ABSTRACT

Title of Dissertation:	QUANTIFYING AND PREDICTING USER REPUTATION IN A NETWORK SECURITY CONTEXT
	Margaret Stephanie Gratian, Doctor of Philosophy, 2019
Dissertation directed by:	Associate Professor Michel Cukier, Department of Mechanical Engineering

Reputation has long been an important factor for establishing trust and evaluating the character of others. Though subjective by definition, it recently emerged in the field of cybersecurity as a metric to quantify and predict the nature of domain names, IP addresses, files, and more. Implicit in the use of reputation to enhance cybersecurity is the assumption that past behaviors and opinions of others provides insight into the expected future behavior of an entity, which can be used to proactively identify potential threats to cybersecurity. Despite the plethora of work in industry and academia on reputation in cyberspace, proposed methods are often presented as black boxes and lack scientific rigor, reproducibility, and validation. Moreover, despite widespread recognition that cybersecurity solutions must consider the human user, there is limited work focusing on user reputation in a security context. This dissertation presents a mathematical interpretation of user cyber reputation and a methodology for evaluating reputation in a network security context. A user’s cyber reputation is defined as the most likely probability the user demonstrates a specific characteristic on the network, based on evidence. The methodology for evaluating user reputation is presented in three phases: characteristic definition and evidence collection; reputation quantification and prediction; and reputation model validation and refinement. The methodology is illustrated through a case study on a large university network, where network traffic data is used as evidence to determine the likelihood a user becomes infected or remains uninfected on the network. A separate case study explores social media as an alternate source of data for evaluating user reputation. User-reported account compromise data is collected from Twitter and used to predict if a user will self-report compromise. This case study uncovers user cybersecurity experiences and victimization trends and emphasizes the feasibility of using social media to enhance understandings of users from a security perspective. Overall, this dissertation presents an exploration into the complicated space of cyber identity. As new threats to security, user privacy, and information integrity continue to manifest, the need for reputation systems and techniques to evaluate and validate online identities will continue to grow.

QUANTIFYING AND PREDICTING USER REPUTATION IN A NETWORK SECURITY CONTEXT

by

Margaret Stephanie Gratian

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, College Park, in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
2019

Advisory Committee:
Associate Professor Michel Cukier, Chair
Professor Jennifer Golbeck, Dean’s Representative
Assistant Professor Mark Fuge
Assistant Professor Monifa Vaughn-Cooke
Assistant Professor Katrina Groth
Josiah Dykstra, PhD

Dedication

To my parents, Joe and Majda, and my brother, Nicholas.

Acknowledgements

First, I would like to express my sincere gratitude to my advisor, Dr. Michel Cukier, for his guidance, encouragement, and support. I never could have imagined working as a teaching assistant at one of your cybersecurity camps would be the catalyst for my academic and professional career. I am incredibly thankful for having met you and look forward to future collaboration.

Thank you to Janelle for taking a chance on me and introducing me to this topic, making this research and degree possible in the first place. Words cannot express how grateful I am for all that you have done for me.

Thank you to Dr. Josiah Dykstra for serving as a mentor, friend, and special member of my committee. You provided a unique perspective that helped shape the direction of this research.

Thank you to the security team in the Division of IT - Amy, Bertrand, Ed, and Gerry. It was a pleasure to collaborate with you all and I am grateful for your support. I will miss seeing you all regularly at our weekly research meetings.

Thank you also to my friends, Angela, Caleb, Jacob, Jake, Kat, Megan, Michael, Michaela, and Sam. You kept me sane, gave me strength, and filled my rare free moments with happiness, laughter, and love. I am truly lucky to know all of you.

Finally, I would like to thank my family for their love and support. To my brother, for providing humor, encouragement, and reminders that it was all going to be easy peasy lemon squeezy (it wasn’t, but I appreciated the vote of confidence nonetheless). And to my parents - you, of course, are the foundation upon which all of this was built. Thank you.

Table of Contents

List of Tables

List of Figures

1 Introduction

1.1 Background and Motivation

Reputation has long been an important factor for establishing trust and evaluating the character and quality of others. Today, efforts to evaluate, quantify, predict, manage, and repair reputation are pervasive in nearly every industry. Companies and institutions are given star ratings and reviews on e-commerce platforms based on customer experiences. Countries are given reputation scores, such as the Reputation Institute’s annual RepTrack score, which indicates the effectiveness of a country’s government, the appeal of its environment, and the state of its economy using a score on a 0 to 100 scale [2]. In the United States, individuals are given financial reputation scores in the form of FICO credit scores on a 300 to 850 scale based on features such as payment history, debt, and length of credit history [1]. In 2015, Ant Financial Services Group launched a similar credit score in China called Zhima Credit, which assigns individuals credit scores between 350 and 950 based on credit history, personality traits, interpersonal relationships, Internet activity, and a variety of other aspects of an individual's identity; by 2020, the Chinese government plans to have fully implemented a far more expansive credit system that will quantify an individual’s social reputation [105]. Implicit in the widely accepted importance and use of reputation is the assumption that insights gleaned from an entity’s past behaviors, based either on evidence or the judgments of others, can provide meaningful insight into the entity’s nature and its expected future behaviors.

Reputation can have substantial impact: whether described in a qualitative manner or scored in a quantitative manner, reputation may be used to reward good behavior or punish bad behavior, enforce desired behaviors, predict future behaviors, or foster new relationships or opportunities. Consumers make heavy use of star ratings when determining whether or not to buy a company's goods and services; research suggests that the difference in a star rating on the review site Yelp can equate to a 5 to 9% difference in yearly revenue [3]. The Reputation Institute claims that having strong reputation, as defined by the RepIndex score, makes a country better poised to attract a highly skilled workforce, enhance tourism, and encourage purchases of the country's exports [2]. FICO credit scores are used to measure a lender's financial risk of giving a loan to an individual; a low credit score can prevent an individual from receiving the credit to make large purchases, which in turn motivates many individuals to exercise financial responsibility and caution. China's social credit system rewards people deemed trustworthy with financial perks, travel opportunities, access to good healthcare and schools, and so on; untrustworthy people face consequences ranging from denial of a loan or a flight to jail time [105].

Over the last couple decades, reputation has emerged as an important factor in the field of cybersecurity. Originating as a vital component for building trust and mitigating dishonest behavior on e-commerce websites [4, 5] and as a metric for identifying spam-sending emails and IP addresses [6, 7, 8], reputation is now quantified and evaluated for a range of cyber entities. There has been extensive work in academia and industry to evaluate the reputation of domain names [9, 10, 11, 12, 13, 14, 15]; IP addresses [16, 17, 18, 19]; infrastructure [20, 21]; files [22], and binaries [23]. These reputation assessments are a valuable component of cybersecurity and serve a variety of purposes, such as supporting identification of suspicious or malicious activity, enabling attribution of cyber attacks or anonymous behaviors to specific entities or advanced persistent threats (APTs), in addition to facilitating a wide range of other decisions and judgments about cyber entities.

Industry and academia have also explored assessing the reputation of users in online communities and social networks [24, 25, 26]. The United States' 2016 Presidential Election brought mainstream media attention to fake, automated, and malicious social media accounts; the concept of user reputation has unsurprisingly had a surge of attention in response. In 2018, the Washington Post reported that Facebook was developing a capability to assign users a reputation score on a zero to one scale indicating trustworthiness [106]; meanwhile, Twitter has been removing tens of millions of suspicious accounts from its platform [107].

Clearly, reputation - whether of users or infrastructure - can play a key role in enhancing cybersecurity on a computer network or online service. However, despite the plethora of available research and tools and the incorporation of reputation systems into many online platforms, reputation evaluation suffers from a lack of scientific rigor. For a given context there is no single recognized definition for reputation or framework for evaluating reputation. There is also little consensus across the many different definitions and scoring methodologies. Moreover, new approaches for assessing reputation are often constructed without regard to existing techniques, representing a major shortcoming in the field of online reputation systems and further preventing reputation scoring from becoming a reliable or scientific process [27]. Finally, reputation scores and evaluation methodologies are often highly reactive in nature, documenting only known incidents and failing to consider behaviors or warning signs that may be indicative of changes in an entity's reputation before these changes have actually manifested. Just as there is a need in the broader field of cybersecurity for foundational research contributing to the 'science of security' [28], in order for reputation to become a reliable tool for improving the security of online networks and services, there is similarly a need for foundational research contributing to the 'science of reputation.' Reputation may be a subjective concept amongst humans, but when used as a metric in a system it should be an objective quantity.

1.2 Research Scope

Though there are many different types of cyber entities that can be evaluated for reputation, this dissertation will focus on quantifying and predicting the reputation of human users on a computer network or Internet service. In this dissertation, a *user* is considered to be an account or collection of accounts on a computer network or Internet service associated with a unique human who exists in the physical world, as shown in Figure 1. Accounts or personas that represent businesses, institutions, or governments or are created for purposes such as entertainment or marketing and do not represent a real individual are not considered users and are therefore outside the scope of this research.

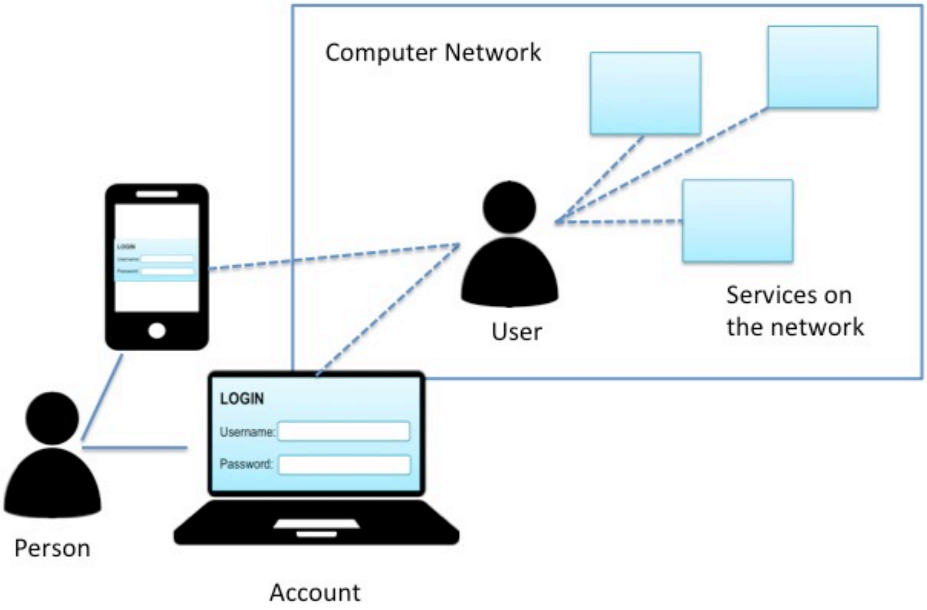

Figure 1: The mapping of a person in the physical world to a user in cyberspace

It is assumed that users must authenticate onto the network or service so there is ground truth regarding the mapping of a user account to activity on the network. However, depending on the nature of the network or service, there may be several challenges associated with defining a user as an account or collection of accounts. First, there is the challenge of correctly associating an account with a user and ensuring it is not another user with a similar or identical name or alias. Second, there is the challenge of ensuring all activity stemming from an account belongs to the true account owner, i.e. the user's credentials have not been stolen or shared and no one is attempting to impersonate the user. Finally, there is the challenge of identifying all the accounts belonging to a user: depending on the type of network or service, users might have multiple accounts and may or may not use their real name or identity on these services. Depending on the challenges associated with a particular use case, there may be some uncertainty regarding the identity of a user; this is a limitation of trying to evaluate users in a cyber environment. However, even if interpreting a user as an account creates some uncertainty, it enables the network administrator to evaluate the point of access used to conduct activity on the network.

A user's online reputation, denoted their *cyber reputation,* is intended to provide insight into the user's propensity for exhibiting some characteristic of interest to a network administrator for the purpose of improving cybersecurity. This cyber reputation is based on the behaviors the user has demonstrated in cyberspace and individual human traits of the user that impact their behaviors in cyberspace, as shown in Figure 2. Cyber reputation is used for two main purposes: quantifying the reputation of users on a computer network and predicting the future reputation i.e. behaviors of users on a computer network.

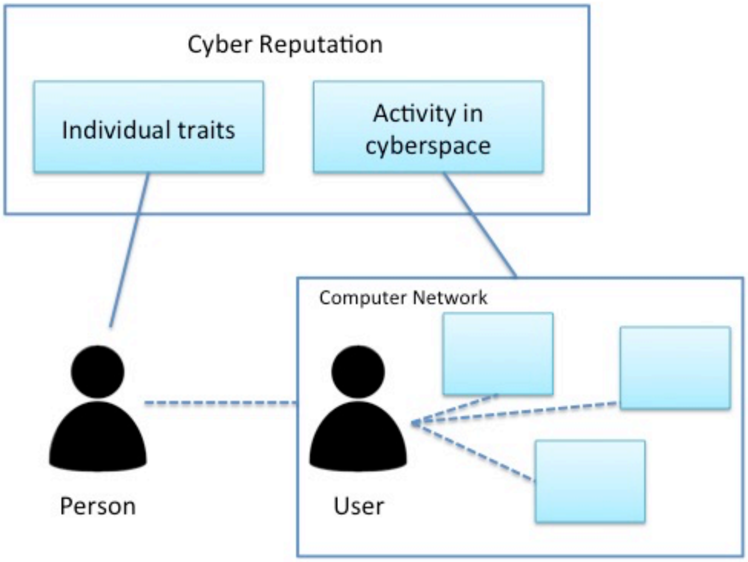

Figure 2: Elements that form user cyber reputation

A user's cyber reputation may be a set of behaviors, represented as numerical features in a vector denoted the *cyber reputation vector*. This vector is intended to provide a comprehensive picture of behaviors a user has demonstrated that are indicative of a certain characteristic that may adversely impact cybersecurity. In order to ensure that the reputation vector is comprehensive, a variety of exploratory data analysis techniques must be conducted and the vector should be periodically reevaluated and updated with new data.

A user's reputation may also be a single number, denoted their *cyber reputation score,* that provides network administrators or security practitioners with a metric to evaluate a user or compare multiple users. Just as a FICO credit score takes a set of observations about a person's financial history and provides lenders with a metric for assessing the risk of granting the person a loan, a cyber reputation score is intended to condense a collection of observations about a user's past behavior into a summary that can easily be used to assess the risk a user might pose to a network.

Ultimately, the goal of both the cyber reputation vector and the cyber reputation score is to help network administrators or security practitioners identify users who are more likely than others to harm or degrade the quality of a network or have already exhibited behaviors that harm or degrade the network. Depending on the network or service, these reputation assessments may translate into actions taken by the network administrators or into privileges or restrictions imposed on users. For example, users with a reputation for being more susceptible to phishing attacks or exhibiting risky cyber behaviors may be given additional security training or may be denied access to more sensitive areas or services on the network. Similarly, users who behave selfishly and have a reputation for consuming more resources than others or users who behave maliciously and have a reputation for propagating malware may incur higher costs to use services on the network or may have their ability to interact with others restricted or their account suspended. This concept has been explored and implemented in both e-commerce and P2P network settings [29, 30].

1.3 Research Questions and Approach

The goal of this research is to develop a quantitative definition of user cyber reputation and a methodology for evaluating user cyber reputation that can be used in a network administration or security setting to objectively quantify how likely a user is to exhibit a characteristic that may threaten the network's security. The proposed methodology is intended to function as a type of *centralized* reputation system, in which there is some key entity, denoted the *central authority,* which has oversight of the network and the users who exist on the network and is responsible for calculating user cyber reputation.

The research questions for this dissertation are as follows:

- **Research Question 1 (RQ1):** What data can be used as evidence of a user's cyber reputation for specific characteristics or behaviors on a computer network?
- **Research Question 2 (RQ2):** How can this evidence be used to define and quantify a user's cyber reputation and develop predictive models of future reputation?

- **Research Question 3 (RQ3):** How can quantification assessments and predictions of a user's cyber reputation be validated and refined?

To address RQ1, we propose an experimental approach to identify correlations between observations about users and the characteristics or behaviors that they exhibit on the network. Observations about users are collected from the network by the central authority and then various feature engineering and dimensionality reduction techniques are conducted to transform this initial set of observations into numerical features stored in a feature vector of size n, denoted the *observation space.* Correlation analysis is performed between the features of the observation space and the characteristic for which user cyber reputation is being evaluated. Strongly correlated features are considered *evidence* and stored in a *cyber reputation vector,* of size $\leq n$.

To address RQ2, we propose interpreting a user cyber reputation score as the probability with which a user demonstrates some characteristic or behavior on the network. This probability is determined through the process of Bayesian Inference, in which an initial probability or probability distribution is formed to capture the central authority's subjective belief about the nature of the user and then updated using the evidence obtained in the experiments for RQ1. In other words, a user's cyber reputation is the probability a user exhibits a characteristic, based on past behaviors on the network and the judgment of the central authority.

Finally, to address RQ3, we propose making initial assessments and predictions about user reputation and then experimentally validating these assessments and predictions by monitoring their accuracy over time. Additionally, addressing this question will require sensitivity analysis on the parameters of any user reputation models; for example, we may vary the amount of data, time period of data collection, or sources of data for which initial assessments and predictions are made and then evaluate if accuracy is impacted.

1.4 Contributions

From a network security and administration perspective, understanding the cyber reputation of users is particularly important due to the now widespread recognition that humans are the weakest link in the cybersecurity chain. In recent years, interest in better understanding users and user behavior to tailor and strengthen cybersecurity has increased. We hypothesize that cyber reputation assessments of users on a network, based on their past activities on the network, may provide a new and important angle from which to study users, similar to how the reputation of technical entities has become an invaluable tool for proactive identification of security threats. By understanding the reputation of users on a network, a network administrator may be able to proactively identify users who are more likely to introduce threats to the network, become victims of cyber attacks, or intentionally harm the network. Understanding the reputation of users in the context of network security also represents a gap in the literature, as the majority of work in the network security domain has focused on the reputation of network and Internet infrastructure, end hosts and devices, and various other technical cyber entities. And, as is the case with cyber reputation for many other types of cyber entities, there is no widespread, accepted definition for user reputation, nor is there an existing scientific methodology for evaluating or quantifying user reputation.

The contributions of this research are as follows:

- We introduce a mathematical interpretation of user cyber reputation for network security settings that is grounded in statistical theory and existing reputation literature.
- We present a methodology for evaluating a user's cyber reputation that enables a network administrator to identify evidence that can be used to quantify cyber reputation, make predictions about future behaviors, assess the uncertainty associated with these predictions, and update cyber reputation assessments and predictions as new evidence becomes available.
- We illustrate and validate the proposed reputation interpretation and evaluation methodology through a case study at the University of Maryland in which we quantify the reputation of users in order to identify those with susceptibility to device infection or exploitation.
- Through a case study of account compromise incidents self-reported by users on Twitter, we explore social media as a source of data for studying users and evaluate the feasibility of using social media to extract insights into user behavior and demographics that can be used as evidence to quantify user reputation.
- Through the case studies of user infection on a network and user-reported compromise on Twitter, we contribute new understandings of users to the field of cybersecurity by identifying and evaluating features in a user's

network traffic and online activity patterns that may indicate an increased likelihood of victimization or propensity toward risky security behaviors. We also uncover trends in user victimization, identify particularly vulnerable populations based on the compromise data, and reveal startling user attitudes and misconceptions about cybersecurity.

1.5 Dissertation Outline

The remainder of this dissertation is organized as follows. Chapter 2 provides the background and related work. Chapter 3 presents a mathematical interpretation of user reputation based on Bayesian probability. Chapter 4 presents the methodology for evaluating and quantifying user reputation with the proposed definition. Chapter 5 presents a case study on a large university network that investigates the feasibility of extracting evidence of user reputation from network traffic and using it to differentiate between two different types of users. Chapter 6 extends this case study by showing how this evidence can be used to evaluate and predict user cyber reputation, where user cyber reputation is considered a measure of how likely a user is to be compromised ("infected") on the network. Chapter 6 also shows how these reputation scores can be validated and refined. Chapter 7 presents a case study exploring social media data as a source of information about user reputation, while Chapter 8 explores the feasibility of using this data to make predictions about users. Chapter 9 concludes this dissertation with a summary and discussion of the completed work, important considerations when quantifying user reputation, the limitations of this dissertation, and areas of future work.

1.6 Summary of Terminology

This section provides definitions for the terms used in the introduction and throughout the remainder of this dissertation.

- *User*: an account or collection of accounts on a computer network or Internet service associated with a unique human who exists in the physical world.
- *User Cyber Reputation*: a user's propensity for exhibiting some characteristic that is of interest to a network administrator for the purpose of improving cybersecurity.
- *Cyber Reputation Score*: a user's propensity for exhibiting some characteristic of interest to a network administrator for the purpose of improving cybersecurity, quantified as a probability reflecting the likelihood a user exhibits the characteristic.
- *Central Authority*: A key entity, in most cases the network administrator, who has oversight of the network and the users who exist on the network and is responsible for calculating and maintaining user cyber reputation scores.
- *Characteristic of Interest*: The quality for which user cyber reputation is being evaluated.
- *Observation*: A numerical feature summarizing some behavior a user has exhibited on the network or a human trait of the user.
- *Observation Space*: An n-dimensional vector containing all the observations for a user.
- *Evidence*: Observations that are strongly correlated with a specific characteristic of interest.

- *Cyber Reputation Vector*: A vector of size $\leq n$ containing all the evidence for a specific characteristic of interest.
- *User Reputation Model*: A function that is used to quantify or predict user reputation, such as a regression model or equation for combining sources of evidence into a final distribution.

2 Background and Related Work

2.1 Introduction

The meaning of reputation is highly context dependent and many different interpretations of reputation can be found in the literature. Though this dissertation is focused on quantifying user cyber reputation on a computer network, the literature review was conducted to understand how reputation is generally understood in the context of a computer network or Internet service and therefore covers existing work to assess the reputation of domain names, IP addresses, peers or nodes on a network, and users in a variety of online communities and services. The majority of work to assess online reputation can be broadly categorized as either an effort to distinguish between malicious and benign entities or an effort to gauge the trustworthiness of an entity, though there is work relating reputation to perceived influence, authority, and quality.

2.2 Reputation as a Measure of Maliciousness

In both academia and industry, it is common practice to assign reputation assessments to domain names and IP addresses. In [17] IP reputation is defined as the quality of content originating from an IP address and indicates whether the IP is sending solicited or unsolicited content. The authors in both [18, 19] applied a similar definition. In [9], the authors used domain reputation as a mechanism for predicting future malicious activity based on evidence from a domain's past behavior and its relationship to other domains based on DNS topology. Similarly, [10] used domain reputation to indicate a domain's history of suspicious activities and opinions held about the domain and used DNS traffic analysis to identify botnet domains. The authors in [31] determined an IP's reputation score based on the number of times it appeared on different blacklists. In [23], the authors created the reputation system EXPOSURE which used 15 different features derived from DNS traffic to categorize domains into categories such as 'spam,' 'risky,' 'adult,' 'Conficker,' 'malware,' 'blacklisted.' In [32], the authors considered domain reputation a binary classification problem to label sites as malicious or benign using lexical and host based features of domain names to detect malicious websites.

Some approaches base the reputation of a domain on its 'social network.' In [11], the authors classified websites with low reputation if they mapped to IP addresses found in IP address blocks commonly used by malicious actors or were registered through registrars commonly used by malicious actors. In [16], the authors applied a similar approach by considering 'bad neighborhoods' on the Internet: domains or IPs located within a certain geographic region or within the same IP address space as known malicious domains are considered to have the same reputation as the malicious domain. In [12], the authors used a graph inference system to predict a domain's likelihood of being malicious based on its connectivity to known malicious or benign domains. Based on this connectivity, domains were assigned numerical reputation scores in the [-127, 127] range, with [50, 127] indicating high-risk domains, [30, 49] indicating suspicious domains, [15, 29] representing neutral domains, and [-127, 14] representing good domains.

Other approaches have combined a domain's history and social network with Domain Name Server (DNS) information. In [13], the authors leveraged DNS registration information, appearance in DNS zone files, and similarity to known malicious domains to proactively blacklist domains. In [15] the authors built Notos, a dynamic reputation scoring tool, that assigns low reputation scores to domains involved in malware spreading, phishing, and spam campaigns and high reputation scores to domains used for legitimate Internet services. Notos used DNS registration information, DNS zone information, blacklist presence, and network behaviors to assign scores between 0 and 1 - with 0 being the lowest possible reputation and 1 being the highest - to unknown domains.

The definition of reputation as an indication of maliciousness is also applied when assessing the reputation of nodes in Peer-to-Peer (P2P) networks in order to determine whether a node has or will introduce malware or corrupted files to the network, behave selfishly by consuming excesive resources, purposely transmit incorrect data, or induce some other form of harm on the network or the entities on the network. In [33], the authors presented SocialTrust, a reputation system for P2P file sharing networks that used both relationships between nodes and ratings to distinguish between reputable nodes and malicious nodes. In [34], the authors introduced a distributed Bayesian-based reputation system that allowed nodes in a vehicular ad-hoc network to label their neighbors as either malicious or benign and predict their neighbors' future behaviors.

2.3 Reputation as a Measure of Trustworthiness

Understanding and evaluating trust on the web is an active area of research going back several decades. Trust is necessary for promoting and enhancing the quality and usefulness of Internet based platforms and is used extensively to evaluate services, content, and users [35]. Reputation and trust are often closely coupled, with reputation systems commonly serving as a method for assessing trust [36]. Measures of reputation for services, content, and users enable other entities to make informed decisions about whether or not to trust or interact with the service, content, or user in question.

The link between reputation and trust is evident in ecommerce and Internet transaction settings, social networks, and various other types of online communities. For example, in the context of e-commerce, the authors in [37] define reputation as "what is generally said or believed about a person's or thing's character or standing" and considers reputation a "collective measure of trustworthiness based on referrals or ratings from members in a community." In [37], the authors observed that trust reflects one entity's subjective view of another entity, while reputation reflects a community's perception of an entity. Similarly, [38] observed that reputation is more objective than trust because it represents the collective opinion of an online community, while [39] defined reputation as a "social notion of trust". Trust and reputation systems are critical in settings such as Ebay, Amazon, and various other commerce and review platforms because an entity's history is condensed into a value that allows others to make a judgment about expected future behavior [37]. In the context of social networks, a user's reputation for being trustworthy is often informed by their behaviors on the social network and the experiences of other users when interacting with the user in question. Reputation allows users to determine how much information to share, who to interact or exchange information with, and how to evaluate the information they are presented with [35, 36]. For example, in question and answer communities such as StackOverflow, reputation is considered “a rough measurement of how much the community trusts you” [40]. Reputation on StackOverflow is represented as a numeric score with no maximum value; users begin with a score of 1 and can earn up to 200 points per day.

Reputation as a measure of trustworthiness and reputation as a measure of maliciousness may also be closely coupled. The interconnected nature of reputation, trust, and predicting maliciousness is evident in many P2P settings. Though a goal of P2P network reputation systems is to identify malicious nodes, these reputation systems often work by collecting information about past behaviors of nodes on the network and then using these behaviors to make judgments of how trustworthy a node is in comparison to other nodes [29]. In some cases, trust may be measured based on the number of relationships and the strength of the relationships a node has with other nodes [29, 33]. Similarly, trust may be based on the ratings of other nodes, as in the well-known EigenTrust algorithm [41]. In cases where nodes are transmitting data or information - such as mobile sensing networks - trust may be determined by evaluating whether data sent by a node in the past turned out to be honest/dishonest or accurate/inaccurate [30, 34, 42].

2.4 Additional Definitions for Reputation

Reputation can also relate to the perceived authority, expertise, or influence of an entity. In academia, the H-index is a type of reputation measure used to “quantify an individual’s scientific research output” [43]. Google’s PageRank - the algorithm used to order webpages returned by a Google search - is also a form of reputation and attempts to objectively rank the relative importance of a web page [44]. This effectively provides users with a measure of the authority of a website on a given topic. Similarly, on the forum StackOverflow, though reputation is used to reflect the trust the community has placed in a user, it is gained by "convincing your peers that you know what you're talking about" and is gained by "posting good questions and useful answers" and therefore conveys notions of authority and expertise [40]. The authors in [39] observed that using reputation to vet the reliability of someone's statements or someone's expertise on a topic must be done in a context-specific manner. Similarly, in [38] reputation reflects a user's expertise on a topic and is considered relative to a specific context since expertise varies depending area of knowledge. By using reputation as a measure of expertise, it is effectively "summarizing the quality of information which a user produces" [38]. This introduces yet another definition of reputation - reputation as a measure of perceived authenticity. An active area of research in academia and industry is mitigating the risk of review fraud either by analyzing changes in an entity's ratings over time or assessing the reputation of the reviewers themselves [45, 46, 47, 48].

2.5 Summary of Reputation Interpretations

As is evident from the related work, reputation is defined and evaluated in a variety of different ways in the literature. Based on the literature, the reputation of a cyber entity may be interpreted in the following ways:

- An indication that an entity will exhibit malicious behaviors that may be damaging to other entities.
- A community's perception of an entity's typical or notable characteristics or behavior.

- A collective assessment of the trustworthiness of an entity.
- An indication of an entity’s perceived expertise, authority, or influence on a specific topic or on a specific platform.
- An indication of the quality of goods or services provided by the entity.

Based on the related work, it can be concluded that reputation never stands alone - it is always relative to some specific characteristic or behavior of interest. Common across all of the interpretations is the use of reputation to make assessments about an entity’s expected future characteristic or behavior within a specific context, based on some evidence. This evidence may take the form of past behaviors, judgments of others, or similarity to others.

3 A Mathematical Interpretation of User Reputation

3.1 Introduction

Regardless of the field or application, an entity's reputation, built from understandings of the entity's past behaviors and the judgments of others, is assumed to provide valuable insight into the expected future behavior of the entity. In a network administration or security setting, a user's reputation should provide a network administrator with insight into the user's past behavior on the network so that the administrator can proactively identify users who may threaten, harm, or degrade the quality of the network. Therefore, a useful mathematical interpretation of reputation is one that provides the central authority with an assessment of the probability a user will exhibit some characteristic of interest in the future and this probability should be based on past evidence. As discussed in Chapter 1, representing user reputation as a score – and now more specifically as a probability – is intended to capture the complex and multifaceted concept of reputation in shorthand notation that can easily be used by the central authority to compare users or make determinations about users. This chapter presents the background and related work that motivated this interpretation and formalizes the definition of user reputation.

3.2 Background and Related Work

Reputation has been interpreted as a probability in a range of diverse settings. In the context of assessing domain names, tools such as Webroot BrightCloud assign domains and IP addresses a Web Reputation Index (WRI) ranging from 1 to 100, with a score of 1 implying a user has a high probability of being exposed to malicious content on the site and a score of 100 implying the site is well known and a user has a very low probability of being exposed to malicious content. WRI is based on features such as site history, age, registration information, links, real-time performance, content, and behavior over time [49]. Similarly, the Zulu URL Risk Analyzer assesses the reputation of domains using risk scores on a 0 to 100 scale, with 0 being the highest score and 100 being the lowest possible score. This risk score is informed by content analysis, host information, suspicious URL patterns, presence on public blacklists and whitelists, historic risk assessments of both top-level domains and subdomains associated with the domain, geographic location, and behavioral features of the host [50]. In the context of Peer-to-Peer (P2P) networks, the authors in [51] defined *soft reputation* as the "average probability of inaccurate- or outright wrong-readings that stem from malicious intelligence" in order to assess the trustworthiness of data generated by a node and therefore the reputation of a node in a mobile sensing environment.

Several authors, including [4, 27, 34, 52], interpreted reputation as the probability derived from a Bayesian Inference process. Not only is this a statistically rigorous approach, but it also provides an intuitive interpretation, since the Bayesian Inference applies Bayes Rule to determine the probability of an event occurring, given some evidence. In the context of assessing reputation, Bayesian Inference is used to determine the probability of an entity exhibiting characteristic c in the future, given some historical evidence e, denoted $p(c|e)$. A prior probability, $p(c)$, is selected to represent the belief about the probability of c before any evidence has been observed. The likelihood function, denoted $p(e|c)$, reflects the probability of observing e given c. It is effectively a model of the probability distribution of the evidence. Finally, $p(e)$ reflects the probability distribution of observing the evidence. The posterior probability, $p(c|e)$, is computed as:

$$p(c|e) = \frac{p(e|c) * p(c)}{p(e)}$$

and represents the probability of a new, unobserved event.

The equations above reflect the process of Bayesian Inference using point estimates, but as is the case with most real-world applications of Bayesian Inference, in the context of assessing reputation the prior, likelihood, and posterior are more useful if they are calculated not as point estimates, but as probability distributions. In this case, the prior reflects the belief about the distribution of probabilities with which characteristic c occurs, the likelihood function represents the distribution of how evidence is observed, and the posterior probability distribution provides the probability distribution of a observing characteristic c given the evidence. More details on the mathematics behind Bayesian Inference can be found in [53] and [54].

In [4], the authors proposed the use of a Bayesian Inference reputation system for e-commerce platforms. In this approach, reputation functions based on the Beta distribution and reputation ratings based on the expectation of probability for the Beta distribution were developed for sellers using information about buyer satisfaction or dissatisfaction with a transaction. The Bayesian approach was motivated by its strong foundation in statistical theory and the Beta distribution was selected due to its flexibility and simplicity. The Beta distribution is a continuous probability distribution denoted $f\ (p|\alpha,\beta)$ with $0\ \leq p\ \leq 1$ and shape parameters $\alpha,\beta > 0$.

Because the Beta distribution forms a conjugate pair with the Binomial distribution, in Bayesian inference the Beta distribution is used to derive the posterior probability distributions for events with binary outcomes. If event A occurred in the past with a frequency of x and event B occurred in the past with a frequency of y, then the probability distribution of observing event A in the future is a Beta distribution of the form $f\ (p|\alpha, \beta)$, where $\alpha = x + 1$ and $\beta = y + 1$, and $\alpha, \beta > 0$. The expectation of probability is computed as $E(p) = {}^{\alpha}/_{(\alpha + \beta)}$ and represents the most likely probability with which event A will occur [4]. In [4], (α, β) are replaced with (r, s), where $r \geq 0$ represents the degree of satisfaction with a transaction and $s \geq 0$ represents the degree of dissatisfaction with a transaction. Reputation is considered subjective and therefore buyer X provides feedback about seller T and T's reputation function defined by X is written as $\varphi(p\ |r_T^X, s_T^X)$. T's reputation rating is then given by $E(\varphi(p\ |r_T^X, s_T^X))$; when scaled between $[0\ , 1]$, a probability expectation of 0.5 indicates a neutral rating. The authors in [4] also presented a process for combining feedback from multiple buyers in order to derive an overarching reputation function and rating for a specific seller that can be maintained by some central authority. This process can be done by simply accumulating all r, s parameters, or allowing buyers with high reputation to carry more weight and discounting the feedback provided by buyers with lower reputation. The authors in [4] also introduced the concept of a forgetting factor, which would enable older feedback to be forgotten, motivated by the fact that buyers may change their behavior over time.

In [27], a modified version of the Beta reputation system was proposed. This version replaced the Beta distribution with the Dirichlet multinomial distribution, allowing a more nuanced representation of reputation. In other words, instead of reputation reflecting binary outcomes such as probability of a seller being bad or probability of a seller being good, the Dirichlet distribution enabled probability estimates for a seller being mediocre, bad, average, good, or excellent.

In 2004, [52] presented a Bayesian based reputation system for Peer-to-Peer (P2P) and mobile sensing networks. In this distributed reputation system, every node maintains a reputation rating for every other node in the network; node i's reputation rating for node j reflects node i's opinion about whether j behaves correctly, either by sending the correct files in the P2P network setting or using the correct routing protocol in the mobile sensing setting. Because this is a distributed reputation system and nodes rely on the opinions and ratings created by other nodes for information, trust ratings are also a key component. Node i's trust rating for node j reflects node i's opinion about whether j is honest when evaluating other nodes. Both the reputation rating and trust rating are Bayesian-based and the reputation (or trustworthiness) of a node is understood to be some probability Θ drawn according to the posterior distribution, indicating that a node misbehaves with probability Θ.

As in [4], the authors in [52] used Beta as the prior and posterior distributions. When a node makes firsthand observations about another node, the distribution is updated by modifying the shape and scale parameters α, β: setting $\alpha = \alpha + s$ when s number of misbehaviors are observed and setting $\beta = \beta + f$ when f correct behaviors are observed. If node i updates its assessment of node j based on node k's rating of j, a model merging approach is taken instead of directly updating α, β. Specifically, the authors in [52] proposed a linear pool model merging method in which node k's model is merged into node i's model using weights determined by how trustworthy node i believes node k is. Finally, [52]'s approach also allows for a discounting factor so that older evidence is given less importance than newer evidence in a reputation rating.

