Frequent Inventory of Network Devices for Incident Response: A Data-driven Approach to Cybersecurity and Network Operations

Philip Delano Kobezak

Thesis submitted to the faculty of the
Virginia Polytechnic Institute and State University
in partial fulfillment of the requirements for the degree of

Master of Science
In
Computer Engineering

Joseph G. Tront, Chair
Randolph C. Marchany
Scott F. Midkiff

May 1, 2018
Blacksburg, Virginia

Keywords: Cybersecurity, Log Analysis, Network Inventory, Host Inventory

Frequent Inventory of Network Devices for Incident Response: A Data-driven Approach to Cybersecurity and Network Operations

Philip Delano Kobezak

(ABSTRACT)

Challenges exist in higher education networks with host inventory and identification. Any student, staff, faculty, or dedicated IT administrator can be the primary responsible personnel for devices on the network. Confounding the problem is that there is also a large mix of personally-owned devices. These network environments are a hybrid of corporate enterprise, federated network, and Internet service provider. This management model has survived for decades based on the ability to identify responsible personnel when a host, system, or user account is suspected to have been compromised or is disrupting network availability for others. Mobile devices, roaming wireless access, and users accessing services from multiple devices has made the task of identification onerous. With increasing numbers of hosts on networks of higher education institutions, strategies such as dynamic addressing and address translation become necessary. The proliferation of the Internet of Things (IoT) makes this identification task even more difficult. Loss of intellectual property, extortion, theft, and reputational damage are all significant risks to research institution networks. Quickly responding to and remediating incidents reduces exposure and risk.

This research evaluates what universities are doing for host inventory and creates a working prototype of a system for associating relevant log events to one or more responsible people. The prototype reduces the need for human-driven updates while enriching the dynamic host inventory with additional information. It also shows the value of associating application and service authentications to hosts. The prototype uses live network data which is de-identified to protect privacy.

Frequent Inventory of Network Devices for Incident Response: A Data-driven Approach to Cybersecurity and Network Operations

Philip Delano Kobezak

(GENERAL AUDIENCE ABSTRACT)

Keeping track of computers or hosts on a network has become increasingly difficult. In the past, most of the hosts were owned by the institution, but now more hosts are owned by the end users. The management of institution networks has become a mix of corporate enterprise, federated network, and Internet service provider. This model has survived for decades based on the ability to identify someone responsible when a host or system is suspected to be infected with malware or is disrupting network availability for others. Mobile devices, roaming wireless access, and users accessing services from multiple devices has made the task of identification more difficult. With increasing numbers of hosts on networks of higher education institutions, strategies such as dynamic addressing and address translation become necessary. The proliferation of the Internet of Things (IoT) makes identification even more difficult. Loss of intellectual property, theft, and reputational damage are all significant risks to institution networks. Quickly responding to and remediating cybersecurity incidents reduces exposure and risk.

This research considers what universities are doing for host inventory and creates a working prototype of a system for associating relevant log events to one or more responsible people. The prototype reduces the need for human-driven updates while incorporating additional information for the dynamic host inventory. It also shows the value of associating application and service authentications to hosts. The prototype uses real network data which is de-identified to protect privacy.

Acknowledgements

I thank all of my committee members for the opportunity to pursue this research. Without the guidance, support, and belief in me of my advisor, Dr. Joseph Tront, I would not have been successful in my undergraduate and graduate studies. Also, I must thank Randy Marchany for his mentorship in the field of cybersecurity and encouragement for many years. Together, Dr. Tront and Randy have supported my individual development for twenty years. I also must thank Dr. Scott Midkiff for being on my committee and the two ECE courses he developed, Network Applications and Wireless Networks, which have been foundational for my research and career.

I must thank Dr. Leslie Pendleton for her advising over the years which has led to this point. More recently, Dr. Scot Ransbottom and Dr. David Raymond have provided mentoring which has encouraged me to refine and complete this research. In addition, I thank Carl Harris for creating the NetRecon system which this work is based on.

I also thank everyone I have worked with in the IT Security Office and Lab over the years to include Brad Tilley, Mark DeYoung, Zach Burch, Mike Cantrell, Chris Morrell, Matt Sherburne, Reese Moore, and Stephen Groat.

I especially must thank my loving wife, Amy, for her understanding and patience in this process. Without her considerate support, this work would not be possible.

Lastly, I thank my sisters and mother: Stephanie, Sarah, and Deborah, for encouragement and guidance to get to where I am today. I dedicate this work to my father, Thomas Kobezak, who passed away shortly after I started graduate school. His commitment to our family and community was never ending. His career in the Air Force and Army was inspirational to many.

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

A pessimist sees the difficulty in every opportunity;
an optimist sees the opportunity in every difficulty.

- Winston S. Churchill

In incident response, the time between initial identification and containment is critical to reducing damage when sensitive or high-risk data is involved [1]. This is particularly true with modern malware moving to mobile devices and evolving to include theft of messages, position data, and banking credentials, all with real-time attacker command and control [2]. The malware or compromised host can be contained by denying network access. This can often be accomplished at the network operator level. However, this cannot always be done given the distributed nature of many university networks and the potential to deny non-compromised hosts. Instead, identifying a responsible contact and motivating that person to contain the compromised host and eradicate any malware or point of entry. This is becoming more common as there are virtual machines, application containers, and non-centralized network address translation (NAT), which can lead to more than the compromised host on the same network connection. Therefore, accurate methods to track hosts on a network, as well as identify a responsible person for each host are necessary.

1.1 Problem Statement

Current host inventory control methods are error prone due to manual, human data entry. There is also a greater number of host types on many networks than previously. For example, there are more embedded or Internet of Things (IoT) devices. There are also more Bring Your Own Devices (BYOD) and mobile or wireless devices. In past decades, a university might own most of the network devices. Those hosts would physically stay in the same place or move very infrequently. The hosts would only have IPv4 addresses and frequently be statically defined. The number of hosts on the network would be a small fraction of what universities see today. The types and numbers of hosts on a university network has changed. To illustrate the trend, Figure 1-1 depicts the shift in ownership (R. Marchany, personal communication, May 1, 2017).

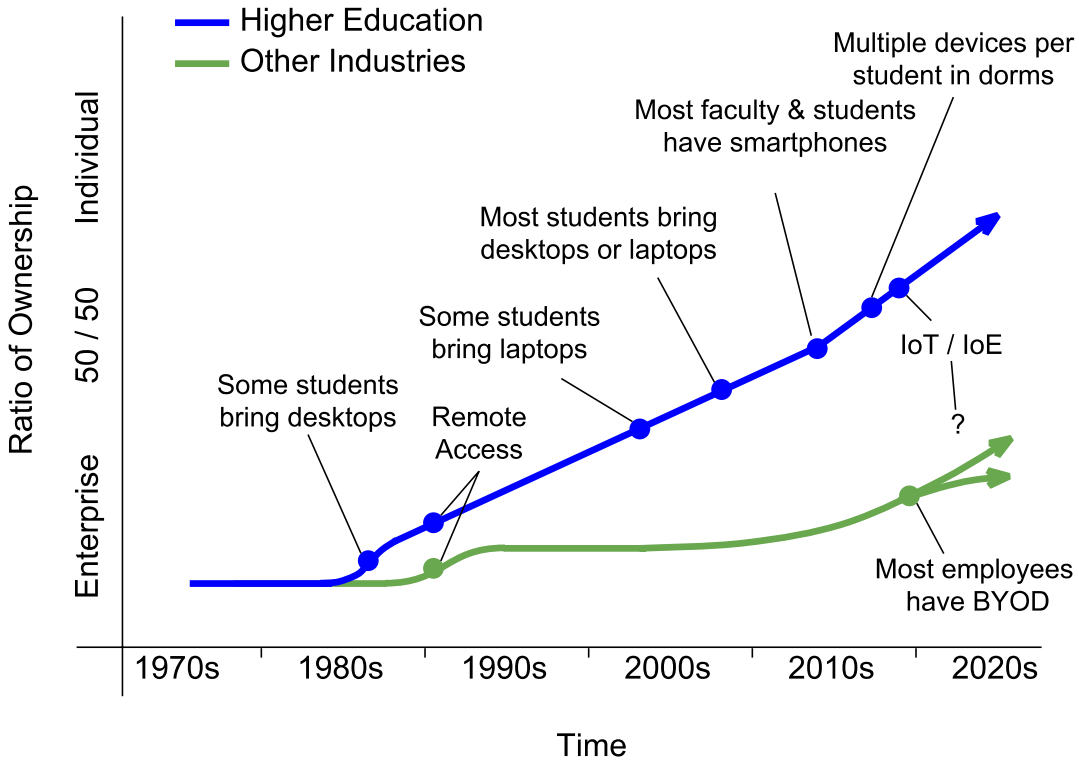

Figure 1-1: Trend in Ownership of Devices Over Time

1.2 Motivation

Previously, attackers went directly after servers and large data sources. Now, they attack endpoints with the assumption that they are not as well maintained or monitored and can then be used to access services with the user's credentials [3]. With the commercialization of the Internet, research institution networks are now being attacked by activists, criminal organizations, and nation states. Portions of the following sections were first published in the proceedings of the 2018 Hawaii International Conference on System Sciences [4].

1.2.1 CIS Critical Security Controls

Organizations must prioritize the application of resources in the defense of cyber-attacks to minimize risk to their networks. Cyber security controls frameworks help with this prioritization, and often recommend specific methods, software, and systems to implement individual controls. Johnson states "all security and corporate managers now need to be concerned with compliance and governance of risks, security, and the information usage in their systems" [5]. This is especially true for higher education institutions that conduct research and must comply with mandates to defend against cyber-attacks or risk losing funding.

The Center for Internet Security (CIS) is a not-for-profit organization "dedicated to enhancing the cyber security readiness and response among public and private sector entities" [6]. The CIS Critical Security Controls (CSC) for Effective Cyber Defense exist as a framework to help organizations improve their information security strategy. The Controls were developed by experts from many different organizations who "pooled their extensive first-hand knowledge from defending against actual cyber-attacks to evolve the consensus list of Controls, representing the

best defensive techniques to prevent or track them" [7]. The twenty Controls are "a prioritized, highly focused set of actions that have a community support network to make them implementable, usable, scalable, and compliant with all industry or government security requirements" [7]. The CSC framework is intended to provide an organization with key areas where they should specifically focus their efforts. Each Control gives example technologies that an organization can implement to help achieve their goal of reducing risk. As no single measure is guaranteed to prevent cyber security incidents, organizations are encouraged to implement all the Controls to have a defense in depth strategy.

The motivation for this research is derived from the first Control outlined in the CSC: inventory of authorized and unauthorized devices. As of version 6.1 of the CSC, six sub controls are defined for the first Control.

1.2.2 Critical Security Control One Sub-controls

CSC 1.1, "deploy an automated asset inventory discovery tool…" [7] is common for Internet Protocol version 4 (IPv4) networks. Organizations can scan their network address space to identify hosts, and even attempt operating system identification. Nmap and other scanning tools can provide this ability [8]. Unfortunately, scanning an Internet Protocol version 6 (IPv6) network is not so straightforward, due to the extremely large address space and time it would take to iterate through each address and send probe packets to solicit a response [9]. In more recent years, passive scanning, or the listening for active hosts on the network, has become more common. This involves "the process of monitoring network traffic at the packet layer to determine topology, services, and vulnerabilities" [10].

CSC 1.2, "deploy dynamic host configuration protocol (DHCP) server logging..." [7] is something that most organizations can easily implement. By simply logging DHCP server events, we can better track hosts on the network. This is commonly used in IPv4 networks; however, depending on the IPv6 deployment, DHCP may or may not be used. IPv6 networks may use Stateless Address Autoconfiguration (SLAAC), DHCPv6, or statically assigned addresses [11], [12].

CSC 1.3, "ensure that all equipment acquisitions automatically update the inventory..." [7] is fundamentally a business process. To comply, organizations must make sure there are automatic updates to the inventory based on new acquisitions. This can be accomplished by integrating an Enterprise Resource Planning (ERP) application with the inventory system. Doing these updates manually becomes problematic for many organizations, requiring data entry in the business and financial applications that is IT-specific. In some organizations there is a fundamental decoupling of business operations from network operations.

CSC 1.4, "maintain an asset inventory of all systems connected to the network…" [7] describes an all-encompassing inventory. This Control seems to be solved by using a database-driven application to track this information. This is common, as are spreadsheets, in many organizations. However, the accuracy of these manual processes usually erodes over time, given the significant effort required by personnel to enter and update each host's details. This technique also does not scale for networks with tens or hundreds of thousands of hosts. Something that is common to research institution networks is the ability to Bring Your Own Device (BYOD) and connect it to the network. While many corporate networks are able to resist BYOD, higher

education has seen this for decades. This means that CSC 1.3 is not applicable in this situation since the owner of the device is not the same as the owner of the network.

CSC 1.5, "deploy network level authentication via 802.1x…" [7] requires every host to be authenticated to the network. This is commonly deployed for wireless and some wired networks. Sometimes it is also deployed to authenticate Voice over IP (VoIP) devices to separate Virtual Local Area Networks (VLAN). Depending on the end points, this may be less feasible to deploy across the entire network of an organization. There may also be limitations on the deployment of IEEE 802.1x with older networking equipment.

Finally, CSC 1.6, "use client certificates…" [7] requires the use of certificates to authenticate each device instead of a username and password. Client certificates are a highly secure method of authentication but do carry significant management overhead.

1.3 Research Objectives and Questions

This research effort evaluated what other universities are doing for host inventory and created a working prototype of a system for associating relevant log events to one or more responsible people. The following questions were the target of this research.

1. What are the current host inventory controls used in higher education networks?
   (a) Can we categorize these methods and controls?
   (b) How do we define a host and is it consistent across universities?
   (c) Are the current host inventory controls effective?
2. Can we improve the accuracy of host inventory controls using existing data sources?
   (a) Can we reduce the manual updating of host information in an inventory?

(b) Can we automate the association of host activity?

3. Is the method of associating an application login with a host beneficial for finding a responsible person?

   (a) How often does it lead to a responsible person without network authentication?

   (b) When does this not help?

1.4 Hypothesis

1. Given information in the form of logs from network infrastructure devices, applications, and services, we can develop a method for dynamically and accurately identifying hosts, associated users, and responsible personnel for each device on a given network. More specifically, for each host on a network we can reach a high level of certainty of the responsible people, in near real time, and without the need for manually updating an inventory database.
2. It is also possible to identify responsible people without the need for network authentication if the network operator also provides applications which constituents utilize regularly.
3. The above statements are possible without requiring changes to hosts, network infrastructure, or services. This includes not installing new software on hosts or implementing network architecture changes.

1.5 Structure of Thesis

The following chapters are organized to provide background, previous work, system design, evaluation, and conclusion. Chapter 2 provides justification for conducting this research

with a survey of other institutions. Chapter 3 provides background in device identification, categorization, network operations, cybersecurity incident response, and log analysis. Chapter 4 looks at previous solutions to the problem including human-driven tools, network discovery, network authentication, Grand Unified Logging Program, and NetRecon. Chapter 5 discusses the design of the prototype solution with requirements, privacy preservation, data flow, and architecture. Chapter 6 details the experimental setup, results, and analysis. Finally, Chapter 7 concludes with a summary of the work, contributions, benefits, limitations, and future work.

Chapter 2

Justification

When I walk along with two others,
from at least one I will be able to learn.

- Kong Fu Zi (Confucius)

The problem, as proposed by the author, was based on observations of one higher education institution network. The author deemed this to not be sufficient to understand the problem, nor adequate to make assumptions about how other institutions have structured their networks. To validate and justify this research direction, a survey of Chief Information Security Officers (CISOs) was conducted to determine how inventory controls were being used in higher education. Results from that survey were published in a paper [4].

2.1 Survey Design

The higher education CISO survey was designed to answer the following high-level questions: Are new technologies changing the accuracy of inventory controls? How quickly can the location of a host and the responsible user be identified? Are current host inventory controls

effective? Have there been changes in effectiveness due to increases in Internet of Things (IoT) and BYOD hosts? Do the responses vary with subsets of the population such as size of the network or number of employees dedicated to information security operations?

The survey attempted to determine correlations between sizes or types of institutions and network architectures. Network architectures will vary with the type of institution to include the amount of research, number of residential students, and user population size. Some questions were based on a survey that identified cybersecurity challenges for higher education institutions [13].

The survey was reviewed by several information technology professionals for the quality of questions and answers. Many questions were modified or removed to reduce ambiguity and improve readability. The survey was tested prior to release and the respondents' test results were also used to improve the questions.

No personal information of the respondents was solicited. The only identifying attribute recorded was the respondent's IP address. This was used to identify whether multiple responses were recorded for the same institution. This also gave the ability to delete a response at the request of a respondent by asking them to verify the IP address they used. Even with this single piece of information that could be linked to the institution, it was removed once the survey closed. This was to encourage honest responses without fear of the respondent being identified.

The target population for the survey was all higher education institutions in the United States. According to the Carnegie Classification, there are approximately 4,600 institutions [14]. The first question of the survey was to identify the respondent's institution's Basic Carnegie Classification. Additionally, questions were asked to determine the size and attributes of each institution to include numbers of students, employees, employees in information security roles,

and estimated research expenditures. This first section of the survey was used to provide a framework for comparison of like institutions only. The second, third, and fourth sections of the survey asked questions pertinent to network size, host identification, and evaluation of controls. Samples of those survey questions are mapped to the appropriate CSC control in Table 2-1 based on [7]. A complete list of survey questions is available in Appendix A.

Table 2-1: Select questions mapped to the CSC

Survey Question	CSC
2.1. What is your best estimate for the peak number of hosts on your network at one time?	1.2, 1.4, 1.5
2.2. What is the average number of BYOD hosts each type of end-user connects to your network?	1.3, 1.4
2.4. What is your best estimate for the number of sub-networks (Local Area Network segments or broadcast domains)?	1.1, 1.4
2.10. Where does your institution allow embedded hosts or Internet of Things (IoT) on your network?	1.3, 1.4
3.1. What is the estimated percentage of each type of host on your network?	1.3, 1.4
3.3. What IP addressing methods do you use?	1.1, 1.4
3.7. How confident are you in your organization's ability to identify hosts with multiple, changing addresses, to include application containers (Docker) and IPv6 privacy extensions (RFC 4941)?	1.1, 1.4
3.8. What percentage of hosts on your network utilize some form of network authentication to connect (IEEE 802.1x, NAC, etc.)?	1.5, 1.6
4.1. For the purposes of your host inventory controls, what types of hosts do you track?	1.3, 1.4
4.2. During a potential security incident or event, how long does it usually take to track down the responsible user or owner of these host types?	1.4, 1.5
4.6. How accurate have you found the following tools and technologies to be in keeping track of hosts in your network?	1.4, 1.5
4.7. Do you consider embedded devices or IoT hosts more difficult to track than other hosts?	1.4

The survey had 42 questions, but some asked the respondent to answer for different cases which results in up to 96 total data points. Only one question had a required response due to validation needed to constrain the sum of the response to one hundred percent.

2.2 Survey Results

The survey was opened for distribution to participants on May 24, 2017. Since the survey was targeting Chief Information Security Officers of higher education, several email lists were used to distribute the anonymous link. Most respondents completed the survey in less than 20 minutes. The following are a subset with the complete results available in Appendix A.

2.2.1 Institution Classification

These survey results cover 51 responses. More than half of the respondents reported their institutions to be R1, R2, or R3 doctoral granting universities with research activity as shown in Figure 2-1.

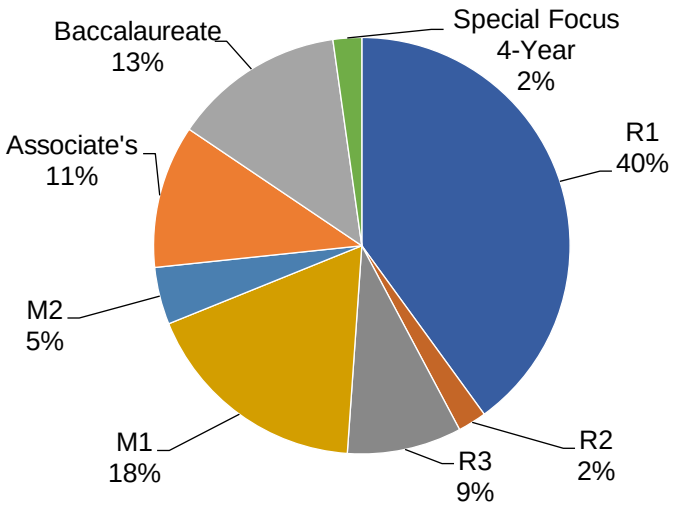

Figure 2-1: Percentage of respondent institutions by Basic Carnegie Classification

Using the high-end of the selected ranges for total employees, employees in IT, and employees in security operations, ratios were calculated. The results showed the ratios of employees to be 5.3 percent for IT to total employees and 5.4 percent for information security to IT employees.

