to the system who will also alter the system. This concept of translation can also be used to describe the transformation of the concept of privacy. Based on the culture of a society, the actors in their respective networks will displace and transform the concept of privacy until a definition that is acceptable to the society is reached.

Method

For my data collection, I plan to utilize surveys, interviews, and document review. I will use surveys to collect information from large amounts of people. I will ask questions pertaining to the topic of personal privacy and data collection. I will target both a younger age group and an older age group to see if different generations have different views on privacy. The answers will either consist of a rating system between 1 and 7 or a short response between 15 to 30 words. The survey will have around 10 to 15 questions to get as many participants as possible. The survey will be done

using Google Forms because it is difficult to distribute physical surveys during this pandemic. The main intent of the surveys is to get the overall consensus on the research topic. I will use interviews to hear arguments from people who support one side strongly. The interviews will have less questions than the survey, but they will have longer answers. I will be asking each person the same questions and will conduct the interviews over phone call. I will only conduct a handful of interviews, likely only 3 to 5, because they are more time consuming than the surveys. I will also be reviewing documents to learn more about the topic. It would be quite difficult for me to get ahold of a representative from a large corporation, so I will be instead be reading official documents that they release that pertain to the research topic. These documents will reveal insights about what these companies think about these topics and will be a valuable source of information for my thesis.

Timeline

I plan to complete my thesis next semester. This timeline is based on the 15-week college semester. For the first 5 weeks I will be collecting information. This will involve using surveys, conducting interviews, and reviewing documents. For the next 5 weeks I will process the collected information write a rough draft of my thesis. I will use the results of the survey to draw conclusions about the general mentality of the people I surveyed. I will also utilize the important parts of my interviews and the document review. Over the last 5 weeks I will refine the thesis using any new information I find and will continue improving it until I submit it. During this entire process I will be working on the technical research. This research will inform me as I work on my thesis.

Conclusion

This prospectus outlines the topic of my thesis and how I will go about writing it. I will be focusing on the issue of data privacy in relation to wearable devices and how that ties into the ongoing COVID-19 pandemic. I will be doing research on prominent schools of thought regarding this topic and will be collecting information from primary sources through the use of surveys and interviews. This topic is also related to my technical research, which I will be partaking in while writing my thesis.

Bibliography

Allam, Z., & Jones, D. (2020). On the Coronavirus (COVID-19) Outbreak and the Smart City Network: Universal Data Sharing Standards Coupled with Artificial Intelligence (AI) to Benefit Urban Health Monitoring and Management. *Healthcare*, *8*(1), 46. https://doi.org/10.3390/healthcare8010046

Bellman, S., Johnson, E., Kobrin, S., & Lohse, G. (2004). International Differences in Information Privacy Concerns: A Global Survey of Consumers. *The Information Society*, *20*(5), 313-324. https://doi.org/10.1080/01972240490507956

Beskorovajnov, W., Dörre, F., Hartung, G., Koch, A., Müller-Quade, J., & Strufe, T. (2020). *ConTra Corona: Contact Tracing against the Coronavirus by Bridging the Centralized-Decentralized Divide for Stronger Privacy* [Ebook]. Retrieved 25 October 2020, from https://eprint.iacr.org/2020/505.pdf.

Cho, H., Ippolito, D., & William Yu, Y. (2020). *Contact Tracing Mobile Apps for COVID-19: Privacy Considerations and Related Trade-Offs* [Ebook]. Retrieved 26 October 2020, from.

Kim, J., Lim, J., Moon, S., & Jang, B. (2019). Collecting Health Lifelog Data From Smartwatch Users in a Privacy-Preserving Manner. *IEEE Transactions On Consumer Electronics*, *65*(3), 369-378. https://doi.org/10.1109/tce.2019.2924466

Liu, J., & Sun, W. (2016). Smart Attacks against Intelligent Wearables in People-Centric Internet of Things. *IEEE Communications Magazine*, *54*(12), 44-49. https://doi.org/10.1109/mcom.2016.1600553cm

Perera, C., Ranjan, R., Wang, L., Khan, S., & Zomaya, A. (2015). Big Data Privacy in the Internet of Things Era. *IT Professional*, *17*(3), 32-39. https://doi.org/10.1109/mitp.2015.34

Perez, A., & Zeadally, S. (2018). Privacy Issues and Solutions for Consumer Wearables. *IT Professional*, *20*(4), 46-56. https://doi.org/10.1109/mitp.2017.265105905

Terzi, D., Terzi, R., & Sagiroglu, S. (2015). A survey on security and privacy issues in big data. *2015 10Th International Conference For Internet Technology And Secured Transactions (ICITST)*. https://doi.org/10.1109/icitst.2015.7412089

Udoh, E., & Alkharashi, A. (2016). Privacy risk awareness and the behavior of smartwatch users: A case study of Indiana University students. *2016 Future Technologies Conference (FTC)*. https://doi.org/10.1109/ftc.2016.7821714