In 2015, the authors in [34] introduced a distributed Bayesian-based reputation system for Vehicular Ad-hoc Networks (VANETs) in which nodes determined the probability of other nodes behaving maliciously by observing packets sent by the nodes over a monitoring period. Nodes were then assigned class labels based on these probabilities; nodes were labeled 'malicious' if their probability of being malicious was over 0.5, otherwise they were labeled 'benign.'

3.3 Defining User Reputation

Based on the related work presented in here and in Chapter 2, it can be concluded that reputation does not stand alone - it is relative to some specific characteristic of interest. Therefore, reputation should be defined and quantified relative to a characteristic; defining a user's reputation without respect to anything else is meaningless. In a network security setting, particular characteristics of interest may be whether a user has a reputation for being malicious - i.e. has the user spread malware, attempted to steal protected information, or launched cyber attacks against others? - or whether a user has a reputation of being infected or being more likely than another user to be infected - i.e. has the user been the victim of a social engineering or phishing campaign or does the user own devices that have been compromised with malware?

Common across all of the interpretations is the use of reputation to make assessments about an entity's expected future characteristic or behavior within a specific context, based on some evidence. This evidence may take the form of past behaviors, judgments of others, or similarity to others. As observed in [4, 27, 34, 52], using a Bayesian Inference framework for assessing reputation produces an intuitive and statistically rigorous mathematical interpretation of reputation because an entity's reputation for a specific characteristic can be understood as a prediction and quantified as a probability distribution shaped by past evidence. This approach also allows for evaluation of the uncertainty surrounding a reputation assessment of a user.

Based on this, the following mathematical interpretation of user reputation as a score in a network security context is proposed:

A user's cyber reputation score is a probability representing the likelihood a user demonstrates a specific characteristic on a network, based on evidence of past behaviors that are considered representative of this characteristic.

Mathematically, a user's reputation score is the probability expectation of a posterior distribution derived from the process of Bayesian Inference, i.e. $E[p(c|e)]$. The posterior distribution is calculated as:

$$p(c|e) = \frac{p(e|c) * p(c)}{\int p(e|c) * p(c)\, dx}$$

[54]. The prior distribution $p\,(c)$ is selected to represent the central authority's subjective belief about the likelihood with which a user exhibits a specific characteristic. The likelihood function $p\,(e|c)$ reflects the probability of observing evidence e given c. Finally, the denominator $\int p(e|c) * p(c)\, dx$ normalizes the

calculation. The expectation value of the resulting posterior distribution is determined either analytically or through a random sampling method.

Chapter 4 provides more detail on how reputation can be quantified in practice using the proposed definition. Additionally, in Chapters 5 and 6 we illustrate the entire methodology through a case study.

4 A Methodology for Evaluating User Reputation

4.1 Introduction

This chapter presents an overview of our methodology for quantifying and predicting user reputation in a network administration or security setting. The methodology objectives and assumptions are outlined and the three major phases of the methodology - characteristic definition and evidence collection; reputation quantification and prediction; and model validation and refinement - are presented in detail. The chapter concludes with an overview of the case studies that will be used to illustrate and validate the methodology.

4.2 Methodology Requirements, Objectives, and Assumptions

A major shortcoming in the field of cybersecurity is the lack of well-defined frameworks for conducting security experiments and deriving understandings from them [55, 56]. Quantifying and predicting the reputation of users on a network is effectively a type of experiment since the goal is to establish and test a hypothesis about the expected future behaviors of a user on a network. Therefore, the proposed methodology for evaluating user reputation should satisfy the requirements of a sound scientific experiment. According to the authors of [57] and [58], a well-structured method - i.e. the process that is followed to reach a correct conclusion - achieves the following criteria:

- **Requirement 1:** *Internal Validity (or correctness)*: "the mechanism under experimentation is of suitable scope to achieve the reported results; the entities and activities of the experimental mechanism are not susceptible to systematic error" [57]. In the context of this research, this implies assessments of user reputation are built from clean datasets and experiments on the dataset are designed such that results of the experiment are fully explained by the variable(s) being tested.

- **Requirement 2:** *External Validity (or realism)*: "the mechanism under experimentation (and therefore the result of the experiment) is not solely an artifact of the laboratory setting" [57]. In the context of this research, this implies that procedures for assessing user reputation generalize to different types of computer networks and user populations.
- **Requirement 3:** *Containment (or safety)*: "no pre-mechanism causes threaten to confound the results, and no post-mechanism effects are a threat to safety" [57]. In the context of this research, safety may be interpreted as a guarantee of privacy of sensitive data about a user.
- **Requirement 4:** *Transparency*: "there are no explanatory gaps in the experimental mechanism; the diagram for the experimental mechanism is complete" [57]. In the context of this research, this implies the procedures for quantifying user reputation should be described clearly and in enough detail that another researcher, network administrator, or security practitioner can replicate the analysis on the same network or reproduce the process on another network and set of users.

Satisfying the four requirements above will ensure the proposed methodology represents a scientifically valid approach to evaluating reputation. To ensure the methodology actually provides a practical and useful framework for assessing user reputation from a network security perspective, the methodology should satisfy the following objectives:

- **Objective 1:** The methodology enables the central authority to be proactive rather than reactive: it provides a process for identifying behaviors that are significantly correlated with or indicative of a characteristic of interest and can be used to predict future characteristics or behaviors for current or new users.
- **Objective 2:** The methodology accounts for uncertainty: it provides a process for identifying and evaluating uncertainty associated with the model for user reputation, the reputation probability score, and the accuracy of the data itself, and a process for propagating this uncertainty into any predictions made about a user.
- **Objective 3:** The methodology is flexible and recognizes that reputation is dynamic: it enables the tracking and updating of user reputation over time and the refinement of models for user reputation based on new evidence.

The proposed methodology for assessing user reputation is based on the following assumptions:

- **Assumption 1:** There is a central authority tasked with calculating and maintaining user cyber reputations. In other words, this is a type of centralized reputation system and is not intended for decentralized settings where users maintain subjective reputation assessments of other users.

- **Assumption 2:** Users have authenticated onto the computer network so there is ground truth about the mapping of an account (or set of accounts owned by the same user) and any traffic or events generated by the account(s) on the network.
- **Assumption 3:** The central authority maintains a record or history of events, incidents, and alerts on the network and can correlate these events with a unique user.

4.3 Methodology Overview

At a high level, the methodology for assessing and quantifying user reputation consists of iterating through the following three phases, as shown in Figure 3: characteristic definition and evidence collection; reputation quantification and prediction; and model validation and refinement.

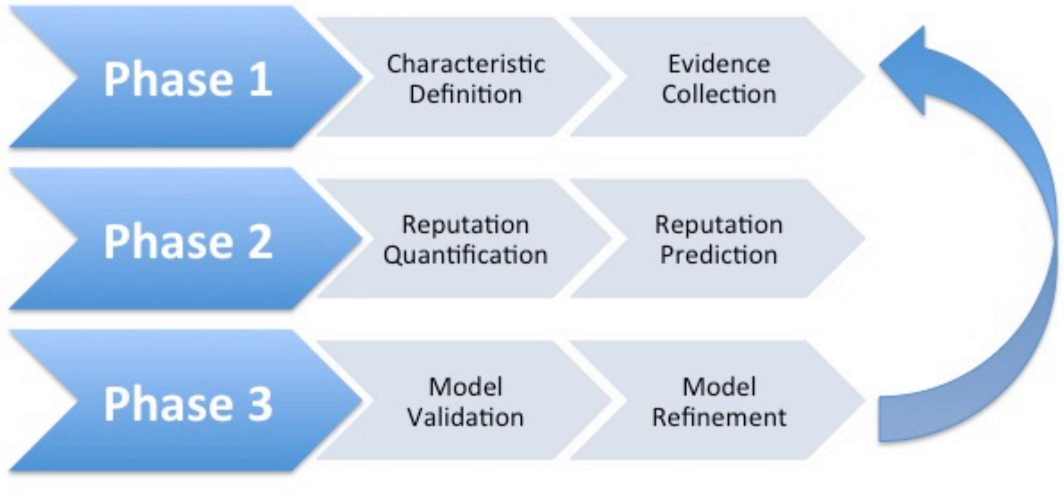

Figure 3: Phases of the methodology

These three phases should enable the central authority to answer the three high-level research questions posed in Chapter 1. First, what data can be used as evidence that user X has a reputation for a particular characteristic c? Second, what is the probability of user X exhibiting characteristic c based on this evidence and how can this probability and/or evidence be used to predict user behavior and reputation on the network in the future? And finally, how can reputation assessments be validated and refined? The following sections describe the steps to answer these questions, while the following chapters of this dissertation present a series of case studies to experimentally evaluate the proposed methodology.

4.4 Phase One: Characteristic Definition and Evidence Collection

Recall that Research Question 1 (RQ1) presented in Chapter 1 asked: What data can be used as evidence of a user's cyber reputation for specific characteristics or behaviors on a computer network? This question is addressed in *Phase One* of the methodology, in which the central authority tasked with quantifying reputation identifies a characteristic of interest, defines how this characteristic manifests on the network, and then collects data that can be used as evidence of the characteristic.

4.4.1 Characteristic Definition

Reputation is always relative to some characteristic, so the central authority must define the perspective from which they are trying to understand user reputation. In this dissertation, a characteristic is some quality or property of a user that is necessary to understand for the purpose of improving cybersecurity and can be used to differentiate between different types of users that may be more or less harmful to the network's security. For example, the characteristic may be whether the user is malicious or benign or whether the user is infected or uninfected with malware. How this characteristic manifests on the network must also be carefully defined. In the realm of domain name and IP reputation scoring, even though the majority of tools are intended to indicate to what degree a domain or IP is malicious, the meaning of malicious is often ambiguous or undefined entirely, resulting in a lack of clarity regarding what is actually being measured. Therefore, it is necessary to define what it actually means to be malicious, infected, and so on in a particular network or service setting. For example, a malicious user may be defined generally as someone who has violated the policies, terms, and conditions of a service, or more specifically as someone who has spread malware to other user devices, while an infected user may be defined as someone who has malware on one or all of their devices or has been flagged by the network's Intrusion Detection/Prevention system or other security tool. Certain definitions may require reliance on a third party algorithm or tool to define whether a characteristic has manifested, but some form of 'ground truth' is necessary up front for initial reputation assessments.

4.4.2 Sources of Data

Once the central authority has identified the characteristic of interest and defined how this characteristic manifests, data about users that can be used to assess reputation must be collected. Recall from Chapter 1 that data about users must be collected and represented in a set of numerical features denoted *observations* that are stored in an n-dimensional vector denoted the *observation space*.

Observations about users may be derived from computer network traffic, hosts and devices associated with the user, online profiles or accounts associated with the user, or services and platforms maintained by the network on which the user has a presence. The nature of these observations will of course depend on data availability, privacy expectations of users on the network, and the context or environment. Additionally, depending on the source of the data, there may be uncertainty in the completeness and accuracy of the data.

According to the methodology assumptions, a ground truth source of data about users can be derived from network traffic generated and collected on the computer network or service on which the user is being evaluated. In other words, the central authority tasked with calculating reputation should be able to match network traffic records with unique users. Observations derived from traffic record header files may include patterns of usage, source and destination IP addresses, source and destination ports, protocol, packets sent, bytes sent, and so on. If full packet capture is performed on the network, then further observations can be obtained from the payload, such as the nature of content of any files sent during the traffic session. If a firewall or antivirus system is employed, any alerts or threats created by users are also valuable observations for assessing user reputation.

Host or device based observations can be derived from any devices associated with the user’s login credentials on the network. Again, depending on the nature of the network and the privacy expectations of the users, the insights gained from user devices will vary. On a network where users do not consent to monitoring of personal devices (for example, a university network or an Internet cafe), host-based observations may be limited to insights such as the number of devices a user has on the network and how frequently they are used. On a network where users do consent to monitoring, observations may include any installed applications, downloads, keystroke analysis, mouse click patterns, security software, regularity of applying patches and other software updates, and so on. Aspects of these observations may still be obtained in the 'non-consent to monitoring' scenario through means such as surveys or focus groups. However, there is no guarantee that users will be willing to participate in such surveys and if they are, there is no guarantee that they will be honest about their applications, updating behaviors, and various other device features of interest.

Additional observations may come from a user's online accounts or profiles. These profiles may exist on platforms or services owned and maintained by the network, or they may exist on public Internet platforms, to include social media sites such as Facebook [59] or Twitter [60], online communities such as Reddit [61], question and answer forums such as Quora [62] or StackOverflow [40], interest or profession specific platforms such as GitHub [63], and more. Observations from these platforms or services may include user activities, patterns of usage, content, interactions with other users, social networks, etc. Insights from Internet platforms may be difficult to obtain however, because there may not be ground truth available regarding profiles associated with a user and data availability may be limited to due privacy expectations of the users, or privacy policies and terms and conditions associated with the service.

It is worth noting that it is not practical to create a finalized list of all the sources where observations about users can be derived. This is due not only to inherent differences that will exist in computer networks, but also due to the rapid pace at which technology evolves. New sources of data about users appear as new technology platforms and devices are introduced. For example, Internet of Things

(IoT) devices are not discussed in this dissertation, but could potentially provide a wealth of information about a user's cyber reputation on a network.

If the central authority has access to data regarding the unique physical world human associated with a cyber account(s), a final important source of data may be individual differences in people. Individual differences encompass a wide range of characteristics that vary between people [64], but the four major categories commonly evaluated in cybersecurity literature include demographic factors, personality traits, risk-taking preferences, and decision-making styles. Demographics include characteristics of a human population such as age, gender, occupation, and so on. While there are many different personality traits, there are five major categories in the widely accepted "Big Five" model: agreeableness, conscientiousness, neuroticism, openness, and extraversion [65]. Agreeableness measures cooperative traits; conscientiousness measures dependable and organized traits; neuroticism, also sometimes categorized as emotional stability, measures insecure and nervous traits; openness measures intellectual and imaginative traits; finally, extraversion measures energetic and outgoing traits [65]. Risk-taking is a measure of risk attitude and shapes decision-making, which is examined in the literature in relation to several forms of risky behavior [66]. There are five dimensions of risk-taking preferences that are commonly studied in relation to cybersecurity and user behaviors: ethical (RTE), financial (RTF), health/ safety (RTH), recreational (RTR), and social (RTS) risk-taking. Finally, decision-making style is the response pattern exhibited by an individual in a decision-making situation [67]. Decision-making styles are generally categorized into five broad categories: rational (DMR), avoidant (DMA), dependent

(DMD), intuitive (DMI), and spontaneous (DMD) decision-making. Rational refers to using logic when making decisions; avoidant refers to delaying decision-making; dependent refers to relying on others for decision-making help; intuitive refers to decision-making based on instincts; lastly, spontaneous refers to quick decision-making [68].

While the availability of such data will depend on the network and the privacy expectations of users, this data can provide a wealth of information for user reputation assessments. For example, research has correlated a user's age with a variety of cybersecurity behaviors and risks, to include phishing susceptibility [108, 109]; password sharing habits [110]; likelihood of malware encounters [111]; and likelihood of malware contamination [112]. As another example, in a survey we conducted to correlate human traits with cybersecurity behavior intentions, we found gender, extraversion, financial risk-taking, and rational decision-making were all significant predictors of good security behavior [113].

4.4.3 Evidence Shaping

Once observations about users have been collected, the central authority must identify a subset of features that is highly indicative or predictive of the characteristic of interest. Each individual feature is considered *evidence* and the evidence is stored as numerical features in the *cyber reputation vector* of size $\leq n$. It is important to distinguish between data that defines a characteristic has manifested and data that is used as evidence to predict whether a characteristic may manifest. For example, if the characteristic of interest is maliciousness and a user with a reputation for being malicious is defined as someone who has spread malware to other users, evidence may include the number of messages or files sent by this user to others and the number and nature of the links and files contained in these messages. Observations that are useful forms of evidence can be identified via various feature engineering, dimensionality reduction, and correlation analysis techniques.

In order to use the observation space in statistical or predictive models, observations should be represented in the feature vector as numeric features. For example, if observations are taken from network traffic data, features may include the number of traffic sessions created by a user per day, the average length of a traffic session, and the number of alerts created by a user per day. As another example, if observations are taken from user devices, numeric features may include the number of devices owned by a user, the frequency that a user applies software updates, and the number of third party applications installed on average on a user's device.

Feature engineering can and should also reflect the preferences of the network administrator and events that are considered deviations from appropriate behavior on the network. For example, in a corporate network environment, an example feature might be the number of non-work related websites visited by a user per day, while on a university network, an example feature might be a Boolean value indicating the presence of file sharing software on the user's device.

Once the observations have been transformed into a set of numerical features, denoted the *observation space*, dimensionality reduction and correlation analysis can be performed to identify a subset of features that best explain user behavior and the features from this subset that are highly correlated with the characteristic of interest.

For example, Principal Component Analysis, a statistical dimensionality reduction technique that converts a set of potentially correlated features into a smaller set of orthogonal linearly uncorrelated features called the principal components, may be used to find combinations of features that explain as much variance in the data as possible [69]. The correlation values between the Principal Components and the original features of the observation space can then be analyzed to identify the features that contribute most to explaining variance in user behavior and therefore may be most important in differentiating between users. Unsupervised clustering techniques can then be used with both the original features of the observation space and the newly formed principal components to identify users or behaviors that are outliers or to understand if there are any natural partitions in the data. Finally, statistical correlation tests or classification algorithms may be used to identify if any of the original features or principal components are correlated to the characteristic of interest. If high correlation or classification accuracy values cannot be obtained from the current observation space, additional data will need to be collected and incorporated into the observation space, and the process of evidence shaping will need to be repeated. Depending on the classification algorithm, it may also be possible to determine feature importance, a ranking indicating which features were used the most of used to make key decisions [114].

Note that in some cases, depending on the size of the network, it may be infeasible to determine the cyber reputation vector using the entire user population. In this case, a random sample of users should be obtained from the network and the previously described process to identify evidence should be applied on this sample.

4.5 Phase Two: Reputation Quantification and Prediction

This section presents *Phase Two* of the methodology in which Research Question 2 (RQ2) is addressed. Recall that RQ2 asked: How can evidence be used to define and quantify a user's cyber reputation and develop predictive models of future reputation? In *Phase Two* the evidence contained in the reputation vector is used to either quantify reputation using the mathematical interpretation presented in Chapter 3 or is used to make predictions about an existing or new user's future reputation or behaviors.

4.5.1 Reputation Quantification

Recall from Chapter 3 that in the context of quantifying user reputation as a single score, Bayesian Inference is used to determine a posterior probability distribution $p\,(c|e)$ providing the likelihood of probabilities with which a user exhibits a characteristic c given some historical evidence e.

The prior probability, $p\,(c)$, is selected to represent the belief about the probability of c before any evidence has been observed. As observed in the work of [4, 52], the Beta distribution is a simple and intuitive distribution that can be used as the prior. Though the information it captures is somewhat limited, the Beta distribution allows for easy updating via counts of observed evidence. Regardless of what type of distribution is used, the prior is shaped using knowledge from previous studies or existing knowledge about the nature of users on the network. If there is no previous work or existing knowledge on which to base the distribution, a non-informative prior, known as Jeffrey's Prior, can be used [54].

While a prior distribution must be shaped independently of any data that is used as evidence, Bayesian Inference can be performed iteratively. In other words, if Bayesian Inference has already been applied to users on the network and a posterior distribution has been formed using evidence, this posterior can be continuously updated and refined as new evidence comes in, effectively serving as the prior.

The likelihood function, denoted $p\ (e|c)$ is effectively a model of the probability distribution of the evidence. If a Beta prior has been selected, the Binomial distribution should be used as the likelihood function since the two are a conjugate pair [54].

The posterior distribution updates the prior distribution using the collected evidence and Bayes Rule. When the Beta-Binomial conjugate pair is used, the Posterior is also a Beta distribution. As discussed previously, Bayesian Inference can easily be conducted iteratively. In the Beta-Binomial distribution scenario, if additional evidence is observed after the posterior has been computed, evidence counts can simply be added in order to reshape the parameters of the posterior Beta; it is mathematically no different than if all of the evidence had been observed at once and the posterior had been calculated afterward.

The discussion on reputation quantification so far has assumed that all evidence is the same and can simply be incorporated into the posterior distribution as a frequency count. This is, however, an oversimplification. How evidence is actually represented and incorporated into the Bayesian model will depend on the nature of the evidence and selected distribution. If there are multiple types of evidence – i.e. different features – the posterior may be formed using techniques such as the method of weighted posteriors, in which the process of Bayesian Inference is applied separately for each source of evidence and then the resulting posteriors are each assigned a weight and combined into a single distribution representing a summation of the weighted posteriors [54]. These weights may be adjusted based on how strongly evidence is correlated with a characteristic of interest.

Once the posterior distribution has been developed, a reputation score can be calculated as the probability expectation of the distribution. In other words, a reputation score defines the most likely probability with which a user will exhibit a characteristic in the future, based on evidence. This probability expectation value may be derived analytically, as done in [4] in which a seller is given a reputation rating in the form of the probability expectation of the Beta distribution, which is calculated directly as $E(p) = {}^{\alpha}/_{(\alpha + \beta)}$. While this approach is effective, a more useful approach may be calculating the score via repeated sampling of the posterior. This enables the central authority to determine a confidence interval around the score. Moreover, in some cases it may be impossible to derive the probability expectation value directly using analytic methods.

Sampling from the posterior to derive a reputation score can be done using techniques such as Monte Carlo Sampling or Latin Hypercube Sampling. Both techniques produce a frequency distribution of the possible outcome probability values by repeatedly recalculating point estimates for the outcome probability using input values that have been sampled from the input distributions. After thousands or tens of thousands of iterations, the sample uncertainty distribution is assumed to approach the true distribution [70]. The two techniques differ in how they sample from the input distributions. The Monte Carlo Method randomly samples from any input value that has an associated uncertainty distribution [70]. The Latin Hypercube Method seeks to improve upon the Monte Carlo Method by sampling from across the entire space of an input distribution, ensuring that calculations are performed even for input values that occur with low probability [70].

The reputation score may be taken as the most frequently occurring probability. The uncertainty in the score can be evaluated by calculating confidence intervals; narrow confidence intervals indicate higher certainty in the score than wide confidence intervals. For example, if the most frequently occurring probability for a user behaving maliciously is 0.80, and the 90% confidence interval is between 0.79 and 0.81, there is low uncertainty in the estimate and the central authority has high confidence in assigning the user a reputation score a reputation value of 0.80. However, if the 90% confidence interval is between 0.50 and 0.90, there is high uncertainty in the estimate and the central authority may have low confidence in the 0.80 reputation score.

Because a user's reputation score is calculated as a probability expectation value, this score may be interpreted differently depending on the preferences of the central authority. One approach to making the reputation score intuitive and useful for decision support is mapping the score to a class label based on a probability threshold. For example, if the characteristic of interest was maliciousness, users may be assigned the label 'malicious' if their reputation score exceeds a certain value; the remaining users who do not have reputation scores that exceed this value may be assigned the label 'benign.' The actual threshold value can be selected based on whether importance is placed on minimizing false negatives or false positives. Setting a high threshold for defining users as malicious - for example, defining malicious users as those who have a probability expectation of behaving maliciously that is over 0.90 - will reduce false positives and may be applicable in network settings where the impact of a user introducing a threat or vulnerability to a network is limited. Setting a lower threshold will increase false positives and reduce false negatives, and may be applicable in a setting where the impact of a user introducing a threat or vulnerability is considered high.

The uncertainty bounds can also be used to determine how users are defined and labeled or may be used to supplement a label with more information. For example, a user may be labeled as malicious with low confidence or labeled as malicious with high confidence. This estimate of confidence in turn can impact any decisions the central authority makes about the user.

4.5.2 Reputation Prediction

While a reputation score provides a probability that a user will exhibit a certain characteristic, in some cases the central authority may want to make predictions about future user reputation or the reputation of new or unknown users. In this case, the cyber reputation vector may be used as input to clustering or supervised learning algorithms.

For users with unknown reputation, unsupervised learning techniques may also be used to make predictions about these users based on their similarity to known users. For example, if the central authority has determined that there is a set of users with a strong reputation for a certain characteristic (based on some probability threshold, as discussed in the section on *Reputation Quantification*), the cyber reputation vectors may be calculated for both the known and unknown users and then used as input to a clustering algorithm. Depending on proximity of the unknown users to the known users based on some distance metric, the central authority may consider these users to have the same reputation. If this is the case, depending on the characteristic, the central authority may choose to restrict or grant privileges, conduct additional monitoring, or require additional training of the unknown users until they can be monitored for enough time to make a reputation determination based actual user behavior on the network. This determination may also be used for reputation quantification - these unknown users may have their prior distributions modeled after the posterior distributions of the known users.

Unsupervised learning can also be applied to make predictions for users whose reputation is already known. Specifically, there may be some subset of users that do not have a reputation for the characteristic of interest based on their past behaviors. However, behavior - and therefore reputation - can change over time. So, if clustering reveals that these users are very similar to users that do have a reputation for this characteristic, again based on some selected similarity metric, the central authority may still choose to restrict or grant privileges, conduct additional monitoring, or require additional training until the user proves otherwise.

If a sample of users is assigned class labels based on their reputation scores, as discussed in the section on *Reputation Quantification*, this sample may be used as training data for supervised learning techniques in order to classify other users.

Supervised learning techniques may also be used to make more nuanced predictions about the behaviors of users. If there is some behavior that the central authority is particularly interested in monitoring, this behavior can be tracked for a period of time and then regression models can be built to analyze how this behavior will manifest in the future. For example, if the number of unverified application downloads is evidence in the cyber reputation vector that is correlated with a reputation for device infection, the central authority may track this value per user on a weekly basis and use these values to predict future suspicious downloads counts. The central authority may then decide that users whose regression models indicate increasing numbers of suspicious downloads are required to take additional security trainings or require download approvals from network administrators.

4.6 Phase 3: Validation and Refinement

This section presents *Phase Three* of the methodology for addressing Research Question 3 (RQ3), which asked: How can quantification assessments and predictions of a user's cyber reputation be validated and refined?

4.6.1 Validation

In order to validate the scores and predictions, a pilot study should be conducted in which a sample of users are evaluated and monitored over a period of time and then any scores or predictions are compared to the actual observed outcomes. Since reputation assessments of users are either probabilities or predictions, forecast verification methods can be used to validate the prediction accuracy or assess the reliability of user reputation models [104]. For measuring the accuracy of probabilities, metrics such as the Brier Score, commonly used in assessing weather predictions, can be used. The Brier Score provides the mean squared error for probabilistic forecasts for two category predictions (in this case 'yes' the user did exhibit the characteristic of interest or 'no' the user did not exhibit the characteristic), with a score of 0 indicating the probability is deterministic and a score of 1 indicating the probability is always incorrect [104]. Other metrics that may be used to validate probabilities (if they are mapped to binary outcomes) include Pearson's Correlation Coefficient or Spearman Rank Correlation, which can both be used to assess the linear association between predictions and outcomes [104]. Reliability diagrams, which plot the frequency with which an event is observed against the probability, may also be used to assess predictions. A "perfectly reliable" probability model is one with a curve that lies on the 45^o diagonal line of the reliability diagram [104].

For predictions of user behavior involving supervised learning algorithms, metrics such as Mean Square Error (MSE), the square of the difference between predictions and actual observations, or Mean Absolute Error (MAE), the average of the absolute difference between predictions and actual observations, may be used [104]. A Receiver Operating Curve (ROC), which plots false positive rates on the x-axis against true positive rates, can also be used to assess prediction models.

4.6.2 Refinement

Depending on the accuracy of the reputation scores and predictions, steps can be taken to refine the user reputation model. First, if all possible sources of data have not yet been exhausted, the data collection and evidence-shaping phase can be revisited. Second, if there are no additional sources of data that the central authority can analyze, the characteristic definition may be revisited and narrowed or broadened. If the original characteristic definition was too broad, it may be difficult to find evidence with strong correlation to the characteristic. Similarly, if the original characteristic definition was too narrow, it may be difficult to find enough sources of data to form evidence.

A third approach to improving user reputation models may be increasing the time period of data collection before any models are constructed. Data may need to be collected and analyzed over time to understand the nature of the network and the users on the network. This is especially important for networks that may experience varying levels of activity depending on time of day, seasons, and various other external features. If the time window of data selected for building reputation models does not actually match the time periods where users are active or fails to capture how external factors influence the network, model accuracy may be hindered.

Closely tied to this idea is a fourth approach to refining user reputation models: increasing the frequency with which the models are updated and new evidence is incorporated. Since reputation is dynamic in nature, again depending on the nature of the network and user, the frequency with which evidence features are recalculated may need to be increased. Taking a random sample of users and recalculating their evidence values at varying intervals of time, such as hourly, daily, weekly, and monthly, can help determine this. If evidence values show no significant difference when calculated hourly or daily, but do show a difference when calculated weekly, then weekly may be selected as the update frequency in the reputation model.

In some cases, evidence update frequency may only need to be increased for users who demonstrate more erratic or unpredictable behavior.

The fifth approach to refining reputation models involves incorporating more nuances into the model and how evidence is handled. For example, the central authority may vary the amount of weight given to evidence when performing Bayesian Inference. The concepts of ‘forgetting’ and ‘forgiveness,’ in which older behavior and evidence is given less importance than newer behaviors and evidence, can be incorporated into the model. The authors in [4] presented the idea of ‘forgetting’ in an e-commerce setting. Motivated by the fact that agents change their behavior over time, in their model, older feedback was given less weight via a *forgetting factor.* A similar concept was presented in [52] as a *discounting factor.* In the context of network security, undesirable behaviors that contribute to a lower user reputation may be given less weight or importance over time (“forgetting”) or may be given less weight or importance if the user takes steps to improve their behavior, such as taking security training (“forgiveness”).

Note that any efforts to refine the model should be conducted incrementally – i.e. only one method at a time should be applied and then user reputation scores and predictions should be recalculated and validated. Multiple improvement techniques may be conducted in parallel on different random samples of users. Refined models can be assessed using the same validation metrics presented previously. Also note that the steps proposed in this section may also be used to reduce the uncertainty bounds around a reputation score.

4.7 Case Study Outline

Chapter 5 illustrates Phase One of the methodology through a case study of users on the wireless network at a large, public university. User infection was selected as the characteristic of interest. A user is considered to be a single account that is used to access and conduct activity on the university's wireless network from an endpoint device or multiple endpoint devices. An infected user is defined as someone who has a record of device compromise in the university's Intrusion Detection/Prevention System logs; an uninfected user is someone who has no record of device compromise in the university's logs. Wireless network traffic for all infected users on the network and random samples of uninfected users is used as the initial source of data for user observations. We conduct a series of feature engineering, dimensionality reduction, clustering, and classification techniques in order to determine the feasibility of using network traffic as evidence of user reputation with respect to infection status.

Chapter 6 expands on this case study by illustrating Phases Two and Three of the methodology. We construct models to represent infected and uninfected users and then take a sample of unknown users and assign them reputation scores indicating their likelihood of infection. We monitor and refine these scores over time. We conclude with a discussion of additional sources of data that could help reduce uncertainty in the reputation assessments.

Chapters 7 and 8 present a separate case study, in which we collect and analyze incidents of account compromise self-reported by users on Twitter. We illustrate the viability of using user-reports on social media to understand people and their cybersecurity attitudes and behaviors. We conclude this case study with an outline of our future work in this area.

5 Identifying Infected Users via Network Traffic: An Empirical Case Study

5.1 Introduction

This chapter presents a case study demonstrating Phase One of the methodology presented in Chapter 4. This chapter investigates the feasibility of extracting evidence from network traffic that can be used to identify users who will be exploited, i.e. become *infected* on the network. The goal of this chapter is to determine whether network traffic provides enough evidence to quantify and prediction whether users on the network have or will have a reputation for infection.

In both research and development there has been extensive work to use network traffic to proactively identify entities that are susceptible to becoming infected and entities that are more likely to exploit others or spread infection. Network traffic analysis of IP addresses, binaries, and devices has become an invaluable process for proactive identification of security threats; anomaly and network intrusion detection systems often boast over 99% accuracy rates for detecting suspicious, malicious, or compromised entities. Here, we explore whether network traffic can be just as useful for evaluating users.

Several authors have observed that users' network traffic contains discernable patterns that can act as user fingerprints [71, 72, 73]. However, work in this area is limited and generally focuses on the goal of uniquely identifying users from their traffic flows. Our research presents a complementary approach: given a corpus of network traffic and ground truth data about the mapping of traffic records to the unique user who generated the traffic, we explore whether features from a user's ‘network fingerprint’ can be used to differentiate between users who are more likely to become compromise victims and users who are less likely to become victims.

We analyze two months of wireless network traffic at a large, public university for all the infected users that appeared in the logs during that time window (1,923 total) and random samples of uninfected users. In this case study, we consider a user to be represented by a single account that is used to access and conduct activity on the wireless network from one or more endpoint devices. We define an infected user as someone who has a record of device compromise in the university’s intrusion detection/protection system logs.

We make the following contributions:

- We extract 36 features from network traffic and apply Principal Component Analysis to identify 13 principal components that explain about 92.8% of the variance in user behavior. Analysis of these 13 components reveals 10 features that best explain user behavior, suggesting 26 features that may be less useful for understanding users.
- Through unsupervised learning techniques, we show that there are differences in users revealed by their network traffic that can be used to partition users into distinct groups, and in some cases these clusters separate the infected and uninfected users, with some clusters having a composition up to 100.0% of only one type of user.
- We apply 10 different supervised learning techniques using both the original 36 features and the 13 principal components and evaluate the accuracy, area under the receiver operating characteristic curve (ROC AUC), false positive

rate, and false negative rate with which we can classify users as either infected or uninfected. We show that that using network traffic alone to classify users, we can achieve accuracy values up to 79.0% and ROC AUC values up to 86.0%. However, the lowest false positive rate is 16.8% and the lowest false negative rate is 21.2%, suggesting additional features or sources of data may be necessary to further refine the models.

- We show there is potential to identify signals of infection status from network traffic patterns without knowledge of the actual content a user is accessing.

The remainder of this chapter is organized as follows. Section 5.2 presents the background and related work and research questions. Section 5.3 presents the methodology for feature engineering, dimensionality reduction, unsupervised clustering, and supervised learning. Section 5.4 presents the results, followed by a discussion of the results in Section 5.5. Section 5.6 presents the limitations and Section 5.7 concludes the chapter.

5.2 Background and Related Work

Network traffic, either raw or summarized in a form such as Cisco's NetFlow, is commonly analyzed for the purpose of anomaly or network intrusion detection [74]. Since hostile traffic generally looks different from benign traffic, these approaches often involve extracting various features about an IP address from the network traffic - such as byte or packet values, temporal patterns, source and destination location information - and then using some statistical, classification, clustering, or rules-based algorithm to distinguish between malicious and benign IP addresses [74-80].

Several authors have observed that grouping and analyzing traffic flows from a human user-centric perspective versus an IP or host-centric perspective provides a more nuanced understanding of a network and may prove useful for identifying threats to the network, especially because users typically generate traffic on multiple devices and a single IP can map to thousands of unique users [71, 72, 81, 82]. In [71], the authors observed that traffic generated by an individual can act as a biometric signature and developed a system that could infer the identity of users from NetFlow records even when thousands of users were hidden behind a few IP addresses. The authors used eight hours of NetFlow records to profile five users on a large metropolitan WiFi network with a total of 200,000 users and an average of 1,000 users whose IPs have been mapped (Network Address Translated) behind two IPs addresses. When attempting to identify these users from a set of 100 million raw NetFlow records collected over a 24 hour time period, the system correctly identified the users with true positive rates over 90% and false positive rates below 0.08.

In [82], the authors examined the possibility of user identification using network packet metadata and neural networks. They collected network traffic consisting of fields such as time stamps, source and destination IP and port information, and packet length for 46 users over a two-month period. They then examined whether or not they could use traffic metadata to identify specific users on nine popular services. Results varied depending on the user and the service. For example, on Google, the neural net achieved a true positive identification rate of 90% for one user, but a true positive identification rate of only 48.3% for another. Similarly, in [81], the authors developed user-behavioral profiles from raw traffic network metadata and found that they could accurately identify users on Skype, Hotmail, and BBC with 98.1%, 96.2%, and 81.8% accuracy, respectively. In [72], the authors examined how to derive high-level behavioral features from low-level network traffic metadata in order to build profiles that indicate the services used by an individual and how they use the service; when the authors evaluated their profiling technique on Facebook, they were able to distinguish between different user activities using the raw traffic alone.

In [83], the authors developed a tool that used HTTP traffic from four users in a domestic household network to identify the user generating the traffic with a true-positive rate of 64% and a false-positive rate of 28%. In [84], the authors profiled 25 synthetic users with 90% accuracy based on web transactions and observed that user web transactions are fairly consistent and exhibit little novelty over time. In [73], the authors examined how effectively they could identify whether a 802.11 traffic sample came from a user and found that 'implicit identifiers' such as network destinations (IP address and port pairs) and SSIDs were highly useful in distinguishing users.