There was a wide variety of reported enrolled students with most reporting between 2,000 and 50,000. There was some variance with the number of reported remote students. However, more than half of the respondents reported 10 percent or fewer remote students.

2.2.2 Network Characteristics

For question 2.1, most of the respondents selected the peak hosts on their network to be 10,000 to 50,000. Eleven respondents said their networks were greater than 50,000. One stated they had more than 500,000 peak hosts on their network.

One important question asked, "How is the respondent's network managed?" The results are shown in Table 2-2.

Table 2-2: CISO's Ability to Deny or Allow Hosts

	Response Option	%
1	The CIO or CISO has the ability to deny or allow all hosts on the network	64
2	The CIO or CISO has the ability to deny or allow most hosts but not all	26
3	The network is mostly federated. Most organizational units control their networking, to include network equipment and hosts	2
4	The network is completely federated. The CIO or CISO has no ability to allow or disallow hosts.	0
5	Other	7

This question may have been interpreted differently than anticipated. The intent was to determine how federated or completely centralized the institution's IT functions were. If the

respondents understood the question, it is possible that their institutions are mostly centralized in terms of managing the network.

Figure 2-2 shows the responses to question 2.2, BYOD host percentages by user type. The differences are particularly pronounced for residential students in the 4 to 6 owned devices range.

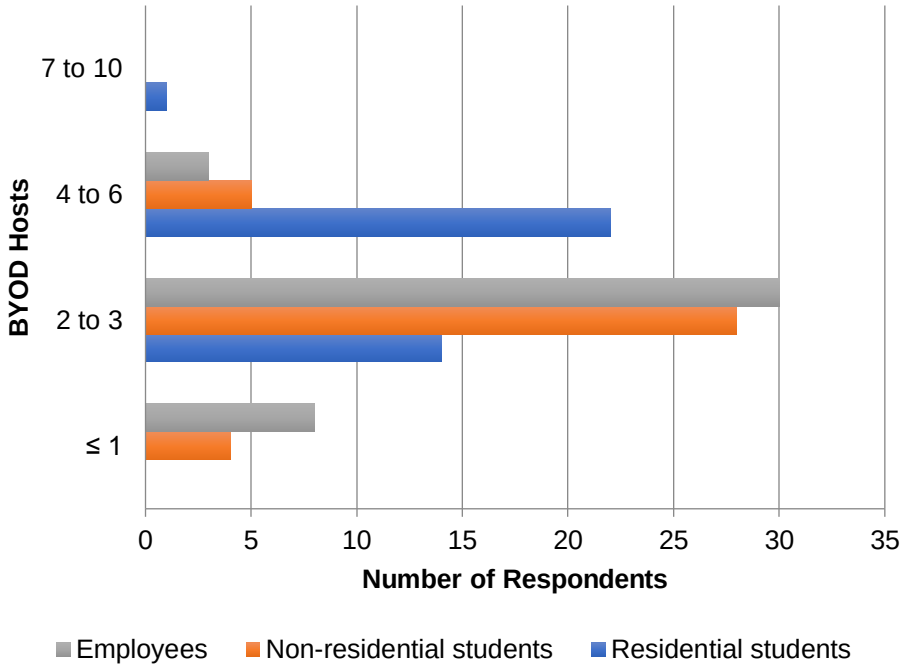

Figure 2-2: BYOD host quantity by user type

For the average number of BYOD hosts connected by non-residential students, 67 percent of respondents said 2 to 3. This differs from the number of BYOD hosts connected by residential students, which was split between 2 to 3 and 4 to 6. This is not surprising, as you can assume that residential students will connect devices in their dorm rooms that they would otherwise keep in an off-campus residence. What we found surprising is that 73 percent of respondents said employees connected 2 to 3 BYOD hosts. This means that most institutions expect employees to connect 2 to 3 personal, BYOD devices that are not institutionally owned.

Questions 2.9 and 2.10 in the survey asked where the institution allowed BYOD and embedded or IoT hosts. The results in Figure 2-3 show that most respondents chose "most logical network zones" for BYOD hosts.

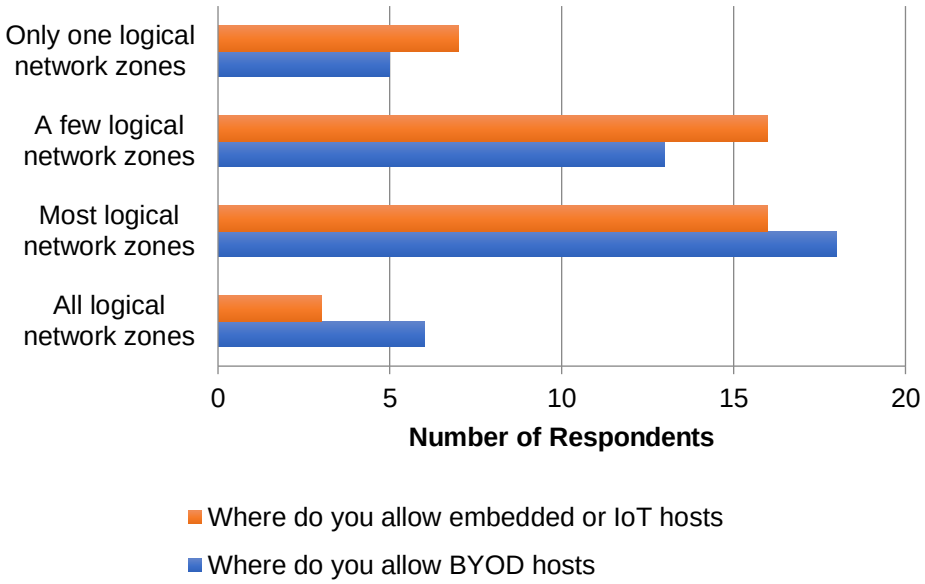

Figure 2-3: Where BYOD, Embedded, and IoT Hosts Are Allowed

For embedded or IoT hosts, the response was evenly distributed. The exception to both is that a few institutions allow BYOD and embedded or IoT hosts on all network zones.

2.2.3 Defining a Host

The authors were interested in understanding what the average distribution of host types are on an institution's network. Question 3.1 asked respondents to provide their estimated percentage of each of four host categories: embedded devices, servers, institutionally owned end-user devices, and BYOD end-user devices. The results in Table 2-3 show that, on average,

embedded devices (IoT, printers, cameras) and BYOD end-user devices make up half of an institutions network.

Table 2-3: Host Type Percentages

Host Type	Min %	Max %	Mean	Standard Deviation
Embedded devices (IoT, printers, cameras, etc.)	1	25	9.11	6.03
Servers with full operating systems (either physical or virtual)	1	80	15.89	14.39
Institution owned end-user devices (desktops, laptops, mobile devices)	4	75	33.56	16.28
BYOD end-user devices (desktops, laptops, mobile devices)	0	92	39.44	20.64
Other	0	11	0.86	2.57

When asked about the percentage of hosts that use statically assigned IP addresses, all but one respondent said 10 or 20 percent. In addition, the respondents were asked what addressing methods they used. The results are shown in Figure 2-4.

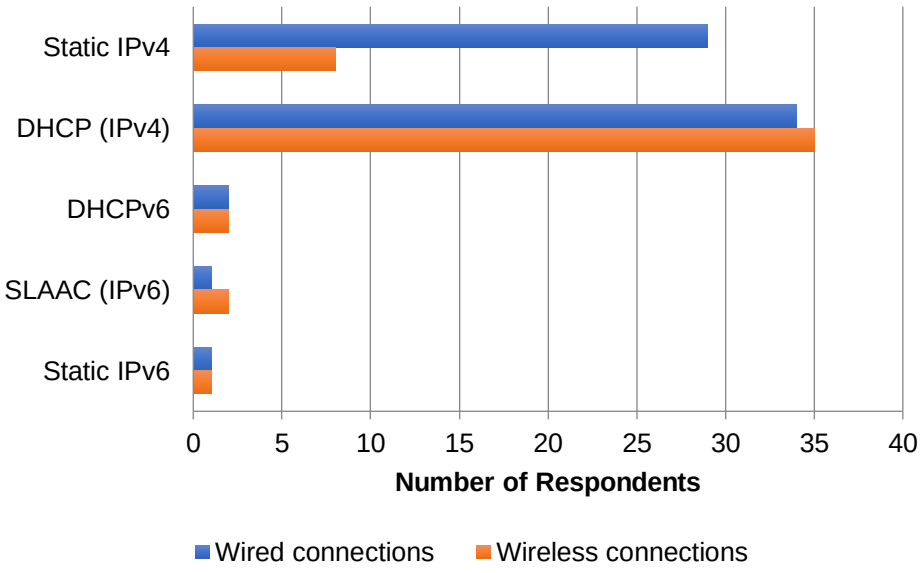

Figure 2-4: Types of Addressing Used on Wired and Wireless Connections

Interestingly, eight respondents stated that they used static IPv4 addressing on wireless connections. These could be embedded devices such as printers or copies using wireless however, the authors would expect DHCP Reservations to be used for wireless devices

Questions 3.6 and 3.7 asked how confident the respondent was in identifying unique individual hosts for virtual machines and application containers. The results are shown in Figure 2-5.

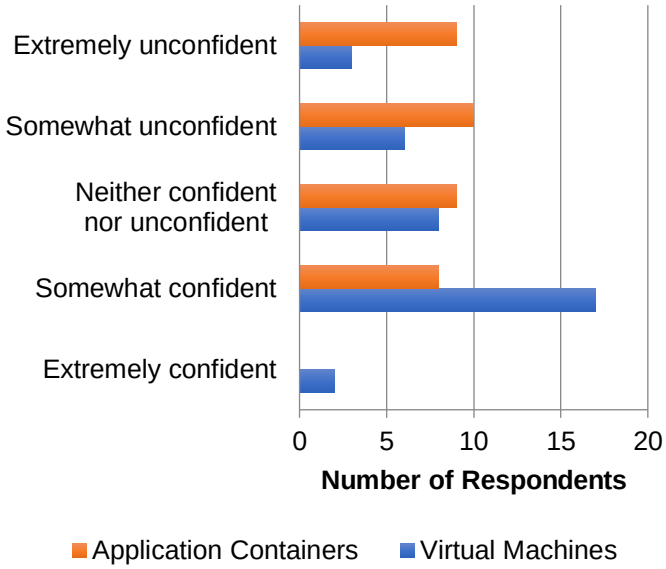

Figure 2-5: Confidence in Identifying Virtual Machines and Application Containers

The last question of this section asked how respondents identified a unique host. Most all stated, in their own words, that a MAC address was the unique identifier.

2.2.4 Evaluation of Inventory Controls

In this section of the survey, the first question asked respondents to identify whether or not a particular host type was tracked. The results are shown in Figure 2-6.

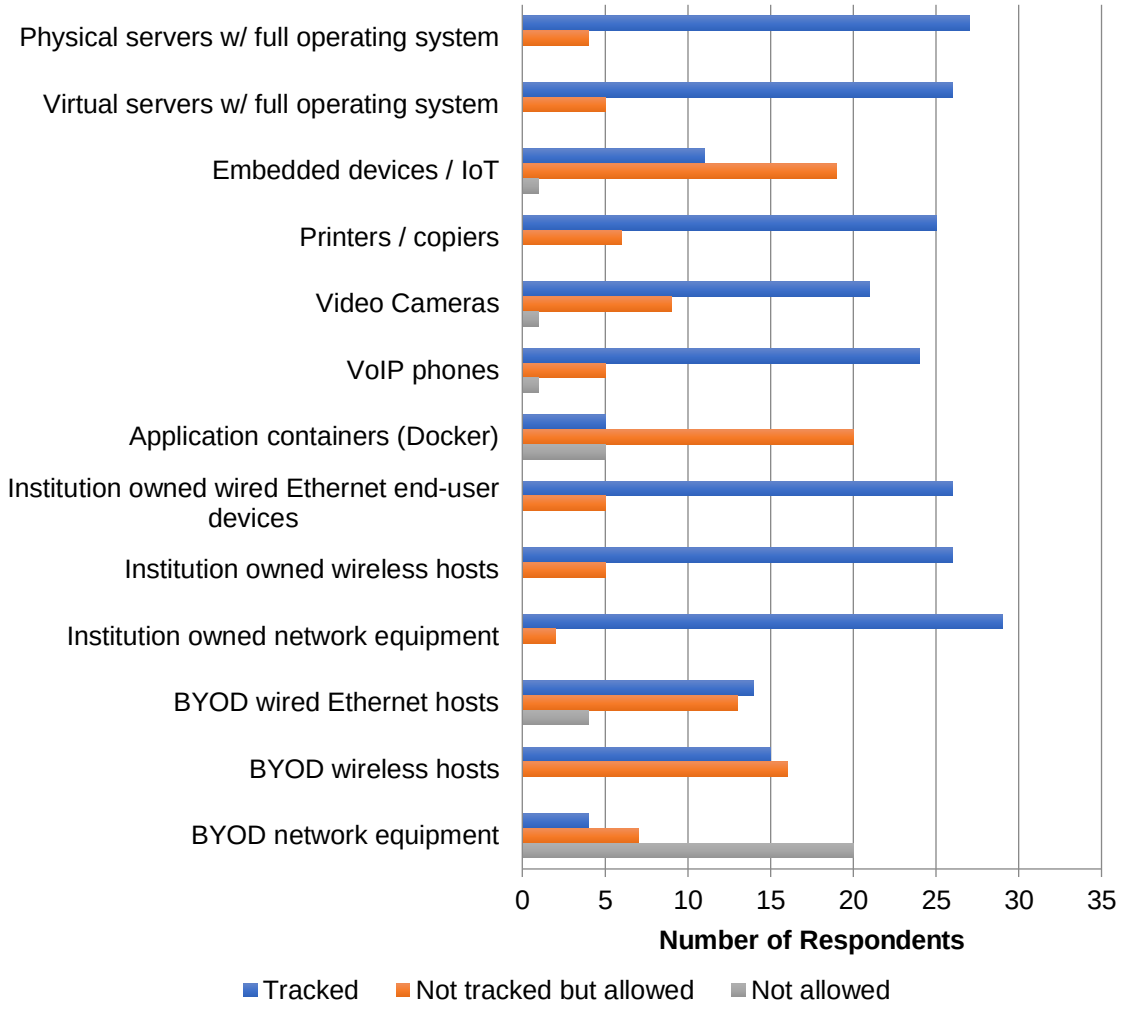

Figure 2-6: Host tracking by type

It is worth noting that fewer respondents said BYOD, embedded or IoT, and application containers were tracked. Most respondents tracked physical and virtual servers, VoIP phones, video cameras, printers, and institutionally owned network equipment.

Figure 2-7 shows the time it takes to track down the physical location of a host for non-research and research institutions. It is worth noting that a greater number of research institutions (R1, R2, and R3) selected the more than 60 minutes option for multiple host types. This could be

due to the larger number of hosts on research institution networks or the distribution of IT responsibility.

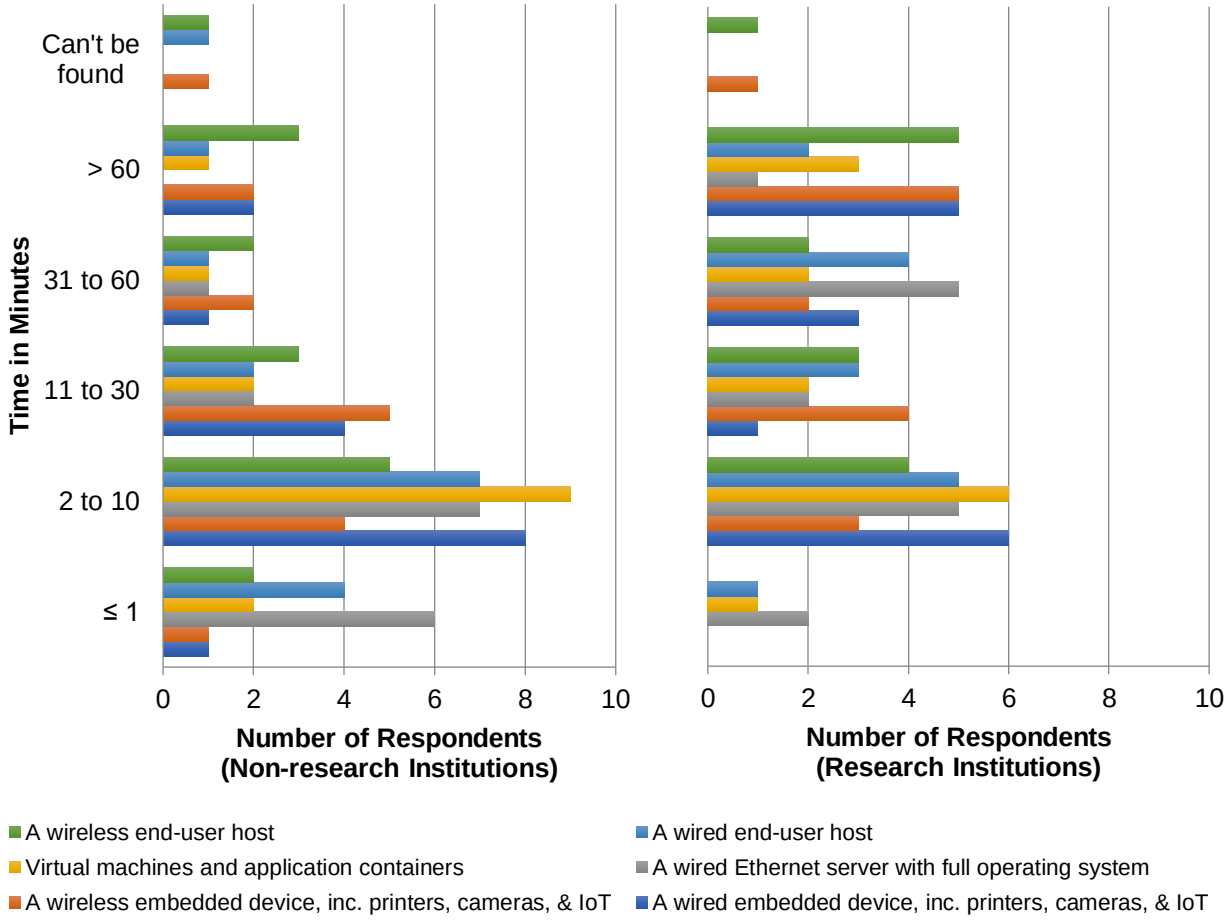

Figure 2-7: Time to Find Physical Location of Different Host Types

For question 4.4, respondents were asked how often their inventory controls and tools lead to someone who is not the current responsible user; most respondents selected a few times a month. Some wrote in that it varies widely and is worse for lab environments.

Question 4.6 asked respondents how accurate they thought various inventory tools were. The results are show in Figure 2-8. It is worth noting that five respondents said that mapping or

scanning of IPv6 was somewhat or very accurate. It would be interesting to know their methods given the large address space.

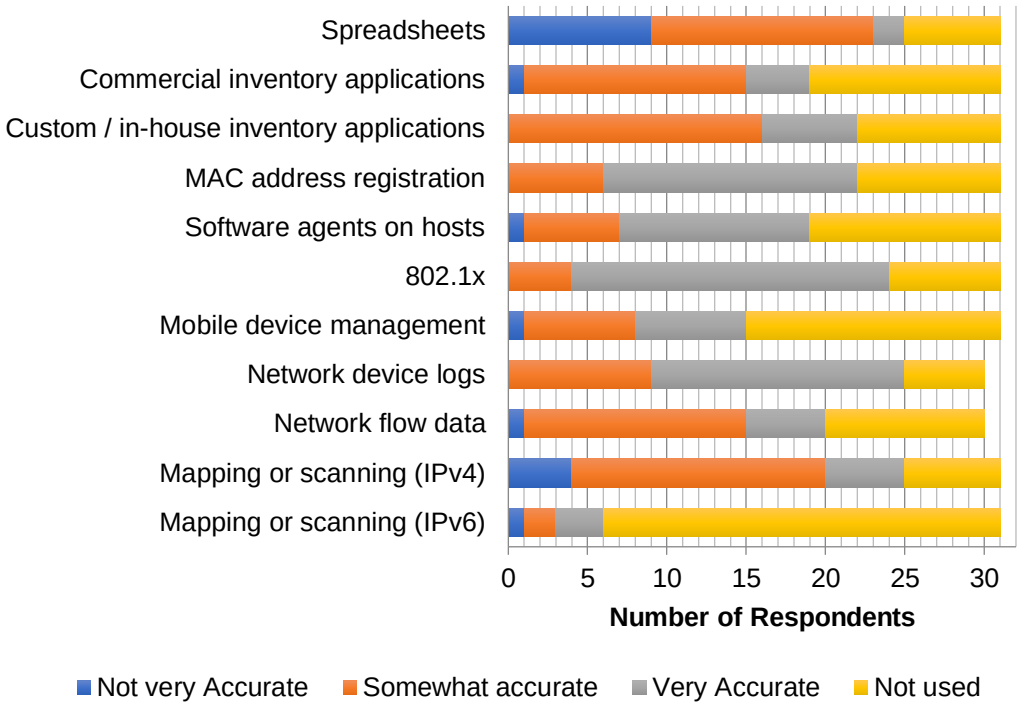

Figure 2-8: Accuracy of Inventory Tools

Questions 4.7, 4.8, and 4.9 asked if certain host types were more difficult to track than others as shown in Figure 10.