Our work is most similar to the work presented in [115], in which the authors analyzed the web browsing behaviors of users to predict who was at risk of being exposed to malware. Using telemetry data from a major AV company containing 100,000 users and millions of URLs visited by the users, the authors extracted 74 features summarizing user behavior, to include volume of activity, temporal patterns of activity, number and type of websites visited, and variability of browser activity. Correlation analysis was performed to understand the relationship between these features and the probability of the user encountering a malicious web page. The authors defined *at risk* users as those who visited at least two distinct malicious URLs or at least three blacklisted domains during the two-month study period; *safe* users never visited malicious URLs or blacklisted domains; and the *uncertain* users were those who did not fall into either category. The authors found that there was a significant increase in malicious URLs visited by users on the weekend; *at risk* users spent more time online at night; and volume of activity was one of the best predictors of being *at risk*. The authors then used insights from their correlation analysis to build logistic regression models; the model classified users as *at risk* with 74% accuracy and an 8% false positive rate.

An important distinction between our work and the work in [115] is that our work is better suited to network administration contexts where an administrator would like to be able to forecast which users will become victims of threats or vulnerabilities as identified by the monitoring tool implemented on the network, but due to the privacy expectations of users they cannot instrument user devices, require specific antivirus products, or know the actual content users are accessing on the web. Our work attempts to identify indicators and make predictions about users in a privacy-preserving manner using network patterns alone. The actual specifics of our data will be discussed in detail in Section 5.3.

Based on the related work that has shown how features can be extracted from network traffic to either uniquely identify or profile different types of users, the first research question for this study is:

Research Question 5.1 (RQ5.1): *Are there features that can be extracted from the network traffic that best explain the differences in user behavior?*

Motivated by the extensive related work that has shown how features extracted from historic network traffic can be used to develop profiles that distinguish between malicious and benign hosts or IP addresses (and in the case of [115], *at risk* users), we examine whether the features we extract to explain user behavior can also be used to distinguish between infected and uninfected users. The second research question for this study is therefore:

> **Research Question 5.2 (RQ5.2):** *How well can we differentiate between infected and uninfected users on the network using features derived from user network traffic?*

We evaluate RQ5.2 using both an unsupervised clustering and supervised learning approach. In the context of unsupervised clustering, we evaluate cluster purity, i.e. the percent of users in a cluster that represent only one type of user (infected or uninfected). In the context of supervised learning, we evaluate accuracy, area under the receiver operating characteristic curve (ROC AUC), false positive rate, and false negative rate. The technique to compare these four attributes is discussed in Section 5.3.

5.3 Methodology

This section presents an overview of the dataset, the steps taken to clean and process the data, and the experiments to test our research questions.

5.3.1 Data Overview

The data for this experiment consists of the wireless network traffic and wireless network threat events collected by a Palo Alto Networks Intrusion

Detection/Prevention System at a large, public university. The security division of the university's Information Technology (IT) department manages this system. From September 26 to October 18, the system version was Pan OS 7.0.10. It was upgraded to version 8.0.4 on October 18th and to 8.0.5 on October 25th. The system is configured to detect threats using the default Palo Alto signatures and detection rules. The IT department modifies the list of domains and IPs that the system's firewall is set to block based on a feed of active hostile threats updated every 15 minutes.

The IT security office provided us with a random sample of traffic records from each hour of each day for the period of September 26, 2017 through November 21, 2017. They provided us with a complete record of the threats during this time period.

Of all the traffic generated by users at the university, 65% is wireless. Users of the wireless network include students, faculty, staff, and guests at the university. Users must authenticate onto the wireless campus network using a unique user ID that is recorded by the Palo Alto system.

To guarantee privacy of the users, the IT security office replaced user IDs with unique hashes. Anonymous sessions or sessions associated with a router or device that did not map back to a specific user were discarded. The resulting data was then pushed to a secure server hosted at the university for our analysis. After the process of discarding and anonymization, the dataset consisted of 66,561,686 wireless traffic records and 14,621 threat records.

The traffic logs record wireless traffic sessions on the network, with a session representing a complete conversation between a source and destination IP address.

That is, a single record in a traffic log captures the request sent from a source IP to a destination IP and the response received by the source IP from the destination IP. The traffic logs record 54 fields associated with the session, including the source user ID, time the session began, source and destination IP, number of bytes sent and received, number of packets sent and received, length of the session, and so on.

The threat logs record events that occurred during wireless traffic sessions that were identified as malicious or suspicious by the Palo Alto system. Therefore, every record in the threat logs can be matched with a record in the traffic logs representing the session during which the threat occurred. If a threat is identified by the system during a wireless traffic session, once the session is complete the associated source and destination IP addresses and ports are recorded in the threat log, along with information about the protocol, time of the threat, type of threat, severity level, and so on. Severity levels are ranked as informational, low, medium, high, or critical. Our threat dataset consisted of the threats ranked medium, high, or critical, as these are the events that trigger a response from the IT department.

More information about the Palo Alto system and its traffic logs and threat logs can be found at [85].

5.3.2 Data Validation and Preprocessing

This section discusses the steps that were taken to validate and prepare the data for our analysis. Validation and preprocessing were done using Python, with the Pandas [86] and NumPy [87] packages.

The first step was to identify the number of unique users in our data. Filtering by unique user hash, we determined there were 53,165 unique users who appeared in the traffic logs. In the threat logs, we identified 1,935 unique users. Since every threat record is associated with a traffic session, any user who appears in the threat logs should also appear in the traffic logs. Comparing the users in the threat logs to the users in the traffic logs, we identified 12 users who never appear in the traffic logs. Because data is taken directly from the IDS/IPS threat logs, we assume these 12 users represent a corner case in the threat recording process. One possible explanation is if a threat is identified at the start of a session and is immediately blocked, a completed traffic session never occurs and therefore is never recorded in the traffic log. These 12 users and their records in the threat logs were dropped from the analysis, resulting in 1,923 unique users in the threat logs. We denote the 1,923 users who appeared in the threat records – in other words users who have created threats assigned a severity level of medium, high, and/or critical – as the *infected users.*

The next step was ensuring that we could identify the traffic sessions associated with the threat records for these 1,923 users. For each infected user threat record, we filtered the traffic logs by the user's unique user hash. From this set of potential traffic sessions, we then filtered by the recorded time of the traffic session and the threat record. Since traffic records only appear in the traffic logs when the conversation between the source and destination IP has been completed or terminated and because some sessions may last several days, we first tested filtering the traffic sessions by a 48-hour window before the threat log was recorded and a 48-hour window after the threat log was recorded. From this filtered set of records, we checked if there was an IP address in the traffic record that matched an IP address in the threat record. After doing this process for all the threat records, there were only 28 records created by a total of 11 unique users that were not matched. By expanding the time window from 48 hours to 96 hours before and after a threat was recorded and then checking for a match on either source or destination IP address, we matched six more records associated with five unique users. By expanding our search to month-long windows and then searching for a match on source or destination IP, we matched three records created by two unique users. By relaxing the condition on matching a source and destination IP, we matched the remaining 19 records created by four unique users by first searching a 96-hour window and then expanding to a month-long window.

After this phase of preprocessing, there were 1,923 unique infected users and 51,242 unique uninfected users. Using the entire set of uninfected users in our analysis would result in artificially high accuracy results when attempting to classify users as infected or uninfected. So, we conducted our analysis on two different ratios of infected and uninfected users. One ratio represented a 50:50 split of infected to uninfected users, so 1,923 infected and 1,923 uninfected. The uninfected users represented a random sample of the total uninfected population. To ensure the results were not the product of the random sample, we produced three different random samples of uninfected users and repeated our analysis of the 50:50 ratio for each sample.

In order to also have a dataset with an imbalance of infected and uninfected users, we also selected a 30:70 ratio of infected to uninfected, so 1,923 infected and 4,487 uninfected. Again, we conducted our analysis on three random samples containing 4,487 uninfected users each.

The final step in our data-preprocessing phase was to analyze and clean the fields in the actual traffic sessions. In the traffic logs, there are 54 fields recorded for a single traffic session. These fields are defined by Palo Alto so not all of them are relevant on the university network. We reduced this list of 54 fields to ten fields after dropping fields that contained missing values, always recorded the same value, or provided only minimal information about a traffic session. The final fields retained for feature engineering were: *Start Time, Elapsed Time, Source IP, Source Port, Destination IP, Destination Port, Bytes Sent, Bytes Received, Packets Sent,* and *Packets Received.*

5.3.3 Feature Engineering

The ten features were then transformed into numerical features that summarized all of a user's traffic records. Feature engineering was driven both by exploratory data analysis and related work that has identified useful features for differentiating between malicious and benign IP addresses [74, 78, 79, 80]. This process resulted in 36 different features, presented in Table 1.

To account for the fact that some features are quantified using different units of measurement (for example, frequency count vs. time) and to ensure each variable received equal weight, feature values were standardized prior to analysis using the process described in [88].

Code	Feature	Code	Feature
NREC	Number of user records in the traffic logs	***0600***	Number of sessions initiated by a user between 0600 and 0659
LEN	Average session length	***0700***	Number of sessions initiated by a user between 0700 and 0759
TDIF	Average time difference between session start times	***0800***	Number of sessions initiated by a user between 0800 and 0859
BYS	Average number of bytes sent	***0900***	Number of sessions initiated by a user between 0900 and 0959
BYR	Average number of bytes received	***1000***	Number of sessions initiated by a user between 1000 and 1059
PAS	Average number of packets sent	***1100***	Number of sessions initiated by a user between 1100 and 1159
PAR	Average number of packets received	***1200***	Number of sessions initiated by a user between 1200 and 1259
SIP	Number of unique source IP addresses	***1300***	Number of sessions initiated by a user between 1300 and 1359
SP	Number of unique source ports	***1400***	Number of sessions initiated by a user between 1400 and 1459
DIP	Number of unique destination IP addresses	***1500***	Number of sessions initiated by a user between 1500 and 1559
DP	Number of unique destination ports	***1600***	Number of sessions initiated by a user between 1600 and 1659
NAPP	Number of applications used	***1700***	Number of sessions initiated by a user between 1700 and 1759
0000	Number of sessions initiated by a user between 0000 and 0059	***1800***	Number of sessions initiated by a user between 1800 and 1859
0100	Number of sessions initiated by a user between 0100 and 0159	***1900***	Number of sessions initiated by a user between 1900 and 1959
0200	Number of sessions initiated by a user between 0200 and 0259	***2000***	Number of sessions initiated by a user between 2000 and 2059
0300	Number of sessions initiated by a user between 0300 and 0359	***2100***	Number of sessions initiated by a user between 2100 and 2159
0400	Number of sessions initiated by a user between 0400 and 0459	***2200***	Number of sessions initiated by a user between 2200 and 2259
0500	Number of sessions initiated by a user between 0500 and 0559	***2300***	Number of sessions initiated by a user between 2300 and 2359

Table 1: Final list of features

5.3.4 Principal Component Analysis

Principal Component Analysis (PCA) is a statistical dimensionality reduction technique that converts a set of potentially correlated features into a smaller set of orthogonal linearly uncorrelated features called the principal components that explain as much variance in the data as possible [69]. PCA has been used in the related work [8] to identify subsets of features from larger sets of features derived from traffic record data that are the most useful for understanding normal IP/host behavior.

PCA was performed using Python's sklearn.decomposition.PCA library [89] on the standardized data. Scree plots were produced for each sample to determine the number of principal components that should be retained for analysis. We examined the Scree plots to see at what number of components the graphs began to level off. For the three samples with a 50:50 ratio, the graphs all leveled off at 13 components, providing us with percent variance explained ratio values of 90.01%, 90.22%, and 98.09%, respectively. Using similar reasoning for the 30:70 ratio samples, we also retained 13 components for all three cases, achieving percent variance explained ratios of 90.27%, 90.39%, and 97.71%.

We then examined how each of the original 36 features contributed to the percent variance explained and the principal components. Figure 4 displays the average percent variance explained by each feature for the three samples in the 50:50 dataset and the average percent variance explained by each feature for the three samples in the 30:70 dataset.

The correlation values between each feature and each of the 13 principal components were also calculated for each sample using the sklearn.decomposition.PCA library. Absolute values of the correlation values were then calculated, because based on the method we used for calculating these variables, magnitude, not direction, is used for determining variable importance [90]. Tables 2 and 3 present the correlation values between features and principal components where the correlation value was greater than 0.5; we discuss these features in more detail in the Section 5.4.

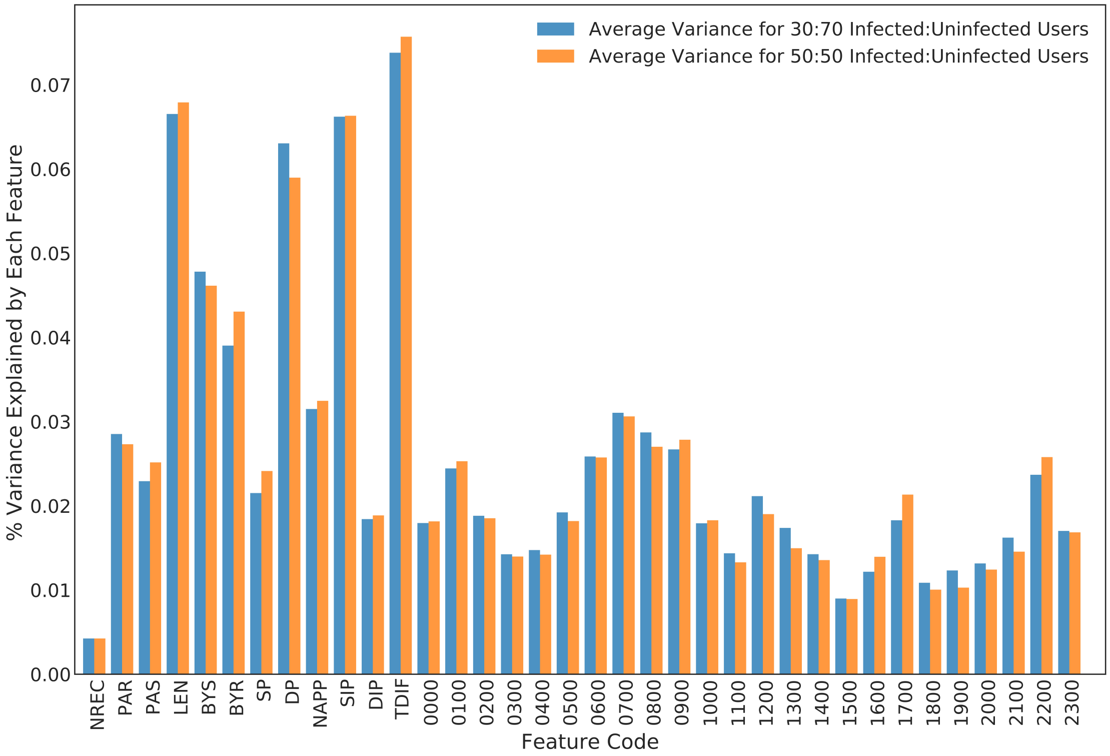

Figure 4: % Average variance explained by each feature for the 13 components for both ratios

.

Feature code	**Principal Components**									
	PC2	***PC5***	***PC6***	***PC7***	***PC8***	***PC9***	***PC10***	***PC11***	***PC12***	***PC13***
PAR	0.557 0.565 --									
PAS	0.557 0.562 --									
LEN			0.578 -- --	0.640 -- --		-- 0.508 --	-- -- 0.765			
BYS				-- 0.811 --	0.871 -- --		-- -- 0.607			-- -- 0.618
BYR	0.548 0.535 --									
DP				0.513 -- 0.775	-- 0.52 --					
SIP							-- 0.519 --			
TDIF			-- -- 0.516		-- 0.514 0.699	0.667 0.507 --				
1400								-- -- 0.546	0.721 -- 0.546	
1500		-- -- 0.501								
1900										-- 0.714 --

Table 2: Correlation values between features and principal components for the three samples of the 30:70 ratio where the correlation value was greater than or equal to 0.5

Feature code	**Principal Components**									
	PC2	***PC5***	***PC6***	***PC7***	***PC8***	***PC9***	***PC10***	***PC11***	***PC12***	***PC13***
PAR	0.556 0.556 --									
PAS	0.553 0.545 --									
LEN					0.763 -- --	-- 0.522 --				
BYS				0.898 0.807 --				-- -- 0.833		
BYR	0.540 0.516 --									
DP		0.515 -- --	-- -- 0.744							
SIP		-- -- 0.499			-- -- 0.500					
TDIF		-- -- 0.557	0.545 -- --	-- -- 0.683	-- 0.601 --	0.700 -- --				
1400							-- -- 0.751		0.779 0.737 --	

Table 3: Correlation values between features and principal components for the three samples of the 50:50 ratio where the correlation value was greater than or equal to 0.5.

5.3.5 Unsupervised Clustering

Motivated by related work [79, 80] that has explored the feasibility of using K-means clustering to differentiate between normal and anomalous traffic, we apply K-means clustering to our six samples using both the 36 original features and the 13 principal components as input. K-means clustering works by grouping entities into a user-specified number of clusters based on similarities between the features of the entities.

We evaluated K-values ranging from [1, 5] and for each K, calculated the sum of squares difference between the cluster center and each observation in the cluster. This value is 0 when all observations in a cluster are identical, so using this approach, the ‘best’ value of K is the one in which the sum of squares distance is minimized [91]. Based on the sum of squares difference, the best value was K=4.

Once the algorithm partitioned users into clusters, we examined how many infected and uninfected users appeared in each cluster. Note that K-means is an unsupervised technique, so the algorithm did not have knowledge of these labels. Below we present the results of the percent and number of infected and uninfected users in each cluster for K=4 for all three samples for each of the 30:70 and 50:50 ratios; Table 4 presents the clustering results for the 36 original features while Table 5 presents the clustering results for the 13 principal components.

% Infected (# infected) in each cluster % Uninfected (# uninfected) in each cluster			
Cluster 1	***Cluster 2***	***Cluster 3***	***Cluster 4***
Original 36 features, 30:70 Ratio			
12.15% (489) 87.85% (3537)	**76.39% (356)** 23.61% (110)	**56.18% (1077)** 43.82% (840)	**100.0% (1)** 0.0% (0)
10.95% (416) 89.05% (3382)	**71.97% (534)** 28.03% (208)	**51.37% (938)** 48.63% (888)	**79.55% (35)** 20.45% (9)
14.47% (635) 85.53% (3754)	**0.0% (0)** 100.0% (1)	**63.79% (1288)** 36.21% (731)	**0.0% (0)** 100.0% (1)
Original 36 features, 50:50 Ratio			
88.54% (340) 11.46% (44)	**25.27% (514)** 74.73% (1520)	**74.84% (1068)** 25.16% (359)	**100.0% (1)** 0.0% (0)
25.01% (513) 74.99% (1538)	**89.09% (343)** 10.91% (42)	**75.66% (1066)** 24.34% (343)	**100.0% (1)** 0.0% (0)
29.49% (681) 70.51% (1628)	**0.0% (0)** 100 (1)	**0.0% (0)** 100% (1)	**80.91% (1242)** 19.09% (293)

Table 4: K-Means clustering result for K=4 for each data sample using the 36 original features.

5.3.6 Supervised Learning

To use our dataset for supervised learning, infected users were labeled with a 1 and uninfected users were labeled with a 0. Each sample in the 30:70 and 50:50 infected to uninfected ratio datasets were both split into training and testing sets. We evaluated two common train/test splits: 70/30 and 80/20.

Eight supervised algorithms were implemented using Python's sklearn libraries: K-Nearest Neighbors [92], Logistic Regression [93], Random Forest [94], AdaBoost classifier [95], Gradient Boost [96], Extra Tree [97], Bagging [98], Voting (Majority Rule) [99]. For all the algorithms except Voting (which is built on the output of the other algorithms), Grid Search [100] was used to tune the hyper-parameters and stratified 10-fold cross validation [101] was conducted to prevent model overfitting. Two neural networks were implemented using the keras.io library [102], one with two layers, denoted Neural Network 1, and one with four layers, grid search to tune parameters, and 10-fold cross validation to prevent overfitting, denoted Neural Network 2. For all models, we calculate the accuracy, false positive rate, false negative rate, and area under the receiver operating characteristic curves produced by varying the decision threshold on balancing the true positive rate with the false positive rate.

% Infected (# infected) in each cluster % Uninfected (# uninfected) in each cluster			
Cluster 1	***Cluster 2***	***Cluster 3***	***Cluster 4***
13 Principal Components, 30:70 Ratio			
12.19% (424) 87.81% (3053)	**80.0% (16)** 20.0% (4)	**51.85% 1483)** 48.15% (1377)	**0.0% (0)** 100.0% (53)
60.07% (1309) 39.93% (870)	**14.19% (595)** 85.81% (3598)	**18.18% (4)** 81.82% (18)	**93.75% (15)** 6.25% (1)
61.03% (1339) 38.97% (855)	**13.86% (584)** 86.14% (3630)	**0.0% (0)** 100.0% (1)	**0.0% (0)** 100.0% (1)
13 Principal Components, 50:50 Ratio			
81.85% (496) 18.15% (110)	**26.33% (465)** 73.67% (1301)	**94.12% (16)** 5.88% (1)	**64.93% (946)** 35.07% (511)
79.28% (1213) 20.72% (317)	**51.28% (20)** 48.72% (19)	**30.32% (690)** 69.68% (1586)	**0.0% (0)** 100.0% (1)
78.46% (1235) 21.54% (339)	**30.31% (688)** 69.69% (1582)	**0.0% (0)** 100.0% (1)	**0.0% (0)** 100.0% (1)

Table 5: K-Means clustering result for K = 4 for each data sample using the principal components

In Table 6 and Table 7 we present the average accuracy, ROC AUC, false positive rate (FPR), and false negative rate (FNR) across all three samples for the 50:50 ratio and 30:70 ratio, respectively. We do not present the results of all classifiers, but rather we present the *non-dominated* classifiers when comparing classifiers on four attributes: accuracy and ROC AUC (both of which we aim to maximize) and false positive rate and false negative rate (both of which we aim to minimize. Results for all classifiers can be found in Appendix A.

Non-dominated classifiers were determined using the dominance choice strategy [116], a technique commonly used in engineering disciplines to select the best alternative using multiple attributes as the criteria. To apply the technique, we eliminated classifiers if there was another classifier that performed better on all four attributes until we were left with a list of classifiers that could no longer be eliminated. For example, when identifying the non-dominated classifiers for the 50:50 ratio using the 36 original features and an 80/20 train/test split, Neural Network 1 was eliminated because Gradient Boost had a higher accuracy, higher ROC AUC, lower false positive rate, and lower false negative rate than it did. Gradient Boost was not eliminated because there was no other classifier that performed better across *all* four attributes. Note that the dominance calculations were made using six significant digits, but the tables present the values rounded to three significant digits for clarity and simplicity.

Classifier	*Accuracy*	*ROC AUC*	*FPR*	*FNR*
30:70 Ratio, 36 original features, 70/30 train/ test split				
KNeighbors	0.784	0.851	0.211	0.229
Gradient Boost	0.784	0.853	0.215	0.217
Extra Tree	0.781	0.832	0.168	0.340
30:70 Ratio, 36 original features, 80/20 train/ test split				
Gradient Boost	0.785	0.850	0.214	0.217
Extra Tree	0.775	0.830	0.169	0.360
30:70 Ratio, 13 principal components, 70/30 train/ test split				
Random Forest	0.781	0.809	0.211	0.238
Ada Boost	0.780	0.845	0.211	0.239
Gradient Boost	0.780	0.849	0.220	0.221
Extra Tree	0.770	0.821	0.176	0.358
30:70 Ratio, 13 principal components, 80/20 train/ test split				
Extra Tree	0.774	0.824	0.196	0.230
Neural Network 2	0.781	0.847	0.218	0.220

Table 6: Supervised results for the 30:70 ratio dataset

Classifier	*Accuracy*	*ROC AUC*	*FPR*	*FNR*
50:50 Ratio, 36 original features, 70/30 train/ test split				
Ada Boost	0.768	0.836	0.214	0.250
Gradient Boost	0.784	0.851	0.215	0.217
50:50 Ratio, 36 original features, 80/20 train/ test split				
Ada Boost	0.775	0.849	0.201	0.249
Gradient Boost	0.784	0.860	0.213	0.218
Bagging	0.784	0.854	0.211	0.221
Neural Network 2	0.788	0.853	0.210	0.214
50:50 Ratio, 13 principal components, 70/30 train/ test split				
Ada Boost	0.737	0.837	0.175	0.351
Gradient Boost	0.783	0.847	0.215	0.218
Extra Tree	0.752	0.810	0.203	0.294
50:50 Ratio, 13 principal components, 80/20 train/ test split				
Gradient Boost	0.787	0.854	0.212	0.214
Extra Tree	0.710	0.822	0.172	0.413
Neural Network 2	0.790	0.852	0.208	0.212

Table 7: Supervised results for the 50:50 ratio dataset

5.4 Results

This section provides a more in depth discussion of the results of Principal Component Analysis (PCA), K-means clustering, and supervised learning.

5.4.1 Principal Component Analysis

For the datasets with a 50:50 ratio of infected to uninfected users, the average variance in the data explained by the 13 principal components was 92.77%. As was shown in Figure 4, for all three samples there were about five different features that contributed the most variance to the overall variance: average session length, average number of bytes sent, number of unique destination ports, number of unique source IP addresses, and average time difference between session start times.

By examining the correlation values between the principal components and the original 36 features, we can further understand the network traffic features that contribute most to explaining the differences in users. Determining the threshold for a correlation value to be large enough for a feature to be considered important is a subjective decision [117]; higher values reduce the number of features considered important, while lower values increase the number of features.

For the 50:50 ratio, if we consider 0.4 to be the threshold on importance we retain 17 features; a threshold of 0.5 retains 10 features; 0.6 retains six features; 0.7 retains two features; and 0.8 retains one feature, a correlation value of 0.885 between average number of bytes sent and principal component 7 (PC7).

Setting the threshold at 0.5, the ten features considered important are: average number of packets received, average number of packets sent, average session length, average number of bytes sent, average number of bytes received, number of unique source ports, number of unique destination ports, number of unique applications used, number of unique source IP addresses, and average difference between session start times. Five of these features overlap with the five features that contribute the most variance to the overall variance.

For the datasets with a 30:70 ratio of infected to uninfected users, the average variance in the data explained by the 13 principal components was 92.79%. The same five features as in the 50:50 ratio dataset contributed the most variance to the overall variance: average session length, average number of bytes sent, number of unique destination ports, number of unique source IP addresses, and average time difference between session start times.

Again, if we consider 0.4 to be the threshold on importance we retain 17 features; a threshold of 0.5 retains nine features; 0.6 retains six features; 0.7 retains four features; and 0.8 retains two features, with the highest correlation value being 0.895 again between average number of bytes sent and PC7.

Setting the threshold at 0.5, the nine features considered important are: average number of packets received, average number of packets sent, average session length, average number of bytes sent, average number of bytes received, number of unique destination ports, number of unique applications used, number of unique source IP addresses, and average difference between session start times. Nine of these features overlap with the ten features presented in the 50:50 ratio; in this ratio, only number of unique source ports is no longer considered important.

In Research Question 5.1, we asked if we could derive features from network traffic that could explain the difference in user behavior, motivated by existing research that has suggested network traffic patterns can function as user fingerprints. Overall, comparing the PCA results of the two different ratios, there is a total of 10 unique features that, either based on the percentage of total variance explained or correlation values above 0.5 with individual principal components, are most important in shaping the final principal components and explaining user behavior.

5.4.2 Unsupervised Clustering

We performed K-means clustering using both the original 36 features as input (see Table 4), and the 13 principal components as input (see Table 5). For all data samples the partitioning of users created by the K-means algorithm often aligned with the infected/ uninfected labels that we assigned to users.

Examining the clusters created using the 36 original features as input, for the 50:50 ratio there are a few instances where clusters consist of a single user; we interpret Cluster 4 in the first and second samples and Clusters 2 and 3 in the third sample as potential outliers. Excluding these instances, we see that cluster purity ranges from about 70.51% (1628 uninfected users in Cluster 1 in the third data sample) to 89.09% (343 infected users in Cluster 2 in the second data sample). For the 30:70 ratio, we similarly see a majority of high cluster purity instances and three outlier instances. We also see instances where the cluster composition is split almost evenly between infected and uninfected users; for example, in the second sample, 51.37% of Cluster 3 represents infected users and 48.63% represents uninfected users.

When we examine the results of K-means using the 13 principal components, for the 50:50 ratio there are three outlier instances and one cluster instance that splits infected and uninfected users evenly. Besides these instances, cluster purity values range from 64.93% (946 infected users in Cluster 4 in the first sample) to 94.12% (16 infected users in Cluster 3 in the first sample). For the 30:70 ratio, there are two outlier instances and one instance where a cluster is evenly split. Besides these instances, cluster purity ranges from 60.07% (1,309 infected users for Cluster 2 in the second sample) to 100.0% (53 uninfected users for Cluster 4 in the first sample).

By the nature of unsupervised clustering, we do not have insight into how the algorithm made its decisions to partition users. Our interest in applying K-means was to understand if there was a signal of infection status present in the network traffic data; since the algorithm had no insight into user infection status, users were partitioned according to whatever features or combinations of features K-means determined provided the strongest signal for separating users into clusters.

Our results suggest that a signal for infection status is present in both the 36 features we extracted from the network traffic records and in the 13 principal components derived from these features. This implies both may have potential for our long-term goal to develop a set of user infection symptoms. However, because some clusters were partitioned almost evenly between infected and uninfected users, to validate the strength of the infection signal derived from network traffic, further analysis over longer periods of time and for additional samples of users is necessary.

Recall that Research Question 5.2 asked what is the highest cluster purity of infected or uninfected users we can achieve when using unsupervised clustering to differentiate between users. Using the network traffic, the highest purity we achieve is 89.09% infected users; using the principal components, the highest purity we achieve is 100.0% infected users.

5.4.3 Supervised Learning

For the 50:50 ratio of infected to uninfected users, classification accuracies of the top performing models ranged from 71.0% to 79.0%, ROC AUCs ranged from 81.0% to 86.0%, false positive rates ranged from 17.2% to 21.5%, and false negative rates ranged from 21.2% to 41.3%. While Bagging is listed as a top performing classifier (see Table 6), its training accuracy was at least 10% higher than its testing accuracy, suggesting that the model may have been overfitting.

For the 30:70 ratio of infected to uninfected users, classification accuracies of the top performing models ranged from 77.0% to 78.5%, ROC AUCs ranged from 80.9% to 85.3%, false positive rates ranged from 16.8% to 22.0%, and false negative rates ranged from 21.7% to 36.0%.

In deciding which algorithm results to include in this case study, we chose to consider the four different attributes of equal importance. In practice of course, a decision-maker would need to make a determination about what attribute(s) is most important. If the goal is to supplement a firewall or anti-virus system, i.e. a disease confirmation test, with indicators of user infection, i.e. warning symptoms that trigger early intervention – as is the intended goal of our work - then high false positive rates may not be as detrimental or concerning as high false negative rates and the number or rate of true positives may be considered most important. In this case, an algorithm like Neural Network 2 for the 50:50 ratio using the 13 principal components and an 80/20 train/test split performs quite well, with an accuracy of 79.0%, an ROC AUC of 85.2%, an FPR of 20.8%, and an FNR of 21.2%.

Recall that Research Question 5.2 asked how well we can classify users as infected or uninfected using supervised learning. For the 50:50 ratio, there is no one model that outperforms all the others on every attribute, but examining the attributes individually and excluding the Bagging classifier because of overfitting: 79.0% accuracy by Neural Network 2; 86.0% ROC AUC by Gradient Boost; 17.2% FPR by Extra Tree; and 21.2% FNR by Neural Network 2. Examining attributes individually for the 30:70 ratio: 78.5% accuracy by Gradient Boost; 85.3% ROC AUC by Gradient Boost; 16.8% FPR by Extra Tree; and 21.7% FNR by Gradient Boost.

5.5 Discussion

We extracted 36 features from the traffic records based on exploratory analysis and an extensive review of the related work. We identified features that summarize the data sent and received during traffic sessions, source and destination IP and port information, and temporal patterns of user network activity. Our features are suited to a network administration context where it is not possible to instrument devices or know the specific content that users are accessing on the web due to the privacy expectations of users. We therefore investigate the feasibility of understanding user infection status in a context-*unaware* manner.

We applied PCA to identify 13 principal components that best explain the variance in user behavior. These 13 components are able to explain 92.77% variance on average for the three data samples with a 50:50 ratio of infected to uninfected users; for the three data samples with the 30:70 ratio, the components explain on average 92.79% of the variance. After analyzing how each of the original 36 features contributed to the overall variance, we identified five features that may be of more importance than others in explaining the difference in user behavior: average session length, average number of bytes sent, number of unique destination ports, number of unique source IP addresses, and average time difference between session start times. After analyzing the correlation values between the original features and the 13 principal components and using a correlation value of 0.5 as a threshold for feature importance, we confirmed these five features and identified five additional features: average number of packets received, average number of packets sent, average number of bytes received, number of unique source ports, and number of unique applications used.

K-means clustering was performed using both the 36 original features and the 13 principal components. After the algorithm partitioned the users in to 4 distinct clusters based on the optimal K value, we identified the number and percent of infected and uninfected users that were partitioned into each cluster. Despite a few clusters consisting of a single user and a few clusters split evenly between infected and uninfected users, cluster purity ranged from 60.07% up to 100%, suggesting a signal of infection status exists in network traffic patterns.

Finally, we also tested an array of supervised learning algorithms for both data ratios and two different training/ testing splits to ensure any results we obtained were not just the product of our selected parameters. Though at a minimum our false positive rates are 16.8% and false negative rates are 21.2% - in other words, quite high - we were able to classify users as either infected or uninfected with accuracies up to 79.0% and ROC AUCs up to 86.0%, further showing that features extracted from the network traffic records of users can differentiate between the infected and uninfected.

5.6 Limitations

Our study has several limitations. First, we used traffic records flagged for containing threat activity by the university's IDS/IPS as ground truth for defining infection status. We do not know the specifics of how the system determines threats, and we also do not know the system's false positive rate or false negative rate. While knowing the labeling process might give us more insight into why users do or do not appear in the threat logs, understanding the nuances of the system is not relevant to our goal, as the IDS/IPS is simply what triggers a response from the network team at the university. Revisiting the disease analogy we presented earlier, developing a method to confirm a disease is a separate problem from identifying symptoms of the disease; the threat labeling process is analogous to a disease confirmation test. Moreover, the system makes infection determinations with insight into the content a user is accessing, to include specific websites and downloads, while we are making determinations using only network traffic patterns.

Another limitation is how we chose to define the infected users. As noted previously, the Palo Alto IDS/IPS tags each threat with a severity level of informational, low, medium, high, or critical. We labeled users as 'infected' if they created medium, high, or critical threats and did not differentiate between infection severities. 1,858 users created medium severity threats only, 41 created high severity threats only, and 11 created critical threats only. If we consider users who created high and/or critical threats as the infected population, the differences between the infected and uninfected may be more pronounced. However, our sample of critical and high severity threat users is currently too small to effectively apply the unsupervised or supervised algorithms presented here.

The next limitation is that we were unable to verify completeness of the traffic data. Though the traffic logs should capture every session that occurs on the university network and this data is stored in a redundant manner by the IT division, due to the sheer volume of data it was infeasible for us to validate that we have every record created by every user during our two-month window of data.

Finally, our study focused on users at a university network. This was motivated by the availability of ground truth data to match users with network traffic flows. Therefore, though our methodology generalizes to other networks, it is possible that the network indicators identified in this experiment do not. However, the advantage of using an open, university network is that there is the possibility of discoveries about users that would not be possible on a closed network. Unlike many corporate networks where extensive security checks and restrictions are in place, at the university users maintain their own devices, there are no restrictions on the type of devices or the software on them, and almost no websites or services are blocked. This enabled us to analyze how users behave on a network on which they have almost complete freedom.

5.7 Conclusion

In this chapter, we conducted an empirical study on the feasibility of extracting indicators from the network traffic of users that can be used as early warning symptoms of infection. Motivated by work that has used insights from network traffic to proactively identify malicious IPs or infected hosts and more recent work that has suggested users have unique network traffic fingerprints, we analyzed two months of wireless network traffic for 1,923 infected users and random samples of uninfected users at a large public university.

Though our classification results are not successful enough to merit use in practice, they do show that network traffic records have the potential to reveal differences in users. In the following chapter, we show how these features and insights can be used to quantify and predict user reputation, keeping in mind the uncertainty that exists in the features themselves and ensuring this uncertainty is reflected in the predictions.

6 Quantifying and Predicting User Reputation via Network Traffic: Demonstrating the Proposed Methodology Through a Case Study

6.1 Introduction

Recall from Chapter 4 that the process of evaluating user reputation consists of three major phases: Phase One - Characteristic Definition and Evidence Collection; Phase Two - Reputation Quantification and Prediction; and Phase Three - Validation and Refinement.

The previous chapter illustrated Phase One of the methodology. We selected *infection* as our characteristic of interest and defined infected users as those who appeared in the threat records of the university's intrusion detection/ prevention system (IDS/IPS). Using wireless traffic data provided by the university's Division of Information Technology, we extracted 36 features and investigated the feasibility of using these features as evidence of infection. Framing this feasibility study as a supervised classification problem in which we tried to differentiate between infected and uninfected users, we achieved AUC ROC values up to 0.86. Based on these results, we hypothesize that network traffic activity of users is an appropriate source of initial evidence for making predictions about user reputation. This chapter focuses on illustrating Phases Two and Three of the methodology through this university network case study. We use data from September 2017 through December 2017 to build reputation models for infected and uninfected users. We experiment with different parametric model building techniques, using individual features and combinations of features.

We then test reputation scoring using these models on samples of users who appeared in the traffic records in 2018. We update the user scores over the course of three months by incorporating new evidence from additional network activity generated by the user. We validate the model by observing whether or not users become infected or remain uninfected.

This chapter makes the following contributions:

- We illustrate Phases Two and Three of the reputation evaluation methodology by quantifying and predicting user reputation scores and validating and refining these scores using the network traffic data and insights presented in the previous chapter.
- We illustrate the process of developing reputation models using several different model-building techniques. We compare scores associated with different types of models and discuss the advantages and challenges associated with different modeling approaches.
- We show the feasibility of using Bayesian-based probabilities as reputation scores and the valuable information regarding confidence in a score gained by using this approach.
- Using the Brier Score as a metric to confirm the accuracy of our predicted reputation scores, we show the potential of using network traffic data to make predictions about user reputation for infection. We obtain a Brier Score of 0.296 by the end of our three-month pilot study (recall from Chapter 4 that 0 is the best possible score while 1 is the worst and indicates a forecast is always incorrect). By recalculating the Brier Score for each month in our pilot

study sample, we also illustrate how adding more evidence can help improve prediction accuracy.

- By analyzing the Brier Scores and 95% confidence intervals associated with our reputation scores, we emphasize the importance of varied sources of evidence for predicting user reputation and motivate future work to obtain additional sources of data that can help to improve the reputation predictions and narrow the confidence intervals.

The remainder of this chapter is organized as follows. Section 6.2 presents the background on this chapter. Section 6.3 presents the methodology, which discusses the process for developing, quantifying, validating, and refining user reputation models and scores. This is followed by the results in Section 6.4 and then a discussion of the case study and the proposed methodology in Section 6.5. Section 6.6 presents the limitations; 6.7 presents areas of future work; and lastly, 6.8 presents concluding comments on this case study.

6.2 Background

This section revisits Phases Two and Three of the methodology presented in Chapter 4, providing additional background on how they will be applied to this case study.

6.2.1 Phase Two: Reputation Quantification and Prediction

Phase Two of the methodology provides an outline for quantifying user cyber reputation and developing predictive models of user reputation. Recall that the methodology and scoring technique are based on a Bayesian Inference process. In

Chapter 4, we presented a simple example of applying Bayesian Inference using a Beta-Binomial conjugate pair to model and update evidence, motivated by related work in the field of e-commerce.

The Beta-Binomial approach works well if the evidence is collected as counts of behaviors that reflect the characteristic of interest; for example, if we are interested in a user's reputation for being trustworthy and are counting how often a user behaves in a trustworthy manner or in an untrustworthy manner. This case study is somewhat more complicated, so using a Beta-Binomial approach to update counts of evidence is not feasible.

Additionally, another nuance to consider is that due to the nature of our evidence and the characteristic of interest, we are not actually just looking at one characteristic, we are looking at two: the likelihood of being infected and the likelihood of being uninfected. Going back to the example of trustworthiness, one could think of being trustworthy and untrustworthy as being on opposite ends of a scale, in which case reputation can measure the degree of trustworthiness based on evidence. Infection status is not relevant on such a scale, especially because our evidence illustrates how the infected and uninfected users behave on the network, not the 'degree' with which a user is infected.

So, when quantifying user reputation in this case study, we are really trying to answer the following question: given evidence about how a user behaves on a network and historical data about how infected and uninfected users behave on the same network, what is the most likely probability the user belongs to either the infected population or the uninfected population?

To answer this question, we must first build models (distributions) that summarize the behavior of known infected users and known uninfected users. We can then determine a user's reputation score by evaluating how likely a user is to belong to either model. The methodology section (6.3) shows the process for model construction and probability calculation.

6.2.2 Phase Three: Validation and Refinement

Phase Three of the methodology provides a process for validating and refining the reputation scores. A pilot study can be conducted in which a sample of users are evaluated and monitored over time and the predicted reputation scores are compared against the actual outcomes. As discussed in Chapter 4, there are various forecast verification techniques that can be employed to assess the success of a prediction. In the context of this case study, validation will involve calculating reputation for a sample of users, monitoring these users over a three month period, observing if the users do actually become infected during the three month period, and calculating the Brier Score, a metric that provides the accuracy associated with a probabilistic forecast [104]. Additionally, we quantify the 95% confidence intervals associated with reputation scores and use these intervals to understand the confidence in a score.

Chapter 4 proposed several options for refining the reputation models and scores. In this case study, we illustrate refining the models by experimenting with different parametric model building techniques, in which we fit distributions to individual features or combinations of features. In addition to exploring different model construction approaches, we also show how scores can be updated with new evidence.

There were other techniques for model and score refinement presented in Chapter 4, such as finding different sources of data to use as evidence, combining multiple sources of data into the same reputation model, and altering how the characteristic of interest is defined. These are all considered areas of future work and are discussed later in this chapter and in the conclusion of this dissertation.

6.3 Methodology

This section presents an overview of the data used in this case study and the techniques to construct reputation models and assign users reputation scores.

6.3.1 Data

As in Chapter 5, the data consists of the wireless network traffic and wireless network threat events collected by a Palo Alto Networks IDS/IPS at the university. We receive random samples of traffic records for each hour of the day and each record is labeled with an anonymized, unique user hash.

6.3.1.1 Reputation Model Data

The data to build the reputation models was taken from the period of September 26, 2017, through December 24, 2017. There were 2,910 total unique users who appeared in the threat logs during this time period, and we defined these users as the infected users. We took a random sample of 2,910 users who appeared in the traffic logs but not in the threat logs during this same time period and defined these users as the uninfected users. We then calculated the 36 network traffic features for each user (refer back to Table 1 in Chapter 5 for a complete list of these features).

6.3.1.2 Pilot Study User Sample

The percentage of infected users on the network is small compared to the total number of users; for example, in Chapter 5 infected users represented about 3.8% of all users over a two-month period. If we were to take a random sample of users to test the reputation evaluation methodology, it is possible that it would result in an imbalanced dataset of infected and uninfected users. To aid us in testing the methodology, we iteratively identified infected users over the course of January, February, and March of 2018. Each month, we also took random samples of uninfected users and assigned these users reputation scores so we could evaluate how well the models perform on assessing uninfected users.

We began with the month of January 2018, during which there appeared 51,123 total unique uninfected users and 764 infected users. We filtered out any users who overlapped with the users used to construct the reputation models. This left us with 46,596 uninfected users and 532 infected users. We took all 532 infected users to be our sample of infected users. We also took a random sample of 532 uninfected users from the total uninfected population active on the network in January.

In February 2018, we updated our sample of 532 infected users and 532 uninfected users with any new network activity conducted by these users. In February, there were an additional 599 infected users on the network. Of the 599 infected users, 584 existed on the network in January but were uninfected. We added the January data associated with these users to our sample of now 1,131 infected users so that we could calculate and analyze their reputation scores prior to infection.

The remaining 15 users of the 599 newly infected users were new to the network in February, so no previous January data existed for these users.

In February there were 50,329 unique uninfected users on the network. We took an additional random sample of 599 users from this population and added it to our sample of uninfected users, bringing our sample of uninfected users to 1,131 total users. If the user existed in January but was not already part of our sample, we included any network activity conducted by this user in our data. 55 of these users did not exist on the network in January, so there was no additional data to add.

We applied the same dataset building approach in March, updating the data for users currently in our infected and uninfected sample with any additional network activity. In March, there were an additional 3,656 infected users on the network. We added these 3,656 infected users to our sample of infected users, bringing the infected user sample total to 4,787. Of these 3,656 users, 3,468 existed on the network in January and February but were uninfected during those months. 94 of the 3,656 users were new to the network in February and only became infected in March. Again, we updated the network traffic features for these users using the data from January and February. An additional 94 of these 3,656 users were new to the network in March and became infected in March, and thus there was no data from the previous months to include.

In March 2018 there were 51,035 unique uninfected users on the network. We again took a random sample of the uninfected users in March, adding an additional 3,656 uninfected users so that our total sample of uninfected users by the end of March was 4,787. We included data for these users from the previous months if it was available; 282 of these users were new to the network in March so there was no additional data to add.

It is reasonable that there might be a disparity in infected user count between January and other months, since in January students and a large portion of faculty and staff are on break and therefore not active on the network. However, the spring semester begins again at the end of January and we saw similar numbers of infection between January and February. The surge in March prompted us to look closer at the March data. There were no significant university network, IDS/IPS, or logging changes that could account for this large increase. Though there were some new threat signatures added, we observed that even existing threat categories, such as spyware, DNS, brute force, and code execution, experienced surges from the previous months. As of January 2019, the number of threat events has never been as high as it was in March of 2018, suggesting the increase may have been due to an outbreak of malicious activity on the wireless network. We discuss this further in Section 6.6 and plan to investigate this in future work.

6.3.2 Reputation Model Construction

Recall that construction of a reputation model entails fitting a distribution to some historical evidence. In this study, we use the 36 network traffic features and ground truth about user infection from the period of September 2017 to December 2017 to build our reputation models. Again, refer back to Table 1 in Chapter 5 to review these features. As discussed in the background (Section 6.2) we build separate reputation models for the infected and uninfected users, as our approach for reputation scoring is to determine which model better describes the user.

We experiment with several different approaches for constructing reputation models. This section presents techniques to build parametric models based on individual features and combinations of features.

To illustrate the concept of reputation scoring, we first apply a naive approach in which we construct individual parametric models for each feature. We use the GaussianNB implementation in Python's scikit learn library to fit Gaussian distributions to the data for each feature [160]. This process results in 72 different distributions, i.e. for each feature, we have a distribution summarizing infected user behavior and a distribution summarizing uninfected user behavior. Note that this approach is essentially implementing the backbone of a Naive Bayes Classifier.

For a stronger parametric approach, we combine all the features into a single model. This time, we use the Bayesian Gaussian Mixture implementation in Python's scikit learn library [161]. This approach creates a mixture of Gaussian distributions, where each individual distribution describes a feature. More technical details about this approach can be found in [162].

6.3.3 Reputation Quantification

The next two sections present the process for obtaining point estimate reputation scores and confidence intervals around these scores.

6.3.3.1 Point Estimates

This section presents the process of using the models to assign point estimate reputation scores to users. First, for each user in our pilot study sample, we calculate the 36 network traffic features to use as evidence using the data from the traffic logs associated with the user.

Next, to calculate the actual reputation scores, we use the *predict_proba* function provided in both the GaussianNB and the Bayesian Gaussian Mixture Python classes. At a high level, the process to determine the probability of the user belonging to the infected or uninfected model requires calling *predict_proba* on the reputation model and passing the user's cyber reputation vector - the evidence formatted as an n-dimensional feature vector - to the *predict_proba* function.

For the individual feature models, we must do this separately for the infected and uninfected models associated with each of the 36 features. In this scenario, the user's cyber reputation vector is 1-dimensional and contains the calculated network traffic feature associated with the feature model. The probability associated with the infected model gives us the user's reputation for infection (i.e. the most likely probability that they are infected), while the probability associated with the uninfected model gives us the user's reputation for uninfection (i.e. the most likely probability that they are uninfected).

For the combined feature model, we do not need to perform probability computations for features or infected/uninfected models separately. In regards to the features, this time we have combined them all into one model. In regards to the models, because the Bayesian Gaussian Mixture is implemented as unsupervised learning model, it uses information about infected and uninfected users together to assign a user a probability of belonging to either class; because of this, the probabilities associated with each model sum to 1. This time, the user's cyber reputation vector is 36-dimensional and contains all the network traffic features. As with the individual models, this computation provides us with reputation scores for both infection and uninfection.

To help illustrate reputation scoring, we present a sample of scores for 10 users identified as infected in January. We use the first 9 digits of the user's unique hash in the traffic data as the user ID in the table. Table 8 shows scores associated with an individual feature model (using the feature NREC, the number of records produced by a user) and with the combined feature model.

User	Reputation Score for Infection		Reputation Score for Uninfection	
	Individual Feature Model (NREC)	*Combined Feature Model*	*Individual Feature Model (NREC)*	*Combined Feature Model*
688e55ee1	23.7	69.2	0	30.8
7a59cd2f0	27.1	100	38.7	0
7cd3fb2d7	39.9	41.9	12.9	58.1
2df53e57c	37.5	0.1	28.6	99.9
fba66479e	31.1	100	0.05	0
af1a7205a	29.5	100	39.9	0
472676714	37.8	100	27.5	0
1087cbd4b	27.3	100	38.9	0
dd49757e6	24.6	100	35.9	0
f4cc6bbb1	25.5	100	37.1	0

Table 8: Sample reputation scores for infection and uninfection for a model based on a single network traffic feature and a model based on a combination of all 36 features. All users presented here were identified as infected in the month of January.

Though Table 8 presents a small sample of the scores for illustrative purposes, an important observation reflected in the table is that scores derived from individual feature models are not very high in comparison to scores derived from the combined models. In fact, across all scores generated by individual feature models, the highest probability ever assigned to a user is 39.9%. In the case of the combined feature models, users are sometimes given maximum reputation scores of 100%. In other words, more evidence enables us to better differentiate between infected and uninfected status. This is intuitive, but shows the importance of varied evidence for reputation scoring.

6.3.3.2 Confidence Intervals

Table 8 showed point estimate-based reputation scores. While a single probability value gives some insight into likelihood (i.e. a reputation score of 100% for infection indicates high probability that a user belongs to the infected class), point estimates do not give us insight into the confidence associated with the reputation score. By adding confidence intervals, we can understand the interval of reputation scores that would include a user's reputation score. Here, we calculate the 95% confidence interval, i.e. the range for which we can be 95% sure contains the user's reputation score. We implement this computation using Python's *numpy.percentile* function [163].

To calculate confidence intervals for the individual feature models, we recalculate probability estimates on 1,500 resamplings of the data for each of the infected and uninfected feature models. To generate these samples, we rerun testing and training using Gaussian NB 1,500 times, each time computing the area under the receiver operating curve (ROC AUC). We pass these ROC AUC values to the *numpy.percentile* function, which allows us to specify the desired confidence interval. In Table 9, we show the confidence intervals associated with the point estimates for the sample of 10 confirmed infected users presented above using the parametric model based on the feature number of records (NREC).

For the Bayesian Gaussian Mixture-based models, we employ a similar repeated sampling approach to generate confidence intervals. Because Bayesian Gaussian Mixtures are implemented as unsupervised models, in order to obtain different samples we take random subsets of the original 2017 data used to generate the models. Specifically, we sample 90% of the 2017 data multiple times to create different models and therefore different probability point estimates. Additionally, because Bayesian Gaussian Mixture uses the infected and uninfected model together to assign probabilities to users, to obtain separate confidence intervals for infected and uninfected users we repeatedly resample the data and generate point estimates such that we have 1,500 samples for users who are predicted to most likely be infected and 1,500 samples for users who are predicted to most likely be uninfected. A sample of these point estimates and associated confidence intervals for the same users presented in the Tables 8 and 9 are shown below in Table 10.

While much lower probability values are associated with scores generated by the individual feature models versus scores generated by the combined feature models, the 95% confidence intervals surrounding the scores derived from the individual feature models are much narrower. In fact, in some cases, as evidenced in Table 10, we see confidence intervals where the lower and upper bound are equivalent to the point estimate. Though in some cases this happens with the combined feature model confidence bounds, with these we sometimes see intervals that span the entire range of probabilities, i.e. (0, 1), indicating very little confidence in the true reputation score. With the individual feature models, the greatest range we observed is 0.079; this corresponds with a user given a reputation score of uninfection of 18.6%, with confidence interval (0.113, 0.192).

User	**Reputation Score for Infection**		**Reputation Score for Uninfection**	
	NREC Feature Model	*Confidence Interval*	*NREC Feature Model*	*Confidence Interval*
688e55ee1	23.7	(0.224, 0.242)	0	(0, 0.001)
7a59cd2f0	27.1	(0.262, 0.273)	38.7	(0.381, 0.388)
7cd3fb2d7	39.9	(0.399, 0.399)	12.9	(0.062, 0.135)
2df53e57c	37.5	(0.373, 0.376)	28.6	(0.229, 0.291)
fba66479e	31.1	(0.302, 0.315)	0.05	(0, 0.006)
af1a7205a	29.5	(0.287, 0.297)	39.9	(0.399, 0.399)
472676714	37.8	(0.377, 0.379)	27.5	(0.214, 0.28)
1087cbd4b	27.3	(0.265, 0.276)	38.9	(0.385, 0.34)
dd49757e6	24.6	(0.256, 0.249)	35.9	(0.34, 0.361)
f4cc6bbb1	25.5	(0.245, 0.258)	37.1	(0.357 ,0.372)

Table 9: Sample reputation scores and confidence intervals using an individual feature model

User	**Reputation Score for Infection**		**Reputation Score for Uninfection**	
	Combined Feature Model	*Confidence Interval*	*Combined Feature Model*	*Confidence Interval*
688e55ee1	69.2	(0, 0.998)	30.8	(0.002,1)
7a59cd2f0	100	(1,1)	0	(0,0)
7cd3fb2d7	41.9	(0.022,0.964)	58.1	(0.036, 0.978)
2df53e57c	0.1	(0,0.018)	99.9	(0.982,1)
fba66479e	100	(1,1)	0	(0,0)
af1a7205a	100	(1,1)	0	(0,0)
472676714	100	(1,1)	0	(0,0)
1087cbd4b	100	(0.991,1)	0	(0,0.009)
dd49757e6	100	(1,1)	0	(0,0)
f4cc6bbb1	100	(1,1)	0	(0,0)

Table 10: Sample reputation scores and confidence intervals using the combined feature model

6.3.4 Reputation Validation

Validation helps ensure the reputation models and evidence used to build these models are useful for assigning user reputation scores. In Chapter 4, we discussed several different ways that scores and models can be validated; here we illustrate a few of these options.

First, by mapping scores to labels, we can assess models using metrics such as True Positive Rate (TPR), False Positive Rate (FPR), True Negative Rate (TNR), and False Negative Rate (FNR). In order to map a score to a label, we first determine whether the infected or uninfected reputation model better describes the user. Just as if we were doing classification, we consider the distribution associated with the larger of the two probability values to indicate the best model for the user and assign the model type (infected or uninfected) as the user's label.

Though this validation technique is useful for a quick assessment of the success of the models with respect to classification, we lose information associated with likelihood and confidence in a score by taking this approach. Because reputation scores are probabilities, determining if a model was successful (or useful) is not really as simple as evaluating if the score was 'right' or 'wrong.' Consider the example of weather: if a model predicts there is a 60% chance of rain and it does not rain, this does not necessarily mean the model was incorrect. The Brier Score, mentioned previously in Chapter 4, is a metric for evaluating the accuracy of probabilistic predictions that can be applied when there are mutually exclusive, discrete outcomes. A Brier Score of 0 indicates a probability is always correct, while a 1 indicates the probability is always incorrect. In other words, the closer the score to 0, the better the prediction. The Brier Score is calculated using the following formula:

$$Brier\ Score = \frac{1}{N}\sum_{1}^{N}(forecast - outcome)^2$$

In the context of this dissertation, *N* is the number of users in the sample, *forecast* is the user's reputation score, and the *outcome* is whether the user became infected or remained uninfected.

We calculate the Brier Score for reputation scores associated with the parametric combined feature models. We calculate the score for each month of our sample (January, February, and March), considering *forecast* to be the user's reputation score for infection and *outcome* to equal 1 if the user was infected by the end of March and 0 if the user remained uninfected at the end of March. Note that if we were to instead assign the forecast value to the user's reputation for uninfection (and in turn, assign outcome to 1 if the user remained uninfected and 0 if the user became infected), we would obtain the same score since probability of infection and probability of uninfection sum to 1 in the context of the Bayesian Gaussian Mixture approach.

Because we quantify user reputation for two different characteristics and because we know that infection is a confirmed outcome, whereas uninfection implies the user could still become infected, we also calculate the Brier Score separately for the confirmed infected and currently uninfected populations. This enables us to understand and compare how well the reputation models work for infected and uninfected users. Note that when we calculate the score for the uninfected users, we consider the forecast to be the user’s score for uninfection and set outcome to 1 if the user remained uninfected.

Lastly, another way to validate and interpret reputation scores is to look at the probability values themselves and the associated confidence intervals. Though a label may categorize all users with reputation scores for infection above 50% as the infected, a probability of 100% indicates a much higher likelihood of a user being infected than a probability of 50%. To assess the scores using confidence intervals, we analyze how wide the 95% confidence interval is around the reputation score. Wide intervals imply less confidence in the true value of a user’s status as either infected or uninfected, and indicate that refining the scores and models to narrow the intervals is necessary.

6.3.5 Reputation Refinement

As mentioned previously, there are different ways a score can be refined. In this chapter, we focus on updating scores on a monthly basis as a technique for refinement. Each month, we recalculate a user’s reputation score by updating the user’s cyber reputation vector with the new network traffic activity generated by the user. Tables 11 and 12 show reputation scores for infection and reputation scores for uninfection updated monthly for the 10 sample users presented in the previous tables.

Example of Updating Reputation Scores for Infection

User	*January Probability*	*February Probability*	*March Probability*
688e55ee1	69.2	100	100
7a59cd2f0	100	100	100
7cd3fb2d7	41.9	100	100
2df53e57c	0.1	100	100
fba66479e	100	100	100
af1a7205a	100	100	100
472676714	100	100	100
1087cbd4b	100	100	100
dd49757e6	100	0.0	0.0
f4cc6bbb1	100	0.0	0.0

Table 11: Sample of monthly reputation scores for infection

Example of Updating Reputation Scores for Uninfection

User	*January Probability*	*February Probability*	*March Probability*
688e55ee1	30.8	0	0
7a59cd2f0	0	0	0
7cd3fb2d7	58.1	0	0
2df53e57c	99.9	0	0
fba66479e	0	0	0
af1a7205a	0	0	0
472676714	0	0	0
1087cbd4b	0	0	0
dd49757e6	0	100	100
f4cc6bbb1	0	100	100

Table 12: Sample of monthly reputation scores for uninfection

6.4 Results

Section 6.3 presented several different techniques for reputation model building and examples of reputation scores associated with these models. In this section, we present a deep dive into the results of scoring users with the parametric, combined feature model. The full results of reputation quantification using the parametric combined feature model, to include user scores for infection and uninfection, the associated confidence intervals, monthly updates to the reputation scores, and the month of infection if applicable, are available upon request and at http://www.terpconnect.umd.edu/~mgratian/.

This section follows the structure of Section 6.3.4. We first assess the results with respect to the mapping of scores to labels, then with respect to the Brier Score, and finally with respect to the likelihood and confidence intervals associated with reputation scores.

6.4.1 Assessing Results by Mapping Scores to Labels

There are several ways we can analyze and interpret the scoring results. First, if we use the approach where we consider the model with the higher probability to be the 'better' prediction and map the user's score to a label, we can interpret the results using a confusion matrix. Tables 13-15 show the confusion matrices for January through March.

	Assigned Label in January	
Ground Truth in January	***Infected***	***Uninfected***
Infected	434	98
Uninfected	526	8

Table 13: Confusion matrix associated with mapping reputation scores to labels in January

	Assigned Label in February	
Ground Truth in February	***Infected***	***Uninfected***
Infected	832	299
Uninfected	220	911

Table 14: Confusion matrix associated with mapping reputation scores to labels in February

	Assigned Label in March	
Ground Truth in March	***Infected***	***Uninfected***
Infected	3088	1699
Uninfected	1154	3633

Table 15: Confusion matrix associated with mapping reputation scores to labels in March

In January, we correctly identify 434 out of 532 infected users (81.6%). In February, we correctly identify 832 of 1131 infected users (73.6%, including the users already identified in January). In March, we correctly identify 3088 of 4787 infected users (64.5%, including the users already identified in January and February). Interestingly, the number of correctly identified infected users based on probability mappings to labels decreases over time.

When looking at the correctly identified uninfected users, we see the opposite trend. In January, we label only 8 of the 532 uninfected users as infected (1.5%). By

February, this jumps to 911 out of 1131 uninfected users (80.5%) and by March drops slight to 3633 out of 4787 users (75.9 %).

Table 16 shows the true positive rate (TPR), true negative rate (TNR), false negative rate (FNR), and false positive rate (FPR) associated with the assigned labels.

	TPR	TNR	FNR	FPR
January	81.6%	1.5%	18.4%	98.5%
February	73.6%	80.5%	26.4%	19.5%
March	64.5 %	75.9%	35.5%	24.1%

Table 16: TPR, TNR, FNR, and FPR associated with the reputation labels each month

While looking at confusion matrices and various prediction rates is useful and provides a quick way to understand the results, the purpose of the reputation score presented in this dissertation is to capture likelihood and confidence. As mentioned previously, by evaluating a score as a label, we are losing the benefits that a probability-based score provides over a label-based score. The next two sections discuss the results of calculating the Brier Scores and present a deeper analysis of the probability values and confidence intervals associated with reputation scores.

6.4.2 Assessing Results by Evaluating Probabilities

In this section we present the Brier Scores associated with the reputation scores. All scores have been rounded to the 3rd decimal place. In Table 17, we present the Brier Scores computed for each month using all users in our sample, just the confirmed infected users, and just the uninfected users. As a reminder, a 0 is the best possible score and a 1 is the worst possible score. Going back to the weather example provided earlier, if a forecast said there was a 90% chance of rain and it rained, the Brier Score would be 0.1; if a forecast said there was a 30% chance of rain and it rained, the score would be 0.7; if a forecast said there was a 0% chance of rain and it rained, the score would be 1.

Month	Brier Score for all users	Brier Score for infected population only	Brier Score for uninfected population only
January	0.582	0.176	0.988
February	0.227	0.261	0.193
March	0.296	0.353	0.24

Table 17: Brier Scores

Looking first at the scores associated with all users, in January, the score of 0.582 does not provide any information about the success of our predictions. In February and March, however, we attain scores of 0.227 and 0.296, indicating that the predictions become more accurate. Looking at the Brier Scores specifically associated with users confirmed as infected, we see that in January, the score is good - 0.176. However, it increases over the following two months. Finally, looking at the Brier Scores associated with users defined as uninfected, we see that in January, we are almost always inaccurate. This was already reflected in the False Positive Rate reported previously; the models are essentially defaulting toward giving all users a greater probability of being infected than uninfected. By February and March, the score drops drastically from what it was in January. However, again, in March we see a slight increase from February.

Overall, in assessing the Brier Scores, we see that scores do appear to improve in February and March in comparison to January. However, in each case the March scores are worse than the February scores, indicating that refining the scores will likely not just be a matter of including more data. This also indicates additional research is necessary to understand the cause of score fluctuations. This finding may also be related to the unexpected surge in infection March; our models, built on infection data from 2017, may not have been equipped to handle this anomaly.

6.4.3 Assessing Results by Evaluating Probabilities and Confidence Intervals

In this section, we present an overview of the likelihood and confidence intervals associated with user reputation scores.

Recall that there were 532 infected users in January. Of these 523 users, we assigned reputation scores for infection of 100% to 284 users. For 262 of these users, the 95% confidence interval was (1, 1), indicating the model had high confidence that the user's most likely probability of infection was 100%. The remaining 22 were given 95% confidence intervals where the lower bound was at least 99%. An additional 118 users were given reputation scores for infection above or equal to 99%; 93 of these users had a lower bound on their 95% confidence interval of at least 99%. All but 10 of the 118 users had lower bounds on their confidence intervals of 80% or more. 19 users of the 532 confirmed infected in January were given reputation scores for infection between [90%, 99%); for these users, the confidence intervals spanned the entire range of probabilities (0,1), indicating that the model had little confidence in the true reputation score.

There were 599 newly infected users in February. 331 of these users were given reputation scores for infection of 100%; 230 had a lower bound of 100% on the 95% confidence interval while an additional 57 had a lower bound of 99%, indicating high confidence in the score. 60 of the 559 newly infected were given reputation scores for infection in the range [99%, 100%). For the majority of these 60 users, their confidence intervals span the entire range of potential probabilities, indicating that the model had little confidence in the true reputation score. Overall however, in February we see similar trends to those of January, where the majority of infected users correctly identified as infected are given very high probabilities of infection and very narrow confidence intervals.

There were 3,656 newly infected users in March. 1,902 of these users were given 100% reputation scores of infection. Only 4 of these users were given 95% confidence intervals of (1, 1). The remaining 1,898 had 95% confidence intervals that spanned the entire range of probabilities, i.e. (0, 1). If we consider the entire pool of confirmed infected users by the end of March – 4,787 users total - we see 2,810 of these users are given 100% reputation scores of infection by the end of March. Only 5 of these users are given 95% confidence intervals between (1, 1). For the remaining 2,805 users, the 95% confidence interval for their score spans the entire probability range of (0, 1). These results present a drastic difference from January or February, where confidence intervals were quite narrow around scores of 100%. As before, we hypothesize that this may be related to the unexplained surge of infected users in March, which may have been due to an outbreak of new malicious activity that was different from the malicious activity captured in our reputation models. We discuss this as a limitation of our data in the Section 6.6 and consider investigating this phenomenon as an area of future work.

Another way to analyze the reputation scores is to assess how many users were assigned higher probabilities of infection when they were uninfected and vice versa, and to look at the 95% confidence intervals for these incorrect predictions.

Recall that of the 4,787 users who were infected by the end of March, 1,699 were given reputation scores of uninfection above 50% by the end of March; these users correspond with the false negatives when we mapped scores to labels. How confident was the model in this incorrect score assignment? Looking at the 95% confidence intervals, a lack of confidence is clear - the intervals span the entire range of probabilities. However, because in some instances we also saw wide confidence intervals when the scores correctly identified infected users, we refrain from over-exaggerating the importance of this finding.

Recall that of the 4,787 users who were uninfected by the end of March, 1,154 were given reputation scores of infection above 50%; these users correspond to the false positives associated with mapping scores to labels. Looking at the 95% confidence intervals, all 1,154 of these users span close to the entire range, (0, 0.99). So, though they were given higher probabilities of infection, the model was not confident in this assessment. Again, however, we also saw wide confidence intervals in several cases where uninfected users were correctly identified as such.

It is also interesting to understand how reputation scores changed over the months as new data was added. Of the 4,787 confirmed infected users at the end of March, the reputation scores of 4,518 of these users showed a monotonically increasing trend from February to March, i.e. their March infection scores were greater than or equal to their February scores. 2,448 had scores that showed a strictly increasing trend from February to March. Looking at scores from January to February to March, 2,096 of the infected 4,787 showed a monotonically increasing trend; 50 scores showed a strictly increasing trend.

Conducting a similar analysis for uninfected users, we find that 935 of the 4,787 uninfected users show a monotonically increasing score trend for uninfection from February to March; only 9 show a strictly increasing trend. Across all three months, only 346 show a monotonically increasing uninfection score trend; only 3 of these show a strictly increasing trend. This result, in comparison to trends seen with infected user scoring, suggests our uninfection model was perhaps not as strong as our infection model.

6.5 Discussion

This section presents a discussion of the results in this case study, observations on the methodology, and some comments on the benefits and challenges of using a Bayesian-based approach to reputation scoring.

Section 6.4.1 first presented an overview of the TPR, TNR, FNR, and FPR associated with assigning users labels of infection or uninfection based on the model that assigned them the largest probability. These results are somewhat difficult to interpret, especially evaluating the trends associated with these metrics over January, February, and March. Though in some cases we see improvements between January and February (e.g. with respect to TNR and FPR), we see slight decreases in performance between February and March. Looking at the March metrics, in which we are including up to three months of data for every user, we see that the TPR and TNR indicate we are doing better than random guessing when assigning reputation scores, but the FNR and FPR are quite high. Overall, if the scores are to be mapped to labels, network-based features alone do not provide enough evidence to make reputation assessments about users. Though in Chapter 5 we framed the feasibility study of using network traffic features as a supervised problem and concluded that the features did provide substantial evidence to differentiate between infected and uninfected users, by attempting to score a larger sample of users over a longer time window, in this chapter we reevaluate our original assessment and conclude network traffic features provide only a starting point for understanding users.

Assessing Brier Scores (6.4.2) provided us with a different perspective from which to understand users. Though we still see a decrease in prediction success between February and March, the Brier Scores for those two months present substantial improvements from the January Brier Scores and suggest there is potential for using reputation scores to make assessments about the future behavior or state of a user. Analyzing the probabilities assigned to users in more depth, we observed that in some cases, a reputation score was able to proactively warn us that a user was ‘trending’ toward infection or already infected before the user was identified as such in the network threat logs. This is exactly the goal of a reputation system - to allow proactive judgments about an entity based on some evidence.

Assessing likelihood, based on the value of a reputation score, and confidence, based on the 95% confidence interval around a reputation score, was even more telling than looking at sensitivity/specificity metrics (6.4.3). We found that in the majority of cases where we correctly identified a user as infected (as in, we gave them a higher reputation score for infection than uninfection), the score was often 100% or 99% and the confidence intervals were quite narrow around these scores. We also found that the majority of infected users had reputation scores for infection that increased each month; this was a promising result and suggests that for infected users at least, evidence of infection can be found in network traffic features.

The current results of reputation scoring users via network traffic are certainly not at the level where they could be used in an operational setting. However, the results do indicate the viability of using a Bayesian-based approach to reputation scoring. First, by comparing understandings gained by looking at classification metrics to understandings gained by analyzing the probabilities and confidence intervals themselves, we emphasize the wealth of information that can be contained in a probability-based score. With metrics such as TPR or FPR - and by extension, scores that are mapped to labels - we do not get a sense of likelihood or confidence. By interpreting user reputation via probabilities, we're able to understand the strength of our predictions and identify areas where we could improve them. Regarding this case study specifically, we see that network traffic can function an initial form of evidence and hypothesize that stronger evidence may help strengthen our score accuracy.

Additionally, by illustrating the process of collecting data, identifying evidence, updating scores over time, and including confidence intervals, we show that the proposed reputation methodology is flexible and capable of generalizing to other networks, environments, data sources, and characteristics of interest. Moreover, the methodology provides a systematic way to create and refine reputation models and scores, which was the intended goal of introducing a formal definition of reputation and evaluation methodology.

As mentioned previously, we hypothesize that adding forms of evidence could help improve the prediction accuracy of the reputation scores and narrow the confidence intervals. We discuss this further in Section 6.7, but here we note that this emphasizes one of the key challenges of a Bayesian approach to reputation scoring - acquiring good data. The accuracy of a reputation score can only be as good as the evidence used to make the prediction. And in the context of understanding users, acquiring data is challenging and requires deep understandings of the nature of users on a particular network or in a specific setting. Therefore, in any attempt to assess user reputation, a major (but feasible to overcome) hurdle will always be identifying and preparing useful sources of data.

6.6 Limitations

There are several limitations of the case study and the reputation evaluation methodology. First, as highlighted previously in Chapter 5, we define infection status using the university's IDS/IPS threat logs. The Palo Alto system is essentially a black box: we know the system rules set by the university, but we do not know the specifics of how threats are identified or the system's false positive or false negative rates. Again, while understanding this may provide more insight into why users are infected, the focus of this case study was on illustrating the reputation evaluation methodology, not on reverse engineering the Palo Alto system. Moreover, since we are scoring reputation of users on a university network, it is appropriate to use the university's definition of infection, i.e. using the threat logs as ground truth.

Second, we do not differentiate between different types of infection. In the IDS/IPS threat records, threats are assigned different severity levels, categories, and signatures. We do not currently separate infected users by these differences, but it would be useful to understand how reputation scores and their associated confidence intervals vary with threat type. This may help us understand why in certain cases were have more or less certainty and confidence in a score. Future work will include exploring more nuanced definitions of infection. We plan to incorporate infection type into our definition and look for other sources of data to corroborate that a user is truly infected. Note that in Chapter 4, we suggested altering how the characteristic of interest is defined as a score refinement technique.