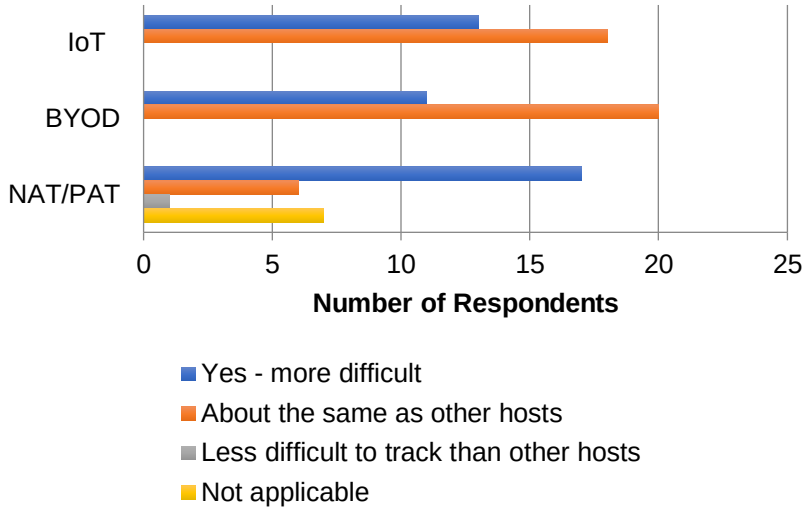

Figure 2-9: Difficulty Tracking Host Types

The third host type, NAT/PAT, was used to determine whether address translation has an impact on inventory controls. Interestingly, seven respondents selected not applicable for NAT/PAT. The authors surmise that these institutions may have enough IPv4 addresses for all hosts, and therefore have no need for address translation.

Questions 4.11 and 4.12 asked if respondents believe the effectiveness of their host inventory controls changed in the past five years for either BYOD and IoT hosts. As shown in Figure 2-10, nearly half of the respondents from research institutions stated that both host types impacted the effectiveness of their inventory controls.

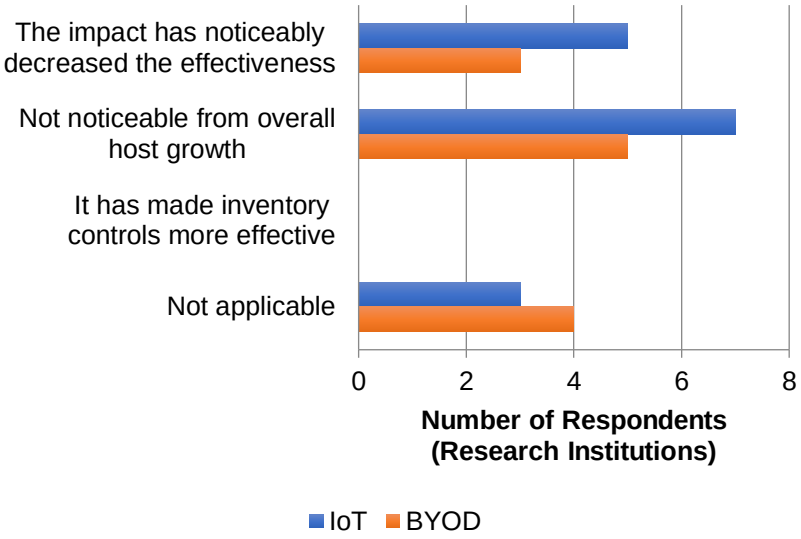

Figure 2-10: Effectiveness of Host Inventory Controls from Impact of IoT and BYOD

Question 4.13 asked respondents how much time they spend updating host inventory control tools. The results, charted in Figure 2-11, show that more than half of institutions spend a moderate to significant amount of time updating records.

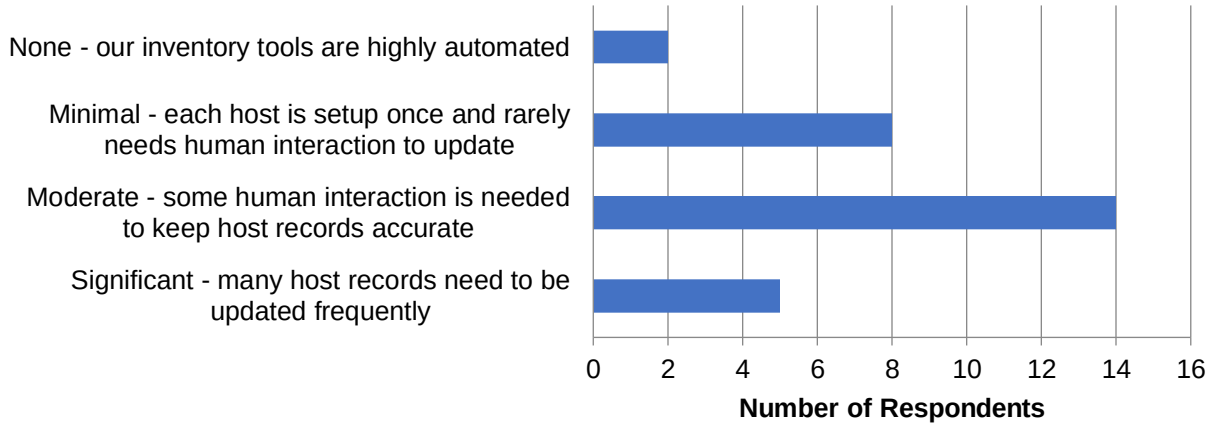

Figure 2-11: Time Spent Updating Inventory Tools

The final question of the survey asked if the respondent had any specific challenges with host inventory controls. A couple of respondents stated that it is difficult to have a unified inventory with a distributed IT responsibility. One respondent also stated that NAT/PAT can be

an issue for their DMCA complaints. Another stated that they would like to raise awareness of keeping inventories current and correct.

2.3 Discussion and Insights

It is worth noting that even though the CSC 1 provides methods for inventory, these are corporate enterprise-centric. Even though a higher education institution network may be a special case, the way it works may actually become more common. With BYOD, wireless, and virtualization, the methods of traditional inventory are becoming more difficult to deploy and scale. Specifically, with BYOD, corporate networks are allowing more personal devices in their environments [15]. Some will segment their wireless networks; however, there is ever growing pressure for these corporate networks to allow personal devices on their more restrictive network segments.

2.3.1 Network Access

We must consider the user-base as we discuss access and authenticating to a network. In a higher education institution's network, there is an expectation that access to the Internet should be unhindered. This is because faculty and students need to complete their work by collaborating with other higher education institutions and industry partners. To them, the network is only a tool to accomplish this. Additionally, many research institutions have multiple campuses along with faculty and students who are frequently traveling around the world. With this culture, there is usually greater emphasis on controlling access at the applications that are globally available.

This leads to the discussion of private versus public networks. Many higher education institutions operate networks, which could be considered hybrids. For legal reasons, most institutions consider themselves private networks, but discussions have been ongoing ever since The Communications Assistance for Law Enforcement Act (CALEA) and the USA PATRIOT Act have come into existence [16]. Even with this designation, most faculty and students expect open access to the Internet. This is a culture that has been around since the early years of the Internet. Larger higher education networks are traditionally operated like Internet Service Providers (ISPs) where their primary focus is to make sure packets are getting from one host to another. In recent years, some higher education institutions have become more limiting on the free flow of traffic in and out of their networks. Nonetheless, these networks remain much more open than other private networks such as those in corporate environments. This cultural tendency makes requiring high assurance authentication to the network, and ultimately the Internet, a challenge.

2.3.2 Host Attributes

In previous decades, static, wired hosts might have used the same IPv4 address for long periods of time, sometimes months or even years. The pace at which IT systems are changing is increasing. The life cycle of an individual host has shortened while the expectations of service availability has increased. This leads to redundancy inside a host's subsystems and to redundancy in entire hosts. With redundancy at the host level, the service may change which hosts are responding to requests. This leads to hosts that are dynamic and taken out of service for maintenance or failure. Virtualization furthers this trend of more difficulty in tracking hosts. Virtualization enables the decoupling of hosts from hardware, thereby allowing movement. The Media Access Control (MAC) addresses, previously considered relatively static, are now created

when new virtual machines (VM) are defined [17]. The ease of creating and moving VMs can be a challenge for traditional host inventory tools.

Some environments are moving to services being deployed in containers by which the operating system or host is considered separate. This leads to even more churn in the traditionally static hosts providing services. For example, Docker is a containerization platform that provides separation of applications from their operating system. Using Linux kernel technology, the containers even have their own network interfaces [18]. These interfaces, like the virtual machines, have their own MAC addresses. Again, this can complicate the issue of how we define a host and what attributes we inventory.

2.3.3 Host Responsibility and Organizational Inefficiencies

Answering the question of who is responsible for a host is core to host inventory. This can be a difficult problem in a federated research institution network. There can be hosts in which the user is the responsible party, as is the case for BYOD. There are also groups of hosts in which an IT professional is responsible. In some instances, the research institution can have both a central IT organization and distributed IT professionals reporting through different leadership. This federated network management model requires more effort to define and track who is responsible for any host. One common method involves the assignment of blocks of addresses to organizational units. The institution assumes that organizational units will track hosts within their assigned block. It is an honor system and can be problematic if the organizational unit has no knowledge of a host using one of its addresses.

Two of the sub controls from CSC 1 are focused on authenticating to the network. If we accept the scenario in which all devices on a network are authenticated, we still have to map the

user to a group or responsible IT professional when the primary user is not. Again, BYOD comes into play whereby the organization may not have a record of who the device belongs to or who should be contacted if there is an incident involving it.

One last consideration is the time involved in maintaining most host inventories. It is simple to keep the inventory of a twenty-host network up to date. The time it takes to maintain the inventory increases steadily with the number of hosts unless efficient tools are used. Even then, there is significant time spent on updating each host entry. This can be a burden on already busy IT personnel and takes them away from solving more high-profile issues. IT professionals can also miscategorize or mistype information. This fundamentally human element makes a tedious tracking process more inaccurate as time goes on.

2.4 Summary

In this chapter, the CISO survey results were presented along with analysis and insights. Evidence supporting the problem statement was found along with answers to the first set of questions in section 1.3. Specifically, question 4.13 of the survey, that most institutions spend at least a moderate amount of time updating host records. Additionally, the time necessary to identify physical locations of hosts as shown in Figure 2-7 is less than desirable. Lastly, overall confidence in the ability to track all host types could be improved.

Chapter 3

Background

The Internet was done so well that most people think of it as a natural resource like the Pacific Ocean, rather than something that was man-made.
- Alan Kay

In order to explore the topic of host inventory, an understanding of Internet Protocol (IP) and Local Area Networks (IEEE 802) is necessary. Most modern networks rely heavily on both for host communications. It is also important to understand the first three layers of the Open Systems Interconnect (OSI) model. Furthermore, a clear understanding of how host inventory is important to network operations and cybersecurity is needed. This scopes the problems being explored and provides greater insight for the reasons that the problems exist. Lastly, this research focuses on the use of preexisting log events for the creation of the inventory. Therefore, current generation storage and processing tools are needed to make use of large amounts of data.

The following sections describe key background concepts which the reader may already be familiar with, but will reinforce their importance to the design.

3.1 Device Identification

For the purposes of this research, network devices are considered to be all nodes that connect to a network. Network equipment are network devices operating in OSI layers 1, 2, and 3. These devices form the infrastructure of the network and are responsible for maintaining the state of how to reach hosts and the timely delivery of packets. Common examples are an Ethernet switch, router, or wireless access point.

A network host is a network device participating in a network by using one or more addresses. It transmits or receives packets which it acts upon for higher layers of the OSI model, frequently initiated by a person. Common examples laptop, mobile device, printer. Note that a piece of network equipment can be a host if it is managed in the same network to which it is part of the infrastructure. For example, if it provides services for management or to facilitate other network device connectivity such as DHCP.

The concept of what a host is has become more complicated as more devices have connected to the Internet. In the origins of the Internet, institutions rarely had more than a few hosts to be concerned with. In 1988, Virginia Tech only had 6 hosts on the Internet according to the combined Stanford / NIC, DoD Host Table – known as the *hosts.txt* file (see

Figure 3-1) [19]. There were only around 8,000 hosts on the Internet in this time period and a single file could reference all.

```
HOST : 128.173.1.4 : VTNET1.CNS.VT.EDU : BRIDGE-CS/100 ::
HOST : 128.173.1.5 : VATECH.CNS.VT.EDU : SUN-2/170 : UNIX :
HOST : 128.173.2.1 : VTOPUS.CS.VT.EDU : VAX-11/785 : UNIX :
HOST : 128.173.2.6 : VTCS1.CS.VT.EDU : VAX-11/785 : VMS :
HOST : 128.173.4.1 : VTVM1.CC.VT.EDU : IBM-3090 : VM/CMS :
HOST : 128.173.4.247 : DCSSVX.CC.VT.EDU,VT.EDU : VAX-2000 : ULTRIX :
```

Figure 3-1: Virginia Tech Hosts in 1988

Today, the concept of a host has changed dramatically. No longer are computers something only a large institution would purchase. There are now network hosts that cost about the same as an inexpensive meal: $10 [20]. These System on Chip (SoC) devices and other forms of IoT have only added to the ever-expanding number of concurrent hosts on the Virginia Tech network.

3.1.1 Data Link Layer and Network Layer

Network inventory is fundamentally the combined, global state of the OSI data link and network layers. At any one point in time, and assuming correct configuration, the state is maintained on each piece of network equipment for the layers it is operating in and only for its portion of the network. For a common Local Area Networks (LAN), the data link layer is frequently IEEE 802.3 (Ethernet) and IEEE 802.11 (Wireless Ethernet) [21].

Many networks utilize NAT [22] to reduce the number of needed global IPv4 addresses. NAT can be implemented with one global IPv4 address mapped to one RFC 1918 IPv4 address or as Port Address Translation (PAT), where multiple RFC 1918 IPv4 addresses use a single global IPv4 address. This is accomplished by mapping a TCP port range of the global address to each internal IPv4 address.

3.1.2 Device Categories and Attributes

IPv4 and IPv6 addresses can be static but most hosts on the Virginia Tech network are dynamic. This is especially true for hosts on the 802.11 network. Therefore, IP addresses are rarely a good identifier for a host. Based on the results of the Q3.9 in the CISO survey, the MAC address of a network interface is the best option. This means that for the purposes of this research, we

consider each unique MAC address to be a separate host. This is valid for considering virtual machines and application containers as separate hosts. This is an issue for hosts that alternate use of a wired and a wireless interface. In this case, the single host would have two MAC addresses.

We can think of a host as a set of dynamic attributes which can, and often do, change for certain time periods. For example, a physical host may be purchased by someone but setup by a responsible IT professional. Someone may also loan their host to someone else to use for a period of time. Some hosts may frequently be used by multiple people. These attributes are shown in Table 3-1 and further explained in the following sections.

Table 3-1: Host Attributes and Values

	Attribute	Possible Values
1	Users *Is the host used by 1 or more people or is it autonomous?*	0, 1, or more persons
2	Owner *Who acquired the host and can control access?*	1, >1 persons, or an institution
3	Responsible person *Who can resolve technical issues?*	1 or more persons
4	Network connection types *What connection types and quantities?*	IEEE 802.3, 802.11, 802.15, and others
5	Network interface address *What is the low-level address of the network interface?*	Frequently a globally unique identifier
6	OS Type *What abilities does the OS have?*	Traditional, Mobile, Embedded
7	Software update frequency *How often will the manufacturer issue updates?*	High, Medium, Low
8	Lifespan of updates *How long will the manufacturer support it?*	Long, Medium, Short
9	Lifespan of host *How long does the host actually get used?*	Long, Medium, Short
10	End-user originating authentication *Can a person authenticate to a service on another host?*	True, False
11	Rate of foreign code execution *Does the host frequently execute code provided by other hosts?*	High, Medium, Low

3.2 Network Operations

The function of network operations is concerned with the measurements of latency, throughput, reliability, and scalability. The design and continued improvement of a network should aim to reduce latency and increase throughput in areas of the network where needed. This should be accomplished in a reliable and scalable manner to reduce operating costs. Many network engineers see their primary responsibility as making sure packets get from one host to another. The emphasis is on providing connectivity for the users of the network.

3.2.1 Large Organizations

For large networks, it is difficult for a single team to be responsible for all the hosts and the network equipment. This can lead the organization to a distributed management model where the network is operated more like an Internet service provider. The network operations team may have visibility down to the hosts yet may not have any access to or management responsibility for them. In a hierarchical, distributed management model, the network operations team can become further separated from the actual host maintainers. However, this model also allows for greater flexibility for different organizational units to choose host configurations which work best for their users. To get a sense of scale for a large network, Figure 3-2 shows a typical university network with approximate number of network equipment connected (L. Whitteker and K. Johnson, personal communication, April 11, 2018).

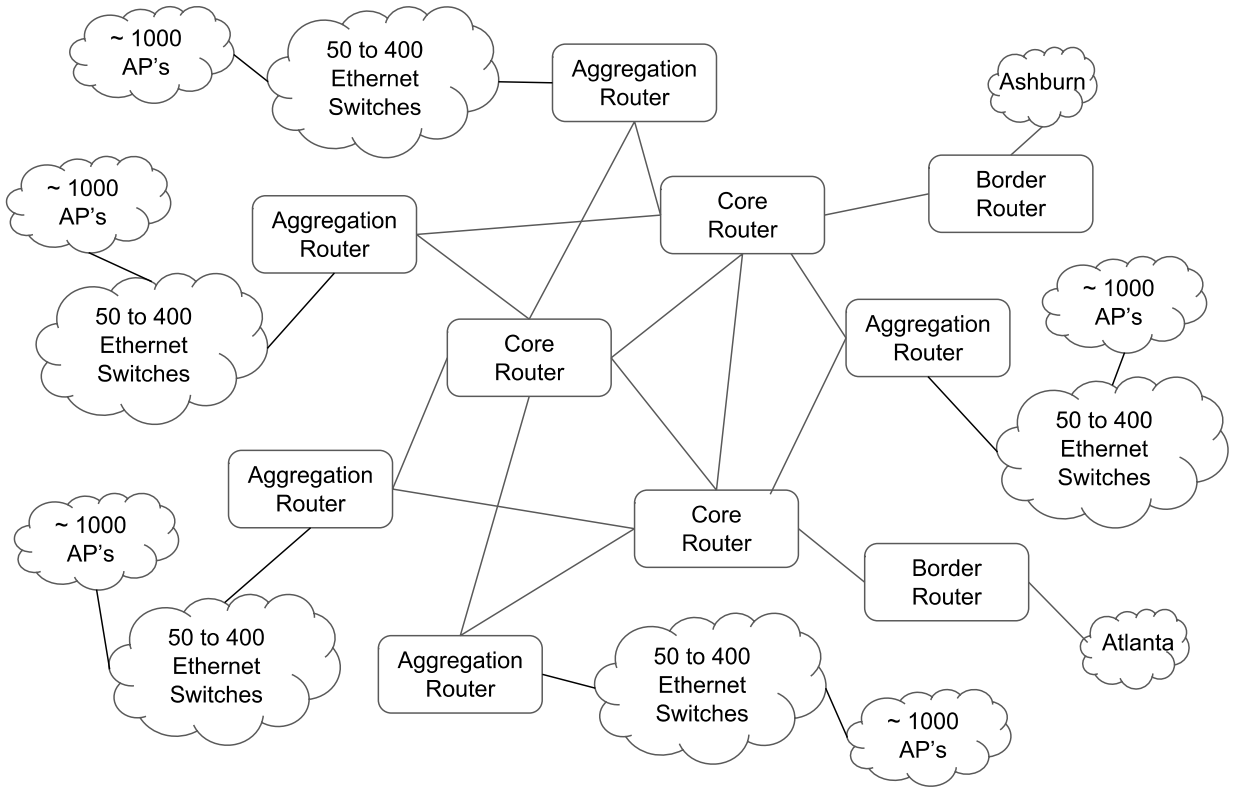

Figure 3-2: Sample University Network High-Level Diagram

3.2.2 Higher Education and Complex Architectures

Higher education networks support many and complex use cases which are disallowed on other enterprise-class networks. These networks support administrative functions, instruction, research, and residential activities. While the administrative functions closely mirror the classic enterprise network needs, they are quite different from the other categories.

Instruction necessitates students being able to explore and actively engage in the use of online technology. Networks that block or disallow certain technology may not permit the students

to learn as much or as well as they might be able to on an open network. Instructors also need flexibility to keep curriculum current and effective.

There are researchers who must use advanced software and transfer large amounts of data around the network. These researchers must also collaborate internally and with external institutions. Computer security researchers may need to operate honey pots to study malicious probing and attempts. These same researchers may configure hosts to purposefully download malware for later analysis. These activities preclude the use of oppressive protocol or content filtering.

3.2.3 Network Authentication and IEEE 802.1x

Authenticating a host to a network can assist in tracking the current state. A commonly used method of this is IEEE 802.1x and is frequently used with 802.11 and for some 802.3 deployments. It allows a host to authenticate against a RADIUS server and to allow or reject access to a specific VLAN. In a wireless deployment, 802.1x allows the host to connect to the access point and have a secure layer 2 session separate from other hosts [21].

Network authentication and IEEE 802.1x should be understood as part of a network inventory solution but incomplete on their own. In larger higher education networks, it is difficult to deploy 802.1x for all wired 802.3 connections. This is especially true for residential segments of these networks and BYOD hosts. Network Access Control (NAC) solutions that require software installed on the host can preclude some host configurations. If this software is not distributed as part of the operating system, like the many 802.1x supplicants used for 802.11, it is less likely to support a diverse set of hosts.

3.3 Cybersecurity Incident Response

Consider the following scenario. At the beginning of a normal workday, an analyst is monitoring for incidents in a security operations center. The analyst is enjoying a slow start, so they are catching up on emails from the previous day. Unfortunately, it does not take long before they see an alert from one of the institution's intrusion detection systems. The analyst is concerned because this alert is for a particularly nefarious type of malware associated with theft of personally identifiable information. As the analyst creates a ticket to begin the response process, another alert comes up. This time for a host identified with a ransomware download. The analyst recognizes the IP address as being in one of the administrative areas of the institution. The analyst knows that if the ransomware executes, it will begin encrypting the user's local files and any folders on a file server. Even if backups of the data are available, either incident could lead to data exfiltration. Now the analyst must work fast to notify responsible individuals quickly. If the tools available to the analyst cannot provide an answer to whom they should contact, or the tools provide the incorrect person, more time will be spent finding the responsible user while the malware is in control and potentially doing harm.

The cybersecurity incident response process is an accepted best practice for security operations. The process's primary objective is to reduce harm to the institution and network from bad actors. The process shown in Figure 3-3 is based on the SANS Institute Incident Handling Step-by-Step [23]

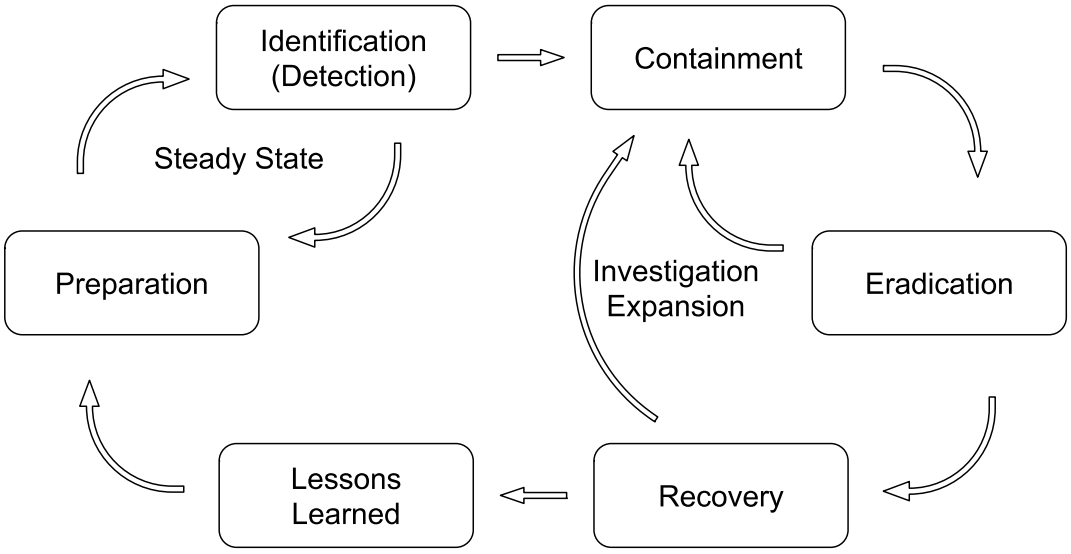

Figure 3-3: Incident Response Process

3.4 Log Analysis for Cybersecurity

The hosts on a network, and the network itself, generate substantial information in the form of log events that can be used to determine many details about host-level activity without actually knowing content. Analyzing this information by correlating across different sources allows a more complete view of the network. As the cost of storage and CPU cycles has declined, we are now in an era where data analytics can be economically applied to a greater range of problem domains.

3.4.1 Events and Sources

Logs are essential to modern network and cybersecurity operations. Most network equipment can emit log events just as hosts can log events from the operating system and applications. The basic form of an event is a timestamp coupled with the result of an action or state change. The timestamp allows alignment of events generated from different systems on a timeline.

Modern services are distributed and are interdependent on other services in the network. This makes having log events from multiple different components in a larger service vital to understanding the big picture.

The timestamp needs to be as precise as can be afforded by the system generating the event. This necessitates using Network Time Protocol (NTP) to synchronize the real-time system clocks of the network equipment.

Log events which can be collected from network infrastructure and supporting services are useful in determining host attributes. Potential sources include layer 2 and 3 equipment, authentication servers, wireless LAN controllers, DHCP servers, and NAT equipment.

3.4.2 Log Analysis Systems

Centralized collection of log events serves three key purposes. First, it archives the event when the system that generated it may not have enough storage space to keep a local copy for long. This allows for historical analysis and forensic investigations. Second, it allows for personnel and automated systems to review events in one location. The centralize collection enables greater access to the events and reduces the overhead of having multiple connections to the source. Lastly, it protects the integrity of the events by having a copy of the event off the device. If the device is compromised, an attacker could modify the local logs. Changing the event in the centralized collection would require compromising that system as well.

Many types of log analysis systems exist but most have some form of three stages. First, there is an initial ingestion and transformation of the events. Second, the events are loaded in a data store as records. Third, there is a query interface for accessing the data store. Log analysis systems can be as simple as Syslog writing to an organized set of files and using Gnu Parallels

with the Grep command. This system meets the needs of many who are already familiar with Unix-like systems. There are also many systems with graphical interfaces and more automated data processing.

Open source packaged solutions, such as Graylog, are based on Elasticsearch for data storage and indexing [24] [25]. Other solutions use either row-based relational databases or other document-oriented data stores. Larger log analysis deployments with Elasticsearch will often include Logstash for ingestion and transformation, and Kibana for visualizations and query interface.

3.5 Summary

This chapter briefly looked at concepts and technology that the reader should be familiar with to understand this research. At least a high-level comprehension of TCP/IP and networks, as deployed in larger organizations, is necessary to understand the problem of host inventory. Operational concerns for these networks, to include cybersecurity incident response, provide context for the problem while log events provide a path to a solution.

Chapter 4

Previous Solutions

If I have seen further it is by standing on ye sholders of Giants.
- Isaac Newton

In this chapter, previous ways that have addressed the problem of host inventory will be explored. Some of these solutions are partial or do not meet the needs of Hypotheses 1, 2, and 3. Some are more accurate in smaller networks or have become less accurate over time. As networks in higher education research institutions have evolved, many of these solutions cannot keep pace with the diversity, quantity, and motion of current hosts.

4.1 Human-driven Tools

For smaller host inventories, a simple record of IP addresses, MAC addresses, and assigned users can be kept in a spreadsheet or text file. However, this method presents issues with scale, accuracy, and data availability. It is relatively easy to manage perhaps a few dozen hosts in this manner, but as the number increases to hundreds keeping up with the changes becomes time consuming and error prone. It is also laborious to the point that the person responsible for keeping

the record up to date may deprioritize the effort causing an ever-increasing proportion of entries. Depending on how the records are stored, there can be accessibility issues for others needing the information in a timely manner. For example, cybersecurity operations may not have access to an inventory file maintained by a departmental systems administrator. Even with a database that can handle thousands of hosts, it still requires human data entry to maintain accuracy.

MAC address registration is also categorized and human-driven. In this system, every host has one or more MAC address recorded in a database with attributes. While providing a form of access control, it does not solve the issue of who to contact. The biggest issue with any existing human-driven inventory system is the lag-time since a record was last touched. In other words, if someone registers a host, it is quite possible that the day-to-day user is someone else. Another likely scenario is that the host has been transferred to someone else or another organization, but the inventory has not been updated. As long as the current user of the host still has network connectivity, they are unlikely to be concerned with manually updating a record. Unless there are constant checks or auditing, the inventory will be inaccurate.

4.2 Network Discovery and Authentication

To address the issue of human data entry, other techniques have been used to identify hosts on a network using the network itself. This removes the initial, human-driven element of entering host identifiers and attributes. These methods are not without limitations and can be part of a solution but are incomplete on their own.

4.2.1 Active IP Scanning

Sequential IP address scanning is easier with IPv4. Software could send a packet to every possible address in an IPv4 subnet in a few seconds. Even if the address was unused, it did not matter much since IPv4 subnets have a relatively small number of possible addresses. This is not the case for IPv6 and as such, sequential scanning will not work. Also, if the scanning is not performed on each subnet, MAC address information of the hosts will not be available.

4.2.2 Traffic Monitoring

When active scanning does not work, monitoring network traffic at routing locations in the network can yield good results. This is also referred to as passive scanning. This method involves recording the IP addresses and associated MAC addresses from traffic as it traverses from the subnet to a router [10]. It is inherently event-driven and can build an inventory in near real-time. It is transparent to the hosts but does require additional equipment and connection. The network device will receive traffic from either a tap of a physical link to the router or an available interface on the router for a mirror port. The traffic monitoring device needs to have capacity to record the frame and packets headers with minimal loss.

Implementing traffic monitoring between subnets would require a significant investment in additional equipment. Also, traffic monitoring alone cannot determine user authentications. If traffic monitoring is only at the edge of a large network, it will not provide enough information to determine the logical location in the network. The exception to this is if the host is using a link local IPv6 address without privacy extensions so that the MAC address is encoded in the Interface Identifier as described in RFC 4291 [26].

4.2.3 Network Authentication

Authenticating hosts, as they join the network, was briefly described in section 3.2.3. This can be considered a previous solution in that it is often recommended by standards, such as the CSC 1.5 [7]. In some enterprise networks or network segments with tight control over the permitted hosts, this can provide an effective solution. However, as a network grows in terms of diverse use, it becomes less feasible. This is particularly true for BYOD hosts and hierarchical, distributed network management. It is also worth noting that this solution, on its own, does not always lead to a current user or a responsible person.

4.3 Specific Implementations

Going beyond the general approaches, the following are three implementations that are noteworthy for solving at least specific parts of the problem. These implementations are foundational to the prototype system in design and as data sources.

4.3.1 Grand Unified Logging Program

Correlation of user authentication with host activity has been implemented in higher education institution networks in the form of the Grand Unified Logging Program (GULP) [27]. This system, developed at Columbia University, demonstrates that it is possible to maintain open access to a network and identify responsible users without preregistration or network authentication.

The basic concept is as follows. A host, with a user, connects to an institution’s network without authentication. The user proceeds to conduct their normal activities such as checking

Email and web browsing. At some point while connected, the user logs into a learning management system or other institutionally provided service. Columbia found that one or more user authentication to application services occurs with a high percentage of connections to the network. So much is the case that correlating the user application authentication to the host was found to be sufficient for notifications of host compromises.

This method will obviously work well for hosts with user interactivity. However, not all hosts on a network have interactive users. For example, this would not provide a potential responsible person for IoT or embedded devices such as lighting controllers.

4.3.2 NetRecon

NetRecon was originally developed at Virginia Tech by Carl Harris. The concept was similar to GULP in that there was a need to associate IP addresses with hosts for a given time. The first version, in 2004, provided mechanisms for scheduling, collecting, and parsing network facts. This was achieved by programmatically logging into network equipment via the Secure Shell (SSH) protocol and issuing commands to retrieve state from Content Addressable Memory (CAM) tables. This periodic sampling of Address Resolution Protocol (ARP) entries, Neighbor Discovery Protocol (NDP) entries, and MAC addresses seen on interfaces, provided the information necessary to trace IP addresses to hosts and physical locations. The high-level data flow is depicted in Figure 4-1.

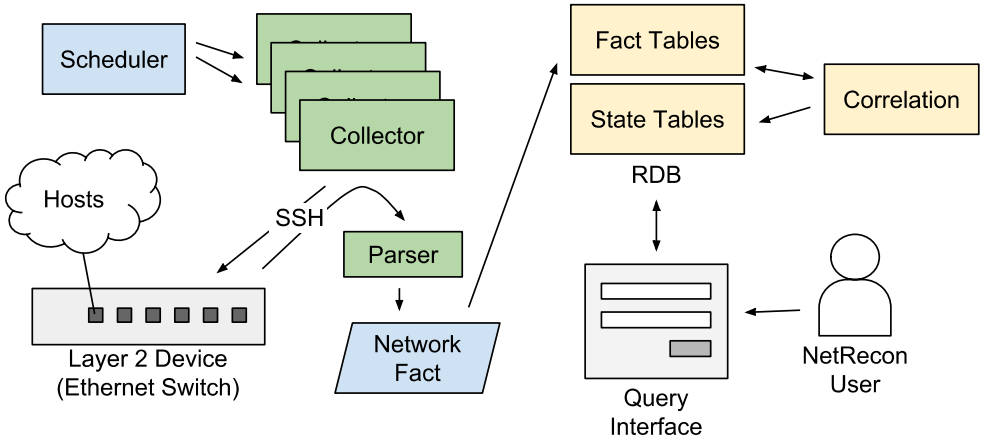

Figure 4-1: Original NetRecon Data Flow

4.3.3 Central Log Service

In 2015, the Log Archiving and Analysis initiative was untaken to centralize the collection of disparate system logs at Virginia Tech. A primary goal was to provide more relevant information to cybersecurity operations for incident response and forensics. The resulting Central Log Service (CLS), is a deployment of Logstash, Kafka, Elasticsearch, and Kibana. The data flow is show in Figure 4-2.

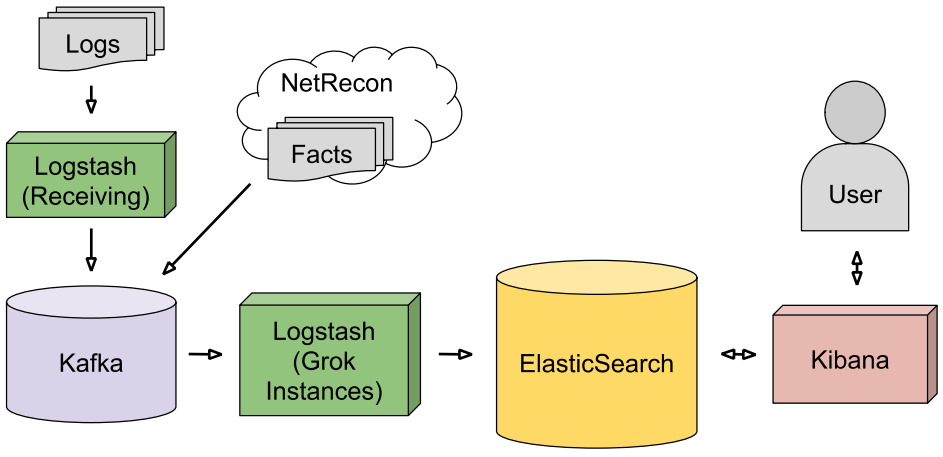

Figure 4-2: CLS Data Flow

Currently, more than 3,000 log sources are feeding the CLS. The log sources include operating system and application logs from hosts. In addition, services such as wireless, DHCP, PAT, Virtual Private Network (VPN), and RADIUS are sending events to the CLS. There are many other log sources including intrusion detection systems, firewalls, and the primary Single Sign On (SSO) service. The SSO is utilized by many applications and services for user authentication.

The current NetRecon service is sending network facts to the CLS for storage and querying. Dashboards in Kibana provide a good way to visualize the records. However, the query language of Elasticsearch does not easily permit join operations. Documents either match query conditions and are returned or they are not. This makes reviewing and querying the relationships of the network facts difficult in the CLS. Multiple queries can be performed by a CLS user to establish the relationships. However, this requires a human operator with knowledge of the logs and does not provide a way to answer queries as shown in chapter 6.

4.4 Summary

In this chapter, multiple solutions were reviewed for their benefits and limitations. On their own, each has advantages or works for a subset of hosts. Active scanning, traffic monitoring, and spreadsheets do not scale well. All of these, except NetRecon, GULP, and CLS, lose accuracy over time due to the human-driven updates required. NetRecon produces the facts to associate IP addresses with a host yet does not have the ability to associate application authentications. GULP has the features to answer the research questions but is specific to services of one university. The

CLS has a vast array of events yet does not store relationships or have a single query to produce them.

As such, the solution proposed in the next chapter is based on a combination of three of these prior works. Specifically, the concepts in GULP, NetRecon, and CLS are utilized to create a solution that answers the second and third research questions.

Chapter 5

Design of FINDIR

It's not just what it looks like and feels like. Design is how it works.
- Steve Jobs

The Frequent Inventory of Network Devices for Incident Response (FINDIR) system is based on previous work and designed to answer the second and third questions outlined in section 1.3. The design must also work with available data and within the constraints of Hypothesis 3, where there are no changes to hosts or network infrastructure, to ultimately prove or disprove it. The design is intended to help answer the research questions given the network environment of Virginia Tech. However, the design is general enough that it could be applied to other networks which are at least partially de-centralized and allow BYOD.

The main goal of the design was to utilize existing log sources to create a data-driven system that does not rely on human interaction to update state as hosts enter or leave the network or move around. The following sections describe the requirements, use cases, inputs, and resulting design.

5.1 Requirements and Constraints

The primary requirement of FINDIR is to provide one or more possible responsible persons given any type of IP address and time period. It should return all applicable association records, as shown in Table 5-1.

Table 5-1: Required Host Associations

	Host Associations
1	User network authentication from the host
2	User application authentication from the host
3	IP addresses and A or AAAA records associated with the host
4	PAT addresses and port ranges associated with the host
5	Location of host to include building information and coordinates
6	Organization name, department name, and organization head name of wired connections
7	Department network admin group contact of host
8	VPN sessions with foreign addresses and approximate locations of host
9	Operating system of host
10	Manufacturer of interface of host
11	Affiliations and department of users

Not all hosts are expected to return all host association types. The system should not expect the query from the user to specific anything other than an address and a time frame.

In addition to the above, FINDIR must not require human data entry. All data should come from log events, polled network facts, and semi-static organizational records. The system should be designed to accept events and network facts from text files. It should also be built to accept a stream of events with minimal modification.

When a record represents a time range, the system should handle a null start or a null end. This is to accommodate lost events or loading from a fixed time period, such as one day.

The technology constraints shown in Table 5-2 are imposed by networks implementing TCP/IP, IEEE 802.3, and 802.11. These constraints guide the record layout and relationships.

Table 5-2: Network Technology Constraints

	Constraint
1	A MAC address could be associated with one or more IP addresses at a time
2	An IP address could be associated with one MAC address at a time
3	A network equipment interface could be associated with one or more MAC addresses at a time
4	A user could authenticate from more than one IP address at the same time
5	A user could authenticate more than once from one IP address at the same time
6	A PAT IPv4 address and TCP port range should be associated with only one RFC 1918 IP address at a time
7	An IEEE 802.11 MAC address should be associated with only one AP at a time
8	An IEEE 802.11 user could authenticate from one MAC address multiple times for one timestamp
9	An IEEE 802.11 user could authenticate from multiple MAC addresses for one timestamp
10	An SSO user could authenticate from one IP address multiple times for one timestamp
11	An SSO user could authenticate from multiple IP addresses for one timestamp

5.2 Use Cases

FINDIR must associate multiple different records to each host. While the general use case is querying for a responsible person given an IP address, there are multiple paths FINDIR must walk to retrieve the related objects. Using Figure 5-1, the reader can trace the relationships of the events and arrive at the results for the following queries. Starting with the givens, trace the associated activity of each host.