Third, when we mark a user as infected, we do not differentiate between users who are intentionally creating threats and those who are victims. Intention is difficult to assess without speaking to a user directly, but tightening the definition of reputation may help create this distinction. For example, if we differentiate by threat type, users who are infected because of 'code execution' on their devices are probably victims and not perpetrators. This, of course, adds some subjectivity to the definition of the characteristic of interest, but the reputation evaluation methodology is structured such that even subjective characteristics can be quantified. Overall, in their current state the reputation scores we present here are useful for identifying accounts that may pose enhanced threats to the network or are in need of human scrutiny or intervention; a score could be made even more useful if there were models to differentiate between different types of infection, which in turn might help one understand *why* an account poses a threat to the network.

Fourth, and connected to the previous limitations, we assume a user is truly responsible for all the network traffic that we use to assess their reputation. This was a useful assumption for a pilot analysis of the reputation evaluation methodology, but this may not necessarily be true. The user may have willingly shared their credentials with someone else. Alternatively, someone may have stolen the user's credentials or hijacked the user's device and/or accounts without the user's knowledge. If so, the reputation score actually reflects the *point of access to the network's* reputation and not a user's reputation.

Verifying that activity truly belongs to a user was outside the scope of this case study, but is something that could be accomplished with a reputation system. Analogous to financial credit monitoring, if a network administrator (or even the user, assuming their reputation score is available to them) observes a change in a reputation score that has generally been consistent over time, this may be an indication that someone besides the user is responsible for the traffic. While this may be expected if the user is switching from an uninfected status to an infected status, it is possible for a user to remain in an uninfected state even if their account has been hijacked; for example, a malicious actor is simply using the user's account to access university resources. In this case, changes in a user's score or differences in the distribution of the user's activity and how they compare to other users may facilitate the discovery of this. Future work to explore this will involve expanding this case study over a longer period of time and conducting deeper analysis of users.

Fifth, there are also limitations associated with our data. We were missing threat log data on January 1, February 5-7, and February 21-27, 2018. So, it is possible that we missed some infected users when building our infected sample each month. Additionally, we do not know what caused the surge of infected users or threat alerts in March. While our initial exploration of this with the Division of IT on campus has indicate there was simply an outbreak of malicious systems and activity in March, we will continue to investigate this surge in future work.

Sixth and lastly, there are also limitations associated with the quality of the evidence used to build our reputation models and evaluate user reputation. While the previous chapter indicated network traffic had potential for differentiating between infected and uninfected users, the false positive and false negative rates were quite high, indicating network traffic alone was unlikely to suffice as a data source. In this chapter, though in some cases we obtained reasonable probability estimates (in that the outcome of infection or uninfection was predicted by the reputation score), it is unsurprising that the network traffic data was not strong enough evidence.

6.7 Future Work

There are several ways we plan to expand the work presented in this chapter and further test the reputation evaluation methodology. Future work will include efforts to improve both the models and the scores themselves and analysis of the reputation scores over longer periods of time.

As a first step to improve the models in future work, we will include more data into the reputation models. Currently, the models were built using three months of data, but we will evaluate how scores are impacted if the models are constructed using more data and are incrementally updated as new evidence becomes available. Additionally, we will also explore the idea of dynamically updating the models each time a new infected user is identified in order to refine the model's 'knowledge' of normalcy for infected users.

Another area of future work to improve the models will be to evaluate different types of distributions. In this chapter, we focus on a parametric approach using the Gaussian distribution. In future work, we will evaluate different distributions and non-parametric approaches and test how reputation scores vary based on the distribution.

Related to this, we will also explore different techniques for incorporating features as evidence. This may include testing different combinations of features; in this chapter, we used features individually or all together. Another technique is to specify the weights of features directly using some knowledge about which features represent stronger evidence; in the approach presented in this chapter, we allowed the Gaussian Mixture model algorithm to choose the weights.

In addition to evaluating different model construction techniques and assessing how they impact user reputation scores, in future work we will also monitor, assign, and update user scores over longer periods of time. This will allow deeper trend analysis and better understandings of how variances in the models and the available data impact a score. Related to this, we will also conduct analysis to understand why and how scores fluctuate and how scores differ based on types of infection, as highlighted already in Section 6.6. Instead of just assigning users reputation scores, we will try to uncover *why* a user is given a particular score.

Finally, we will also look for other sources of data to incorporate as evidence. While network traffic enabled us to establish initial reputation scores for users, we will attempt to find additional evidence that may help improve scores and narrow the confidence intervals. In particular, related work has identified correlations between demographics and likelihood of cyber victimization. We are working with the Division of IT at the university to pair demographics with the network traffic data. We discuss the importance of demographics and other user traits in the following chapters (Chapters 7 and 8) and in the conclusion of this dissertation (Chapter 9).

6.8 Conclusion

This chapter expanded on the case study from Chapter 5. Using the network traffic data from the university, we built reputation models for infected and uninfected users and conducted a pilot study in which we assigned 9574 users reputation scores over the course of three months. Though there is certainly work that can be done to improve these predictions, overall this chapter illustrated that probability-based scores are feasible to generate, useful for making proactive reputation assessments, and informative due to their ability to capture the certainty, confidence, and subjectivity associated with a score.

This chapter also emphasized the importance of obtaining strong evidence and multiple sources of data for creating reputation scores with high certainty and narrow confidence intervals. Recall that individual feature models were never able to assign users reputation scores higher than 39.9%, while combined feature models were able to assign users scores of 100% (and in many cases, this prediction was correct). Though it may be obvious that more data leads to stronger predictions, finding substantial data to use to make predictions about users in a cybersecurity context is challenging. The next two chapters present a case study where social media is explored as a potential source of evidence.

7 Social Media as Source of Data: Constructing a Dataset of Self-Reported Cyber Victimization Cases on Twitter

7.1 Introduction

Data is key to any reputation system; without substantial evidence, reputation assessments are simply guesswork. A major challenge when it comes to evaluating user reputation in a cybersecurity context is obtaining meaningful data. As was illustrated in the previous case study, quantifying user reputation for cybersecurity requires understanding how various human traits and behaviors are connected to user cybersecurity experiences. From a broader perspective, understanding user behavior is also critical for developing secure systems and effective cybersecurity solutions. Unfortunately, research in this space is still limited: a majority of studies evaluate users in the context of phishing attacks using surveys or controlled experiments. These studies are certainly valuable, but phishing attacks represent just one context for understanding users, while surveys and controlled experiments are resource intensive endeavors.

Motivated by work in public health that has tapped social media to identify trends and develop new theories about people and the medical ailments that affect them, in this chapter we investigate the feasibility of using Twitter to understand user cybersecurity experiences. We hypothesize that user-reported cyber victimization experiences will provide a valuable new source of data that can both supplement traditional forms of data collection and offer previously unknown insights about users.

While other researchers have analyzed Twitter to detect vulnerabilities or to identify malicious or fake accounts, we explore a distinct problem: identifying the victims of cybercrime. Observing that users will often use social media to self-report cyber victimization (*Don't click the link, I've been hacked!*) we search Twitter for users who tweeted about everything from falling for social engineering scams to downloading malware on their personal devices during the period of January 1, 2018 through March 13, 2018. We apply an intensive manual review and labeling process to these Tweets and the context surrounding them in order to build a clean dataset of cases of self-reported victimization. Our final corpus consists of 2,910 users, each representing what we have determined to be an authentic cyber victimization case.

We conduct exploratory analysis of our dataset to understand the nature of users who self-report victimization. We identify the accounts, devices, and consequences associated with victimization, as explained by users themselves. We also identify users who have repeatedly been victims of cybercrime. Initial analysis reveals Twitter is a key source for understanding user attitudes toward cybersecurity and cyber victimization: users openly and explicitly Tweet about everything from illegal downloading, financial and reputational loss, and compromised corporate infrastructures. This analysis also hints at the prevalence of victimization on various online services: of the posts that specify an affected account, about 43% report victimization on social media and about 32% report victimization in online games.

The contributions are as follows:

- To the best of our knowledge, we present the first Twitter dataset of user-reported cyber victimization cases for which every case has been manually reviewed in order to validate its authenticity.
- We illustrate the viability of using Twitter as a source of unsolicited user feedback on user cybersecurity experiences. We propose using this data to supplement and address some of the challenges associated with traditional means of data collection such as surveys and controlled-experiments, which are resource intensive, limited in the breadth of the population they can cover, and can suffer from response bias and inauthentic user behavior.
- We begin to uncover various victimization trends and user attitudes that can help security practitioners understand the attack vectors they must defend against and the factors they must contend with when developing secure systems and security policy.
- We identify social media as an additional source of data that can be used to colllect evidence for user reputation assessments.

The remainder of this chapter is organized as follows. Section 7.2 presents the related work. Section 7.3 follows with our data collection methodology. We then present our final dataset and observations made during data collection in Section 7.4. We then provide a broader discussion of our dataset, additional observations, and an overview of the advantages and disadvantages of using Twitter in Section 7.5. Section 7.6 concludes the chapter concludes with a discussion of the limitations of our dataset. Chapter 8 further expands on this case study.

7.2 Background and Related Work

Twitter data is used in many fields to collect user self-reports and self-diagnoses for a variety of conditions. The field of public health in particular takes advantage of self-reported data. In [118], researchers monitored posts on Twitter that mentioned the flu ("down with swine flu") in order to predict emergence of the flu in a population. While individual pieces of data were noisy, in aggregate they captured the pattern of the epidemic over time and across geographic regions; their Twitter-based dataset had a 0.985 Pearson Correlation Coefficient with the US Center for Disease Control flu trend data. The authors in [119] conducted a similar study and also found that Twitter was highly effective for tracking influenza levels. In [120], researchers searched for Tweets reporting diagnosis of various mental health issues ("I have been diagnosed with depression);" this data was used to discover previously unknown linguistic signals associated with the disease. In [121], the researchers identified Tweets where users reported dental pain ("toothache" or "tooth pain"); their dataset consisted of 772 Tweets that were studied to understand the nature of the dental pain and how people reacted to it. The authors in [122] conducted a much more expansive study of user health reports on Twitter: they searched Twitter for Tweets containing words and phrases related to medications (e.g. "ibuprofen"), ailments (e.g. "cancer"), and more. In [123], the author identified Twitter users beginning an Alcoholics Anonymous program ("first AA meeting") and identified users who either maintained sobriety ("I've officially been sober for 4 months") or relapsed ("Taking 5 shots of vodka after I left work tonight was not a good idea").

Twitter has also been used in the field of cybersecurity to monitor emerging threats. In [124], the authors presented one of the first Twitter-based methods for early detection of software exploits. The authors monitored Twitter over a one-year period for the keyword "CVE" (Common Vulnerability and Exposures) and then conducted quantitative and qualitative analysis of the 5,865 CVEs collected to understand the nature of vulnerability disclosure on Twitter. They also developed a 'whitelist' of 4,335 Twitter users who post 'information-rich' security messages and advocated use of these for real-time vulnerability and exploit discovery. Similarly, the authors in [125] developed a framework to issue vulnerability alerts to users. Their system searches Twitter for keywords such as "XSS," "spoofing," and "buffer" and then automatically filters and labels the Tweets to identify relevant information for a security analyst.

Other security researchers have studied Twitter to identify the spread of malicious, suspicious, or spam-sending URLs [126, 127, 128]. There is also extensive work to identify malicious, spam, or fake accounts on Twitter [129, 130, 131, 132]. While this work is valuable, especially because social media sites are well established as lucrative platforms for criminals, understanding the victims of malicious cyber activity is a distinct problem.

Our work is most similar to that of [133], in which the authors analyzed how users responded to "hacked" Twitter accounts. They collected Tweets containing the keywords '"hacked" or "compromised" and "account"' and then trained a Support Vector Machine to classify Tweets with different types of user responses to victimization, such as apologizing to followers or creating entirely new accounts.

Our work is distinct from [133] for several reasons. First, we employ a thorough manual review of our victimization cases; the methodology for this is discussed later in this chapter. The only data cleaning and validation performed by the authors of [133] is to separate self-reports from peer-reports of victimization (i.e. “I’ve been hacked” vs. “I think you’ve been hacked”). While their resulting dataset of 358,639 users is impressive, it is likely that many of these do not represent true cases of victimization and should be excluded from analysis. To emphasize this point, we replicate the author’s search of ‘“hacked” or “compromised” and “account”’ on the period of January 1, 2018 through March 13, 2018 using the Twitter API. The search returns 81,866 Tweets. Manually reviewing these Tweets reveals results such as:

- *Sorry for the posts last night...alcohol hacked my account* ☹
- *The Russians have managed to hack me and cause many of my Twitter accounts to be suspended because they did not like what I said...*

These are either clearly not real victimization cases or are difficult to validate. Therefore, while fully automated searches may return a high volume of Tweets, user-generated content can be noisy so we opted for a manual labeling process for this initial investigation into Twitter as a data source for security research.

The next major distinction is the focus of the work in [133]. The authors identify the percent of users who create a new account in response to victimization and the percent of users who either apologize or state they’ve been hacked. The authors only retain victimization cases related to Twitter. We extract a more extensive range of insights from the Tweets and retain Tweets related to victimization on a variety of other platforms.

7.3 Methodology

This section outlines the methodology to construct our dataset. We collected cases of self-reported victimization by first developing a set of cybercrime related keywords, searching Twitter for posts containing these keywords, and then manually reviewing these posts to confirm if they represented a true case of victimization. The final dataset consists only of users with Tweets and profiles publicly available through Twitter at the time of data collection. Example Tweets are paraphrased.

7.3.1 Search Term Development

Search terms were developed using a Grounded Theory Sampling approach [134]. To apply this approach, we first identified a few Tweets demonstrating characteristics relevant to our study and then iteratively added new cases as we gained insight into user reporting behavior. Through our personal experiences with social media, we knew that people often announce 'hacked' accounts in order to warn or apologize for account behavior to friends or followers. So, we began with the keyword 'hacked.' We searched this term directly on https://twitter.com/search. Manually reviewing the results that appeared in the *Top* category, we confirmed that 'hacked' was a word often employed by users to self-report on hijacked accounts or credentials. However, because using the word 'hacked' returned many Tweets that were not relevant to our study, we also developed more refined search rules that would help us filter Tweets more efficiently.

This included creating a search rule that excluded words related to the US Presidential Election, as this was a very popular topic of discussion on Twitter at the time of data collection. Under this rule, we searched for Tweets containing the word ‘hacked’ and not containing the words “Trump,” “Hillary,” “Clinton,” ”RNC,” “DNC,” or “Russia” using the following format: *hacked -Trump -Hillary -Clinton -RNC -DNC –Russia*.

As an alternate to search by excluding keywords, we also examined ways we could narrow our searches by including additional keywords. As already noted, many users Tweet about victimization for the purpose of notifying their social media community, i.e. *Don’t click the link I sent out!*, so the term ‘hacked’ was often paired with terms such as ‘link’ or ‘click.’ So, we created another rule resulting in Tweets containing both “hacked” and “link” and one term from the set [“don’t”, “ignore”, “click”]. This was searched using the following format: *hacked link don't OR ignore OR click.*

In addition to self-reporting in order to alert followers, we observed many users seeking help from companies with a Twitter presence, i.e. @*instagram please help my account has been hacked*. This lead to a search rule for Tweets containing both the words “hacked” and “help,” searched using the following format: *hacked help*.

Finally, we also observed a few cases of the term “virus.” This term alone resulted in a high volume of physical health-related Tweets, so we narrowed the results by searching “virus” and either “computer” or “phone” using the format *virus computer OR phone.* We tried variants of the word “virus,” such as “malware," but found this term was mainly used by security professionals or security-focused Twitter accounts and not by the general Twitter population. Based on this observation, we chose not to use specific cybersecurity words, as was done in [124] and [125].

7.3.2 Data Collection

We searched Twitter with our keywords over the period of January 1, 2018 through March 13, 2018. We experimented with using both the Twitter Search API and Twitter's website interface. All cases for January were obtained via the API; all cases for February and March were obtained using direct search on Twitter's website. While the API helped us acquire a greater volume of Tweets, direct search helped us realize the importance of the context surrounding the Tweet, such as screenshots, gifs, emojis, and the responses of others, in order to make a judgment about the authenticity of the victimization claim. Direct search also enabled us to make additional qualitative observations about user experiences and attitudes towards cybersecurity, which we present in the Results (7.4) and Discussion (7.5) sections.

Two reviewers manually reviewed the search results and either labeled a result as a true case of victimization or discarded it. We only retained Tweets as cases if they provided some proof of the victimization, either through a written description of the incident, requests for company or customer support assistance, or screenshots of the hacked account. Additional proof was sometimes found in the replies or reactions of others, which could include everything from condolences to advice for dealing with the incident. Since the API provided only a text representation of a Tweet, API-derived Tweets were searched directly on Twitter to view the surrounding context if proof was not found in the Tweet text.

We also only retained Tweets if a user self-reported on victimization. Like the authors in [133], we observed that users will sometimes notify others that they've been victimized, i.e. *I think you've been hacked, your account is sending a lot of* *spam messages.* We discarded these Tweets since we were only interested in self-reporting instances.

During our period of data collection there was a trending hashtag on Twitter called "#TweetLikeThe2000s." Many users self-reported on victimization in response to this hashtag, i.e. *I downloaded a virus using Limewire* or *That game always gave the family computer a virus!* We did not include any Tweets associated with this hashtag because there was no way to validate if users were reflecting on a real past experience of victimization or simply responding to a trending topic. Similarly, there were also many Tweets in which users reported on a specific adult-content site leading to computer or phone viruses. The exact same language was often used in regards to the same site by many different users or accompanied by laughing or smiling emojis, so we interpreted this as a trend and/or attempt at humor and did not include any of these Tweets in our dataset.

After this process, we had 2,387 Tweets from the month of January, 579 Tweets from February, and 502 Tweets from March. Users sometimes tweeted multiple times about an incident of victimization, so we filtered for unique usernames, which resulted in 3,459 users.

7.3.3 Data Cleaning and Labeling

Since determining if a Tweet represented a true case of victimization was a subjective decision, once we collected our initial sample of users we had additional reviewers read each Tweet and decide whether to keep it or discard it. The Tweets of our initial 3,398 users were split across four reviewers; we ensured the two original reviewers did not re-read the same Tweets they collected. If a reviewer had any doubt about the authenticity of the self-report, either due to a lack of specifics about the event, grammatical errors making the Tweet difficult to understand (a common problem), contradictions or discrepancies created by the user when explaining the problem, or the context surrounding the Tweet when viewed directly on Twitter's site, the user was discarded. While general approaches to reviewing the Tweets were discussed among the reviewers, each reviewer made a determination about whether to keep or discard independently. This process left us with 2,910 true cases.

During the review process, reviewers also added categories and labels to the more informative posts in the sample. We chose not to apply an open coding analysis here, as the goal of our study was to investigate the feasibility of using Twitter to understand user cybersecurity experiences. Instead, we focused on answering three high-level questions: *What device or account was affected, what were the associated consequences,* and *has this user been victimized before?* We frame these as research questions as follows:

- **Research Question 7.1 (RQ7.1):** *What types of devices or accounts are compromised (i.e. what can we learn about the experiences of users on various devices and services)?*
- **Research Question 7.2 (RQ7.2)**: *What consequences do victims report (i.e., what can we learn about victim experiences from user-reported data)?*
- **Research Question 7.3 (RQ7.3)**: *Can we identify repeat victims (i.e., how prevalent is repeat victimization according to this data)?*

If a user specifically named a compromised device type or service, reviewers added a label indicating this. For example, for the ‘affected device’ category, reviewers assigned labels such as ‘phone’ or ‘computer’ if the user specifically named a device that was compromised by the victimization. Similarly, for the ‘affected account’ category, reviewers added details about the compromised online service if the user explicitly named the service. While many users reported victimization on high profile platforms, there were also many platforms unfamiliar to us; in these cases background research on the specified platform was necessary to understand the user’s post. Once this process was complete, the labels were verified and cleaned. Each category was cleaned and verified by a single reviewer and then confirmed by a second reviewer to ensure consistency within a category. For the affected device category, labeling resulted in six different labels: computer, which included desktops and laptops, phone, USB, iPad, iPod, and memory card.

For the affected account category, labeling resulted in 223 unique services and platforms. To help us understand the nature of these accounts, we sorted them into 17 high-level categories. These high level categories are as follows: social media and chat applications which included sites such as Twitter and Facebook and applications such as WhatsApp and Telegram; online games which included gaming platforms such as Steam and specific games such as Fortnite; media and entertainment which included services such as Netflix and Spotify; technology, electronic, and telecommunication companies which included Amazon, Apple, Microsoft, and Sony; cryptocurrency which included exchanges such as Binance and wallets such as My Ether Wallet; email services such as Gmail and Yahoo; ride sharing applications, which only included Uber and Olacabs; payment applications such as PayPal and Splitwise; traditional banks and financial services such as Bank of America; ecommerce, delivery, and retail services such as eBay; housing and hospitality services such as AirBnB and Booking.com; website and file hosting services such as WordPress; online dating sites such as Match.com; airline websites such as Delta and Emirates; school systems; online betting websites; and miscellaneous, such as a URL shortener site. Each platform was assigned only one category, though we acknowledge there are platforms that could fit into multiple categories.

The consequences category was less objective so we went through several iterations of labeling before finalizing categories of consequences to assign to Tweets. As was done for the device and account categories, each reviewer had a fourth of the total corpus to label. Reviewers assigned descriptive tags only if users made explicit comments about a consequence. For example, *@AskPlaystation my account's been hacked will you please help they changed my email :(* was labeled with 'altered settings,' while *I spent $100 to fix my computer because I accidentally downloaded a virus* was labeled with 'financial loss.' Since reviewers did not follow any specific codebook, sometimes labels differed (e.g. "financial loss" vs. "loss of money").

After this process was complete, two reviewers independently assessed all the labels and began to combine similar consequences into the same category. For example, 'financial loss,' 'monetary loss,' and 'loss of money' were all grouped together as the same type of consequence. The reviewers then compared their results and came to a final consensus on a set of 13 categories; these categories are described in Table 18. Then, one reviewer relabeled all 2,910 Tweets with these new categories to ensure consistency; a second reviewer reviewed and validated the results. Tweets were labeled with multiple consequences if multiple consequences were reported. Again, Tweets were only assigned a consequence if a user explicitly stated the consequence occurred.

For our final question – has a user been victimized before? – we assigned 'repeated victimization' labels to posts if a user made a comment about being victimized 'again' or indicated the number of times they've been victimized.

Consequence	***Description***
Financial loss	Any monetary loss; includes traditional currencies and cryptocurrency
Data loss	Any loss of digital items, such as files and media
Account loss	Any loss of access, account suspension, or deletion; user must explicitly state they've lost access
Device damage	Any damage done to a physical device; includes damage that renders a device slow or frozen
Altered settings	Unauthorized changes to user account settings, including account login information
Spam	Indicates a hijacked account or device was used to spread spam/ unsolicited content
Reputation loss	Any loss of authority or influence or a damaged identity; includes a loss of followers or subscribers
Identity theft	Any cases of identity theft, impersonation, use of hijacked account to masquerade as the user
Threats	Any form of threats, blackmailing, or ransom
Professional issues	Compromise resulting in professional or workplace trouble for the user, their colleagues, or employer
Access vector	Cases where one hijacked account has lead to the compromise of other accounts
Unauthorized account activity	Forms of account activity that are distinct from spam, impersonation, and altered settings
Miscellaneous	Other reported consequences that occurred only once and did not easily fit into another category

Table 18: Victim consequence categories

7.4 Results

This section provides an overview of the affected devices, affected accounts, associated consequences, repeat victims, and observations made during the data collection process.

7.4.1 Affected Devices

Of our 2,910 cases, 472 specified an affected device. The majority of users experienced victimization on a computer (277 total) or phone (189 total). Two users experienced a compromised USB; two a compromised iPad; one a compromised iPod; and one a compromised memory card.

7.4.2 Affected Accounts

Table 19 presents an overview of all the affected accounts that were reported by users with a frequency greater than 10.

Since data collection took place on Twitter, it is unsurprising that Twitter has the highest number of cases. Of the 382 Twitter cases, 102 were posts reaching out to Twitter (i.e. *@Twitter why is my account hacked pls help*) and 56 were posts reaching out to Twitter Support (*@twittersupport I've been hacked, how can you help me?*). Though almost half of all Twitter cases attempted to reach out to Twitter for help, many did not seem optimistic about the possibility of recovering an account or receiving help from Twitter: *Friends: someone seems to have hacked my account so I'm using my secondary once since I have very little faith in Twitter's interest in solving my problem. If you've ever succeeded in getting help from Twitter, please advise.*

Service/ Platform Name	*Frequency*	*Service/ Platform Name*	*Frequency*
Twitter	382	Netflix	30
Roblox	258	Steam	27
Instagram	177	Amazon	27
PlayStation	161	Airbnb	26
Facebook	160	Mojang	25
Spotify	90	Gmail	23
GroupMe	76	Discord	21
Jagex	73	Apple	21
Snapchat	70	Youtube	20
EA	60	PayPal	18
Uber	44	eBay	15
Rockstar Games	36	Xbox	14
Email	35	Bizzard	13
Linkedin	34	Binance	12
Yahoo	33	Google	11

Table 19: Affected account overview

The next highest occurring platform was Roblox, a massive multiplayer online gaming platform in which users can design their own games, socialize with other players, and buy and sell virtual items using Robux, an in-game currency. Originally released in 2006 and mainly marketed towards children and teenagers, the game has over 64 million active players per month, as of November 2017 [135]. Similar to Twitter, we observed many users reaching out to Roblox or to well-known game moderators and developers for help: *@Roblox help please someone hacked my account.*

In addition to Twitter, our sample also contained reports about a variety of other social media and communication platforms, including Instagram (frequency 177), Facebook (frequency 160), Skype (frequency 9), WhatsApp (frequency 6) and Tumblr (frequency 5). The pizza chain Dominos occurred with a frequency of 6; users reported hacked accounts resulting in unauthorized pizza orders and use of account rewards points and free pizza offers.

On some services, we were able to observe a common attack vector against users during our time window of data collection. For example, of the 76 reported cases of compromised GroupMe accounts, 72 users reported their accounts were used to send "Happy New Year" messages that contained either links to weight loss pills or requests for help via bitcoins.

There were 130 platforms and services that occurred with a frequency of one in our sample. These included large retailers like Walmart and BestBuy, online dating platforms, and 17 different cryptocurrency wallets and exchange websites.

Table 20 presents an overview of the 17 high level categories that accounts were grouped into, their frequency of occurrence, and what percent of the total number of reported affected accounts they represent.

Category	Frequency	Percent of Total
Social media + chat	992	42.6
Online games	743	31.6
Media and entertainment	130	5.5
Tech, electronics, and telecommunications	118	5
Cryptocurrency	72	3.1
Email	62	2.6
Ride sharing	45	1.9
Payment and money transfer	26	1.1
Traditional banks and financial services	40	1.7
E-commerce, delivery, and retail	39	1.7
Housing and hospitality	31	1.3
Website and file hosting	13	0.6
Online dating	5	0.2
Miscellaneous	5	0.2
Airlines	4	0.2
School systems	3	0.1
Online betting	3	0.1

Table 20: Categories of affected accounts

Social media and chat applications were the most commonly occurring type of account for which users self-reported compromise. It is unclear if this is because users are simply more likely to discuss social media victimization on social media or because social media sites actually experience a higher volume of cybercrime. Both are valid possibilities. Research has shown that the click through rate for phishing links is higher on social media than in email and that 8% of all URLs posted on Twitter are actually spam, phishing, or malicious links [129]. And since users likely share similar networks of friends on different services, it makes sense that they would post about compromise on one site on another: *Hello friends! My GroupMe was hacked so don't click anything that was sent.*

7.4.3 Consequences

Of our 2,910 cases, 1,884 specified a total of 2,220 consequences. Table 21 presents an overview of the major categories of consequences reported by users and the frequency and percentage with which they occurred. Loss of account access was the most frequently reported: *@Microsoft my account was suspended since it was hacked and now I am locked out of all my services.* Altered setting was the next most frequently reported consequence: *@DavorCoinHELP somebody hacked my account and turned on two factor authentication.* Spam – *My Twitter was hacked don't try to buy fake Ray Bans from the link I sent* – and financial loss – *@AirBnBHelp my account was hacked for $1800 in fraudulent charges* – were also frequently reported.

When we examined how consequences related to account types, we found that social media victims experienced every type of consequence except device damage, with 30.2% reporting a social media account was used to send spam and 20.8% reporting they lost access to the account. For online games, 29% of users experienced account loss, 20.8% experienced data loss, usually in the form of in-game items, and 20.2% experienced altered settings on their accounts, such as username or password changes. 37.7% of those who reported a compromised media or entertainment service experienced altered settings and 32.3% reported account loss. 9.2% observed unauthorized activity on their account, which often entailed the attacker using the entertainment service: *whoever hacked into my @Spotify account, your music taste is awful. I wish @SpotifyCares would let me see your IP address or login info.* For victimization reported regarding technology and electronics companies, almost all consequence categories were reported except reputational loss and professional issues. 67.6% of users who reported victimization involving cryptocurrency experienced financial loss.

We also observed nine users who reported victimization that led to professional issues varying from compromised business accounts to damaged workplace infrastructures:

- *@LinkedInHelp This is my 3rd try to get help. My account's been hacked and I cannot reset my password or login. I'm getting calls at work saying that my connections have been scammed by me. This is beyond embarrassing...*
- *Apparently my computer had a virus so when I connected it to the WIFI at work it wrecked the whole server.*

Consequence	Frequency	Percent of total
Account loss	592	26.7
Altered settings	406	18.3
Spam	359	16.2
Financial loss	267	12
Data loss	247	11.1
Unauthorized account activity	87	3.9
Access vector	65	2.9
Reputation loss	61	2.7
Device damage	59	2.7
Professional issues	29	1.3
Identity theft	22	1
Threats	17	0.8
Miscellaneous	9	0.4

Table 21: Consequences of compromise

7.4.4 Repeated Victims

Some users made generic comments about repeat victimization: *To everyone who got a DM from me with a Instagram link, ignore. I got hacked again.* But others revealed that victimization was a common occurrence:

- *[Facebook] is hacked daily...@facebook you need to do something about this. This is the 13th time this year...*
- *Yesterday I get an email form Facebook that someone was trying to hack my account and today Twitter!? Leave me alone, I lead a normal life people!*
- *Lol my first Instagram was hacked millions of times and now my new Insta was hacked and IG won't help me so I'm done with Insta.*

Criminology research indicates victimization is concentrated on a few repeated cases and that prior victimization is often an indicator of enhanced risk of future victimization [136, 137]. Though only about 3.4% of our sample reported repeat victimization, we do not know how many users chose not to report repeat victimization. Regardless, our data may be useful for evaluating theories on repeat victimization in the cyber domain. Moreover, these results suggest there is a pool of users who are especially vulnerable to victimization.

7.4.5 Observations about Users

While collecting data and labeling our final dataset, it quickly became apparent that Twitter is incredibly valuable for understanding the poor security decisions and risks users choose to take online. Users are regularly reminded not to share passwords and yet our data collection phase revealed that these practices still occur. For example, we observed many users who claimed to have "been hacked," when in reality a friend of the user had temporary access to the user's account or device and posted as the user. Many of these occurrences appeared to be good-natured and were viewed as humorous by the user: *lol I left my account logged in and* @<username> *hacked me!* However, some users experienced the more sinister consequences of shared accounts and passwords:

- *Someone scammed me!!! I gave them my password and now they hacked my account and changed the password :(.*
- *@AirennorGAMES Hi Airennor I was hacked by 18 people because someone said my password in one of their videos and I lost all my things...*

User reports may also be useful for understanding how and why users fall for victimization, as we observed several phishing victims who seemed to recall when they were compromised:

- *I received a DM from a trustworthy account with a link. I clicked it but it was some sort of hack. So if you get a link from me in a DM please don't click.*

- *...I was hacked I stupidly clicked that link and it sent spam to all my followers with a dumb card saying click here.*

We observed some users who not only rejected common security advice but also actively chose to risk viruses or hijacked accounts. Typically, this behavior was reported in relation to accessing online content freely or illegally: *99% sure I gave my computer a virus trying to watch the show from a sketchy site*. Even more troubling is that we observed several users who acknowledged having this attitude while at work: *Got a virus on my computer at work trying to click a link to take a quiz...how was I supposed to resist that clickbait though?*

Users willing to risk illegal downloading sometimes expressed amusement at the outcome:

- *I tried to download an apk file to get Spotify for free and now I have a phone virus LMAO.*
- *LMAO the worse situation I had was when I tried to download photoshop, I caught a virus and my computer wouldn't turn on lol.*
- *My dumb self got a virus on my computer trying to watch a movie lol.*

While it was sometimes difficult to tell if humor was being used in a self-report sarcastically or as a coping mechanism, it was nonetheless expressed in some form quite often: *Hahaha I haven't been able to access my Twitter since early Feb because I was hacked and today someone said I was abusive lol.*

Many users expressed attitudes we expected to see, such as fear and distress: *My computer was hacked by a virus...I miss everything and I'm scared and depressed* *now.* Some even reported having nightmares about viruses: *I've been having the same nightmare over and over where my computer gets a horrible virus and does something personal and its creeping me out.* Note that we did not perform automated sentiment analysis to assess emotional responses; this is considered an area of future work.

Online gamers in particular expressed desperation to recover hijacked accounts, which may be due to the fact that the population skews younger than the population on other services and many people take their online games quite seriously. This was often evident in game-victimization related posts: *Roblox I have been hacked please help me...I am crying so much please help me....* Desperation was also evident in the content users were willing to share to recover an account: *My account has been hacked and they changed my phone number so I can't get back in please help. My username was:* <username> *and my password was*: <password>. For the game Roblox we observed a total of nine users who publicly shared their usernames and passwords on Twitter when requesting help for hacked accounts.

While those who do not play online games may not understand the value of game accounts or in-game items, cybercriminals certainly seem to understand that games can be lucrative platforms: *I got hacked yesterday...the guy that hacked me has it and he said I need to pay $15 to get it back.* In addition to hijacking online game accounts for ransom, game accounts and in-game items can be highly valuable. There are many online services that facilitate buying and selling game accounts. On the game trading website Player Up, we found accounts for the game Roblox, the service that occurs in our sample with the second highest frequency, selling for as much as

$450 [138]. We were also able to find marketplaces for other games in our sample where accounts and digital items sell for monetary value. On one service, we found accounts for the game Fortnite selling for $2000 [139].

During data collection, we also observed many misconceptions about cybersecurity propagated on Twitter. We observed several users who seemed to believe that avoiding viruses simply requires common sense: *If you need anti virus software you probably shouldn't be using a computer.* Others still believe that it is not possible to get viruses on phones or Apple devices: *iPhone is impervious to viruses.*

7.4.6 Additional Observations

Our data collection process also resulted in many observations that were not directly tied to understanding user cybersecurity experiences. Here, we share two observations that may be of interest to other security researchers.

First, our dataset hinted at problems in how account recovery procedures are designed and implemented: *@LinkedInHelp I can't follow up on the case I made because obviously I don't have access to my hacked account. The link you sent when I created the ticket about my hacked account requires log in. VERY SMART YOU GUYS.* We observed many users expressing frustration at the lack of real customer service support: *@instagram my account's been hacked twice in 24 hours. I've tried to call but your voicemail clearly says you don't speak to people. I already changed my password 3 times. Any help?* Others expressed frustration at how long it takes to recover an account: *My Facebook account was hacked January 1st...almost 1.5 months passed and I didn't get any solution.* While most services received criticism for how they handle compromise, one company regularly received praise: *my*

@Netflix account was hacked. The nice guy from the help desk took less than two minutes to recover it. Great service! ☺

Second, our data collection process revealed that we were not the only ones interested in identifying cybercrime victims. In February, we observed eight instances where Kaspersky Lab, the official Twitter account of the anti-virus company Kaspersky, responded to cybercrime victims offering free versions of their software. Kaspersky Lab's responses were all slightly different but generally followed a similar format: "*So sorry to hear that. Let us help you remove the viruses on your computer. Grab a free version at <link> Have a nice night!*

7.5 Discussion

This section presents a discussion of our dataset and observations made during data collection in addition to an overview of the benefits and challenges associated with using Twitter.

7.5.1 Discussion of Results

There is a wealth of information contained in the self-reports of users. A six-week period of data collection provided us with hundreds of examples of infected or hijacked devices and accounts. Real victimization data is generally difficult to obtain, as organizations typically keep a close hold on such information. While this is understandable given an organization's need to protect user privacy, security, and brand reputation, it limits the data available to security researchers. As such, understandings of users often come from limited sources of data: for example, there is substantial work to understanding phishing in university environments. Our work suggests there are opportunities to use user-reported data to study different attacks and different populations of users.