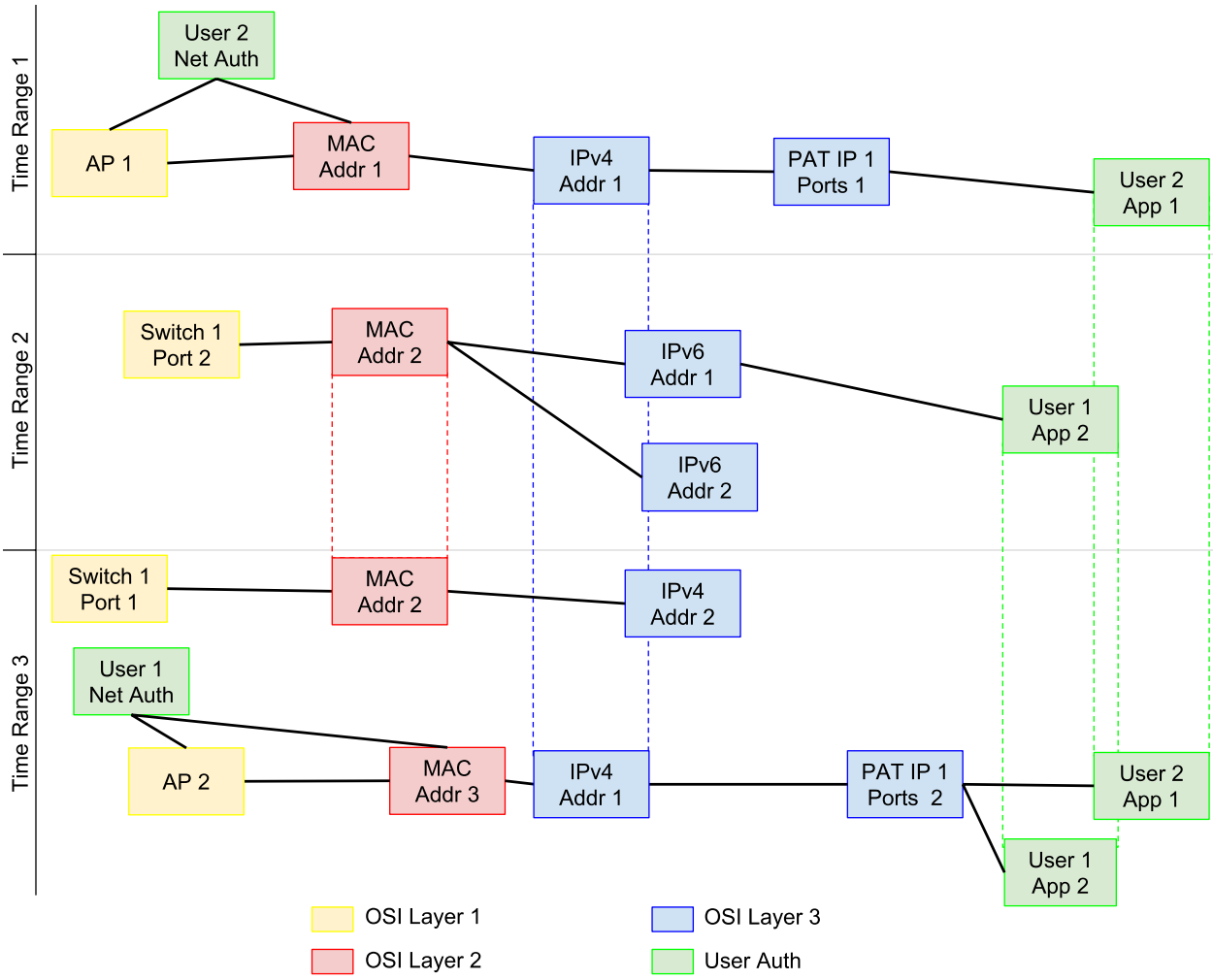

Figure 5-1: Graph Diagram of Associated Events for Tracking Hosts

1. **Query:** Given PAT Address 1 & Port (in Ports 2) & Time (in Time Range 3), what machine and who is using / responsible?

   Result: MAC Address 3, User 1 (Net) & User 2 (App)

2. **Query:** Given IPv6 Address 2 & Time (in Time Range 2), what machine, where, and who is using / responsible?

   Result: MAC Address 2, Switch 1 Port 2 location, User 1

3. **Query:** Given PAT Address 1 & Port (in Ports 1) & Time (in Time Range 1), where is the machine and who is using / responsible?

   Result: Access Point 1 location, User 2

5.3 System Inputs

In order to associate hosts with IP addresses and users, the required network events and organizational data needs to be loaded into the data store (see Table 5-3).

Table 5-3: FINDIR Input Records

	Description	Type	Generated From
1	MAC Address to Equipment Port	Association: L1 to L2	Switch CAM tables
2	IPv4 Address to MAC Address	Association: L2 to L3	Router ARP tables
3	IPv6 Address to MAC Address	Association: L2 to L3	Router NDP tables
4	PAT to IP Address Allocation	Association: L3 to L3	NAT device logs
5	PAT to IP Address Release	Association: L3 to L3	NAT device logs
6	User App Auth via SSO (CAS)	Association: L3 to user	Shibboleth logs
7	User Wireless Net Auth via 802.1x and MAC to Access Point	Association: L3 to user and L2 to L3	Radius server, Authmgr, and Station Manager logs
8	VPN authentication	Association: L3 to user	VPN logs
9	VPN tunnel session start	Association: L3 to L3	VPN logs
10	VPN tunnel session end	Association: L3 to L3	VPN logs
11	IP Address Types	Organization, semi-static	Organization webpage
12	Unit Assigned IP Ranges	Organization, semi-static	Local database extract
13	DNS A and AAAA records	Organization, semi-static	DNS zone transfer
14	Network Outlet to Department	Organization, semi-static	Local database extract
15	Equipment Interface to Outlet	Organization, semi-static	Local database extract
16	Network Equipment Location	Organization, semi-static	Local database extract
17	User Affiliations and Department	Organization, semi-static	Enterprise directory extract
18	Organization Name and Head	Organization, semi-static	Local database extract
19	Manufacturer OUI	Public, semi-static	Wireshark
20	IP Address Geolocation	Public, semi-static	MaxMind GeoLite2

The association records are generated from either events or polling of equipment state. The organization data is semi-static in that it only needs to be updated once a day. This allows the loading process to retain those records in the database to avoid the latency of using multiple external lookups for every association record.

All the associations are collected in the CLS from the NetRecon service. An extract from the CLS was used in the design and testing of FINDIR. These records could be sent in near real time by connecting FINDIR directly to the Kafka topic for each. The local databases are organizationally specific and custom. These databases are also human updated, hence the semi-static nature. A very small percentage of the total records are updated or added each day to these data sources.

Geolocation of IP addresses is useful to incident responders to help determine suspicious activity. For example, if two authentications in different locations occur are both physically possible? Public address geolocation information was provided by the GeoLite2 City dataset created by MaxMind [28]. Network equipment and connection locations provide more accurate location for the local` organization. An extract from a network operations database was used for the proof of concept.

The manufacturer OUI listing is public information maintained by the IEEE and can help categorize hosts [29]. The maintainers of Wireshark, a network traffic analysis tool, provides a list which provides consistent short names for manufacturers [30]. Knowing the manufacturer can give an idea of the type of host. It is the only piece of information we can derive from the MAC address directly.

5.3.1 Assumptions

The design makes assumptions which are necessary. As noted by the CISO survey responses, we assume that a MAC address is a one-to-one relationship with a host. It is used as the unique host identifier in the FINDIR database. If an actual host has more than one MAC address it will show up in FINDIR as more than one host.

In addition, the input data was limited to one day to reduce the time required for data loading during development and testing. The design should work with a greater time period in the database tables or with a separate set of tables for each day. The next chapter discusses the evaluation using this dataset.

5.3.2 Privacy Preservation

Privacy is not often considered when operational tools are developed. In this case, FINDIR was developed using the privacy preserving architecture defined by DeYoung, et al. [31]. In this architecture, FINDIR was developed using de-identified sample data. The last half of every MAC addresses was replaced with a pseudonym. All usernames were replaced with pseudonyms. This process was accomplished with symmetric encryption and a consistent key across all records. The result is that users of FINDIR do not know actual users or MAC addresses. However, in actual operational use, analysts with appropriate access could re-identify as necessary.

5.3.3 Built on Previous Work

NetRecon currently exists as more of a network state data collection system than a complete tool for querying hosts. It polls network equipment, collects facts, and transmits them to

the CLS. This provides an intuitive interface for ad-hoc queries which match a condition but does not establish the appropriate relationships. FINDIR builds on NetRecon by adding enrichment data useful to incident responders. GULP established the concept of correlating a user application authentication with a known host. This is expected to be beneficial to networks which operate applications in addition to the network. The proof of concept of FINDIR is designed based on the ideas of these three systems.

5.4 Data Flow and Architecture

FINDIR has two input types: events and initialization data. The events include network facts generated by NetRecon. The initialization data is loaded from the semi-static sources before any events are loaded. This data is intended to be refreshed daily or at some interval appropriate to the organization and network environment.

Through a combination of direct SQL queries and stored procedures, the initialization data is loaded prior to events being inserted into associative tables. The stored procedures are necessary to conditionally update multiple tables (see Figure 5-2). Object tables represent unique entities such as IP addresses, hosts (MAC addresses), and locations. A simple query interface is provided to execute stored procedures for returning records with matching time period and are associated.

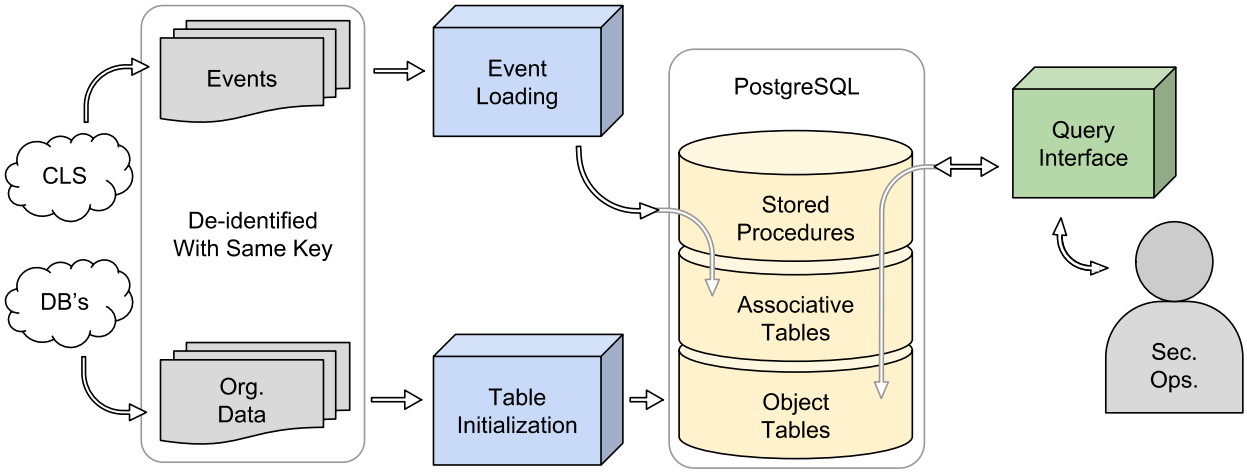

Figure 5-2: Data Flow

5.5 Data Storage

Multiple ways of storing this data was researched to include document stores, distributed columnar stores, and relational databases [32]. Additionally, keeping all the records in key value pairs as Python dictionaries was also considered. This idea was not pursued based on the additional development and concerns for exceeding the scope of the research. The design choice to use a relational database was based on avoiding the limitations of other technologies and unnecessary complexity. Many decades of development have produced relational database systems which are mature and utilize a convenient interface such as Structured Query Language (SQL). SQL allows for a greater focus on the logic of associating the appropriate events and not the indexing, storage, or data access.

Opting for a relational design means that the association records from Table 5-3 are associative or junction records. The design followed the best practices for database normalization as described by Coronel and Morris in [33]. The tables were structured as shown in Figure 5-3.

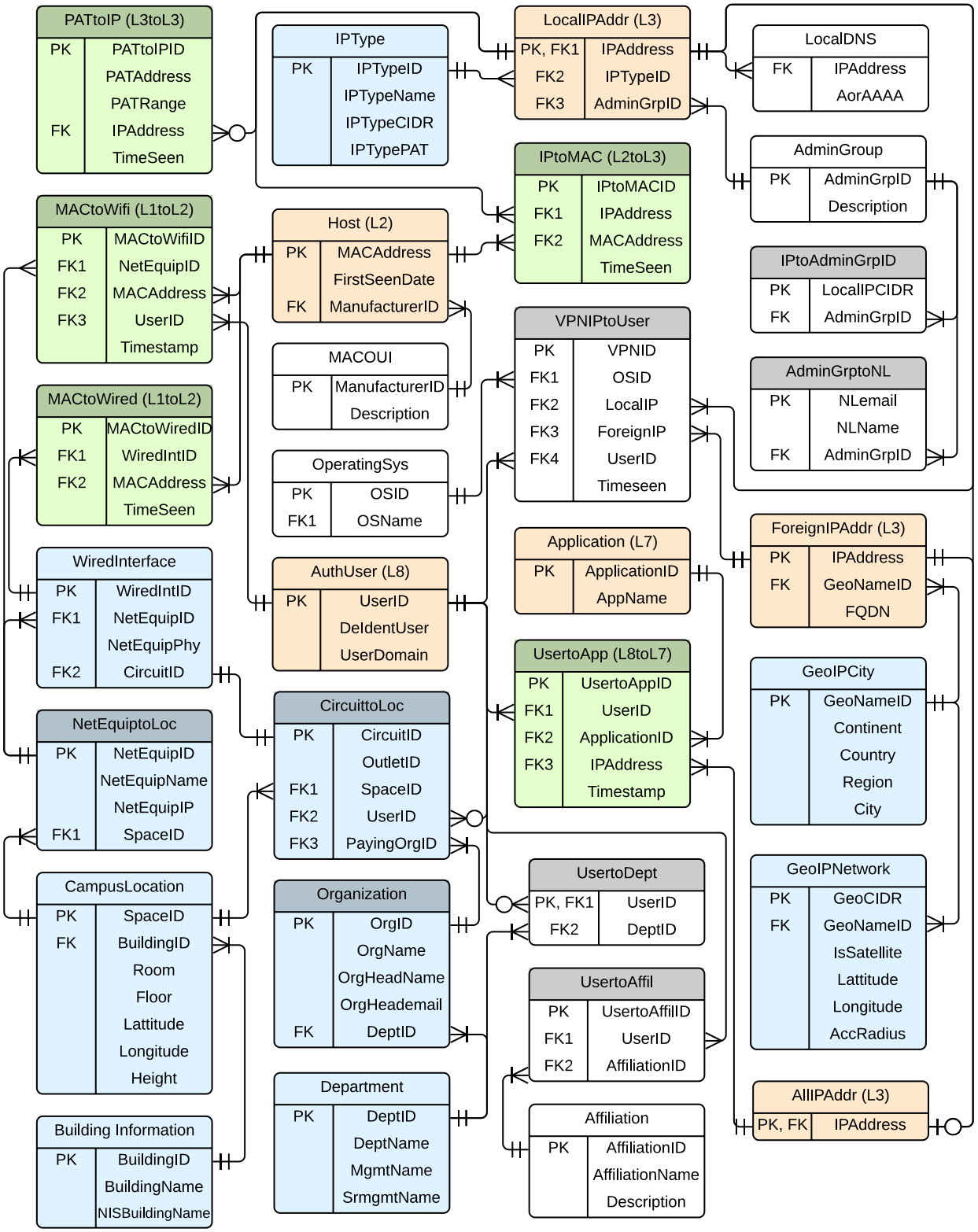

Figure 5-3: FINDIR Entity Relationship Diagram

In the diagram, green tables represent the input records while orange are generated. Blue tables are semi-static records which were loaded infrequently. White tables are VPN records and additional semi-static records for further enrichment which were not implemented for the proof of concept.

PostgreSQL was used as the relational database due to its built-in types for IP addresses, MAC addresses, and time ranges. The latter being important when querying events with overlapping time periods. The concept of ranges inside PostgreSQL provides operators for determining points contained within a time period and overlapping periods. PostgreSQL is an open source project with an emphasis on standards compliance and two decades of development [34].

5.6 Programming Environment

The FINDIR proof of concept was implemented in Python 3.6.1 using minimal external libraries [35]. Psycopg 2.7.4 was used as the driver for PostgreSQL [36]. Python enabled an easily modifiable and concise code base. The language PL/pgSQL was used for stored procedures in the FINDIR database. This is the native stored procedure language in PostgreSQL.

The complete Python code for FINDIR along with the input data structure is provided in Appendix B. The design and code are based on some organizationally specific data, although effort was taken to generalize the implementation.

5.7 Summary

FINDIR, as a concept, is based on NetRecon and GULP. The design and implementation are intended to extend these prior solutions with the inclusion of additional data. It fits within the technology and Hypothesis 3 constraints. It is also data-driven, in that the system only needs the input events and semi-static data. As we will see in chapter 6, FINDIR removes the human-driven element in other solutions and ultimately is designed to be a tool for answering research questions in section 1.3.

Chapter 6

Evaluation

We have to do the best we know how at the moment . . .;
If it doesn't turn out right, we can modify it as we go along.

- Franklin Delano Roosevelt

The FINDIR system was tested for functional use, accuracy, and the authentication associations. Each of these, impact the potential for the system to be used in real world operations. The second and third research questions stated in 1.3 are the most important aspects of the evaluation. The two core questions are does FINDIR improve the accuracy of a host inventory and is the method of associating application authentications with a host beneficial in finding a responsible person.

The experimental setup outlines the nine evaluation tests, two environments, and input data. The initial load times and table counts provides a baseline for future use. The last section provides the evaluation results with insights for interpretation.

6.1 Experimental Setup

Ten evaluations were chosen to answer the questions: does the system work as intended and does the system provide benefits to a host inventory. The following table outlines each of these and the groups them by type (see Table 6-1).

Table 6-1: Evaluations and Types

#	Type	Test	Dataset
1	Functional	Which hosts are associated with a given PAT IP address and time range?	2017-11-15
2	Functional	Where is the host physically located with a given IPv4 address and time range?	2017-11-15
3	Functional	Who is associated with a host with a given IPv6 address and time range?	2017-11-15
4	Authentication Association	What is the percentage of hosts without an application authentication association?	2017-11-15
5	Authentication Association	What is the average number of application users associated with each host by OUI?	2017-11-15
6	Authentication Association	What is the average number of users associated with each host by connection type?	2017-11-15
7	Authentication Association	What is the number of hosts with an application user association but without a wireless user association?	2017-11-15
8	Authentication Association	What is the number of hosts with either an application user, wireless user, or organization association?	2017-11-15
9	Accuracy	What is the number of control set hosts associated with the correct department?	2017-11-15, Control Set

6.1.1 Test Environments

Two consistent environments were setup and used for testing. The FINDIR proof of concept was tested on an Apple MacBook Air and a virtual machine on a Dell PowerEdge server. Specifications are shown in Table 6-2.

Table 6-2: Test Machine Specifications

Specifications		Environment 1 (MacBook Air)	Environment 2 (VM on Dell PowerEdge R620)
CPU	Manufacturer	Intel	Intel
	Model	Core i7 4650U	Xeon E5-2670
	Frequency	1.7 to 3.3 GHz	2.6 to 3.3 GHz
	Processors	1	2
	Cores	2	8
	Threads	4	16 per processor (16 for VM)
	L3 Cache	4 MB shared	20 MB shared
Memory	Type	DDR3	DDR3
	Capacity	2 x 4 GB, 8 GB total	12x 8GB (32GB for VM)
	Frequency	1600 MHz	1600 MHz
Storage	Manufacturer	Apple	Dell
	Model	SD0256F	H710P
	Type	PCIe SSD	SAS RAID 5
	Capacity	256 GB	6 x 600GB 10k (20 GB VM)
Operating System	Version	Mac OS 10.12.6	XenServer 7.2 (CentOS 7 VM)
	File system	Journaled HFS+	XFS VM

6.1.2 Test Data Sources

The tests were performed with previously de-identified data that were extracted from the CLS, organization specific databases, and publicly accessible websites. The datasets cover one day of actual network generated data and are shown in Table 6-3.

Table 6-3: Test Data Sources

	Input Source	Input Type	Number of Records	CSV File Size (MB)
1	Virginia Tech Buildings	Semi-static	398	0.010
2	Virginia Tech Interior Spaces	Semi-static	26,679	1.438
3	Virginia Tech Department Listing	Semi-static	1,559	0.256
4	Virginia Tech IP Address Blocks	Semi-static	52	0.002
5	Virginia Tech Circuits and Outlets	Semi-static	38,623	3.674
6	Virginia Tech Access Point Locations	Semi-static	5,043	1.115
7	GeoLite2 City Locations English	Semi-static	103,009	9.084
8	GeoLite2 City Blocks IPv4	Semi-static	2,662,733	162.710
9	GeoLite2 City Blocks IPv6	Semi-static	2,039,865	130.241
10	MAC Address to Wired Interfaces	Association	8,314,947	642.471
11	MAC Address to AP and User	Association	3,266,438	402.041
12	IPv4 Address to MAC Address	Association	8,710,565	691.733
13	IPv6 Address to MAC Address	Association	4,998,397	488.320
14	PAT Address and Ports Allocated	Association	427,359	26.336
15	PAT Address and Ports Released	Association	427,582	26.350
		Totals:	31,023,249	2,585.781

There are approximately 26 million association records. That averages to 301 events per second across the entire day for this sample dataset. This is useful for comparing the rate at which events can be processed. Even with buffering, events must be processed at this rate with a margin for surges in event flow.

6.2 Initial Data Load and Record Counts

FINDIR was initially evaluated by loading the datasets individually. Any load errors were corrected. These errors were due to missing values or invalid data for a specific type. For example, a record might have an invalid MAC address or IP address which appeared truncated. These were rare in the 26 million records but the FINDIR code had to be corrected to handle these invalid inputs.

6.2.1 Load Times

A batch loading function was included as part of FINDIR. This function also recorded wall time for starts and ends of file loads. This allows for a benchmark to be established for each environment. The load times for Environment 1 and Environment 2 are shown in Table 6-4 and Table 6-5 respectively.

Table 6-4: FINDIR Load Times for Environment 1

	Input Source	Load Time (Seconds)	Number of Records	Events Per Second
1	Virginia Tech Buildings	0.092826	398	4,287.591838
2	Virginia Tech Interior Spaces	8.910464	26,679	2,994.120172
3	Virginia Tech Department Listing	0.658690	1,559	2,366.818989
4	Virginia Tech IP Address Blocks	0.030767	52	1,690.122534
5	Virginia Tech Circuits and Outlets	17.163446	38,623	2,250.305679
6	Virginia Tech Access Point Locations	1.671039	5,043	3,017.882886
7	GeoLite2 City Locations English	30.981187	103,009	3,324.888746
8	GeoLite2 City Blocks IPv4	817.091365	2,662,733	3,258.794688
9	GeoLite2 City Blocks IPv6	650.421038	2,039,865	3,136.222356
10	MAC Address to Wired Interfaces	4,971.800949	8,314,947	1,672.421540
11	MAC Address to AP and User	1,500.798424	3,266,438	2,176.466838
12	IPv4 Address to MAC Address	7,620.210574	8,710,565	1,143.087178
13	IPv6 Address to MAC Address	2,827.100715	4,998,397	1,768.029336
14	PAT Address and Ports Allocated	178.262756	427,359	2,397.354386
15	PAT Address and Ports Released	181.548700	427,582	2,355.191747
		Total: 18,806.74294 (~ 5 hours, 13 minutes)		Average: 2,522.619928

Table 6-5: FINDIR Load Times for Environment 2

	Input Source	Load Time (Seconds)	Number of Records	Events Per Second
1	Virginia Tech Buildings	0.060689	398	6,558.025342
2	Virginia Tech Interior Spaces	6.135022	26,679	4,348.639663
3	Virginia Tech Department Listing	0.401336	1,559	3,884.525684
4	Virginia Tech IP Address Blocks	0.022258	52	2,336.238656
5	Virginia Tech Circuits and Outlets	15.665444	38,623	2,465.490285
6	Virginia Tech Access Point Locations	1.228470	5,043	4,105.106352
7	GeoLite2 City Locations English	19.119697	103,009	5,387.585379
8	GeoLite2 City Blocks IPv4	529.895985	2,662,733	5,025.010710
9	GeoLite2 City Blocks IPv6	395.465825	2,039,865	5,158.132185
10	MAC Address to Wired Interfaces	3,841.291458	8,314,947	2,164.622781
11	MAC Address to AP and User	1,116.268704	3,266,438	2,926.211214
12	IPv4 Address to MAC Address	5,045.071100	8,710,565	1,726.549503
13	IPv6 Address to MAC Address	5,244.415058	4,998,397	953.089514
14	PAT Address and Ports Allocated	113.409740	427,359	3,768.274224
15	PAT Address and Ports Released	121.477921	427,582	3,519.833040
		Total: 16,449.9287 (~ 4 hours, 34 minutes)		Average: 3,621.822302

In the loading tests, Environment 2 was 25% faster than Environment 1 when averaged across all records. Using the dataset average mentioned in section 6.1.2 (301 events per second), it is safe to assume that a single processor machine with reasonable specifications (2,265) can keep up with potential surges in events.

6.2.2 Table Counts

After the initial load, the FINDIR database contained 137,319 hosts (MAC addresses) and 302,191 IP addresses. The total on-disk size was 4,459 MB as reported by PostgreSQL. This was 72% larger than the original text files, but includes the indexes used for building time ranges and faster queries.

It is worth noting the difference for using time ranges rather than the individual polled records in the database. For the three polled association types, there is roughly a 70 percent reduction in the number of needed records to be stored.

Table 6-6: Records Needed for Polled Inputs

	Input Source (Polled Every 5 Minutes)	Number of Polled Records	Number of Time Range Records	Percentage of Polled Records
10	MAC Address to Wired Interfaces	8,314,947	2,499,804	30.0%
12	IPv4 Address to MAC Address	8,710,565	3,854,342 (combined)	28.1%
13	IPv6 Address to MAC Address	4,998,397		

The other tables in the database either match the input records or are slightly reduced due to missing values and those records being discarded. Examples include a missing IP address, MAC address, or equipment name. Without these key pieces of information, the rest of the record is not useful.

It is also worth noting the types of locally seen, non-global IP addresses (see Table 6-7).

Table 6-7: Locally Seen, Non-Global IP Addresses by Type

CIDR	Description	Number of Addresses
fe80::/10	Link Local	35321
172.30.0.0/16	General Wireless (On-campus use only)	32444
172.29.0.0/16	General Wireless (On-campus use only)	32431
172.31.0.0/17	Residential wireless (On-campus use only)	27763
172.16.0.0/12	Undefined Unique Local Addresses	11731
172.18.0.0/16	Unified Communications (On-campus use only)	7593
172.24.0.0/16	Residential wired (On-campus use only)	1546
172.21.0.0/16	General campus wired (On-campus use only)	1083
172.27.0.0/16	Remote access - VPN (VT traffic over VPN) service	365
172.26.0.0/16	RLAN (On-campus use only)	148
10.0.0.0/8	Undefined Unique Local Addresses	48
169.254.0.0/16	Link Local	17
192.168.0.0/16	Undefined Unique Local Addresses	4
No Specific CIDR	Foreign Address Seen Locally (misconfiguration)	26

Most of the IP addresses in these types are previously defined. The exception to this includes Undefined Unique Local Addresses that are in use but may or may not be managed centrally by network operations. Nonetheless, these are associated with hosts and contacts in FINDIR. These will be explored in the next section. The last row in Table 6-7 is purposefully a misnomer. These are IP addresses that are not part of the local network. These hosts are likely misconfigured or interacting only within a subnet and the users have no need for actual Internet connectivity. These are also associated with hosts and responsible people.

In addition to the local IP addresses, 14,309 foreign addresses were collected from user authentications to the SSO. Of these, 98.4% are from the United States. The remaining 1.6% (230 IP addresses) are from 56 countries, with 125 IP addresses from North American and European countries.

6.3 Results and Insights

The following sections are the results of the evaluation scenarios in Table 6-1. If a scenario used an IP address, it was randomly selected from the appropriate table.

6.3.1 Hosts Associated with a PAT IP Address

An SQL query was created to randomly select a PAT IP address and return all the associated local IP addresses and hosts. This was intended to show that the database could return the RFC 1918 addresses and associated hosts. This query was not time bounded given that there was only one day's worth of data in the database. The results of the query are shown in Table 6-8.

Table 6-8: Hosts Associated with PAT IP Address

PAT TCP Port Ranges	Local IP Address	Hosts (MAC Addresses) Associated
[1537,1793)	172.21.5.175	10:9a:dd:xx:xx:05
[1025,1281)	172.29.112.199	e0:5f:45:xx:xx:23
[1025,1281)	172.29.117.231	30:63:6b:xx:xx:09
[2049,2305)	172.29.127.226	10:f1:f2:xx:xx:06
[2049,2305)	172.29.13.15	98:fe:94:xx:xx:16, ec:0e:c4:xx:xx:28
[1281,1537)	172.29.25.235	30:59:b7:xx:xx:08
[1537,1793)	172.29.40.128	50:82:d5:xx:xx:0c, e4:b3:18:xx:xx:27
[1281,1537)	172.29.41.129	00:db:70:xx:xx:02
[2049,2305)	172.29.63.136	e4:b3:18:xx:xx:25
[1025,1281)	172.29.84.140	c0:33:5e:xx:xx:1d, cc:20:e8:xx:xx:1f, d0:25:98:xx:xx:20
[1025,1281)	172.29.84.140	cc:20:e8:xx:xx:1f
[1793,2049)	172.30.1.90	80:e6:50:xx:xx:14, 90:8d:6c:xx:xx:15
[1281,1537)	172.30.106.145	0c:51:01:xx:xx:04, ac:bc:32:xx:xx:1a, b8:8a:60:xx:xx:1c
[1025,1281)	172.30.18.6	28:a0:2b:xx:xx:07
[1537,1793)	172.30.33.42	c0:33:5e:xx:xx:1e
[1793,2049)	172.30.4.234	78:9f:70:xx:xx:13, f8:59:71:xx:xx:2a
[1025,1281)	172.30.45.233	64:b0:a6:xx:xx:0f, 70:81:eb:xx:xx:10, d0:a6:37:xx:xx:21
[1025,1281), [1281,1537), [1793,2049)	172.30.61.249	60:92:17:xx:xx:0e
[1537,1793)	172.30.74.203	ac:37:43:xx:xx:19, e4:9a:79:xx:xx:24
[1281,1537)	172.30.74.84	a8:be:27:xx:xx:18, f4:0f:24:xx:xx:29
[1281,1537)	172.30.79.81	a4:b8:05:xx:xx:17, dc:0c:5c:xx:xx:22
[2561,2817)	172.30.82.249	78:31:c1:xx:xx:11
[2305,2561)	172.30.84.149	00:ae:fa:xx:xx:01, 5c:f7:e6:xx:xx:0d
[1025,1281)	172.30.91.224	48:3b:38:xx:xx:0a, ac:fd:ce:xx:xx:1b, e4:b3:18:xx:xx:26
[1793,2049)	172.30.92.231	4c:7c:5f:xx:xx:0b
[1281,1537)	172.31.208.53	04:4b:ed:xx:xx:03

The results show 26 different local (RFC 1918) IP addresses associated with the PAT IP address and PAT TCP port ranges assigned. There were 42 different hosts, and, in many instances, hosts used the same local IP address at a different time in the day. In one instance, a single host with one local IP address was associated with three PAT TCP port ranges.

The functional test of associating a PAT IP address with hosts is shown to work with FINDIR. As Table 6-8 shows, knowing the PAT TCP port of interest and a time window will narrow down the host of interest.

6.3.2 Hosts Physically Located with an IPv4 Address

The randomly selected IP addresses were selected from ranges used for wired and wireless connections. The results are shown in Table 6-9 and Table 6-10. The organization (a subunit of a department), the organization head name, room number, Latitude, and Longitude were omitted from Table 6-9 to save space and preserve privacy. The room number was omitted for the wireless hosts.

Table 6-9: Wired Host Physical Location Given an IP Address

IP Address	Host (MAC Address)	Division	Department	Building
128.173.xx.1	78:2b:cb:xx:xx:01	College of Eng.	CS	Knowledgeworks
128.173.xx.2	10:dd:b1:xx:xx:02	Vice President - IT	TLOS	Litton-Reaves
128.173.xx.3	f4:8e:38:xx:xx:03	College of Science	Physics	Derring Hall

Table 6-10: Wireless Host Physical Location Given an IP Address

IP Address	Host (MAC Address)	Building	Latitude	Longitude
172.29.xx.1	50:82:d5:xx:xx:01	Torgersen Hall	-80.420xxxxx	37.229xxxxx
172.29.xx.2	74:8d:08:xx:xx:02	Squires Student Center	-80.418xxxxx	37.229xxxxx
172.29.xx.3	bc:9f:ef:xx:xx:03	Robeson Hall	-80.425xxxxx	37.228xxxxx

Even though the test data was de-identified, the results shown here were further obfuscated to eliminate any potential re-identification with an actual person. These results show that given an IP address, we can return a location that includes geolocation and for wired, a department.

The design includes the user affiliations to include, a department for faculty and staff. However, this was not implemented for the proof of concept. If implemented, the department attribute would be included for wireless authentication of faculty and staff.

6.3.3 Users Associated with a Host and an IPv6 Address

Using one randomly selected IPv6 address, the database was queried for associated hosts and users. The results are shown in Table 6-11.

Table 6-11: User Associated with a Host and IPv6 Address

IP Address	Host (MAC Address)	Network User	Application User
2607:b400:26:7fc:x:x:x:1	f8:59:71:xx:xx:01	7c2_VEkIlzjOl3YI9a	oapW4UrLLtltaJr5C

In running this evaluation, it was determined that the test data did not include all IPv6 to MAC address associations for wireless. Only one association was found.

6.3.4 Hosts without an Authentication Association

There were only 11,011 unique hosts with an SSO application authentication. Of those hosts, 1,815 were wired. With 34,102 total wired hosts, that is only 5.3 percent. However, with the assumption that there are missing IP to MAC address associations, the number of reported hosts without an authentication association is not accurate.

To better determine the extent of the missing associations, the number of IP addresses seen in the IP to MAC address association table was compared to the local IP address table. 218,622 unique IP addresses were found in the IP to MAC address association table.

27,978 IP addresses were found in the IP to MAC table that were also in the MAC address to wired interface tables. That is a difference of 6,124 missing associations for wired hosts. That leads to the conclusion there are at least 18 percent missing associations.

59,644 of the IP to MAC associations are seen with the wireless hosts. The MAC address to wireless associations have 60,492 unique hosts. That is a difference of 848 or 1.4 percent.

However, we would expect a modern wireless host to have multiple IP addresses. From individual observations, a wireless host will usually have an IPv6 link local address, IPv6 global address, and an IPv4 DHCP address. Sometimes additional IPv6 addresses will be used if privacy extensions are enabled. This would also lead to the conclusion that IPv6 to MAC address associations are missing from the test data set.

6.3.5 Application Users Associated with each Host by OUI

This evaluation allows the comparison of hosts by the manufacturer OUI that have application user associations. The 2,705 OUI values in the test data are registered by 423 manufacturers. Only the top 15 manufacturers were chosen since they represent 91 percent of the hosts. The results are shown in Table 6-12.

Table 6-12: Top 15 Host OUI with Unique Application Users

Manufacturer	All Hosts	≥ 1 User / Host	% ≥ 1 User / Host	Unique Application Users per Host				
				Min	Max	Mean	Std Dev	Var
Apple, Inc.	58998	3311	6%	0	5	0.058	0.242	0.058
Intel Corporate	14325	2324	16%	0	27	0.170	0.469	0.220
Dell Inc.	11613	2566	22%	0	19	0.250	0.559	0.313
Aruba Networks	9270	0	0%	0	0	0.000	0.000	0.000
Avaya Inc.	7836	0	0%	0	0	0.000	0.000	0.000
Microsoft Corporation	5460	1449	27%	0	28	0.286	0.629	0.396
Samsung Electronics Co.	3643	135	4%	0	2	0.038	0.195	0.038
Murata Manufacturing Co.	3611	162	4%	0	2	0.046	0.217	0.047
Hon Hai Precision Ind. Co.	1931	243	13%	0	2	0.046	0.217	0.047
Cisco Systems, Inc	1669	9	1%	0	4	0.012	0.183	0.033
Liteon Technology Corp.	1433	172	12%	0	3	0.125	0.347	0.120
LG Electronics Inc	1383	76	5%	0	2	0.057	0.241	0.058
Motorola Mobility LLC, a Lenovo Company	1362	52	4%	0	2	0.040	0.208	0.043
HTC Corporation	989	41	4%	0	1	0.041	0.199	0.040
Hewlett Packard	968	9	1%	0	1	0.009	0.096	0.009

6.3.6 Wireless Users Associated with each Host by OUI

This evaluation allows the comparison of hosts by the manufacturer OUI that have wireless user associations. The results are shown in Table 6-13.

Table 6-13: Top 15 Host OUI with Unique Wireless Users

Manufacturer	All Hosts	≥ 1 User / Host	% ≥ 1 User / Host	Unique Wifi Users per Host				
				Min	Max	Mean	Std Dev	Var
Apple, Inc.	58998	36977	63%	0	3	0.628	0.486	0.236
Intel Corporate	14325	9369	65%	0	2	0.657	0.481	0.231
Dell Inc.	11613	0	0%	0	0	0.000	0.000	0.000
Aruba Networks	9270	0	0%	0	0	0.000	0.000	0.000
Avaya Inc.	7836	0	0%	0	0	0.000	0.000	0.000
Microsoft Corporation	5460	3014	55%	0	3	0.558	0.508	0.258
Samsung Electronics Co.	3643	2222	61%	0	2	0.612	0.491	0.241
Murata Manufacturing Co.	3611	2324	64%	0	2	0.645	0.481	0.232
Hon Hai Precision Ind. Co.	1931	1195	62%	0	2	0.621	0.489	0.240
Cisco Systems, Inc	1669	0	0%	0	0	0.000	0.000	0.000
Liteon Technology Corp.	1433	993	69%	0	2	0.696	0.467	0.218
LG Electronics Inc	1383	843	61%	0	2	0.610	0.489	0.239
Motorola Mobility LLC, a Lenovo Company	1362	874	64%	0	1	0.642	0.479	0.230
HTC Corporation	989	649	66%	0	1	0.656	0.475	0.226
Hewlett Packard	968	0	0%	0	0	0.000	0.000	0.000

6.3.7 Hosts with an Application User and without a Wireless User

This evaluation provides the number of hosts which we can associate with a current, active user without a wireless authentication (e.g., wired hosts). Out of the 137,319 hosts, 11,011 had an application user associated, and 60,492 had a wireless user associated. 6,183 hosts were found to have an application user associated but not wireless user associated. This means that 6,183 hosts would have a current, active user as a possible responsible contact. These hosts are a 4.5 percent increase in the total number of hosts with a user association.

6.3.8 Hosts with any User or Organization Association

This evaluation determines how many hosts have any of the potential responsible person associations that were implemented in the proof of concept. These include application users, wireless users, and the organization that pays for wired network access. The result was 90,930 hosts with any one of the three associations. Given that there are missing IP to MAC address associations, this result would likely be higher with more complete input data.

The network equipment manufactured by Aruba, Cisco, and Avaya can be excluded from the total hosts. Even with the missing data, the 90,930 associated hosts are 76.7 percent of the 118,544 hosts (without known network equipment).

6.3.9 Control Set Hosts Associated with User and Location

To validate that the associations represent reality, a control set of five hosts was used. Each known host MAC address and known associated user needed to be de-identified with the same key as the existing dataset. The queries returned the results shown in Table 6-14

Table 6-14: Control Set Hosts Associated with Users and Locations

Host (MAC Address)	IP Address	Application User	Wireless User	Building
ac:d1:b8:xx:xx:01	172.30.x.1	kZ1y0pJ	kZ1y0pJ	Torgersen
ac:d1:b8:xx:xx:01	2607:b400:25:x:x:x:x:2	kZ1y0pJ		
00:24:d6:xx:xx:02	2001:468:c80:x:x:x:x:3	saGG9Eh		
64:76:ba:xx:xx:03	172.29.x.4	pTtuX_S	pTtuX_S	AISB
64:76:ba:xx:xx:03	172.29.x.4	pTtuX_S	pTtuX_S	Torgersen
64:76:ba:xx:xx:03	2600:1003:b86d:x:x:x:x:5	pTtuX_S		
ec:1f:72:xx:xx:04	2001:468:c80:x:x:x:x:6	f2Tn5P6		
64:76:ba:xx:xx:05	172.30.x.7		uy1dlYR	Torgersen

The exact locations were omitted but used to verify that the hosts were in the correct location. In addition, the user associations were verified to be correct based on the known users. As previously stated, missing records in the IPv6 to MAC address associations has limited the results. However, as shown in Table 6-14, three out of the five hosts had records which corresponded with the known location and users.

6.4 Summary

The evaluations conducted were intended to test the FINDIR proof of concept and help answer the research questions. The results show that the FINDIR system is functional for the given test data. The results also show that the system is accurate for the given control set.

It was also found that there are data quality issues with the test set used. This affected the ability to have reliable results for supporting Hypothesis 2. Even so, there were small yet significant benefits to associating hosts with application authentications. This was due to the increased number of hosts with a potential responsible person.

Chapter 7

Conclusion

This is not the end, this is not even the beginning of the end, this is just perhaps the end of the beginning.

- Winston S. Churchill

Researchers and students of higher education institutions perceive access to the Internet as fundamental to the pursuit of knowledge. This continues to be threatened by unnecessary cybersecurity policies that limit access due to external and internal threats. If higher education institutions are to maintain open access for their personnel, methods are needed to ensure expedient incident response. This includes being able to notify personnel of threats to their hosts and data.

The cybersecurity incident response process requires timely action. This includes detection of a potential incident and beginning the response process. The time between detection and containment can be longer than desired if the host inventory is incomplete or inaccurate. If a higher education institution desires to maintain open access to the Internet, it is necessary to quickly identify a responsible person for each host.