Our data revealed that users most frequently self-report cyber compromise on social media, online games, media and entertainment platforms, technology and telecommunication services, and cryptocurrency wallets and exchanges. It also provided insight into how users react to compromise and the consequences they (and sometimes those around them) face, such as frustration and embarrassment, financial loss, and even compromised workplace infrastructures. User reactions to compromise also alerted us to flaws in how fraud/abuse detection and account recovery are handled on various platforms. Understanding the services that experience high levels of compromise and the experiences and reactions of users can help security researchers and developers chose how to focus limited resources.

Our work also highlights that there are some online communities that may be especially vulnerable to cybercrime and may not be aware of the resources and strategies available for avoiding or handling victimization. This was the case for the 99 users who experienced repeat compromise. It is also true for users of online games, which represented the second highest type of account that users self-reported victimization about. While many Internet services require that users be at least 13 or even 18 years of age, online games are often specifically marketed toward and intended for children. Such accounts may be easier to hijack or compromise because young Internet users may not be experienced in recognizing social engineering attempts and may exhibit a variety of poor security practices, such as password sharing.

User reports can also point to platforms and services that may not seem like they would be targeted by cybercriminals. For example, while we were unsurprised to see Twitter and GroupMe hijacked in order to spread phishing links, we were surprised to see that fast food companies were also targeted for rewards points and free food. We observed some users who were equally surprised by this: *Looks like I was hacked on @dunkindonuts account. Need to have strong passwords everywhere even in applications like this. Thanks for the help, hopefully I'll get my new DD card soon.* And as previously discussed in the context of online games, though some reported platforms may not seem as serious or important as bank accounts or financial applications and therefore not worth the attention, there may be entire online economies that do find real financial value on these platforms.

Additionally, some platforms may simply be gateways to other platforms or accounts. This may certainly be the case for games, which are often linked to credit cards, banks, or other payments systems that were used to purchase the game: *@electronicarts my account has been hacked and my password changed. EA support isn't responding. Online pages don't help. I'm worried I have a payment card on my account.*

As another example, some user posts demonstrated how infections and malicious content spread across platforms: *Hi everyone, please don't open any links sent by me in the last 24 hours. My [Twitter] account was hacked through a spam WhatsApp link but I've recovered my password with Twitter support's help.* Monitoring the flow of threats across channels and platforms may help security practitioners better defend their systems; spending time and resources to build a secure service is not useful if a less secure service acts as a gateway.

Our data collection process and dataset also reveal that Twitter is very valuable for understanding user security behaviors and attitudes. Twitter can help security practitioners identify the users who may be particularly vulnerable or susceptible to cyber victimization. This may be especially important on networks where users are allowed to connect their personal devices to the network; note that we had users in our sample who posted about accidentally harming their workplace information technology with their personal devices.

Additionally, users who post, recommend, or otherwise react to inaccurate cybersecurity advice may pose a high risk to an institution's network because they may be more likely than others to practice poor security behaviors or reject security advice. Other concerning users are those who react to victimization with humor or are willing to risk viruses or compromise in order to freely or illegally access content. If a user does not care about downloading a virus when using a certain device or platform, security engineers must find ways to develop systems that nudge the users toward safe practices or make unsafe practices more difficult to accomplish. Furthermore, human error is often credited as the cause of successful phishing campaigns. Cybersecurity training is often presented as a phishing mitigation strategy, but training users to recognize phishing attempts is only effective if human error is truly the reason that people click on malicious links. If a user does not care about the risk associated with victimization, training a user to recognize phishing links is unlikely to be useful. Intentional subversion of security is a separate problem that may need to be handled in a different manner.

7.5.2 Benefits of Using Twitter

Understandings of users and their cybersecurity behaviors and experiences are often derived from surveys or controlled experiments. Twitter - and social media in general - is still a largely underused source of data. Though survey and experiment-based work has revealed valuable correlations between individual differences in users such as gender and age and likelihood of cybercrime victimization [108, 109, 113, 140, 141], these methods have shortcomings. These methods are cost intensive, but social media data can be gathered cheaply and automatically [120, 142]. They also suffer from response bias [143], but social media content is unsolicited and abundantly available [120, 142]. Surveys and controlled experiments are also generally scoped to a specific population, but social media data can provide broad geographic coverage of many different demographic groups [120, 142]. Finally, participants cannot always be trusted to answer questions honestly and may adjust their behavior when being studied, but on social media many users actively, openly, and publicly share everything from their emotions to their personal medical issues [120, 142]. In the field of public health, Twitter has been used to quickly and cheaply assess populations across diverse geographic regions and to uncover findings that would have otherwise been infeasible to study via surveys or experiments [120, 142].

As mentioned previously, social media data may also help fill gaps in attack or victim data available to researchers to study. User-reported cases of victimization can also help with identifying the prevalence of cybercrime. Metrics on the different types of cybercrime and the platforms they occur on are difficult to obtain. Our work indicates that obtaining such estimates may be possible through Twitter data. This may be useful for directing security research and engineering efforts, which is especially important since cybersecurity is expensive and resources to secure a system are limited. Moreover, such metrics may help users evaluate the risk they are taking when they use a particular platform. For example, knowing that accounts are frequently hijacked on Twitter may prompt users to develop stronger passwords or exercise more caution when using Twitter than they otherwise would, just as how warning labels on food and medications allow people to take more informed risks when they consume them.

Twitter is also commonly studied because it mirrors trends and events that occur in the physical world in almost real time. The security field already makes use of Twitter to keep track of new vulnerabilities, exploits, and threats; real-time analysis of user security experiences through Twitter may also facilitate a shorter turnaround time for a company or institution to respond to security threats against users. For example, we observed many users who blamed their hijacked Twitter accounts on an application that supposedly would let them see who viewed their profiles: *Looks like my account was hacked. Don't click on the 'check who has visited your profile' app.* If network administrators are actively monitoring user reports, seeing posts such as this could prompt blacklisting of malicious applications or websites before other users are affected.

7.5.3 Challenges with Using Twitter

There are challenges associated with using Twitter data to understand users. Here we highlight two major challenges that will likely generalize to any Twitter-based user analysis. In the following section we discuss limitations specific to our study.

First, it is difficult to know how representative Twitter users are of the general population. According to a Pew Research survey conducted in January 2018, of the adults in the U.S, 24% of men and 23% of women use Twitter; 40% of 18-29 year olds, 27% of 30-49 year olds, 19% of 50-64 year olds, and 8% of 65+ year olds use Twitter; and 24% of white people, 26% of black people, and 20% of Hispanic people use Twitter [144]. A more detailed breakdown of demographic categories of Twitter users in the U.S is not readily available. Additionally, an estimated 79% of Twitter accounts are actually geographically located outside the U.S [145] and again, it is unclear what the demographic breakdown is like.

Second, there are privacy concerns associated with using Twitter to analyze users. The ethics of using social media data to study users is an active area of debate, brought to mainstream media attention most recently by the Facebook and Cambridge Analytica scandal [146]. Our dataset contained only the users who chose to make their accounts public and is intended to equip security researchers and developers with a resource that may be helpful for understanding users. We very intentionally do not attempt to connect user accounts with real people. So, while our initial analysis suggests that Twitter data is valuable for understanding users, the benefits to users should outweigh the risks. Furthermore, the privacy expectations of users should be carefully considered: for example, is it within the rights of a company or university to monitor any public Tweets of their population in order to identify the users who may pose an additional security risk to the network?

7.6 Limitations

Our final dataset has several limitations. First, our dataset only consists of users who reported victimization during a six-week period at the start of 2018. While this is a relatively small window of time in comparison to how long Twitter has existed (the service was created March 2006), it still provided us with a large sample of users. In future work, we plan to collect users over a longer period of time.

Second, the decision about which Tweets to retain and which Tweets to discard was subjective. While Tweets were only retained if two different reviewers agreed that it reflected a true case of victimization, subjectivity was hard to fully avoid because we were passively inferring victimization and unable to directly verify the victimization claim with the user.

We also cannot guarantee that users are being honest about victimization in this experiment. While we tried to remedy this by requiring proof in the form of details about the victimization, users can of course fabricate such stories. However, we believe most users are being honest, especially because our collected Tweets usually represent warnings to friends and family or outreach to a company. We assume that most users do not want their followers to be victimized by their account if it has been hijacked. Additionally, we observe that many users seem to recognize Twitter as an alternative to email or phone for accessing company customer support:

@AirBnB my account has been hacked and there is no way to contact you since I can't log in with my email. Please help! Similarly, some users turn to Twitter because they are either frustrated with the customer support process - *@Amazon My account was hacked and my email changed...Your customer service hung up on me twice* - or because they cannot find any other customer support resources - *I can't get on Facebook because my account was hacked...Facebook does not have customer support. Any help?* A Sprout Social survey from 2016 suggests 34.5% of users actually prefer to reach customer service via social media [145], further suggesting the legitimacy of Twitter reports.

In addition to the difficulty of guaranteeing the veracity of user reports, there may also be cases where users genuinely believe they have been victims when in reality they have not (i.e. an incorrect self-diagnosis). We cannot assume that users have the security or computer expertise to accurately and consistently distinguish between technology or connectivity issues and cyber victimization. For example, during our data collection phase we observed a user who reached out to a company about a hacked account and posted a screenshot of their browser as evidence. However, the browser simply displayed a blank webpage with a "This page cannot be displayed" error banner. There are a variety of reasons a user might see such an error, many of which are not related to malicious activity. While this case was easy to discard, there may be cases where the user-provided context falsely corroborates the victimization claim.

There is also the slight chance that we captured cases that do not represent real users. According to a 2017 study [147], between 9% and 15% of Twitter accounts are bots. While our manual review process was intended to mitigate the risk of including inauthentic cases in our final dataset, we do not make any additional attempts to filter out fake accounts. This is an entirely separate research problem that was considered out of scope.

The next major limitation is the manual labor necessary to review and confirm our sample of self-reported cases. Though we chose this approach to ensure the quality of our dataset, our methodology is not something that could easily be applied or replicated in an operational, non-research environment. In future work we will explore automating the review process.

Our final limitations are directly connected to the challenges associated with using Twitter data presented in the previous section. We do not have insight into the representativeness of our dataset. Additionally, our sample contains only users who chose to self-report victimization. While data on victimization cases is abundant, we have no insight into how many users choose not to report victimization. Similarly, while we derived some metrics about the prevalence of victimization across various platforms, we do not know how self-reporting behavior varies across users of different platforms. For example, while our results suggest that users of social media and online games are affected the most by hijacked or hacked accounts, these users may simply be more active on Twitter or members of this community may have collectively established that reaching out for help via Twitter is normal and/or effective.

7.7 Conclusion

This chapter presents work to construct one of the first Twitter datasets of user-reported cases of cyber victimization. We search Twitter during January 1, 2018 through March 13, 2018 for various cybercrime related keywords and employ a thorough manual review process of the search results to build a corpus of 2,910 cybercrime victims. Recall that our research questions centered on what we could learn from self-reported data about the experiences of cybercrime victims.

RQ7.1 asked what types of devices or accounts were reported compromised. Of the 472 users who specify an affected device, 58.7% report compromised computers and about 40.0% report compromised phones. Of the 2,331 users who specify an affected account, 42.6% report compromised social media accounts while 31.6% report compromised online game accounts. Media and entertainment applications such as Netflix and Spotify represent the third most reported affected account types, with 5.5% of users reporting victimization on these platforms. Technology, electronics, and telecommunication services (5.0%) and cryptocurrency wallets and exchange sites (3.1%) follow as the fourth and fifth top occurring account types, respectively. RQ7.2 asked what consequences users reported. Of the 1,884 users who specify a consequence, 26.7% report a loss of account access, 18.3% report altered account settings, 16.2% report accounts used to send spam, and 12% report a financial loss. Finally, RQ7.3 asked if we could identify repeat victims in the data. We determine that at least 3.4% of our sample has previously been victimized.

Our quantitative analysis of user reports revealed many interesting insights about users and user victimization. Many users revealed poor security practices and a surprising amount of disregard for cybersecurity. For some, free content or entertainment was often worth the risk of compromise. For others, compromise could be both financially and emotionally devastating. User victimization occurred on a range of platforms – even fast-food rewards points were targeted. The account recovery process was burdensome and sometimes impossible.

Based on the findings presented in this chapter, we hypothesize that mining social media for unsolicited user feedback on security experiences is not only useful for uncovering insights to support user reputation assessments, but can also help security researchers and developers gain new understandings of the user attitudes, behaviors, and misconceptions that they must contend with when designing security solutions. Institutions – and in particular their network administrators - may find Twitter data a valuable resource for understanding their network user populations and identifying the users who may present an enhanced risk to the network. In the next chapter, we extract features from the Twitter profiles of the victims identified in this study, construct a dataset of 'controls' (users who did not report victimization), and evaluate whether these features can be used to predict user-reported cyber compromise.

8 Social Media as a Source of Data: Predicting User Disclosure of Victimization

8.1 Introduction

Social media has facilitated and encouraged a culture of constant communication and oversharing. Research in a variety of fields - from politics to public health - takes advantage of this to study users and develop theories and predictions about human behavior.

In the field of cybersecurity, the use of social media data to understand and make predictions about people is still largely unexplored. In the previous chapter, we observed that some people turn to Twitter to disclose personal incidents of cyber victimization (*Help, I've been hacked!*), often divulging everything from the illegal downloading behaviors that led to account compromise to the usernames and passwords associated with compromised accounts.

As we did with the network traffic case study presented in Chapters 5, in this chapter we test the feasibility of using our dataset to make predictions about users. We use 1,903 cases of compromise - users who publicly tweeted about victimization on Twitter during the period of January 1, 2018 through March 13, 2018 – and identify an additional 7,056 controls - users randomly selected if they were publicly active on Twitter during the same time period and did not self-report victimization in their past Tweets.

Through surveys and controlled experiments, related work has correlated user Internet activity and demographics with susceptibility and likelihood of exposure to various forms of cyber compromise. Motivated by these findings, we investigate whether similar features can be extracted from Twitter data and used to differentiate between cases of compromise and controls. We hypothesize that publicly available Twitter profile data can be used to predict which users will turn to social media to disclose victimization.

We extract features from a user's Twitter profile summarizing the user's activity patterns and volume, Tweet content, and social network. We use predictive lexica to determine demographic factors of age and gender from past user Tweets. We then perform supervised learning and evaluate feature importance relative to each classifier.

Our best performing classifier achieves an accuracy of 79.1% and AUC ROC of 87.4% when differentiating between cases and controls. Examining feature importance to this classifier (a measure of how often and valuable a feature is for making key decisions), we identify account age and number of Tweets as being the most important indicators of whether a user will disclose victimization on Twitter. Predicted gender and age, demographics commonly linked to likelihood of victimization, are ranked as some of the least important features in comparison to other features.

We make the following contributions:

- To the best of our knowledge, we present one of the first studies that makes use of unsolicited, user-reported data to understand online victimization.

- We identify features that can be used to predict if a user will self-report victimization. When considering self-reported victimization as a proxy for ground truth victimization data, our results corroborate related work that suggests online activity is linked to likelihood of victimization.

- We find demographic factors of age and gender play almost no role in a classifier’s decision to label a user as a case or a control. This contrasts related survey-based work that often identifies age and gender as key factors.

- We illustrate the viability of using social media data for user-focused cybersecurity research and show it is possible to make predictions about users with publicly available profile data.

This remainder of this chapter is organized as follows. Section 8.2 presents the background and related work, followed by the methodology for dataset construction, feature engineering, supervised learning, and feature importance assessment in Section 8.3. Section 8.4 presents the results, followed by a discussion of these results in Section 8.5. We provide an overview of the limitations in Section 8.6 and conclude the chapter with future work to refine and expand this study in Section 8.7.

8.2 Background and Related Work

This section presents related work on using Twitter to understand users and related work on assessing the factors that make users more susceptible to victimization and cybercrime.

8.2.1 Using Twitter Data to Understand People

In Chapter 7 we identified papers in the field of public health that relied on self-reported data to make predictions about people. Using Twitter data to make predictions proved highly effective in these studies. In several cases, social media results were similar to clinical results: in [118], the Twitter-based trend predictions had a 0.985 Pearson Correlation Coefficient with the US Center for Disease Control's flu trend predictions. In [123], using features extracted from Twitter data motivated by the existing alcoholism literature, the author was able to predict if a user would maintain 90 days of sobriety with 80% accuracy. The resulting alcoholism recovery rate in this study was predicted to be 22.8%, while the US National Institutes of Health predicts long-term recovery rates of 18.2%.

8.2.2 Understanding the Factors Tied to Online Victimization

In [123], the author used previous alcoholism recovery research to motivate the Twitter-based feature set. In this section, we provide an overview of related work that has correlated Internet behaviors and demographic factors with user susceptibility to cybercrime and likelihood of exposure to malicious activity or content.

In [115], the authors analyzed the web browsing behaviors of users to predict who was more at risk than others to being exposed to malware. Using telemetry data from a major anti-virus company containing 100,000 users and millions of URLs visited by the users, the authors extracted 74 features summarizing user behavior, to include volume of activity, temporal patterns of activity, number and type of websites visited, and variability of browser activity. The authors found that there was a significant increase in malicious URLs visited by users on the weekend; those at risk of encountering malware also spent more time at night and had higher volumes of activity than other users. The authors used insights from their correlation analysis to build logistic regression models; the model predicted users who would visit malicious or blacklisted domains with 74% accuracy and an 8% false positive rate.

In [111], the authors investigated age and gender as risk factors for malware infection using data from 3 million Windows 10 devices running Microsoft's Windows Defender anti-virus tool. While the role of age and gender on malware encounter frequency varied depending on the type of malware, younger users were more likely to encounter malware than older users (users under 24 were twice as like than users over 50 to encounter malware) and men were 1.4 more times likely than women to encounter malware.

In [112], the authors conducted a four-month study of 50 users in which they attempted to understand the risk factors associated with malware infection. The 50 participants were provided with laptops instrumented with several different anti-virus, monitoring, and diagnostic software tools. The participants used the laptops over the four-month period and attended several in person sessions during which their laptops were analyzed for infection. At the end of the study, 38% of users were exposed to malware. Of the users with at least one infection during the four-month period, 61% were male and the number of unique malware detections was higher for the 25 to 35 age group than the age groups of 18 to 24 and 36+. When the authors examined cyber behaviors, those who installed many applications, frequently visited websites regardless of website type, and visited certain categories of websites (such as streaming, adult, and gambling sites) were all significantly more likely to have a malware infection than others.

In [148] the authors used insights from [112] to build predictive models of user malware infection. Factors of age, computer expertise level, most commonly used web browser, total number of hours online, total number of websites visited, total number of files downloaded, and total number of websites visited for specific categories (such as peer-to-peer applications or social networking) were used to construct a neural net that achieved a testing accuracy of 88.89% for classifying users as either *at-risk* or *low-risk* for malware exposure.

Survey-based work has also revealed correlations between demographic factors of age and gender and susceptibility to phishing and poor security practices. The authors in [108] found women and people aged 18–25 to be more susceptible to phishing attacks than men and other age groups, respectively; the authors in [109] found a correlation between people aged 18–25 and phishing attack susceptibility; the authors in [110] found that younger people were significantly more likely to engage in the poor security practice of password sharing; we found correlations between gender and security behaviors such as strong password creation, proactive awareness of security threats, and device updating [113].

Based on the related work that has found correlations between Internet activity and demographic factors and increased likelihood of cyber victimization, we explore the following three research questions:

- **Research Question 8.1 (RQ8.1):** *Can features extracted from the Twitter profiles of users differentiate between cases of self-reported cyber victimization and controls?*

- **Research Question 8.2 (RQ8.2):** *What types of user activity on Twitter are predictive of disclosure of cyber victimization?*

- **Research Question 8.3 (RQ8.3):** *Are user demographics of age and gender predictive of disclosure of cyber victimization?*

8.3 Methodology

This section presents an overview of the process to build our dataset, extract features from the data, and evaluate our three research questions using supervised learning and feature importance calculation.

8.3.1 Dataset

Our sample consisted of 1,905 cases and 7,059 controls. As in the previous chapter, any example Tweets in the following sections have been rephrased for anonymity purposes.

8.3.1.1 Cases

Cases were drawn from the pool of users collected in the previous chapter. As a reminder, these represent the users who publicly self-reported cyber victimization on Twitter during the period of January 1, 2018 through March 13, 2018. To construct this dataset, we developed a set of Twitter search rules based on keywords commonly used to discuss cybercrime on Twitter, searched Twitter using these rules, and manually reviewed returned Tweets to determine if they represented true cases of victimization.

This first round of data cleaning resulted in 3,459 total victimization cases. For each case, we used the Twitter API to collect metadata from user profiles such as number of Tweets, number of followers, number of accounts a user is following, and number of Tweets the user marked as a ‘favorite.’ We also collected up to 3,200 past Tweets for each of these users, the maximum allowed by the Twitter API. In some cases the API could not access accounts or past Tweets, either due to an account switching from public or private or being deleted or banned; if this was the case, we discarded the user and any associated data. We also filtered out any users who did not have their profile language set to English.

For the purpose of this study, we performed a final step of data cleaning in which we discarded any users who had less than 1,000 total words in their past Tweets. At least 1,000 words were necessary to conduct lexica-based age and gender predictions [149]; 1,000 words is also generally accepted as the minimum numbers of words necessary to conduct text analysis [123]. After discarding users with less than 1,000 words in their combined collection of past Tweets, we had 1,905 cases.

8.3.1.2 Controls

The control group consisted of a random sample of Twitter users who were active on Twitter (i.e. they publicly posted on Twitter) during the same time window that our cases self-reported victimization. Because the Twitter API does not actually provide a functionality to collect a random sample of users, we made use of the unique user IDs that Twitter assigns to every user [150]. User IDs are numeric values assigned to users when they join Twitter; these IDs began at the number one (the first few IDs notably belonging to the founders of Twitter) and have followed an increasing order. At the time of writing this paper, there was no authoritative source indicating the maximum user ID number or confirming if IDs are consecutively incremented. So, we selected a range of [10, 10,000,000,000], assuming the first few IDs represent Twitter founders and knowing the estimated number of monthly users on Twitter in Quarter 1 of 2018 was 336 million [151].

Each randomly generated number was used as a user ID value to search for English Tweets during the period of January 1, 2018 through March 13, 2018 using the Twitter Search API. If the randomly generated user ID resulted in a returned Tweet, we added the user associated with this user ID to our sample of random users. This process left us with 9,382 users.

As we did with the cases, we collected various pieces of information from the user's Twitter profile and up to 3,200 past public Tweets. Again, we filtered out any users if their profiles were not public or not set to English.

For each user, we searched their past Tweets for the keywords "hacked" and "virus", the two words at the core of every rule we used to build our case dataset. If a Tweet contained a keyword, we manually reviewed it to see if it was a self-report of victimization. For example, *the CDC reports a new virus is hitting the midwest* is clearly not a self-report of cyber victimization, while *@yahoomail I can't log in I think I've been hacked* does seem to be a real report. If a user disclosed cyber victimization, they were discarded from the control dataset. This left us 7,188 users. Again, we then discarded any users with less than 1000 words in their combined past Tweets. This resulted in 7,059 users.

8.3.2 Feature Engineering

Next, we developed a set of 11 features. At a high level, these features were intended to summarize a Twitter user's activity, content, social network, and demographics. The features are presented in Table 22.

The majority of these features was embedded in user profiles and could therefore be easily extracted. Links and mentions in Tweets were found using Python regular expressions to search for character strings that matched a mention ('@') or link format.

Feature	*Description*
Tweet count	The user's total number of Tweets.
Account age in months	The age of the user's account, based on account creation date, in months.
Favorites count	The number of Tweets a user has marked as a favorite.
Listed count	The number of times the user appears in a Twitter list.
Average number of links	The average number of links shared by a user per Tweet.
Average number of words	The average number of words per user Tweet.
Friends count	The number of users that a user is following and is following the user in return.
Followers count	The number of users following the user's account.
Average number of mentions	The number of times the user has mentioned another user in a Tweet.
Age	The predicted age of the user.
Gender	The predicted gender (male or female) of the user.

Table 22: Features and their descriptions

8.3.2.1 Age and Gender Prediction

Demographics are not indicated on Twitter profiles. We therefore had to use the information available to make predictions about user demographics. For this study, we chose to focus on age and gender, in part because they are frequently correlated with cybersecurity attitudes and behaviors in the related work and because of the availability of predictive lexica to calculate these features using writing samples of users. In particular, we used the age and gender predictive lexicas established by the University of Pennsylvania's World Well-Being Project [152]. The authors established these lexicas by analyzing the language of 97,000 Facebook, Blogger, and Twitter users [149]. The age and gender lexicas consist of collections of words and weights associated with each word. Each lexica also contains an intercept value that is used to correct for model bias. Age and gender are calculated using the following formula:

$$\left[\sum_{word \in lexica} \left(word_{weight} * \frac{word_{frequency}}{total_{words}}\right)\right] + intercept$$

In the equation, $word_{frequency}$ is the frequency with which a word from the lexica appears in a user's text sample, $total_{words}$ is the number of words total in a user's text sample, $word_{weigh}$t is the weight associated with the word in the lexica, and *intercept* is the intercept value associated with the lexica. When using the age lexica, the value of the user is the predicted age. When using the gender lexica, if the value for a user is positive, the user is predicted to be female, otherwise the user is predicted to be male [152].

For each user in both our case and control samples, we used all the publicly available Twitter posts for each user to predict age and gender. There were a few cases where predicted age was less than 0. We discarded 2 users in the case dataset and 3 users in the control dataset who had a predicted age less than 0, bringing out total case dataset to 1,903 users and our control dataset to 7,056 users. The breakdown of age and gender is presented in detail in the following section.

8.3.3 Supervised Learning

We performed supervised classification on our dataset using decision tree-based algorithms. We selected tree-based algorithms to enable us to assess the importance of each feature, discussed further in the following section. We implemented the following algorithms using Python's sklearn library: Random Forest [94], Ada Boost [95], Gradient Boost [96], and Extra Tree [97]. Grid Search [100] was used to tune the hyper-parameters and stratified 10-fold cross validation [101] was conducted to prevent model overfitting.

We performed classification using two common train/test splits: 70/30 and 80/20. Since we do not know the ratio of victims to non-victims on Twitter, we tested two different ratios of cases and controls: 30:70 and 50:50. To obtain these ratios, we took random samples from the control dataset, such that the 30:70 ratio had 1,903 cases and 4,440 controls and the 50:50 ratio had 1,903 cases and 1,903 controls.

Figure 5 presents the breakdown for gender for the 30:70 ratio while Figure 6 presents the distribution of age. Figure 7 presents the breakdown of gender for the 50:50 ratio while Figure 8 presents the distribution of age for this ratio.

Both samples contain more males than females. This is in line with estimates of the age distribution of worldwide Twitter users: a Statista report predicts that Twitter is 66% male and 33% female as of October 2018 [153]. Predicted age seems to be distributed in line with predicted estimates of the Twitter population as well: a

Pew Research survey conducted in January 2018 reports that Twitter is used by 40% of 18-29 year olds, 27% of 30-49 year olds, 19% of 50-64 year olds, and 8% of 65+ year olds [154]. While we do not have ground truth on the actual ages or genders of users (and for privacy reasons we intentionally make no attempt to connect users with real people), these results give us reasonable confidence in the prediction results and the representativeness of our sample compared to the Twitter population.

For each classifier, we calculate accuracy, false positive rate, false negative rate, and area under the receiver operating characteristic (AUC ROC) curve, recognizing that depending on the use case, different metrics can be appropriate for judging the success of a classifier. These results are presented in Tables 23 and 24 for the four different testing scenarios. Values are rounded to the third decimal place.

Gradient Boost had the highest accuracy and highest AUC ROC in every scenario; it almost always had the lowest false positive rate (FPR in the table) and false negative rate (FNR in the table). Random Forest performed quite poorly; in some cases the accuracy and AUC ROC values suggest the algorithm was not performing much better than random guessing.

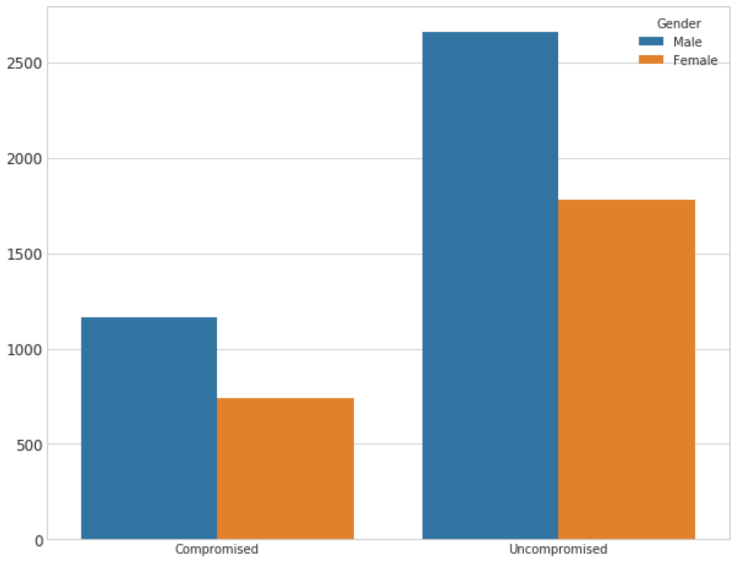

Figure 5: Number of predicted males and females in the cases (compromised) and controls (uncompromised) datasets for the 30:70 ratio

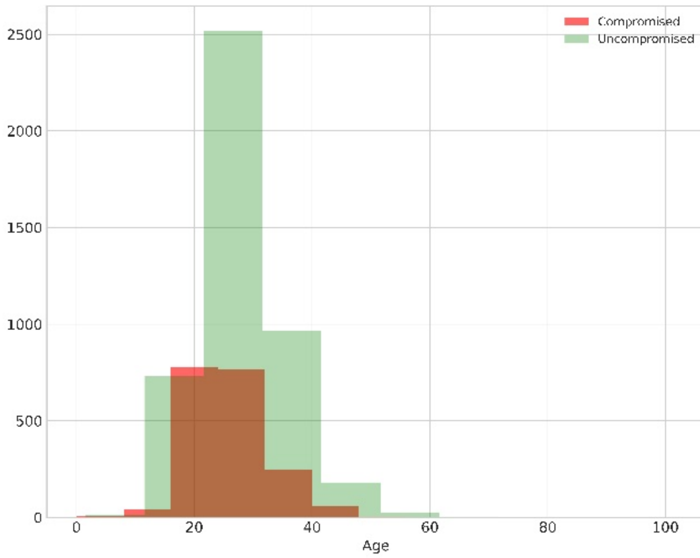

Figure 6: Distribution of predicted age for the 30:70 ratio

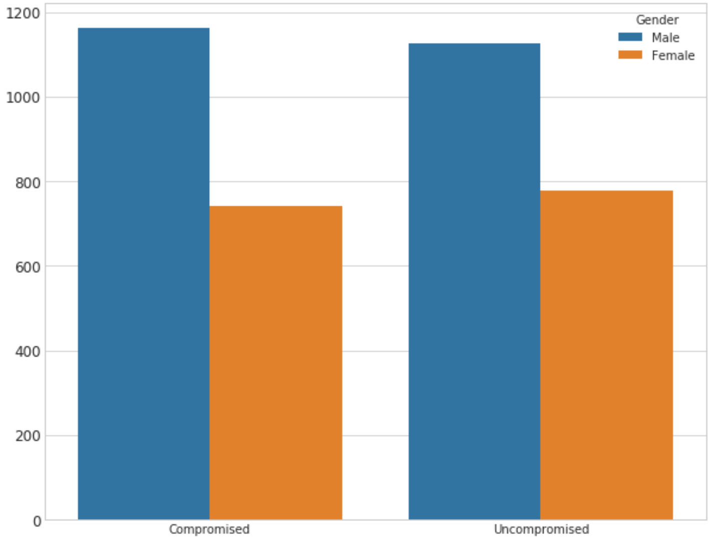

Figure 6: Number of predicted males and females in the cases (compromised) and controls (uncompromised) datasets for the 50:50 ratio

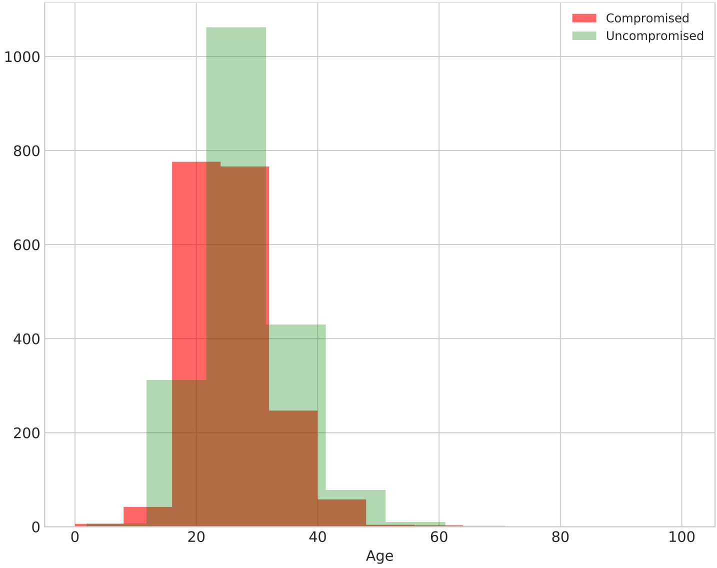

Figure 7: Distribution of predicted age for the 50:50 ratio

8.3.4 Feature Importance

For tree-based algorithms, feature importance is a value indicating how useful a feature was in the construction of the decision tree; features can be ranked based on importance values to understand which features were used the most to make decisions [114]. Feature importance can be calculated directly using the *feature_importances* function provided in each of the Python sklearn classifier libraries.

Tables 25 and 26 present the feature importance determined by the Gradient Boost algorithm. Feature importance of Ada Boost and Extra Tree are presented in Appendix B. Random Forest, our worst performing classifier, gave each feature an importance of 0.0 in almost every ratio and testing/training scenario and so its results are not presented.

30:70 Case to Control Ratio				
Classifier	***Accuracy***	***FPR***	***FNR***	***AUC ROC***
80/20 Train/Test Split				
Random Forest	0.696	0.696	0.304	0.5
Ada Boost	0.768	0.231	0.233	0.854
Gradient Boost	0.782	0.217	0.22	0.863
Extra Tree	0.76	0.216	0.295	0.803
70/30 Train/Test Split				
Random Forest	0.699	0.699	0.301	0.5
Ada Boost	0.774	0.226	0.227	0.854
Gradient Boost	0.788	0.212	0.213	0.866
Extra Tree	0.757	0.232	0.271	0.807

Table 23: Supervised classification results for the 30:70 case to control ratio

50:50 Case to Control Ratio				
Classifier	***Accuracy***	***FPR***	***FNR***	***AUC ROC***
80/20 Train/Test Split				
Random Forest	0.719	0.061	0.499	0.731
Ada Boost	0.753	0.243	0.251	0.836
Gradient Boost	0.768	0.232	0.232	0.867
Extra Tree	0.745	0.243	0.266	0.817
70/30 Train/Test Split				
Random Forest	0.632	0	0.729	0.635
Ada Boost	0.758	0.24	0.243	0.845
Gradient Boost	0.791	0.208	0.21	0.874
Extra Tree	0.739	0.216	0.306	0.805

Table 24: Supervised classification results for the 50:50 case to control ratio

8.4 Results

This section presents an overview of the results.

8.4.1 Research Question 8.1

Research Question 8.1 (RQ8.1) asked if features extracted from the Twitter profiles of users can be used to differentiate between cases of self-reported cyber victimization and controls. Gradient Boost was the best performing classifier, achieving accuracy values ranging from 76.8% to 79.1% and AUC ROC values ranging from 86.3% to 87.4%.

Though our accuracy results are comparable to related work predicting user behavior from Twitter data [123] and related work predicting user encounters with malicious domains [115] or malware [148], our false positive and false negative rates are high (around 21 to 23% for both FPR and FNR). Despite this, our results suggest there are opportunities to use data from a user's Twitter profile to predict if a user will disclose victimization.

8.4.2 Research Question 8.2

Research Question 8.2 (RQ8.2) asked what type of user activity on Twitter could be used to predict disclosure of cyber victimization. We focus on assessing feature importance associated with the Gradient Boost algorithm, our best performing classifier, to answer this question.

For the 30:70 case to control ratio and both training/testing splits, account age is ranked as the most important feature for classification, followed by number of tweets, and then the number of favorites. For the 50:50 case to control ratio and both training/testing splits, number of Tweets ranks first, followed by account age, and then number of favorites again. Overall, features related to Tweet content, such as average number of words and shared links, rank toward the middle of the list. Features related to a user's social network, such as number of followers, number of friends, and list count, rank toward the bottom of the feature importance list.