7.1 Summary of Research

In this research, three main questions were explored. Each of these questions are addressed in the following sections.

7.1.1 Current Host Inventory Controls in Higher Education

What are the current host inventory controls used in higher education networks?

By surveying CISO's of higher education institutions, it was found that others have concerns about their ability to identify locations and responsible persons for each host. There are also concerns about accuracy of their inventories (see Chapter 2). The CISOs also provided their thoughts on current methods in use at their institutions. These methods were further explored in Chapter 4 by identifying benefits and limitations.

The concept of what a host is and how they can be categorized in a network was also provided in Section 3.1. The definition of a host was derived from the survey results and literature review. In summary, a host is a subtype of network devices and accepted to be identified by one or more MAC addresses.

7.1.2 Improve the Accuracy of Host Inventory Controls

Can we improve the accuracy of host inventory controls using existing data sources?

The evaluation of FINDIR showed that there are benefits to enriching a host inventory with organizational data, such as location information. This can be in the form of known locations for network equipment and physical network connections. In addition, associating these locations with departments provided another potential contact.

While the FINDIR system was implemented as a proof of concept and evaluated with only one day's worth of events, it could be deployed with no manual updates. The system was designed to be driven from existing data sources and provides the benefit of utilizing up-to-date information. The system also handles the different events individually so that the database is updated immediately when they arrive. This removes the need to batch process incoming events and provides a data-driven host inventory.

7.1.3 Associating an Application Login with a Host

Is the method of associating an application login with a host beneficial for finding a responsible person?

As shown in the evaluation, there is a benefit to associating an application authentication with hosts. A greater number of potential contacts allows for better visibility. This is particularly applicable with institutionally owned hosts which may be using sensitive data. It is also beneficial when a host does not have network authentication.

7.2 Contributions and Benefits of FINDIR

FINDIR is an extension of the previous work accomplished with NetRecon and GULP. The design is outlined in Chapter 5 along with the code in Appendix Appendix B. This design can be used for a production deployment with a few modifications.

Based on the previously addressed research questions, there are multiple benefits to FINDIR. The system allows the integration of application authentications and additional organizational information. This increases the number of potential responsible persons for each

host. This allows for greater awareness of an incident and increased probability that an actual responsible person will be contacted. It also increases the number of hosts with at least one potential responsible person.

FINDIR also reduces the storage required for records and allows for quicker queries when compared to the CLS. An operational implementation of FINDIR would allow a reduction of documents stored and indexed in the CLS.

This research provides insight into the issues of host inventory in higher education networks. While there are other solutions, the complexities of large and diverse networks need careful consideration in order to meet the needs of all constituents.

7.3 Limitations

There are limitations to FINDIR as identified in the evaluation. Specifically, the system is data-driven and must rely on having complete, or mostly complete, records to provide the most benefit.

In addition, we assumed that a MAC address is a one-to-one relationship with a host. For example, it is possible for someone to modify their transmitted frames to use a different MAC address. This is commonly known as MAC address spoofing. While this is possible, it is also possible that the host will show up in FINDIR with its manufacturer defined MAC address. It is also not uncommon for hosts to have more than one network interface such as a wired and wireless connection. This type of host will have two MAC addresses and will be seen in FINDIR as two hosts. This could be when a laptop uses a wired connection in an office but wireless connection

when attending a meeting. However, this is not necessarily an issue if we assume similar use patterns when alternating which connection is used.

The polling of network state data was set at five-minute intervals. This was defined as part of the NetRecon project and not easily changed due to the resources needed to query all network equipment in the time interval. Ideally, this interval would be set to a minute or less to reduce the chances of missing a state change. Alternatively, the same information could be collected in an event-driven manner when each state change occurs. This may or may not be feasible depending on the network architecture and equipment.

The GULP method works for network devices which have user interaction. Some embedded and IoT devices may not have a user authenticating to an application. Therefore, this method does not produce a possible responsible person.

7.4 Future Work

FINDIR needs to be connected to real-time data sources to evaluate how it will perform in an operational environment. This could be accomplished with separate Kafka topics or another message broker. FINDIR, as implemented with PostgreSQL, should be compared to other methods such as map-reduce. It is possible a similar system could be built in the Hadoop platform and Apache Spark. A distributed approach could allow FINDIR to scale beyond the current implementation and work with even larger networks.

FINDIR also needs additional input validation to handle incomplete or missing values in events. The network data from one day contained records with null, missing, or invalid fields. This was a very small percentage of records. However, in future use there may be other conditions that

need to be validated. This leads to the need for additional testing and further work to generalize the implementation. In addition, the Period Delta variable which represents the allowable variance in the Period Interval needs further evaluation. The initial value of 30 seconds was used for testing the proof of concept. The Period Interval is the given sampling rate for inputs 1, 2, and 3 shown in Table 5-3.

Additional data sources need to be implemented from the design. The proof of concept used for testing in this research did not include DNS, user affiliations, manufacturer OUIs, or distributed IT contacts. In addition, many networks utilize a VPN service for remote hosts to tunnel traffic. The tunneled traffic makes the remote host an extension of the network. As such, these logs should also be considered for incorporating into this concept.

While FINDIR is intended for cybersecurity and network operations, it is possible it could have value for other uses. For example, in a cybersecurity portal, an institution user could login and see hosts they are associated with. This is already being explored by other institutions such as Stanford [37]. There is also an open source project started by Netflix that provides similar device registration and evaluation call Stethoscope [38]. It would seem that the future of validating application and resource access lies in better understanding hosts and automatically correlating current activity.

GULP is also continuing to evolve with the recent inclusion of physical door access control events [39]. This concept could be incorporated into FINDIR or effort could be undertaken to form a consortium for these higher education user access and host inventory tools.

7.5 Concluding Thoughts

It is now possible to take advantage of logged events to understand the state of large networks given the cost savings with commodity computing and storage. While large networks are complex, the benefits to using this log data for cybersecurity are great. At the same time, access to some of this information must be kept private. A solution that maintains privacy in the storage of records reduces the chance of unintended disclosure. It also provides the verification that a system is working with the need to see actual human-identifiable information.

FINDIR is a step in the direction to make better use of data and reduce the burden of mundane tasks by humans. Utilizing data which is already in the possession of higher education institutions, and not requiring changes to networks, will provide great value to cybersecurity operations.

Bibliography

[1] T. M. T. G. ,. a. K. S. P. Cichonski, *Computer security incident handling guide,* NIST Special Publication 800-61, 2012.

[2] C. a. Y. K. W. Hsieh, "A study of android malware detection technology evolution," in *Security Technology (ICCST), 2015 International Carnahan Conference on*, 2015.

[3] B. Morrow, "BYOD security challenges: control and protect your most sensitive data," *Network Security,* vol. 2012, no. 12, pp. 5-8, 2012.

[4] R. M. D. R. J. T. Philip Kobezak, "Host Inventory Controls and Systems Survey: Evaluating the CIS Critical Security Control One in Higher Education Networks," in *Proceedings of the 51st Hawaii International Conference on System Sciences*, 2018.

[5] L. Johnson, Security Controls Evaluation, Testing, and Assessment Handbook, Elsevier Science, 2015.

[6] "About Us - Center for Internet Security," [Online]. Available: https://www.cisecurity.org/about-us/. [Accessed 10 May 2016].

[7] "CIS Critical Security Controls for Effective Cyber Defense Version 6.1," [Online]. Available: https://www.cisecurity.org/controls/. [Accessed 20 May 2017].

[8] G. Lyon, "The Art of Port Scanning.," 1997. [Online]. Available: https://nmap.org/nmap_doc.html.

[9] T. Chown, "RFC 5157 IPv6 Network Scanning," IETF Network Working Group, 2008.

[10] R. G. T. H. R. Deraison, "Passive vulnerability scanning: Introduction to NeVO," Tenable Network Security, 2004.

[11] e. a. J. Bound, "RFC 3315 Dynamic host configuration protocol for IPv6 (DHCPv6)," IETF Networking Group, 2003.

[12] T. N. a. T. J. S. Thomson, "RFC 4862 Stateless Address Autoconfiguration," IETF Networking Group, 2007.

[13] R. Marchany, "Higher Education: Open or Secure?," SANS Reading Room, 2014.

[14] "The Carnegie Classification of Institutions of Higher Education," Indiana University Center for Postsecondary Research (n.d.), 2015.

[15] J. V. G. H. K. Miller, "BYOD: Security and Privacy Considerations," *IT Professional,* pp. 53-55, Sept. 2012.

[16] M. F. a. D. Walker, "Acts Influencing How Higher Education Deals With Information," *Journal of Higher Education Management,* vol. 28, no. 1, 2013.

[17] e. a. C. Clark, "Live migration of virtual machines," in *Proceedings of the 2nd conference on Symposium on Networked Systems Design & Implementation-Volume 2*, 2005.

[18] D. Merkel, "Docker: lightweight linux containers for consistent development and deployment," *Linux Journal,* 2014.

[19] "DoD Internet Host Table," 4 Oct 1988. [Online]. Available: https://emaillab.jp/pub/hosts/19881003/HOSTS.TXT. [Accessed 23 2 2018].

[20] "Raspberry Pi Zero W," [Online]. Available: https://www.raspberrypi.org/products/raspberry-pi-zero-w/. [Accessed 26 2 2018].

[21] A. S. Tanenbaum and D. J. Wetherall, Computer Networks, Fifth Edition, Prentice Hall , 2010.

[22] P. S. M. Holdrege, "RFC 2663 IP Network Address Translator (NAT) Terminology and Considerations," IETF Network Working Group, 1999.

[23] S. Northcutt, "SANS Institute Incident Handling Step-by-Step," The SANS Institute, 1998.

[24] "Graylog2 Server," [Online]. Available: https://github.com/Graylog2/graylog2-server. [Accessed 30 3 2018].

[25] "Elasticsearch," [Online]. Available: https://github.com/elastic/elasticsearch. [Accessed 31 3 2018].

[26] S. D. R. Hinden, "RFC 4291 IP Version 6 Addressing Architecture," IETF Networking Group, 2006.

[27] D. M. M. Selsky, "GULP: A Unified Logging Architecture for Authentication Data," in *Large Installation System Administration Conference*, 2005.

[28] "GeoLite2 Free Downloadable Databases," MaxMind, [Online]. Available: https://dev.maxmind.com/geoip/geoip2/geolite2/. [Accessed 18 3 2018].

[29] I. S. A. R. Authority. [Online]. Available: https://regauth.standards.ieee.org/standards-ra-web/pub/view.html. [Accessed 30 3 2018].

[30] "Wireshark OUI Listing," [Online]. Available: https://code.wireshark.org/review/gitweb?p=wireshark.git;a=blob_plain;f=manuf. [Accessed 3 4 2018].

[31] P. K. D. R. R. M. J. T. Mark E. DeYoung, "Privacy Preserving Network Security Data Analytics: Architectures and System Design," in *Proceedings of the 51st Hawaii International Conference on System Sciences*, 2018.

[32] M. Kleppmann, Designing Data-Intensive Applications: The Big Ideas Behind Reliable, Scalable, and Maintainable Systems, O'Reilly Media, Inc., 2017.

[33] C. C. a. S. Morris, Database systems: design, implementation, & management, Cengage Learning, 2016.

[34] "PostgreSQL," [Online]. Available: https://www.postgresql.org/. [Accessed 28 10 2017].

[35] "Python," [Online]. Available: https://www.python.org/. [Accessed 28 10 2017].

[36] "Psycopg," [Online]. Available: http://initd.org/psycopg/. [Accessed 28 10 2017].

[37] M. Duff, "Going Passwordless at Stanford," [Online]. Available: https://events.educause.edu/~/media/files/events/user-uploads-folder/sec18/sess10/educause-spc--going-passwordless--11apr2018.pdf. [Accessed 11 4 2018].

[38] "Stethoscope," [Online]. Available: https://github.com/Netflix/Stethoscope. [Accessed 11 4 2018].

[39] J. Rosenblatt. [Online]. Available: https://events.educause.edu/~/media/files/events/user-uploads-folder/sec18/sess36/spc-2018-gulp.pptx. [Accessed 12 4 2018].

List of Acronyms

ARP – Address Resolution Protocol

BYOD – Bring your own device

CALEA – Communications Assistance for Law Enforcement Act

CAM – Content Addressable Memory

CAS – Central Authentication Service

CIS – Center for Internet Security

CISO – Chief Information Security Officer

CLS – Central Log Service

CPU – Central Processing Unit

CSC – Critical Security Controls

CSV – Comma Separated Values

DHCP – Dynamic Host Configuration Protocol

DHCPv6 – Dynamic Host Configuration Protocol version 6

DMCA – Digital Millennia Copyright Act

ER – Entity Relationship

ERP – Enterprise Resource Planning

FINDIR – Frequent Inventory of Devices for Incident Response

GULP – Grand Unified Logging Program

IEEE – Institution of Electrical and Electronics Engineers

IoT – Internet of Things

IPv4 – Internet Protocol version 4

IPv6 – Internet Protocol version 6

ISP – Internet Service Provider

LAN – Local Area Network

MAC – Media Access Control

NAC – Network Access Control

NAT – Network Address Translation

NDP – Neighbor Discovery Protocol

NTP – Network Time Protocol

OS – Operating System

OSI – Open Systems Interconnect

OUI – Organizationally Unique Identifier

PAT – Port Address Translation

R1 – Carnegie Basic Classification, Doctoral Universities, Highest research activity

R2 – Carnegie Basic Classification, Doctoral Universities, Higher research activity

R3 – Carnegie Basic Classification, Doctoral Universities, Moderate research activity

RADIUS – Remote Authentication Dial-In User Service

RFC – Request for comment

SLAAC – Stateless Address Autoconfiguration

SoC – System on Chip

SSH – Secure Shell

SSO – Single Sign On

VLAN – Virtual Local Area Network

VM – Virtual Machine

VoIP – Voice over Internet Protocol

Appendix A

CISO Survey Results

Survey Introduction
This survey is part of research being conducted by the Virginia Tech IT Security Lab to better understand host inventory controls utilized in networks of higher education institutions. The motivation is to determine what challenges higher education institutions face with implementing the first CIS Critical Security Control (previously known as the SANS Top 20 Critical Security Controls). The key principle of CSC #1 is, "actively manage (inventory, track, and correct) all hardware devices on the network so that only authorized devices are given access, and unauthorized and unmanaged devices are found and prevented from gaining access." Implementing this control helps identify the owner and location of a machine which is a target or source of an attack. The intended respondent is someone with broad knowledge of an institution's network and security controls such as a CIO, CISO, or their designee. Additionally, the intended respondent is at the institution rather than system level, if applicable. For the purposes of this survey, a host is defined as anything which communicates on a network which the institution controls.

Q1.1 - What is your institution's Basic Carnegie Classification? Look up your institution here: http://carnegieclassifications.iu.edu/lookup/lookup.php		
#	Answer	Count
1	R1: Doctoral Universities: Highest Research Activity	18
2	R2: Doctoral Universities: Higher Research Activity	1
3	R3: Doctoral Universities: Moderate Research Activity	4

4	M1: Master's Colleges & Universities: Larger Programs	8
5	M2: Master's Colleges & Universities: Medium Programs	2
6	M3: Master's Colleges & Universities: Small Programs	0
7	Baccalaureate Colleges: Arts & Sciences Focus	6
8	Baccalaureate Colleges: Diverse Fields	1
9	Baccalaureate/Associate's Colleges: Mixed Baccalaureate/Associate's	0
10	Baccalaureate/Associate's Colleges: Associate's Dominant	0
11	Associate's Colleges: High Transfer-High Traditional	1
12	Associate's Colleges: High Transfer-Mixed Traditional/Nontraditional	0
13	Associate's Colleges: High Transfer-High Nontraditional	0
14	Associate's Colleges: Mixed Transfer/Career & Technical-High Traditional	2
15	Associate's Colleges: Mixed Transfer/Career & Technical-Mixed Traditional/Nontraditional	1
16	Associate's Colleges: Mixed Transfer/Career & Technical-High Nontraditional	1
17	Associate's Colleges: High Career & Technical-High Traditional	0
18	Associate's Colleges: High Career & Technical-Mixed Traditional/Nontraditional	0
19	Associate's Colleges: High Career & Technical-High Nontraditional	0
20	Special Focus Two-Year: Health Professions	0
21	Special Focus Two-Year: Technical Professions	0
22	Special Focus Two-Year: Arts & Design	0
23	Special Focus Two-Year: Other Fields	0
24	Special Focus Four-Year: Faith-Related Institutions	0
25	Special Focus Four-Year: Medical Schools & Centers	0
26	Special Focus Four-Year: Other Health Professions Schools	0
27	Special Focus Four-Year: Engineering Schools	0
28	Special Focus Four-Year: Other Technology-Related Schools	0
29	Special Focus Four-Year: Business & Management Schools	0
30	Special Focus Four-Year: Arts, Music & Design Schools	1
31	Special Focus Four-Year: Law Schools	0
32	Special Focus Four-Year: Other Special Focus Institutions	0
33	Tribal Colleges	0

Q1.2 - How is your institution controlled?		
#	Answer	Count
1	Public	34
2	Private not-for-profit	18
3	Private for-profit	0

4	Other	0

Q1.3 - How much does your institution expend on research annually in dollars?		
#	Answer	Count
1	0 to 1,000,000	10
2	1,000,001 to 2,000,000	2
3	2,000,001 to 5,000,000	2
4	5,000,001 to 10,000,000	2
5	10,000,001 to 20,000,000	1
6	20,000,001 to 50,000,000	0
7	50,000,001 to 100,000,000	1
8	100,000,001 to 200,000,000	4
9	200,000,001 to 500,000,000	7
10	500,000,001 to 1,000,000,000	4
11	More than 1,000,000,000	3
12	Not applicable or no research conducted	11

Q1.4 - How many employees work at your institution?		
#	Answer	Count
1	Less than or equal to 100	0
2	101 to 200	1
3	201 to 500	3
4	501 to 1,000	9
5	1,001 to 2,000	7
6	2,001 to 5,000	15
7	5,001 to 10,000	5
8	10,001 to 20,000	4
9	More than 20,000	7

Q1.5 - How many employees at your institution spend most of their time in Information Technology roles?		
#	Answer	Count
1	Less than or equal to 10	0
2	11 to 20	3

3	21 to 50	14
4	51 to 100	12
5	101 to 200	8
6	201 to 500	3
7	501 to 1,000	5
8	1,001 to 2,000	4
9	More than 2,000	1

Q1.6 - Of those employees in Information Technology roles, how many are involved in network architecture, engineering, and operations?		
#	Answer	Count
1	Less than or equal to 1	1
2	2 to 5	17
3	6 to 10	11
4	11 to 20	7
5	21 to 50	6
6	51 to 100	2
7	More than 100	5

Q1.7 - Of those employees in Information Technology roles, how many are involved in information security architecture, engineering, and operations?		
#	Answer	Count
1	Less than or equal to 1	10
2	2 to 5	24
3	6 to 10	4
4	11 to 20	7
5	21 to 50	1
6	51 to 100	1
7	More than 100	1

Q1.8 - How large is your enrolled student population?		
#	Answer	Count
1	Less than or equal to 1000	3
2	1,001 to 2,000	4
3	2,001 to 5,000	8

4	5,001 to 10,000	7
5	10,001 to 20,000	12
6	20,001 to 50,000	9
7	50,001 to 100,000	3
8	100,001 to 200,000	2
9	200,001 to 500,000	1
10	More than 500,000	0

Q1.9 - How many of your enrolled students are considered remote?		
#	Answer	Count
1	Less than or equal to 1000	27
2	1,001 to 2,000	10
3	2,001 to 5,000	5
4	5,001 to 10,000	4
5	10,001 to 20,000	1
6	20,001 to 50,000	2
7	50,001 to 100,000	0
8	100,001 to 200,000	0
9	200,001 to 500,000	0
10	More than 500,000	0

Q2.1 - What is your best estimate for the peak number of hosts on your network at one time? Include physical and virtual machines, embedded devices, IoT, BYOD, wired, and wireless.		
#	Answer	Count
1	Less than or equal to 1,000	0
2	1,001 to 2,000	2
3	2,001 to 5,000	5
4	5,001 to 10,000	5
5	10,001 to 20,000	12
6	20,001 to 50,000	6
7	50,001 to 100,000	8
8	100,001 to 200,000	0
9	200,001 to 500,000	2
10	More than 500,000	1
11	Unknown	2

Q2.2 - What is the average number of BYOD hosts each type of end-user connects to your network?							
#	Question	Less than or equal 1	2 to 3	4 to 6	7 to 10	More than 10	Unknown
1	Residential students	0	15	23	1	0	4
2	Non-residential students	5	29	5	0	0	5
3	Employees	9	31	3	0	0	1

Q2.3 - For your remote students, are they permitted to connect to the network via VPN?		
#	Answer	Count
7	Yes	18
1	No	20
2	Not applicable	6

Q2.4 - What is your best estimate for the number of sub-networks (Local Area Network segments or broadcast domains)?		
#	Answer	Count
1	Less than or equal to 10	2
2	11 to 20	1
3	21 to 50	7
4	51 to 100	4
5	101 to 200	6
6	201 to 500	8
7	501 to 1,000	5
8	1,001 to 2,000	3
9	2,001 to 5,000	3
10	More than 5,000	2
11	Unknown	1

Q2.5 - How many logical network zones or groups do you have in your network? Examples include Voice over IP, residential networks, Personally Identifiable Information, Payment Card Industry, and data centers.