As discussed in the Background and Related Work section, other researchers found volume and timing of online activity to be important predictors of malware exposure or infection [112, 115, 148]. We similarly find high volume of activity and length of time on Twitter to be predictors of self-reported victimization.

30:70 Case to Control Ratio, Gradient Boost Feature Importance			
80/20 Train/Test Split		**70/30 Train/Test Split**	
Feature	***Importance***	***Feature***	***Importance***
Account age in months	0.18	Account age in months	0.217
Tweet count	0.171	Tweet count	0.162
Favorite count	0.141	Favorite count	0.15
Average number of links	0.107	Average number of links	0.1
Average number of words	0.1	Average number of words	0.08
Average number of mentions	0.078	Average number of mentions	0.079
Age	0.077	Followers count	0.076
Followers count	0.062	Age	0.06
Friends count	0.043	Friends count	0.04
Listed count	0.034	Listed count	0.028
Gender	0.007	Gender	0.008

Table 25: Supervised classification results for the 30:70 case to control ratio

50:50 Case to Control Ratio, Gradient Boost Feature Importance			
80/20 Train/Test Split		**70/30 Train/Test Split**	
Feature	***Importance***	***Feature***	***Importance***
Tweet count	0.183	Tweet count	0.201
Account age in months	0.157	Account age in months	0.165
Favorite count	0.144	Favorite count	0.147
Average number of words	0.113	Average number of links	0.123
Average number of links	0.098	Average number of words	0.119
Average number of mentions	0.087	Average number of mentions	0.075
Age	0.087	Followers count	0.06
Friends count	0.056	Friends count	0.04
Followers count	0.044	Age	0.032
Gender	0.019	Listed count	0.02
Listed count	0.012	Gender	0.016

Table 26: Supervised classification results for the 50:50 case to control ratio

8.4.3 Research Question 8.3

Research Question 8.3 (RQ8.3) asked if demographic factors of age and gender could be used to predict disclosure of cyber victimization. Again, we focus on assessing feature importance associated with the Gradient Boost algorithm, our best performing classifier, to answer this question.

In the four different ratios and train/test split scenarios, age is ranked in the bottom half of all features: it is ranked as 7 out of 11 twice, 8 out of 11 once, and 9 out of 11 once. Gender is ranked as the least important feature three times and as the second to last most important feature once.

Unlike the related work discussed previously which found relationships between age and gender and likelihood of victimization or malware encounters, our results suggest these demographics are not so useful for understanding self-reporting behaviors and self-reported victims.

8.5 Discussion

This section provides a discussion and potential applications of our results.

Our results suggest account age and number of Tweets are some of the most important features for differentiating between users who publicly disclose cyber compromise and those who do not. This suggests a few theories that can be explored in future work. The first is that longer ‘exposure’ may be linked to a higher chance of victimization, i.e. more time online simply means more opportunities for a user to be victimized. By the nature of our dataset, all users we evaluated have public accounts, so another potential theory is users who frequently and freely share information publicly have weaker security practices or are more likely to be noticed and therefore targeted by attackers. Alternatively, users who post more frequently and have been on

Twitter longer than others may simply be more comfortable turning to the platform to express their problems.

Our results also suggest that age and gender, frequently reported in related work as important demographic factors for identifying vulnerable users or users with poor security practices, are not so important when it comes to understanding victimization cases reported on Twitter. The impact of age may not be obvious in our data because Twitter skews young [154] and victimization is generally reported to be more common in the 18-24 age range [108, 110, 111, 113]. It is unclear why gender has little to no importance. This may simply indicate that men and women have similar victimization-reporting habits. Alternatively, the lack of importance of both age and gender in our study may have broader implications when discussing the relationship between demographics and cybersecurity practices.

Most user-focused work in cybersecurity studies a limited range of security incidents and behaviors. As discussed in our related work, correlations between demographics and likelihood of victimization are often specifically related to phishing; a few papers connect demographics to malware encounters. Therefore, it may not be surprising that we found no relationship between demographics and victimization reports, since the Twitter data was not limited to evaluating one type of attack or one type of device (and it is reasonable to assume that vulnerability may be context and platform specific).

Overall, work evaluating a broader range of attacks and user security behaviors and attitudes is limited. We hypothesize that this is largely because real victimization data is often hard to obtain. Most organizations keep a close hold on incident data for privacy, security, and brand reputation-related reasons.

Using user-reported Twitter data as a source of ground truth on victimization may provide an opportunity for researchers to develop theories and understandings about previously unstudied platforms or attacks. While using social media data for this purpose relies on user-reported data, we assume that most users who self-report victimization are being honest based on the nature of their reports.

In our dataset, most users report victimization to apologize to their followers for content propagated by a compromised account, e.g. *Sorry for the spam, been hacked;* to warn their followers away from malicious content, e.g. *If you got a private message from me, don't click the link!*; or to get help from the service provider associated with the compromised account e.g. *@Facebook someone hijacked my account and I can't find your customer service information.* We therefore assume we can use self-reported incidents as a proxy for ground truth data on user compromise.

In addition to enabling broader coverage of security incidents, Twitter data can also potentially be used to proactively identify users who will become cybercrime victims. In the field of public health, identifying risk factors in people that make them more susceptible to certain diseases enables health practitioners to recommend early interventions that can reduce the likelihood of the disease. Similarly, identifying users who are more likely to become cybercrime victims through relatively easy to obtain data may be a way for network administrators or service providers to identify the more vulnerable users in their populations and step in with early interventions or additional protections for these users.

8.6 Limitations

The previous chapter highlighted several limitations of our data collection process and dataset. The following presents a few more limitations specifically relevant to the classification problem addressed in this chapter.

Just as there was uncertainty in our case dataset, there is uncertainty in our control dataset. For our controls, we do not know how many users who did publicly report victimization we failed to identify in our data cleaning process. We also do not know how many controls have been victims of cybercrime but have chosen not to publicly share this on Twitter. Because of this, we can only use our immediate results to analyze user-reporting behaviors. In our discussion section we suggest the possibility of using our results to predict users who will become compromised, but work to understand what percent of people who are victims choose to report victimization is necessary first. Again, we also do not know the mapping of users to real people. Therefore, it is possible that some ‘users’ are multiple people or that multiple user accounts are owned by one person; it is also possible that some accounts are fake. Identifying fake or duplicate identities is a challenging and distinct problem that was outside the scope of this paper.

The next limitation is the lack of ground truth for the gender and age of users. We used University of Pennsylvania's publicly available predictive lexica, which was considered state-of-the-art in age and gender prediction at the time it was published in

2014 [149]. We know it is likely that some predictions are incorrect; in some cases we observed obvious problems, such as users with negative predicted ages (these were discarded from the final sample). However, to the best of our knowledge there was no other readily available or more recently developed lexicon that we could apply to our study. Additionally, our predictions were in line with estimated Twitter demographics in 2018, so we felt confident moving forward with this as our selected approach.

Our final limitation is the high false positive and false negative rates of the classifiers. This may suggest our current features are not strong enough to fully differentiate between cases and controls despite relatively successful accuracy and AUC ROC values. In future work we intend to add additional features to see if we can reduce the FPR and FNR.

8.7 Conclusion

To the best of our knowledge, this chapter presents one of the first research efforts that takes advantage of unsolicited, user-reported victimization data to study and make predictions about users in a cybersecurity context. We identify a set of keywords commonly used by people on Twitter to discuss various forms of cyber victimization and use these keywords and a thorough manual review process to iteratively construct a dataset of 1,903 cases of self-reported compromise. We also build a dataset of 7,056 controls.

Motivated by related work that has uncovered relationships between Internet activity and user demographics and likelihood of compromise, we develop a set of eleven features from the case and control user profile data. These features include number of Tweets, account age, number of followers and more, in addition to demographic factors of age and gender as determined by predictive lexica.

We perform supervised learning using four different tree-based algorithms and evalute feature importance for each classifier. Our best performing classifier achieves an accuracy of 79.1% and AUC ROC of 87.4%. Account age and number of Tweets rank as the top two most important features, while age and gender rank as some of the least important features. This corroborates work that has tied Internet activity to victimization but contrasts work that has found age and gender to be important predictors of victimization.

User-focused security research is limited despite widespread recognition of users as the weakest link in cybersecurity. Research focused on users is generally studied in the context of phishing attacks or malware infections, while understandings of users are typically derived from surveys or controlled experiments. There is a need for additional methods and sources of data that can supplement existing techniques and knowledge.

Our work shows the advantages of tapping social media data for such reseach. Considering user-reports as a proxy for ground truth data on victimization, our work hints at the possibility of predicting which users will become victims using publicly available data alone. More generally, our study shows the feasibility of tapping social media data to obtain unsolicited user feedback on a broader range of attacks and experiences.

Based on the results of this study, in future work we plan to refine and expand the dataset presented here. This will include automating the data collection and cleaning process where possible; including a broader range of features; and evaluating additional techniques to extract human traits and demographics from the Twitter data, in particular focusing on attributes linked to cybersecurity behaviors in related work. We will then apply the reputation quantification methodology and monitor these users over time, as we did in the network traffic case study.

9 Conclusion

9.1 Introduction

Reputation has a long history as a crucial factor for establishing trust and making predictions about an entity's character or quality. Though subjective by definition, in recent years it has emerged as a quantifiable concept in cybersecurity: represented as a numerical score, ranking, or label, it is often used as the basis of decisions such as which IP address to blacklist or which users to grant or deny privileged access. Its value is indisputable, but the methods to quantify it are often ad-hoc, arbitrary, and difficult to replicate or validate. Motivated by this, the goal of this dissertation was to develop a quantitative definition of cyber user reputation and a methodology for evaluating cyber user reputation for use in a network administration or security setting. This chapter concludes this dissertation with a summary of the contributions and completed work (9.2), a discussion of the proposed reputation evaluation methodology and case studies (9.3), an overview of the limitations (9.4), areas of future work (9.5), and concluding comments (9.6).

9.2 Summary of Contributions and Completed Work

This section presents a summary of the completed work, contributions, and peer-reviewed work associated with this dissertation.

9.2.1 Summary of Dissertation

In this dissertation, we defined a user as an account or collection of accounts on a computer network or Internet service associated with a unique human who exists in the physical world. There were three research questions presented in Chapter 1:

- **RQ1:** What data can be used as evidence of a user's cyber reputation for specific characteristics or behaviors on a computer network?
- **RQ2:** How can this evidence be used to define and quantify a user's cyber reputation and develop predictive models of future reputation?
- **RQ3:** How can quantification assessments and predictions of a user's cyber reputation be validated and refined?

Chapter 2 presented the background and related work on reputation systems. In the cyber domain, reputation is primarily used to differentiate between the malicious and the benign, the trustworthy and the untrustworthy. Reputation is also used to quantify an entity's perceived authority, expertise, or influence. Examining these different definitions, it became clear that 1) reputation is always relative to some specific characteristic or behavior and 2) some form of evidence is crucial for any reputation assessment.

Chapter 3 delved further into the background and related work, examining how various definitions of reputation are quantified. Based on related work that represents reputation as a probability, an intuitive approach to quantifying a subjective and uncertain quality, Chapter 3 presented a mathematical interpretation of user reputation:

A user's cyber reputation score is a probability representing the likelihood that a user demonstrates a specific characteristic on a network, based on evidence of past behaviors that are considered representative of this characteristic.

Chapter 4 presented a general methodology for answering the three main research questions of this dissertation. The methodology consists of three phases for evaluating user reputation and developing mathematical user cyber reputation scores. Phase One addresses RQ1 by exploring how evidence can be collected to evaluate some characteristic of interest. In Chapter 4, we presented a variety of different data sources that may be useful for evaluating users, including network traffic, Internet activity, and user demographics. We also presented techniques for identifying, extracting, and validating evidence from these observations.

Phase Two of the methodology addresses RQ2 by exploring how this evidence can be used to mathematically quantify cyber reputation. Since a user's cyber reputation score is a probability representing the likelihood a user demonstrates a specific characteristic based on evidence, a Bayesian Inference process can be used. Chapter 4 presented a simple example of using a Beta-Binomial conjugate pair to develop a prior distribution for a user and updating it with evidence to form the posterior, which in turn can be sampled from to determine the most likely probability a user displays some behavior or characteristic.

Lastly, Phase Three addresses RQ3 by exploring how reputation assessments can be validated and refined through pilot studies, forecast verification techniques, additional data, and variations on the reputation definition.

Chapters 5 and 6 presented a case study at a university illustrating the reputation evaluation methodology. In Chapter 5, we first explored the feasibility of using a user's network traffic as a source of evidence for reputation, where the characteristic of interest was infection status, as defined by the university's IDS/IPS threat logs. The observation space consisted of a sample of wireless network traffic for each user. We extracted 36 features from the network traffic. We evaluated their feasibility as evidence for reputation through a supervised classification problem, in which the 36 features were used to differentiate between infected and uninfected users.

While Chapter 5 served to illustrate Phase One of the reputation evaluation methodology (and therefore RQ1), Chapter 6 focused on Phases Two and Three (and therefore RQ2 and RQ3). Based on the results of the supervised classification study, we determined network traffic provided enough initial evidence to differentiate between infected and uninfected users. Motivated by this, in Chapter 6 we first illustrated the process of developing reputation models, creating various distributions using combinations of the network traffic features for a sample of users. We then used these reputation models to quantify the reputation of 4,787 infected users and 4,787 uninfected users over a three-month period. We analyzed the reputation scores and evaluated them based on various prediction/forecast verification metrics. Overall, Chapter 6 illustrated the feasibility of using a Bayesian probability-based approach for developing user reputation scores.

Chapters 7 and 8 further explored RQ1 through a case study of user-reported compromise on Twitter. Observing that users will frequently turn to Twitter to self-report compromise (*Help, I've been hacked!*), in Chapter 7 we focused on constructing a dataset of user-reported cases of compromise. An analysis of user-reported compromised devices and accounts and the associated consequences revealed many interesting observations about users and their cybersecurity attitudes, behaviors, misconceptions about security, and propensity toward risky cyber decisions.

Using self-reports as ground truth for compromised users, Chapter 8 explored whether Twitter data can be used to make predictions about users and whether enough evidence can actually be extracted from social media data to evaluate user reputation. We framed this feasibility study as a supervised learning problem, similar to the approach taken in Chapter 5. The observation space for this study consisted of up to 3,200 of a user's past Tweets and any publicly available profile data. We extracted various features from the Twitter data, such as the user's number of Tweets, followers, and favorites. We used predictive lexica to extract demographics of age and gender from user Tweets, motivated by the correlations identified in the literature between these features and likelihood of victimization and poor security practices. The study revealed the potential for using self-reported data from social media to understand users and collect evidence for quantifying user reputation. The study did not, however, find a connection between demographics and user-reported compromise. Based on the results, this study will be improved and expanded; this is discussed in more detail in the Future Work section of this chapter.

9.2.2 Summary of Contributions

The contributions of this dissertation are as follows:

- We present an intuitive, mathematical interpretation of user cyber reputation for network security settings grounded in Bayesian statistical theory and motivated by existing reputation literature.
- We present a general methodology for evaluating a user's cyber reputation that enables a network administrator to identify evidence that can be used to quantify cyber reputation, make predictions about future behaviors, assess the uncertainty associated with these predictions, and update cyber reputation assessments and predictions as new evidence becomes available.
- We contribute to the science of security by framing reputation quantification as a scientific experiment to test and validate a hypothesis, improving upon existing techniques for quantifying reputation that are often ad-hoc and arbitrary.
- We illustrate and validate the entire reputation evaluation methodology through a case study at a large, public university in which we use network traffic to develop reputation models for infection and uninfection and then apply these models to assign Bayesian probability-based reputation scores to a sample of users.
- Through this case study of network traffic and user infection, we contribute new understandings of users to the field of cybersecurity by identifying features that can be extracted from a user's network traffic to identify users who may be more susceptible to infection than others.

- We present one of the first Twitter datasets of user-reported cases of device or account compromise.
- Through our analysis of this dataset, we uncover additional insight into user cybersecurity attitudes and behaviors, in addition to illustrating both the viability and importance of using social media data to understand users from a cybersecurity perspective.

9.2.3 Summary of Published Work

The case study presented in Chapter 5 maps to a peer-reviewed paper published in the January 2019 issue of the journal *Computers & Security* [155].

The case study presented in Chapter 7 was submitted to the 2019 Web Conference, while the expansion of this case study presented in Chapter 8 was submitted to the 40th IEEE Symposium on Security and Privacy. The papers were not accepted and revisions are being made to both based on reviewer feedback. The papers will be revised and resubmitted post-defense.

An additional paper related to research presented in this dissertation but not included as a chapter was published in the March 2018 issue of *Computers & Security*. The paper, titled *Correlating Human Traits and Cybersecurity Behavior Intentions*, investigated how differences in demographics, personality, risk-taking preferences, and decision-making styles impact user cybersecurity behavior intentions through a survey of 369 students, faculty, and staff at the University of Maryland. Individual differences accounted for 5%–23% of the variance in cybersecurity behavior intentions. Characteristics such as financial risk-taking, rational decision-making, extraversion, and gender were found to be significant unique predictors of good security behaviors.

A conference paper, not included in this dissertation research, was presented at the International Conference on Security and Privacy in Communication Systems (SECURECOMM, 27% acceptance rate) in October 2015. The paper, titled *An Improved Method for Anomaly-Based Network Scan Detection* [156], investigated the use of machine learning-based approaches for network scan detection and showed the approach on a case study on the University of Maryland's network.

9.3 Discussion

This section presents a discussion of the reputation evaluation methodology, potential applications of user cyber reputation, and some general security and ethical considerations when quantifying user reputation.

9.3.1 Comments on the Reputation Scoring and Evaluation Techniques

Related work in the fields of e-commerce and engineering risk assessment in addition to scientific experiment design motivated the proposed mathematical quantification of user cyber reputation and the evaluation methodology. This section provides some additional justification for the reasoning behind this approach.

Framing a reputation score as a probability is an intuitive way to quantify a subjective concept and allow uncertainty to be transparent in the reputation score. However, as evidenced in Chapter 2, there are many ways reputation can be quantified besides probabilities. Nominal and ordinal metrics are commonly seen in both the academic literature and in reputation systems used in practice: consider star ratings for products or movies; likes and dislikes or ‘upvotes’ and ‘downvotes’ on social media; rankings for institutions; and numeric scores or labels for IP addresses and domain names.

Nominal scales can certainly serve a purpose for making a score more actionable. In the network traffic case study for example, we mapped likelihoods to the labels ‘infected’ or ‘uninfected,’ with the intention that this label could be used to help the network administrator decide where to focus security training and resources. The problem with using nominal scales alone is that they fail to capture uncertainty and can be highly subjective, especially when considering characteristics that cannot be as objectively defined as infection status or compromise. By using them in combination with probability, the degree of confidence in the assigned label can be understood.

Some reputation systems make use of ordinal scales; the FICO score and China’s Zhima Credit are examples of this. The reality is, however, that these scales are often arbitrary. Though a FICO score can certainly help banks understand which people have ‘better’ financial behavior than others, without context a score of 732 does not mean much. Even with context these scores are confusing; exactly how much better is a 733 than a 732?

Interval scales and ratio scales can be appealing but are somewhat problematic, since it is challenging to define fixed intervals between reputation values. Additionally, it may be difficult to determine what a score of 0 (i.e. the “true zero” in the interval scale) actually is in the context of reputation. Overall, using probability values provided a way to ensure reputation scores were not arbitrary numeric values or labels.

The proposed reputation evaluation methodology assumes some central authority is tasked with evaluating and maintaining user reputation scores and has insight into the identities and activities of users on the network. In other words, the methodology is intended for use in a centralized reputation system.

In Chapter 4, we proposed that the reputation evaluation methodology should essentially follow the same process as designing and validating a scientific experiment. Making a judgment about someone's reputation and observing if the judgment is correct can be equated to forming and testing a hypothesis. Chapter 4 outlined several requirements for a reputation evaluation methodology, based on requirements for sound experiments established previously in the scientific literature: datasets must be clean and the results of the reputation evaluation should make sense given the data; the reputation evaluation procedures should generalize to other networks and user populations; sensitive user data must be carefully and securely stored; and finally, other researchers and practitioners should be able to replicate the methodology. We also established several objectives for the methodology: it must allow the central authority to be proactive when assigning scores; it must enable quantification of uncertainty; and finally, it must enable updating of reputation scores over time.

The proposed methodology accomplished these requirements and objectives. Phase One captured hypothesis development and initial experiment design; Phase Two captured experiment testing; Phase Three captured experiment validation and improvement. The two different case studies illustrated the processes for developing clean datasets, practicing ethical and secure data collection and storage, proactively assigning and updating of reputation scores, and incorporating uncertainty in the reputation assessments.

9.3.2 Benefits and Applications of Quantifying User Cyber Reputation

As mentioned previously, one of the major motivations for focusing on reputation quantification in this dissertation was the application of reputation to cybersecurity. The network traffic case study and the social media case study both focused on the idea of using evidence to make some determination about how likely a user is to become compromised on a network or online platform. What can actually be done with a reputation score in practice? This section provides a deeper discussion of the applications and benefits of quantifying the reputation of users on a network.

The phrase 'the human is the weakest link in cybersecurity' echoes throughout most human-focused cybersecurity research. Some researchers have focused on the problem of understanding who is most vulnerable to poor security practices or online victimization; others have tried to uncover why they are vulnerable. In studying online reputation, we pull insights from both these areas of research to proactively identify the vulnerable users. Assigning reputation scores to users on a network can help administrators identify users in need of interventions that could prevent them from being victimized. Users whose scores reveal they are more likely to be compromised than other users may be in need of additional security training, resources, and support.

Reputation can, of course, be evaluated for many other characteristics and behaviors besides vulnerability or likelihood of compromise. Cyber reputation may also be useful for identifying the users who pose an enhanced threat to a network or platform. This has applications in the realm of insider threat detection: who are the users conducting suspicious activities on the network?; how similar are their behaviors to past insiders?; and so on. Knowing the reputation of a user may also be useful for assessing the content produced by the user. This is crucial in social media settings and can be used to address the most notorious social media problems of the last couple of years: “fake news” and fake accounts. With reputation, we can evaluate how likely the user is to be producing authentic or factual content, given past content from the user or similar users.

Discussed briefly in Chapter 6, another application of evaluating and tracking reputation is to identify hijacked or compromised accounts. Similar to financial credit scores, abrupt changes in a user’s behavior and/or reputation score may indicate someone else is making use of the user’s account. If users are aware of their reputation scores, they can monitor their scores just as they might monitor their financial credit reports.

Making reputation scores known to users may also be a technique to enforce ‘good’ behavior on the network or platform and discourage ‘bad’ behavior. Just as stars for product ratings or likes on Instagram and Facebook encourage quality products and content, reputation scores might motivate users to comply with network rules and policies. This is especially the case if reputation scores are used to make decisions about permissions and access on the network or platform.

Regardless of what cyber reputation is used for, it is crucial that it is based on evidence and can be validated. Security decisions about users should not be arbitrary and subjective; a formalized method for evaluating reputation (or any security metric for that matter) can ensure decisions are rooted in data and any uncertainty in the prediction or decision is transparent.

9.3.3 Unintended Consequences of Quantifying User Cyber Reputation

There are drawbacks to assigning reputation scores to users and creating a centralized reputation system for a network. This section discusses some important considerations related to ethics and security.

9.3.3.1 Ethical Considerations

There are ethical concerns associated with any system designed to track and score people. A system that catalogs vulnerable users or ties user demographics to some behavior can intentionally or unintentionally lead to harmful profiling of certain populations. Additionally, biases in the data or the scoring algorithm can result in reputation scores biased toward or against certain populations.

For example, as discussed previously in this dissertation, research including our own has identified age and gender as demographic factors linked to increased likelihood of compromise. If a network administrator identifies people with certain demographic traits to present enhanced risks to their network, are they simply allowed to restrict or deny access and privileges of these users for the sake of protecting the network?

As another example, in the US the FICO credit score has unintentionally led to high-interest rates and denied loans among specific demographic groups. For example, length of credit history is a key part of a FICO score, meaning that younger people often have trouble qualifying for housing mortgages [157]. Recognizing the discrimination that FICO has unintentionally enabled, in 2019 Fair Isaac Corporation intends to roll out the UltraFICO score, an opt-in score intended to address some of the shortcomings of the current score by incorporating additional factors, such as a person's savings and regular bill payments. UltraFICO will also provide a score for the 15 million consumers who do not have a FICO score because they simply do not have a credit history [158]. In the case of these 15 million users, not having a score was as damaging as having a low score because it indicated there was no knowledge about these users and therefore providing them with a loan was risky. In other words, the absence of a score can still imply reputation. Therefore, it is important to consider not only how a score is assigned but also who receives a score.

Another important consideration is ensuring the system is not abused or used for malicious purposes. Questions that arise include: who has access to the reputation system?; who has the authority to assign scores or make decisions based on scores?; and how are the powers of a reputation system's central authority or a network's administrator kept in check? Take, for instance, China's financial and social credit system, which is intended to prevent dishonesty and encourage socially acceptable behaviors. It is solely up to the government to decide what behaviors are acceptable; the system has already been used to punish people for behaviors deemed undesirable through canceled flights, bans from top educational institutions, and a variety of other repercussions [159].

Shortcomings of existing reputation systems will be key to understand in the development of any cyber reputation system. As with any research or system, benefits to users should outweigh the risks. However, as the FICO example illustrated, it is sometimes difficult to know what the risks and problems are in advance.

9.3.3.2 Security Considerations

There are several important security considerations if a user reputation system is implemented in an operational setting. The primary security consideration centers on protecting user privacy. The data used to develop reputation scores should stay confidential. The central authority and/or reputation system will likely aggregate or access a variety of data sources containing sensitive information about users, to include their real identities, demographics, online activities, and more. If the reputation system is breached or the data is somehow exposed, this could cause serious harm to the users. The reputation scores themselves should also be confidential; access to reputation scores could provide an adversary with a detailed list of vulnerable users on a network. Another important consideration is ensuring the integrity of both the data used to develop reputation scores and the reputation scores themselves; a malicious actor should not be able to alter a user’s score by contributing garbage data or altering the score through some other means. Related to this, the people with access to the reputation system and its data must somehow be vetted and trusted not to share or alter the data.

Overall, the security considerations listed here are challenging to implement and enforce (one need only to turn on the news to learn about the latest breach), but they are important to consider and dedicate time and resources toward before establishing a reputation system. Otherwise, the system may do more harm than good on the network.

9.4 Limitations

This section focuses on the limitations of the reputation evaluation methodology and the two case studies presented in this dissertation.

First, in its current state, the reputation evaluation methodology is not resilient to reputation manipulation. If a user tries to somehow manipulate their score through fake data, altered behaviors, or some other means, there is nothing in the methodology to detect or correct for this. This is particularly concerning if the reputation system is transparent to users and there is some penalty or reward associated with a user's reputation score. This was not an issue for the case studies presented in this dissertation, but will be a concern to address in future work.

On a similar note, the methodology does not validate whether all data feeding a reputation score truly belongs to a particular user. Though we assume there is some central authority such as a network administrator with insight into the connection between a user account and network activity, if the user's account is hijacked or if the user willingly allows another person to access the network with their credentials, the scoring system will still assume this data is associated with the true account owner. In theory, if user activity is tracked and scored over time, abrupt changes in behavior or the reputation score may be used as an indicator that a different user is behind the account. Regardless, this limitation is in need of further exploration in future work.

Another limitation of the methodology is its dependence on data. Insubstantial data or weak evidence will produce scores with high uncertainty. Unfortunately, this limitation is extraordinarily difficult to avoid; garbage in will equal garbage out.

There are also limitations associated with the case studies and their ability to illustrate the reputation evaluation methodology. The first limitation is that the two case studies did not illustrate all the nuances and complexities of the presented methodology. While the network traffic case study showed the full methodology applied to users on a university network, only one source of data was used for reputation evaluation. The reputation evaluation methodology provided guidance for combining evidence from multiple sources of data in order to strengthen confidence in scores and predictions; privacy restrictions on user data prevented us from incorporating other sources of data, such as demographics, into this case study at the time of writing this dissertation.

Lastly, both case studies focused on a similar characteristic of interest that was fairly objective in nature i.e. is the user infected or uninfected, compromised or uncompromised. While this dissertation was focused on applying reputation to a network security setting, the selection of these characteristics does not show how a more subjective characteristic can be handled by the evaluation methodology. More subjective qualities, such as maliciousness or trustworthiness, were discussed in this dissertation but were not fully explored through a case study. As with the other limitations, this will be explored and addressed in future work that is addressed in more detail in the following section.

Note that there are also limitations specific to each case studies' design and results; these were discussed previously in the case study chapters.

9.5 Future Work

This section discusses immediate future work to expand the two case studies and refine the reputation evaluation methodology. This is followed by a discussion of long-term areas of research in the cyber reputation space.

9.5.1 Expanding the University Network Case Study

Chapters 5 and 6 presented a case study at a large, public university in which network traffic was used to quantify the most likely probability that a user becomes infected. Infection implied a traffic session generated by a user's device created a threat alert on the university's Intrusion Detection/Prevention System. Features extracted from network traffic were used as the evidence to make the reputation assessments. The results of any reputation assessment are only as good as the evidence, and in this study, using network traffic alone still left uncertainty in the predictions. As discussed previously in this dissertation, research has tied individual differences in people - such as demographics, personality, and so on - to things such as poor security practices, susceptibility to phishing victimization, the likelihood of malware encounters or infections, and more. So, as a next step to strengthen our reputation predictions, we are working with the Division of Information and Technology and various human resource groups to pair user network traffic with university-collected demographics, to include age, gender, place of birth, status at the university (student, faculty staff), and more. We will examine if the incorporation of these features can reduce the uncertainty in our predictions.

9.5.2 Expanding the Social Media Case Study

Chapters 7 and 8 presented a case study motivated by work in public health that leverages user-reported social media data to study and make predictions about people. We collected user reports about device and account compromise and then extracted various activity-based features and demographics of age and gender from the Twitter data. As in the network traffic case study, we first tested the feasibility of using these features as evidence through a supervised classification problem. Our long-term goal for this study is to apply the full reputation evaluation methodology once we have refined and automated our data collection and cleaning process, expanded our feature set, and addressed reviewer feedback to the conference paper versions of these chapters.

There are several additional research problems nested within the effort to expand the social media case study. This includes human-focused cybersecurity research to develop deeper understandings of the relationship between various human traits and victimization. It also includes some timely social media research problems: conducting bot and fake account detection in order to better filter and clean our dataset and making text-based predictions about users in order extract additional demographics and other human traits from the Twitter data.

9.5.3 Improving the Reputation Evaluation Methodology

Additional future work includes addressing the limitations and determining how the methodology can be used in practice.

First, as discussed in our limitations section, in its current state the proposed methodology does not address the issue of reputation manipulation. For the two case studies presented in this dissertation, this was not necessarily an issue since users were unaware their reputation was being evaluated and therefore could not actively attempt to manipulate any score. However, in a setting where a user knows their score, the methodology must be able to handle the fact that reputation fraud may be occurring. Detecting cheaters in reputation systems has been explored in settings such as e-commerce; this problem was outside the scope of this dissertation but will be key in future work.

Second, the next major area of improvement will be ensuring that the reputation system is fair and does not unintentionally favor or punish certain users. We will explore how to ensure no behavior or trait associated with a user can automatically or substantially hurt or help a score; this may also help prevent reputation manipulation. This is a challenging problem, especially if demographics are used as evidence of a propensity toward certain behaviors or characteristics. This was highlighted previously in the discussion section of this chapter: how do we ensure we are not accidentally profiling users in a harmful manner or using biases (either our own or ones in the data) to make determinations about users?

A third area of improvement is using the absence of data as input to a reputation score. In some cases, the absence of data about a particular user may be just as telling as actual data. For example, consider the problem of bot or fake account detection on social media: a lack of user information or activity on a profile may signal an account is not actually authentic. For certain applications of reputation scoring, looking at the absence of data may help address the system's dependency on 'good' data, as discussed previously in the limitations section.

Another immediate area of future work to improve the methodology is exploring how it can be used in an operational setting. Reputation can obviously change over time, so a system used in practice must be able to dynamically update user reputation scores. What triggers this update? How much importance is newer evidence given over older evidence? Additionally, how can reputation scores (and the uncertainty associated with them) be used to make decisions?

Finally, we will also explore additional case studies. In the two case studies presented here, the focus was primarily on reputation related to victimization. Testing our methodology's applicability to other characteristics may help uncover a need for additional improvements or alterations to the methodology. We will also explore if the methodology generalizes to settings not immediately pertinent to cybersecurity. For example, we may apply the methodology to online social networks to assess the amount of influence or authority an entity has in an online community.

9.5.4 Long Term Cyber Reputation Research

Long-term research will include exploring the applications of user cyber reputation, applying concepts of reputation to other cyber entities, and exploring additional properties of reputation. One potential application is using reputation to combat the spread of misinformation. Misinformation is a major problem in many online communities. Whether it is used to deepen political divides between people or propagate inaccurate health information, it can have devastating real world implications. Just as in the physical world where one might consider a professional’s credentials before trusting their advice or actions, attaching reputation scores to the entity propagating information or the information itself can provide an accessible way to quickly vet the source or quality of information. While techniques for assigning reputation scores to users were explored in this dissertation, assigning reputation scores to information is a different, complicated problem. How is a score assigned to a new piece of information? How does the reputation of the information change as it moves between people and across online platforms? How do slight changes in the wording of the information, redactions, or additions alter the reputation? This touches on the idea of exploring additional properties of reputation. For example, can/how does an entity’s reputation change when interacting with different entities or on different platforms? Is reputation, like trust, transitive, i.e. does entity A’s reputation tell us something about entity B’s reputation if the two are somehow connected? As technology and concepts of identity in cyberspace evolve, it is likely that many new questions and areas of research in this space will arise.

9.6 Conclusion

Reputation has emerged as an important factor in the field of cybersecurity and is used to identify domain names, IP addresses, files, binaries, and users that may threaten or harm a network. However, in order for reputation to be used as a reliable tool in cybersecurity, there is a need for foundational research that transforms reputation quantification and prediction into a science. This dissertation presented a mathematical interpretation of user reputation in a network security context, a methodology to evaluate user cyber reputation, and case studies to illustrate the proposed mathematical definition and scoring methodology. From a broader perspective, this dissertation contributed to the science of security, proposed new techniques to understand and evaluate users, and emphasized the importance of human-focused cybersecurity research.

The Internet is a hostile landscape. Online fraud, identity theft, bots and fake accounts, and digital influence operations are just a few of the many threats to cybersecurity, user privacy, and information integrity. The need for reputation systems and techniques to evaluate and validate online identities will likely continue to grow; this dissertation represents an initial exploration of this complicated space.

Appendices

Appendix A

This appendix presents additional tables from Chapter 5. Values have been rounded to the third decimal place.

30:70 infected to uninfected ratio; 36 original features								
Classifier	**80/20 Train/Test Split**				**70/30 Train/Test Split**			
	Accuracy	***FPR***	***FNR***	***ROC AUC***	***Accuracy***	***FPR***	***FNR***	***ROC AUC***
KNeighbors	0.798	0.202	0.202	0.87	0.79	0.203	0.225	0.858
Logistic Regression	0.793	0.206	0.21	0.86	0.779	0.221	0.222	0.853
Random Forest	0.793	0.205	0.21	0.848	0.783	0.217	0.218	0.836
Ada Boost	0.786	0.211	0.22	0.858	0.781	0.217	0.225	0.846
Gradient Boost	0.801	0.199	0.199	0.868	0.784	0.216	0.218	0.858
Extra Tree	0.787	0.186	0.275	0.848	0.794	0.156	0.326	0.842
Bagging	0.795	0.203	0.21	0.861	0.784	0.215	0.22	0.856
Voting	0.79	0.21	0.21	0.854	0.781	0.219	0.22	0.843
Neural Network 1	0.713	0.286	0.29	0.773	0.685	0.315	0.316	0.738
Neural Network 2	0.79	0.21	0.21	0.864	0.785	0.215	0.216	0.854

Table 27: Supervised classification results for the 30:70 ratio using the 36 original features for classification

30:70 infected to uninfected ratio; principal components								
Classifier	**80/20 Train/Test Split**				**70/30 Train/Test Split**			
	Accuracy	***FPR***	***FNR***	***ROC AUC***	***Accuracy***	***FPR***	***FNR***	***ROC AUC***
KNeighbors	0.787	0.212	0.215	0.859	0.781	0.219	0.22	0.852
Logistic Regression	0.787	0.212	0.215	0.858	0.782	0.218	0.22	0.853
Random Forest	0.796	0.203	0.207	0.834	0.782	0.216	0.222	0.827
Ada Boost	0.791	0.208	0.212	0.861	0.785	0.213	0.22	0.853
Gradient Boost	0.795	0.204	0.207	0.864	0.785	0.215	0.216	0.856
Extra Tree	0.806	0.166	0.259	0.853	0.796	0.165	0.298	0.844
Bagging	0.795	0.204	0.207	0.86	0.781	0.217	0.225	0.855
Voting	0.785	0.215	0.215	0.849	0.778	0.222	0.22	0.842
Neural Network 1	0.776	0.223	0.225	0.848	0.753	0.247	0.246	0.831
Neural Network 2	0.795	0.204	0.207	0.863	0.787	0.213	0.215	0.851

Table 28: Supervised classification results for the 30:70 ratio using the principal components for classification

50:50 infected to uninfected ratio; 36 original features								
Classifier	**80/20 Train/Test Split**				**70/30 Train/Test Split**			
	Accuracy	***FPR***	***FNR***	***ROC AUC***	***Accuracy***	***FPR***	***FNR***	***ROC AUC***
KNeighbors	0.788	0.208	0.216	0.857	0.776	0.223	0.225	0.847
Logistic Regression	0.778	0.221	0.223	0.857	0.763	0.237	0.238	0.839
Random Forest	0.778	0.218	0.226	0.852	0.769	0.23	0.231	0.837
Ada Boost	0.768	0.164	0.301	0.856	0.767	0.183	0.282	0.842
Gradient Boost	0.782	0.216	0.221	0.868	0.782	0.216	0.219	0.853
Extra Tree	0.5	0.5	0.5	0.775	0.497	0.497	0.503	0.764
Bagging	0.79	0.203	0.218	0.865	0.776	0.222	0.225	0.85
Voting	0.783	0.216	0.218	0.861	0.774	0.229	0.224	0.847
Neural Network 1	0.757	0.242	0.244	0.827	0.739	0.26	0.262	0.795
Neural Network 2	0.796	0.203	0.205	0.868	0.779	0.22	0.222	0.847

Table 29: Supervised classification results for the 50:50 ratio using the 36 original features for classification

50:50 infected to uninfected ratio; principal components								
Classifier	**80/20 Train/Test Split**				**70/30 Train/Test Split**			
	Accuracy	*FPR*	*FNR*	*ROC AUC*	*Accuracy*	*FPR*	*FNR*	*ROC AUC*
KNeighbors	0.786	0.216	0.213	0.859	0.775	0.225	0.225	0.84
Logistic Regression	0.781	0.218	0.221	0.858	0.766	0.234	0.234	0.845
Random Forest	0.791	0.205	0.213	0.835	0.773	0.225	0.229	0.819
Ada Boost	0.791	0.208	0.21	0.865	0.782	0.216	0.219	0.843
Gradient Boost	0.796	0.203	0.205	0.865	0.786	0.213	0.215	0.848
Extra Tree	0.788	0.197	0.226	0.85	0.769	0.159	0.303	0.83
Bagging	0.795	0.203	0.208	0.859	0.776	0.215	0.232	0.846
Voting	0.773	0.226	0.229	0.845	0.763	0.236	0.239	0.836
Neural Network 1	0.786	0.213	0.216	0.85	0.734	0.264	0.269	0.811
Neural Network 2	0.797	0.2	0.205	0.864	0.771	0.229	0.229	0.842

Table 30: Supervised classification results for the 50:50 ratio using the principal components for classification

Appendix B

This appendix presents additional tables from Chapter 8. Values have been rounded to the third decimal place.

Ada Boost, 30:70 Case to Control Ratio Results,			
80/20 Train/Test Split		**70/30 Train/Test Split**	
Feature	***Importance***	***Feature***	***Importance***
Average number of words	0.18	Account age in months	0.43
Favorite count	0.16	Tweet count	0.14
Average number of links	0.14	Average number of links	0.13
Tweet count	0.12	Favorite count	0.11
Average number of mentions	0.1	Average number of words	0.07
Followers count	0.08	Followers count	0.05
Account age in months	0.08	Average number of mentions	0.04
Age	0.08	Age	0.03
Friends count	0.02	Friends count	0.01
Listed count	0.02	Gender	0.01
Gender	0.02	Listed count	0.00

Table 31: Feature importance values for the Ada Boost classifier for the 30:70 ratio

Ada Boost, 50:50 Case to Control Ratio Results,			
80/20 Train/Test Split		**70/30 Train/Test Split**	
Feature	***Importance***	***Feature***	***Importance***
Account age in months	0.625	Account age in months	0.526
Tweet count	0.121	Average number of words	0.138
Favorite count	0.094	Tweet count	0.111
Average number of links	0.064	Favorite count	0.107
Average number of words	0.051	Average number of links	0.071
Average number of mentions	0.022	Average number of mentions	0.021
Followers count	0.022	Followers count	0.015
Friends count	0.0	Gender	0.005
Listed count	0.0	Age	0.005
Age	0.0	Friends count	0.0
Listed count	0.0	Listed count	0.0

Table 32: Feature importance values for the Ada Boost classifier for the 50:50 ratio

Extra Tree, 30:70 Case to Control Ratio Results,			
80/20 Train/Test Split		**70/30 Train/Test Split**	
Feature	***Importance***	***Feature***	***Importance***
Account age in months	0.33	Account age in months	0.48
Favorite count	0.22	Favorite count	0.22
Followers count	0.13	Tweet count	0.09
Tweet count	0.11	Average number of links	0.08
Average number of words	0.07	Average number of words	0.04
Average number of links	0.06	Average number of mentions	0.02
Average number of mentions	0.03	Listed count	0.02
Age	0.02	Age	0.02
Friends count	0.02	Followers count	0.02
Listed count	0.01	Friends count	0.01
Gender	0.0	Gender	0.00

Table 33: Feature importance values for the Extra Tree classifier for the 30:70 ratio

Extra Tree, 50:50 Case to Control Ratio Results,			
80/20 Train/Test Split		**70/30 Train/Test Split**	
Feature	***Importance***	***Feature***	***Importance***
Account age in months	0.44	Account age in months	0.406
Favorite count	0.267	Favorite count	0.266
Tweet count	0.103	Tweet count	0.117
Average number of words	0.072	Average number of words	0.088
Average number of links	0.037	Followers count	0.03
Listed count	0.024	Average number of links	0.028
Age	0.018	Average number of mentions	0.021
Average number of mentions	0.017	Age	0.017
Gender	0.009	Listed count	0.016
Friends count	0.006	Friends count	0.011
Followers count	0.006	Gender	0.0

Table 34: Feature importance values for the Extra Tree classifier for the 50:50 ratio

Bibliography

[1] "How Credit History Impacts Your Credit Score." *MyFICO*, Fair Isaac Corporation, 2017, www.myfico.com/credit-education/whats-in-your-credit-score/.

[2] "About RepTrak®." *Reputation Institute*, Reputation Institute, 2017, www.reputationinstitute.com/reptrak-framework.aspx.

[3] Luca, Michael. "Reviews, reputation, and revenue: The case of Yelp. com." (2016).

[4] Jøsang, Audun, and Roslan Ismail. "The beta reputation system." *Proceedings of the 15th bled electronic commerce conference*. Vol. 5. 2002.

[5] Xiong, Li, and Ling Liu. "A reputation-based trust model for peer-to-peer e-commerce communities." *E-Commerce, 2003. CEC 2003. IEEE International Conference on*. IEEE, 2003.

[6]Golbeck, Jennifer, and James A. Hendler. "Reputation Network Analysis for Email Filtering." *CEAS*. 2004.

[7] Taylor, Bradley. "Sender Reputation in a Large Webmail Service." *CEAS*. Vol. 27. 2006.

[8] Esquivel, Holly, Aditya Akella, and Tatsuya Mori. "On the effectiveness of IP reputation for spam filtering." *Communication Systems and Networks (COMSNETS), 2010 Second International Conference on*. IEEE, 2010.

[9] Mishsky, Igor, Nurit Gal-Oz, and Ehud Gudes. "A topology based flow model for computing domain reputation." *IFIP Annual Conference on Data and Applications Security and Privacy*. Springer, Cham, 2015.

[10] Sharifnya, Reza, and Mahdi Abadi. "A novel reputation system to detect DGA-based botnets." *Computer and Knowledge Engineering (ICCKE), 2013 3th International eConference on*. IEEE, 2013.

[11] Fukushima, Yoshiro, Yoshiaki Hori, and Kouichi Sakurai. "Proactive blacklisting for malicious web sites by reputation evaluation based on domain and ip address registration." *Trust, Security and Privacy in Computing and Communications (TrustCom), 2011 IEEE 10th International Conference on*. IEEE, 2011.

[12] Huang, Yonghong, and Paula Greve. "Large scale graph mining for web reputation inference." *Machine Learning for Signal Processing (MLSP), 2015 IEEE 25th International Workshop on*. IEEE, 2015.

[13] Felegyhazi, Mark, Christian Kreibich, and Vern Paxson. "On the Potential of Proactive Domain Blacklisting." *LEET* 10 (2010): 6-6.

[14] Bilge, Leyla, et al. "EXPOSURE: Finding Malicious Domains Using Passive DNS Analysis." *Ndss*. 2011.

[15] Antonakakis, Manos, et al. "Building a Dynamic Reputation System for DNS." *USENIX security symposium*. 2010.

[16] Moura, Giovane, Ramin Sadre, and Aiko Pras. "Bad neighborhoods on the internet." *IEEE Communications Magazine* 52.7 (2014): 132-139.

[17] Porenta, Jernej, and Mojca Ciglarič. "Empirical comparison of ip reputation databases." *Proceedings of the 8th Annual Collaboration, Electronic messaging, Anti-Abuse and Spam Conference*. ACM, 2011.

[18] Sinha, Sushant, Michael Bailey, and Farnam Jahanian. "Shades of Grey: On the effectiveness of reputation-based "blacklists"." *Malicious and Unwanted Software, 2008. MALWARE 2008. 3rd International Conference on*. IEEE, 2008.

[19] Esquivel, Holly, Aditya Akella, and Tatsuya Mori. "On the effectiveness of IP reputation for spam filtering." *Communication Systems and Networks (COMSNETS), 2010 Second International Conference on*. IEEE, 2010.

[20] Kamvar, Sepandar D., Mario T. Schlosser, and Hector Garcia-Molina. "The eigentrust algorithm for reputation management in p2p networks." *Proceedings of the 12th international conference on World Wide Web*. ACM, 2003.

[21] Chen, Yizheng, et al. "Measuring Network Reputation in the Ad-Bidding Process." *International Conference on Detection of Intrusions and Malware, and Vulnerability Assessment*. Springer, Cham, 2017.

[22] Walsh, Kevin, and Emin Gün Sirer. "Fighting peer-to-peer spam and decoys with object reputation." *Proceedings of the 2005 ACM SIGCOMM workshop on Economics of peer-to-peer systems*. ACM, 2005.

[23] Bilge, Leyla, and Tudor Dumitras. "Before we knew it: an empirical study of zero-day attacks in the real world." *Proceedings of the 2012 ACM conference on Computer and communications security*. ACM, 2012.

[24] Movshovitz-Attias, Dana, et al. "Analysis of the reputation system and user contributions on a question answering website: Stackoverflow." *Proceedings of the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*. ACM, 2013.

[25] Wang, Alex Hai. "Don't follow me: Spam detection in twitter." *Security and Cryptography (SECRYPT), Proceedings of the 2010 International Conference on*. IEEE, 2010.

[26] Golbeck, Jennifer, and James Hendler. "Accuracy of metrics for inferring trust and reputation in semantic web-based social networks." *Engineering knowledge in the age of the semantic web* (2004): 116-131.

[27] Josang, Audun, and Jochen Haller. "Dirichlet reputation systems." *Availability, Reliability and Security, 2007. ARES 2007. The Second International Conference on*. IEEE, 2007.

[28] “Science of Security.” *Science of Security*, National Security Agency, 21 June 2016, www.nsa.gov/what-we-do/research/science-of-security/

[29] Marti, Sergio, and Hector Garcia-Molina. "Limited reputation sharing in P2P systems." *Proceedings of the 5th ACM conference on Electronic commerce*. ACM, 2004.

[30] Marti, Sergio, and Hector Garcia-Molina. "Taxonomy of trust: Categorizing P2P reputation systems." *Computer Networks*50.4 (2006): 472-484.

[31] Oro, David, et al. "Benchmarking IP blacklists for financial botnet detection." *Information Assurance and Security (IAS), 2010 Sixth International Conference on*. IEEE, 2010.

[32] Ma, Justin, et al. "Beyond blacklists: learning to detect malicious web sites from suspicious URLs." *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2009.

[33] Chen, Kang, et al. "A social network based reputation system for cooperative P2P file sharing." *IEEE Transactions on Parallel and Distributed Systems* 26.8 (2015): 2140-2153.

[34] Begriche, Youcef, et al. "Bayesian-based model for a reputation system in vehicular networks." *Cyber Security of Smart Cities, Industrial Control System and Communications (SSIC), 2015 International Conference on*. IEEE, 2015.

[35] Golbeck, Jennifer. "Trust on the world wide web: a survey." *Foundations and Trends® in Web Science* 1.2 (2008): 131-197.

[36] Sherchan, Wanita, Surya Nepal, and Cecile Paris. "A survey of trust in social networks." *ACM Computing Surveys (CSUR)*45.4 (2013): 47.

[37] Jøsang, Audun, Roslan Ismail, and Colin Boyd. "A survey of trust and reputation systems for online service provision." *Decision support systems* 43.2 (2007): 618-644.

[38] Kim, Young Ae, et al. "Building a web of trust without explicit trust ratings." *Data Engineering Workshop, 2008. ICDEW 2008. IEEE 24th International Conference on*. IEEE, 2008.

[39] Golbeck, Jennifer, and James Hendler. "Inferring reputation on the semantic web." *Proceedings of the 13th International World Wide Web Conference*. Vol. 316. 2004.

[40] *Stack Overflow - Where Developers Learn, Share, & Build Careers*, Stack Exchange Inc., 2017, stackoverflow.com/.

[41] Kamvar, Sepandar D., Mario T. Schlosser, and Hector Garcia-Molina. "The eigentrust algorithm for reputation management in p2p networks." *Proceedings of the 12th international conference on World Wide Web*. ACM, 2003.

[42] Gilbert, Peter, et al. "Toward trustworthy mobile sensing." *Proceedings of the Eleventh Workshop on Mobile Computing Systems & Applications*. ACM, 2010.

[43] Hirsch, Jorge E. "An index to quantify an individual's scientific research output." *Proceedings of the National academy of Sciences of the United States of America* 102.46 (2005): 16569.

[44] Page, Lawrence, et al. *The PageRank citation ranking: Bringing order to the web*. Stanford InfoLab, 1999.

[45] Viswanath, Bimal, et al. "Strength in numbers: Robust tamper detection in crowd computations." *Proceedings of the 2015 ACM on Conference on Online Social Networks*. ACM, 2015.

[46] Molavi Kakhki, Arash, Chloe Kliman-Silver, and Alan Mislove. "Iolaus: Securing online content rating systems." *Proceedings of the 22nd international conference on World Wide Web*. ACM, 2013.

[47] Günnemann, Nikou, Stephan Günnemann, and Christos Faloutsos. "Robust multivariate autoregression for anomaly detection in dynamic product ratings." *Proceedings of the 23rd international conference on World wide web*. ACM, 2014.

[48] Sandulescu, Vlad, and Martin Ester. "Detecting singleton review spammers using semantic similarity." *Proceedings of the 24th international conference on World Wide Web*. ACM, 2015.

[49] "URL / IP Lookup." *URL/IP Lookup | Webroot BrightCloud*, Webroot Inc. , 2017, www.brightcloud.com/tools/url-ip-lookup.php.

[50] "How Safe Is Your Web Destination?" *Zscaler | Zulu - URL Risk Analyzer*, Zscaler, Inc., 2017, zulu.zscaler.com/.

[51] Pouryazdan, Maryam, et al. "Quantifying User Reputation Scores, Data Trustworthiness, and User Incentives in Mobile Crowd-Sensing." *IEEE Access* 5 (2017): 1382-1397.

[52] Buchegger, Sonja, and Jean-Yves Le Boudec. *A robust reputation system for mobile ad-hoc networks*. No. LCA-REPORT-2003-006. 2003.

[53] Friedman, Jerome, Trevor Hastie, and Robert Tibshirani. *The elements of statistical learning*. Vol. 1. New York: Springer series in statistics, 2001.

[54] Stamatelatos, Michael, et al. "Probabilistic risk assessment procedures guide for NASA managers and practitioners." (2011).

[55] Spring, Jonathan. "Avoiding Pseudoscience in the Science of Security: A Case Study of Malware Indicator Analysis." IFIP Working Group 10.4: Dependable Computing and Fault Tolerance. Workshop on the Science of Cyber-Security, Jan. 2014, Tortworth, Bristol, UK, webhost.laas.fr/TSF/IFIPWG/Workshops&Meetings/67/Workshop-regularPapers/Spring-avoiding_pseudoscience_flocon.pdf.

[56] Maxion, Roy. "Making experiments dependable." *Dependable and Historic Computing* (2011): 344-357.

[57] Hatleback, Eric, and Jonathan M. Spring. "Exploring a mechanistic approach to experimentation in computing." *Philosophy & Technology* 27.3 (2014): 441-459.

[58] Rossow, Christian, et al. "Prudent practices for designing malware experiments: Status quo and outlook." *Security and Privacy (SP), 2012 IEEE Symposium on*. IEEE, 2012.

[59] *Facebook - Connect with Friends and the World Around You On Facebook,* Facebook, 2017, www.facebook.com/.

[60] *Twitter - It's What's Happening,* Twitter, 2017, www.twitter.com/.

[61] *Reddit: The Front Page of the Internet,* Reddit, Inc., 2017, www.reddit.com/.

[62] *Quora - A Place to Share Knowledge and Better Understand the World,* Quora, Inc. 2017, www.quora.com/.

[63] *GitHub - The World's Leading Software Development Platform*, GitHub, 2017, www.github.com.

[64] Egelman, Serge, and Eyal Peer. "Scaling the security wall: Developing a security behavior intentions scale (sebis)." *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*. ACM, 2015.

[65] John, Oliver P., and Sanjay Srivastava. "The Big Five trait taxonomy: History, measurement, and theoretical perspectives." *Handbook of personality: Theory and research* 2.1999 (1999): 102-138.

[66] Arnett, Jeffrey Jensen. "Sensation seeking, aggressiveness, and adolescent reckless behavior." *Personality and individual differences* 20.6 (1996): 693-702.

[67] Thunholm, Peter. "Decision-making style: habit, style or both?." *Personality and individual differences* 36.4 (2004): 931-944.

[68] Scott, Susanne G., and Reginald A. Bruce. "Decision-making style: The development and assessment of a new measure." *Educational and psychological measurement* 55.5 (1995): 818-831.

[69] Jolliffe, Ian T. "Principal Component Analysis and Factor Analysis." *Principal component analysis*. Springer New York, 1986. 115-128.

[70] "Monte Carlo Simulation." *Palisade*, Palisade Corporation, 2017, www.palisade.com/risk/monte_carlo_simulation.asp.

[71] Verde, Nino Vincenzo, et al. "No NAT'd user left behind: Fingerprinting users behind NAT from NetFlow records alone." *Distributed Computing Systems (ICDCS), 2014 IEEE 34th International Conference on*. IEEE, 2014.

[72] Alotibi, Gaseb, et al. "Behavioral-based feature abstraction from network traffic." *Iccws 2015-The Proceedings of the 10th International Conference on Cyber Warfare and Security: ICCWS2015*. Academic Conferences Limited, 2015.

[73] Pang, Jeffrey, et al. "802.11 user fingerprinting." *Proceedings of the 13th annual ACM international conference on Mobile computing and networking*. ACM, 2007.

[74] Bhuyan, Monowar H., Dhruba Kumar Bhattacharyya, and Jugal K. Kalita. "Network anomaly detection: methods, systems and tools." *Ieee communications surveys & tutorials*16.1 (2014): 303-336.

[75] Mahoney, Matthew V. "Network traffic anomaly detection based on packet bytes." *Proceedings of the 2003 ACM symposium on Applied computing*. ACM, 2003.

[76] Hammerschmidt, Christian, et al. "Efficient learning of communication profiles from IP flow records." *Local Computer Networks (LCN), 2016 IEEE 41st Conference on*. IEEE, 2016.

[77] Hammerschmidt, Christian, et al. "Behavioral clustering of non-stationary IP flow record data." *Network and Service Management (CNSM), 2016 12th International Conference on*. IEEE, 2016.

[78] Evangelou, Marina, and Niall M. Adams. "Predictability of NetFlow data." *Intelligence and Security Informatics (ISI), 2016 IEEE Conference on*. IEEE, 2016.

[79] Münz, Gerhard, Sa Li, and Georg Carle. "Traffic anomaly detection using k-means clustering." *GI/ITG Workshop MMBnet*. 2007.

[80] Tjhai, Gina C., et al. "A preliminary two-stage alarm correlation and filtering system using SOM neural network and K-means algorithm." *Computers & Security* 29.6 (2010): 712-723.

[81] Alotibi, Gaseb, et al. "User profiling from network traffic via novel application-level interactions." *Internet Technology and Secured Transactions (ICITST), 2016 11th International Conference for*. IEEE, 2016.

[82] Clarke, Nathan, Fudong Li, and Steven Furnell. "A novel privacy preserving user identification approach for network traffic." *computers & security* 70 (2017): 335-350.

[83] Brown, Anthony, Richard Mortier, and Tom Rodden. "An exploration of user recognition on domestic networks using NetFlow records." *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct Publication*. ACM, 2014.

[84] Tomsu, Radek, Samuel Marchal, and N. Asokan. "Profiling users by modeling web transactions." *Proceedings of the 2017 IEEE 37th International Conference on Distributed Computing Systems* (ICDCS), IEEE, 2017.

[85] "Syslog Field Descriptions." *Palo Alto Networks*, Palo Alto Networks, 2017, www.paloaltonetworks.com/documentation/61/pan-os/pan-os/reports-and-logging/syslog-field-descriptions#_41809.

[86] "Pandas Data Analysis Library." *Pandas*, Num Focus , 2017, pandas.pydata.org/.

[87] "NumPy." *NumPy*, NumPy developers, 2017, numpy.org.

[88] "Alternative: Standardize the Variables." *STAT 505: Applied Multivariate Statistical Analysis,* The Pennsylvania State University, 2017, https://onlinecourses.science.psu.edu/stat505/node/55

[89] "Decomposing signals in components (matrix factorization problems)," *scikit learn,* scikit-learn developers, 2017, http://scikit-learn.org/stable/modules/decomposition.html#pca

[90] "Interpret the Key Results for Principal Component Analysis," *Minitab 18 Support,* Minitab Inc. 2017, https://support.minitab.com/en-us/minitab/18/help-and-how-to/modeling-statistics/multivariate/how-to/principal-components/interpret-the-results/key-results

[91] "Error Sum of Squares (SSE)," Stanford University, https://hlab.stanford.edu/brian/error_sum_of_squares.html.

[92] "sklearn.neighbors.KNeighborsClassifier", *scikit learn,* scikit-learn developers, 2017, http://scikit-learn.org/stable/modules/generated/sklearn.neighbors.KNeighborsClassifier.html.

[93] "sklearn.linear_model.LogisticRegression", *scikit learn,* scikit-learn developers, 2017, http://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html

[94] "sklearn.ensemble.RandomForestClassifier", *scikit learn,* scikit-learn developers, 2017, http://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html

[95] "sklearn.ensemble.AdaBoostClassifier", *scikit learn,* scikit-learn developers, 2017, http://scikit-learn.org/stable/modules/generated/sklearn.ensemble.AdaBoostClassifier.html

[96] "sklearn.ensemble.GradientBoostingClassifier", *scikit learn,* scikit-learn developers, 2017, http://scikit-learn.org/stable/modules/generated/sklearn.ensemble.GradientBoostingClassifier.html

[97] "sklearn.tree.ExtraTreeClassifier", *scikit learn,* scikit-learn developers, 2017,http://scikit-learn.org/stable/modules/generated/sklearn.tree.ExtraTreeClassifier.html

[98] "sklearn.ensemble.BaggingClassifier", *scikit learn,* scikit-learn developers, 2017, http://scikit-learn.org/stable/modules/generated/sklearn.ensemble.BaggingClassifier.html

[99] "sklearn.ensemble.VotingClassifier", *scikit learn,* scikit-learn developers, 2017, http://scikit-learn.org/stable/modules/generated/sklearn.ensemble.VotingClassifier.html

[100] "sklearn.model_selection.GridSearchCV," *scikit learn,* scikit-learn developers, 2017, http://scikit-learn.org/stable/modules/generated/sklearn.model_selection.GridSearchCV.html

[101] "sklearn.model_selection.StratifiedKFold," *scikit learn,* scikit-learn developers, 2017, http://scikit-learn.org/stable/modules/generated/sklearn.model_selection.StratifiedKFold.html

[102] "Keras: The Python Deep Learning library,"*Keras Documentation,* https://keras.io/.

[103] Breiman, Leo, et al. *Classification and regression trees*. CRC press, 1984.

[104] "Glossary of Forecast Verification Metrics." National Oceanic and Atmospheric Administration.

[105] Hvistendahl, Mara. "In China, a Three-Digit Score Could Dictate Your Place in Society." *Wired*, Conde Nast, 5 Dec. 2018, www.wired.com/story/age-of-social-credit/.

[106] Dwoskin, Elizabeth. "Facebook Is Rating the Trustworthiness of Its Users on a Scale from Zero to 1." *The Washington Post*, WP Company, 21 Aug. 2018, www.washingtonpost.com/technology/2018/08/21/facebook-is-rating-trustworthiness-its-users-scale-zero-one/?utm_term=.28169f913712.

[107] Confessore, Nicholas and Gabriel J.X. Dance. "Battling Fake Accounts, Twitter to Slash Millions of Followers." *The New York Times,* The New York Times Company, 11 July 2018, https://www.nytimes.com/2018/07/11/technology/twitter-fake-followers.html.

[108] Sheng, Steve, et al. "Who Falls for Phish?" *Proceedings of the 28th International Conference on Human Factors in Computing Systems - CHI '10*, 2010, doi:10.1145/1753326.1753383.

[109] Parrish Jr, James L., Janet L. Bailey, and James F. Courtney. "A personality based model for determining susceptibility to phishing attacks." *Little Rock: University of Arkansas* (2009): 285-296.

[110] Whitty, Monica, et al. "Individual differences in cyber security behaviors: an examination of who is sharing passwords." *Cyberpsychology, Behavior, and Social Networking* 18.1 (2015): 3-7.

[111] Lévesque, Fanny Lalonde, José M. Fernandez, and Dennis Batchelder. "Age and gender as independent risk factors for malware victimisation." *Proceedings of the 31st British Computer Society Human Computer Interaction Conference*. BCS Learning & Development Ltd., 2017.

[112] Lalonde Levesque, Fanny, et al. "A clinical study of risk factors related to malware infections." *Proceedings of the 2013 ACM SIGSAC conference on Computer & communications security*. ACM, 2013.

[113] Gratian, Margaret, et al. "Correlating human traits and cyber security behavior intentions." *computers & security* 73 (2018): 345-358.

[114] Friedman, Jerome, Trevor Hastie, and Robert Tibshirani. *The elements of statistical learning*. Vol. 1. No. 10. New York, NY, USA:: Springer series in statistics, 2001.

[115] Canali, Davide, Leyla Bilge, and Davide Balzarotti. "On the effectiveness of risk prediction based on users browsing behavior." *Proceedings of the 9th ACM symposium on Information, computer and communications security*. ACM, 2014.

[116] Herrmann, Jeffrey W. *Engineering decision making and risk management*. John Wiley & Sons, 2015.

[117] "Interpret the Key Results for Principal Components Analysis." *Minitab 18 Support*, Minitab Inc, 2017, support.minitab.com/en-us/minitab/18/help-and-how-to/modeling-statistics/multivariate/how-to/principal-components/interpret-the-results/key-results.

[118] Achrekar, Harshavardhan, et al. "Predicting flu trends using twitter data." *Computer Communications Workshops (INFOCOM WKSHPS), 2011 IEEE Conference on*. IEEE, 2011.

[119] Signorini, Alessio, Alberto Maria Segre, and Philip M. Polgreen. "The use of Twitter to track levels of disease activity and public concern in the US during the influenza A H1N1 pandemic." *PloS one* 6.5 (2011): e19467.

[120] Coppersmith, Glen, Mark Dredze, and Craig Harman. "Quantifying mental health signals in Twitter." *Proceedings of the Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*. 2014.

[121] Heaivilin, N., et al. "Public health surveillance of dental pain via Twitter." *Journal of dental research* 90.9 (2011): 1047-1051.

[122] Paul, Michael J., and Mark Dredze. "You are what you Tweet: Analyzing Twitter for public health." *Icwsm* 20 (2011): 265-272.

[123] Golbeck, Jennifer. "Predicting Alcoholism Recovery from Twitter." *International Conference on Social Computing, Behavioral-Cultural Modeling and Prediction and Behavior Representation in Modeling and Simulation*. Springer, Cham, 2018.

[124] Sabottke, Carl, Octavian Suciu, and Tudor Dumitras. "Vulnerability Disclosure in the Age of Social Media: Exploiting Twitter for Predicting Real-World Exploits." *USENIX Security Symposium*. 2015.

[125] Mittal, Sudip, et al. "Cybertwitter: Using twitter to generate alerts for cybersecurity threats and vulnerabilities." *Proceedings of the 2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*. IEEE Press, 2016.

[126] Lee, Sangho, and Jong Kim. "Warningbird: A near real-time detection system for suspicious urls in twitter stream." *IEEE transactions on dependable and secure computing* 10.3 (2013): 183-195.

[127] Aggarwal, Anupama, Ashwin Rajadesingan, and Ponnurangam Kumaraguru. "PhishAri: Automatic realtime phishing detection on twitter." *2012 eCrime Researchers Summit*. IEEE, 2012.

[128] Wang, De, et al. "Click traffic analysis of short url spam on twitter." *Collaborative Computing: Networking, Applications and Worksharing (Collaboratecom), 2013 9th International Conference Conference on*. IEEE, 2013.

[129] Grier, Chris, et al. "@ spam: the underground on 140 characters or less." *Proceedings of the 17th ACM conference on Computer and communications security*. ACM, 2010.

[130] Wang, Alex Hai. "Don't follow me: Spam detection in twitter." *Security and cryptography (SECRYPT), proceedings of the 2010 international conference on*. IEEE, 2010.

[131] Chu, Zi, et al. "Who is tweeting on Twitter: human, bot, or cyborg?." *Proceedings of the 26th annual computer security applications conference*. ACM, 2010.

[132] Thomas, Kurt, et al. "Trafficking Fraudulent Accounts: The Role of the Underground Market in Twitter Spam and Abuse." *USENIX Security Symposium*. 2013.

[133] Zangerle, Eva, and Günther Specht. "Sorry, I was hacked: a classification of compromised twitter accounts." *Proceedings of the 29th annual acm symposium on applied computing*. ACM, 2014.

[134] Foley, Geraldine, and Virpi Timonen. "Using grounded theory method to capture and analyze health care experiences." *Health services research* 50.4 (2015): 1195-1210.

[135] Bort, Julie. "50 Startups That Will Boom in 2018, According to VCs." *Business Insider*, Business Insider, 25 Nov. 2017, www.businessinsider.com/50-startups-to-boom-in-2018-according-to-vcs-2017-11.

[136] Lauritsen, Janet L., and Kenna F. Davis Quinet. "Repeat victimization among adolescents and young adults." *Journal of Quantitative Criminology* 11.2 (1995): 143-166.

[137] Polvi, Natalie, et al. "The time course of repeat burglary victimization." *The British Journal of Criminology* 31.4 (1991): 411-414.

[138] "PlayerUp Accounts Marketplace. Player 2 Player Secure Platform." *PlayerUp Accounts Marketplace. Player 2 Player Secure Platform.*, www.playerup.com/.

[139] “Fortnite Accounts for Sale - Stacked OG Account Marketplace.” *How to Sell Your Game Assets Fast for Cash | PlayerAuctions*, www.playerauctions.com/fortnite-account/.

[140] Darwish, Ali, Ahmed El Zarka, and Fadi Aloul. "Towards understanding phishing victims' profile." *Computer Systems and Industrial Informatics (ICCSII), 2012 International Conference on*. IEEE, 2012.

[141] Mohebzada, Jamshaid G., et al. "Phishing in a university community: Two large scale phishing experiments." *2012 International Conference on Innovations in Information Technology (IIT)*. IEEE, 2012.

[142] Eichstaedt, Johannes C., et al. "Psychological language on Twitter predicts county-level heart disease mortality." *Psychological science* 26.2 (2015): 159-169.

[143] Sax, Linda J., Shannon K. Gilmartin, and Alyssa N. Bryant. "Assessing response rates and nonresponse bias in web and paper surveys." *Research in higher education* 44.4 (2003): 409-432.

[144] “Demographics of Social Media Users and Adoption in the United States.” *Pew Research Center: Internet, Science & Tech*, Pew Research Center, 5 Feb. 2018, www.pewinternet.org/fact-sheet/social-media/.

[145] York, Alex. “Social Media Demographics to Inform a Better Segmentation Strategy.” *Sprout Social*, Sprout Social Inc, 7 June 2018, sproutsocial.com/insights/new-social-media-demographics/#twitter.

[146] Granville, Kevin. “Facebook and Cambridge Analytica: What You Need to Know as Fallout Widens.” *The New York Times*, The New York Times Company, 19 Mar. 2018, www.nytimes.com/2018/03/19/technology/facebook-cambridge-analytica-explained.html.

[147] Varol, Onur, et al. "Online human-bot interactions: Detection, estimation, and characterization." *arXiv preprint arXiv:1703.03107* (2017).

[148] Lévesque, Fanny Lalonde, José M. Fernandez, and Anil Somayaji. "Risk prediction of malware victimization based on user behavior." *Malicious and Unwanted Software: The Americas (MALWARE), 2014 9th International Conference on*. IEEE, 2014.

[149] Sap, Maarten, et al. "Developing age and gender predictive lexica over social media." *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 2014.

[150] “User Object - Twitter Developers.” *Twitter*, Twitter, developer.twitter.com/en/docs/tweets/data-dictionary/overview/user-object.

[151] “Twitter: Number of Active Users 2010-2018.” *Statista*, Statista, www.statista.com/statistics/282087/number-of-monthly-active-twitter-users/.

[152] “Lexica.” *Penn World Well Being Project*, Penn Positive Psychology Center , www.wwbp.org/lexica.html.

[153] “Global Twitter User Distribution by Gender 2018 | Statistic.” *Statista*, Statista, www.statista.com/statistics/828092/distribution-of-users-on-twitter-worldwide-gender/.

[154] “Demographics of Social Media Users and Adoption in the United States.” *Pew Research Center: Internet, Science & Tech*, Pew Research Center, 5 Feb. 2018, www.pewinternet.org/fact-sheet/social-media/.

[155] Gratian, Margaret, et al. "Identifying infected users via network traffic." *Computers & Security* 80 (2019): 306-316.

[156] Webster, Ashton, et al. "An Improved Method for Anomaly-Based Network Scan Detection." *International Conference on Security and Privacy in Communication Systems*. Springer, Cham, 2015.

[157] Yale, Aly J. “New Credit Score System Might Make It Easier to Get A Mortgage.” *Forbes*, Forbes Magazine, 6 Nov. 2018, www.forbes.com/sites/alyyale/2018/11/01/new-credit-score-system-might-make-it-easier-to-get-a-mortgage/#4b8062f45a80.

[158] “UltraFICO™ Score .” *Introducing the UltraFICO Score*, Fair Isaac Corporation, 2018, www.fico.com/ultrafico/.

[159] Nittle, Nadra. “Spend ‘Frivolously’ and Be Penalized under China's New Social Credit System.” *Vox.com*, Vox Media, 2 Nov. 2018, www.vox.com/the-goods/2018/11/2/18057450/china-social-credit-score-spend-frivolously-video-games.

[160] “sklearn.naive_bayes.GaussianNB,” s*cikit learn,* scikit-learn developers, 2018, https://scikit-learn.org/stable/modules/generated/sklearn.naive_bayes.GaussianNB.html

[161] “sklearn.mixture.BayesianGaussianMixture,” *scikit learn,* scikit-learn developers, 2018, https://scikit-learn.org/stable/modules/generated/sklearn.mixture.BayesianGaussianMixture.html#sklearn.mixture.BayesianGaussianMixture

[162] “2.1.2 Variational Bayesian Gaussian Mixture,” *scikit learn*, scikit-learn developers, 2018, https://scikit-learn.org/stable/modules/mixture.html#bgmm

[163] “numpy.percentile,” *Scipy.org*, The SciPy Community, 2018, https://docs.scipy.org/doc/numpy-1.15.0/reference/generated/numpy.percentile.html