#	Answer	Count
1	0	0
2	1 to 10	10
3	11 to 20	16
4	21 to 50	4
5	51 to 100	4
6	101 to 200	2
7	201 to 500	1
8	500 to 1,000	2
9	More than 1,000	1
10	Unknown	2

Q2.6 - What speed and quantity of upstream network connections to the Internet and other internets do you have from your main campus?

#	Question	0	1	2	3	5	More than 5
1	Less than 1 Gbps	3	6	5	0	0	2
2	1 Gbps	4	3	10	1	1	2
3	10 Gbps	4	6	11	4	0	1
4	40 Gbps	7	4	1	0	0	0
5	100 Gbps	7	2	2	0	2	0
6	Other	2	1	3	0	0	0

Other (text entry):
“3 gig and 5 gig”
“2”
“KINBER cache 1 Gbps”
“3Gbps”

Q2.7 - How is the network managed at your institution? For the purposes of this question, who has the responsibility to track hosts and allow or disallow hosts on the network?

#	Answer	Count
1	The CIO or CISO has the ability to deny or allow all hosts on the network	28
2	The CIO or CISO has the ability to deny or allow most hosts but not all	12
3	The network is mostly federated. Most organizational units control their networking, to include network equipment and hosts	1
4	The network is completely federated. The CIO or CISO has no ability to allow or disallow hosts.	0

5	Other	3

Other (text entry):
"Managed by networking, no controls"
"Network Admin can deny or allow all hosts"
"Networking does what it wants"

Q2.8 - How many different vendors supply networking equipment, including wireless infrastructure, for your institution?		
#	Answer	Count
1	Exclusively one vendor	5
2	Mostly one vendor	13
3	A mix of up to four vendors	20
4	Five or more vendors	6
5	Unknown	0

Q2.9 - Where does your institution allow Bring Your Own Devices (BYOD) hosts on your network?		
#	Answer	Count
1	All logical network zones	6
2	Most logical network zones	18
3	A few logical network zones	15
4	Only one logical network zones	5
5	BYOD is not allowed	0

Q2.10 - Where does your institution allow embedded hosts or Internet of Things (IoT) on your network?		
#	Answer	Count
1	All logical network zones	3
2	Most logical network zones	16
3	A few logical network zones	17
4	Only one logical network zones	8
5	Embedded devices or IoT is not allowed	0

Q3.1 - What is the estimated percentage of each type of host on your network? The sum should be 100%.

#	Field	Min	Max	Mean	Std Deviation	Variance	Count
1	Embedded devices (IoT, printers, cameras, etc.)	0	25	8.92	6.06	36.67	37
2	Servers with full operating systems (either physical or virtual)	1	80	15.81	14.2	201.56	37
3	Institution owned end-user devices (desktops, laptops, mobile devices)	4	75	34	16.28	264.92	37
4	BYOD end-user devices (desktops, laptops, mobile devices)	0	92	39.24	20.39	415.81	37
5	Other	0	11	0.84	2.54	6.46	37

Other (text entry):
"LAN printers"
"VOIP Phones"

Q3.2 - What percentage of all hosts on your network use a statically assigned IP address?

#	Field	Min	Max	Mean	Std Deviation	Variance	Count
1	What percentage of all hosts on your network use a statically assigned IP address?	0	90	21.11	17.12	293.21	36

Q3.3 - What IP addressing methods do you use?

#	Question	Static IPv4	DHCP (IPv4)	DHCPv6	SLAAC (IPv6)	Static IPv6
1	Wired connections	29	35	2	1	1
2	Wireless connections	8	36	2	2	1

Q3.4 - How long do you have DHCP lease times set? Time is measured in hours.

#	Question	Less than or equal to 1	2 to 4	5 to 10	11 to 24	More than 24	Unknown
1	Wired connections	1	2	2	12	16	3
2	Wireless connections	8	8	5	7	4	4

Q3.5 - How confident are you in your ability to identify unique, individual hosts which have multiple network connections, such as Ethernet and WiFi? An example might be a laptop with a wired Ethernet and WiFi connection, where each is used at different times. In this example, it is a single host but seen by the network at different times using different connections and IP addresses.		
#	Answer	Count
1	Extremely confident	8
2	Somewhat confident	16
3	Neither confident nor unconfident	2
4	Somewhat unconfident	5
5	Extremely unconfident	6

Q3.6 - How confident are you in your ability to identify unique, individual hosts which are virtual machines?		
#	Answer	Count
1	Extremely confident	2
2	Somewhat confident	17
3	Neither confident nor unconfident	8
4	Somewhat unconfident	7
5	Extremely unconfident	3

Q3.7 - How confident are you in your organization's ability to identify hosts with multiple, changing addresses, to include application containers (Docker) and IPv6 privacy extensions (RFC 4941)?		
#	Answer	Count
1	Extremely confident	0
2	Somewhat confident	8
3	Neither confident nor unconfident	9
4	Somewhat unconfident	10
5	Extremely unconfident	10

Q3.8 - What percentage of hosts on your network utilize some form of network authentication to connect (IEEE 802.1x, NAC, etc.)?							
#	Field	Min	Max	Mean	Std Deviation	Variance	Count
1	What percentage of hosts on your network utilize some form of	0	100	44.44	32.1	1030.25	36

	network authentication to connect (IEEE 802.1x, NAC, etc.)?						

Q3.9 - For the purpose of inventory control, how does your institution define what a host is? For example, is a host defined by one or more MAC addresses, a unique certificate, or other mechanism?
“Mac address”
“MAC address”
“MAC”
“MAC address”
“one or more MAC addresses”
“unique mac address, and/or mac+other device fingerprints”
“MAC”
“One or more MAC addresses”
“machine name and mac address”
“MAC address”
“Primary MAC address”
“Mac”
“mac address”
“defined by one or more MAC addresses”
“Generally by MAC address, understanding that this can count some hosts more than once (wired/wireless) and can miss some hosts (virtual machines using NAT) “
“MAC”
“One or more MAC addresses”
“No specific definition :(”
“mac address“
“MAC address, typically”
“mac address”
“MAC address. VM's are included too as they have IP addresses. “
“IP or MAC”
“In the server room it's currently by OS (used to include Solaris Zones) and/or physical hosts - User Devices - by physical host”
“unique MAC address”
“MAC addresses”
“IPv4 address, DNS name or both”
“MAC addresses”
“One Mac Address. ”
“MAC address (though physical tags and dollar limits are a factor too) ”

“mac”
“Varies greatly across the system - generally visible IP addresses. ”
“MAC ”
“MAC address and/or AD membership”
“host has one or more MAC addresses”
“MAC Addresses”

Q4.1 - For the purposes of your host inventory controls, what types of hosts do you track? Select any which apply.

#	Question	Tracked	Not tracked but allowed	Not allowed
1	Physical servers w/ full operating system	28	4	0
2	Virtual servers w/ full operating system	27	5	0
3	Embedded devices / IoT	12	19	1
4	Printers / copiers	26	6	0
5	Video Cameras	21	9	2
6	VoIP phones	25	5	1
7	Application containers (Docker)	5	20	5
8	Institution owned wired Ethernet end-user devices	27	5	0
9	Institution owned wireless hosts	27	5	0
10	Institution owned network equipment	30	2	0
11	BYOD wired Ethernet hosts	14	14	4
12	BYOD wireless hosts	15	17	0
13	BYOD network equipment	4	7	21
14	Other	1	0	0

Other (text entry):
“Only hosts and gear managed by central IT are tracked routinely.”

Q4.2 - During a potential security incident or event, how long does it usually take to track down the responsible user or owner of these host types? Time is measured in minutes.

#	Question	Less than or equal to 1	2 to 10	11 to 30	31 to 60	More than 60	Can't be found	Not applicable
1	A wired embedded device, including printers, cameras, and IoT	3	10	10	5	4	0	0

2	A wireless embedded device, including printers, cameras, and IoT	2	9	10	5	4	1	1
3	A wired Ethernet server with full operating system	7	16	4	2	3	0	0
4	Virtual machines and application containers	2	15	7	4	2	1	1
5	A wired end-user host	4	14	9	2	2	0	1
6	A wireless end-user host	4	12	7	2	5	2	0

Q4.3 - During a potential security incident or event, how long does it usually take to track down the physical location of these host types? Time is measured in minutes.

#	Question	Less than or equal to 1	2 to 10	11 to 30	31 to 60	More than 60	Can't be found	Not applicable
1	A wired embedded device, including printers, cameras, and IoT	1	14	5	4	8	0	0
2	A wireless embedded device, including printers, cameras, and IoT	1	7	9	4	8	2	1
3	A wired Ethernet server with full operating system	8	12	4	6	2	0	0
4	Virtual machines and application containers	3	15	4	3	5	0	2
5	A wired end-user host	5	12	5	5	4	1	0
6	A wireless end-user host	2	9	6	4	9	2	0

Q4.4 - How often do your inventory controls and tools lead to someone who is not the current responsible user? In this situation, you have to ask that individual if they know who is the responsible user or use other information sources to find the actual responsible user.

#	Answer	Count
1	Never	5
2	A few time a month	14
3	A few times week	4
4	A few times a day	1
5	Other	8

Other (text entry):

“A few times per year”
“Never say never, but it's rare”
“annual check”
“A few a year”
“Staff - Never; Labs almost always”
“varies widely”
“rarely”

Q4.5 - How does your institution procure and prepare for first use a host on the network?		
#	Answer	Count
1	All new equipment inventoried and each must be setup to institutional baseline before given to end user	14
2	All new equipment inventoried but only some equipment is setup prior to being given to end user	6
3	Most new equipment inventoried and only some equipment is setup prior to being given to end user	7
4	Other	4
5	Unknown	1

Other (text entry):
“Highly distributed - in central IT, equipment is inventoried and setup to baseline prior to use by user. Other departments may do the same, but it is not yet tracked at the central level.”
“End users do their own machines”
“Some is inventoried and less baselined.”

Q4.6 - How accurate have you found the following tools and technologies to be in keeping track of hosts in your network?					
#	Question	Not very Accurate	Somewhat accurate	Very Accurate	Not used
1	Spreadsheets	9	15	2	6
2	Commercial inventory applications	1	15	4	12
3	Custom / in-house inventory applications (to include simple databases)	0	16	6	10
4	MAC address registration	0	7	16	9
5	Software agents on hosts	1	6	12	13
6	802.1x	0	5	20	7
7	Mobile device management	1	7	7	17
8	Network device logs	0	9	16	6

9	Network flow data	1	14	5	11
10	Mapping or scanning (IPv4)	4	17	5	6
11	Mapping or scanning (IPv6)	1	2	3	26
12	Other	0	1	1	5

Other (text entry):
"Network Access Control"
"we log ip and mac address pairings over time, which is helpful on the wired nets"

Q4.7 - Do you consider embedded devices or IoT hosts more difficult to track than other hosts?		
#	Answer	Count
1	Yes - more difficult	13
2	About the same as other hosts	19
3	Less difficult to track than other hosts	0
4	Not applicable	0

Q4.8 - Do you consider BYOD (non-institutionally owned) hosts more difficult to track than institutionally owned hosts?		
#	Answer	Count
1	Yes - more difficult	11
2	About the same as other hosts	21
3	Less difficult to track than other hosts	0
4	Not applicable	0

Q4.9 - Does Network Address Transition (NAT) or Port Address Translation (PAT) make host identification location more difficult to track?		
#	Answer	Count
1	Yes - more difficult	18
2	About the same as other hosts	6
3	Less difficult to track than other hosts	1
4	Not applicable	7

Q4.10 - Has the effectiveness of your network's host inventory controls changed with increases in the total number of hosts in the past five years?

#	Answer	Count
1	It has decreased significantly	7
2	It has decreased slightly	6
3	No change	9
4	Increased slightly	7
5	Increased significantly	2
6	Not applicable	1

Q4.11 - Has the effectiveness of your network's host inventory controls been impacted with increases in the number of embedded devices or IoT hosts in the past five years?		
#	Answer	Count
1	The impact has noticeably decreased the effectiveness of controls	8
2	Not noticeable from overall host growth	19
3	It has made inventory controls more effective	1
4	Not applicable	3

Q4.12 - Has the effectiveness of your network's host inventory controls been impacted with increases in the number of BYOD (non-institutionally owned) hosts in the past five years?		
#	Answer	Count
1	The impact has noticeably decreased the effectiveness of controls	11
2	Not noticeable from overall host growth	16
3	It has made inventory controls more effective	1
4	Not applicable	4

Q4.13 - How much time does your institution spend on initially entering and updating Host Inventory Control tools?		
#	Answer	Count
1	None - our inventory tools are highly automated	2
2	Minimal - each host is setup once and rarely needs human interaction to update	9
3	Moderate - some human interaction is needed to keep host records accurate	14
4	Significant - many host records need to be updated frequently	5
5	Not applicable	2

Q5.14 - Does your institution have any specific challenges with host inventory controls which you'd like to share?
“none”
“distributed nature of hosts leads to a lack of central inventory”
“We do a very poor job at host inventory”
“Time to audit is impossible to find.”
“We have asset inventories to track items >$25k, but nothing for systems below that--basically considered supplies. There's no way for us to easily differentiate between university owned and byod.”
“There is no overarching inventory control in place, so many of these questions aren't very applicable. Even Institute-owned devices may be purchased by a specific department or lab, with little Institute-level oversight.”
“Consistency and enforcing the use of standard prefixes for different classes of connected devices.Raising awareness of having current and correct inventory.”
“Central Computing operates separately from other campus orgs, thus `institution' inventory controls is difficult to provide”
“We can accurately track user-MAC on the wireless LAN, Time and PAT traversal make responding to DMCA complaints that report the gateway, time and port only, very time consuming if not impossible.”
“Multiple systems that are not correlated, inventories controlled by different groups, no central database, no cam logs, no budget.”
“No”

Appendix B

Data Sources and Code

The code for FINDIR is available at: “https://code.vt.edu/ITSL/LAARG/FINDIR”.

The following tables represent the data sources as stored in CSV files used by FINDIR.

Table 7-1: Organization IP Address Blocks

Field Description	Data Type / Represented As
Description of IP Address Range	String
IP Address CIDR	CIDR, String
Is PAT Address	Boolean, “Y” or “N”

Table 7-2: Department and Organization

Field Description	Data Type / Represented As
Senior Management	String
Management	String
Department ID	Integer
Department Name	String
Organization ID	Integer
Organization Name	String
Organization Head Name	String
Organization Head Email	Email, String

Table 7-3: Institution Buildings

Field Description	Data Type / Represented As
Building ID	Integer
Building Name	String

Table 7-4: Building Interior Spaces

Field Description	Data Type / Represented As
Building Number	Integer
Building Floor Number	Integer
Building and Room	String
Space ID	String
Latitude	Double Precision
Longitude	Double Precision
Height	Integer

Table 7-5: Wireshark OUI Listing

Field Description	Data Type / Represented As
OUI	String, Hex colon delimited 3 bytes
Short Name	String
Long Name	String

Table 7-6: Access Point to Location

Field Description	Data Type / Represented As
Access Point Name	String
LAN IP Address	Number, IPv4 dotted decimal
Building ID	Integer
Building Name	String
Room	Integer
Latitude	Double Precision
Longitude	Double Precision
Height	Integer

Table 7-7: NetRecon Network Equipment Interface to Circuit

Field Description	Data Type / Represented As
Ethernet device (switch) port	String
Ethernet device name	String
Building ID	Integer
Room	Integer
Circuit ID	String
Outlet ID	String
Paying Organization ID	Integer

Table 7-8: MAC Address to Network Equipment Interface

Field Description	Data Type / Represented As
Timestamp when the state was polled	Datetime, ISO 8601 format
MAC address seen on the port	Number, hex colon delimited bytes
Ethernet device (switch) port	String
Ethernet device (switch) IPv4 address	Number, IPv4 dotted decimal
Ethernet device name	String

Table 7-9: IP to MAC Address

Field Description	Data Type / Represented As
Timestamp when the state was polled	Datetime, ISO 8601 format
IP Address	IPv4 dotted decimal or IPv6 hex colon delimited, String
MAC address	Number, hex colon delimited bytes
Network Equipment Name	String

Table 7-10: Wireless Association and Authentication

Field Description	Data Type / Represented As
Timestamp	Datetime, ISO 8601 format
Access Point Name	String
MAC Address of Host	Number, hex colon delimited bytes
SSID	String
De-identified Username	String

Table 7-11: PAT Allocation

Field Description	Data Type / Represented As
Timestamp when allocation begins	Datetime, ISO 8601 format
Internal IP Address	Number, IPv4 dotted decimal
Public PAT IP Address	Number, IPv4 dotted decimal
Start TCP Port	Integer
End TCP Port	Integer

Table 7-12: PAT Release

Field Description	Data Type / Represented As
Timestamp when allocation is released	Datetime, ISO 8601 format
Internal IP Address	Number, IPv4 dotted decimal
Public PAT IP Address	Number, IPv4 dotted decimal
Start TCP Port	Integer
End TCP Port	Integer

Table 7-13: Application SSO

Field Description	Data Type / Represented As
Timestamp	Datetime, ISO 8601 format
Host IP Address	IPv4 dotted decimal or IPv6 hex colon delimited, String
Application URL	String
De-identified Username	String

Appendix C

IRB Protocol and Approval

This research was conducted under the Virginia Tech Institutional Review Board (IRB) protocol number 17-375. The following pages are the submitted protocol and the approval.

VirginiaTech

Institutional Review Board
Existing Data Research Protocol

Note: complete this application only if this research project ***only*** involves the collection or study of **existing data**. Once complete, upload this form as a Word document to the IRB Protocol Management System: https://secure.research.vt.edu/irb

1. DO ANY OF THE INVESTIGATORS OF THIS PROJECT HAVE A REPORTABLE CONFLICT OF INTEREST? (http://www.irb.vt.edu/pages/researchers.htm#conflict)

☒ **No**
☐ **Yes,** explain:

2. IS THIS RESEARCH SPONSORED OR SEEKING SPONSORED FUNDS?

☒ **No,** go to question 3
☐ **Yes,** answer questions within table

IF YES
Provide the name of the sponsor [if NIH, specify department]:
Is this project receiving or seeking federal funds? ☐ No ☐ Yes **If yes,** **Does the grant application, OSP proposal, or "statement of work" related to this project include activities involving human subjects that are <u>not</u> covered within this IRB application?** ☐ No, all human subject activities are covered in this IRB application ☐ Yes, however these activities will be covered in future VT IRB applications, these activities include: ☐ Yes, however these activities have been covered in past VT IRB applications, the IRB number(s) are as follows: ☐ Yes, however these activities have been or will be reviewed by another institution's IRB, the name of this institution is as follows: ☐ Other, explain: **Is Virginia Tech the primary awardee or the coordinating center of this grant?** ☐ No, provide the name of the primary institution: ☐ Yes

3. DESCRIBE THE BACKGROUND, PURPOSE, AND ANTICIPATED FINDINGS OF THIS STUDY:

Co-investigators from the IT Security Office (ITSO) and IT Security Lab (ITSL) will use observational data collected from Virginia Tech's operational network to conduct research in network security data analytics. Personally identifying information will be de-identifed prior to research use. The purpose of this research is:

1) Network Security Data Analytics - Develop means to detect and report anomalous behaviours generated by peoples' interactions with information systems. In this area we anticipate findings that enhance network

security and improve users situational awareness.

2) Privacy Preserving Data Publication (PPDP) - Develop means that support Privacy Preserving Data Publication (PPDP) of de-identified observational data. The two primary avenues of research are anonymization techniques and synthesis techniques that add additional data privacy protections beyond those provided by de-identification processes. We anticipate that we will discover adversarial methods to attack PPDP techniques and develop defensive countermeasures that defeat or delay adversarial attempts to re-identify data.

4. EXPLAIN WHAT THE RESEARCH TEAM PLANS TO DO WITH THE STUDY RESULTS:
For example - publish or use for dissertation

Study results will be published as generalizable knowledge. While we do not have specific venues targeted at this time we plan to publish results in journal articles and conference proceedings.

5. WILL PERSONALLY IDENTIFYING STUDY RESULTS OR DATA BE RELEASED TO ANYONE OUTSIDE OF THE RESEARCH TEAM?
For example – to the funding agency or outside data analyst, or participants identified in publications with individual consent

☒ **No**
☐ **Yes,** to whom will identifying data be released?

6. WILL THE RESEARCH TEAM COLLECT AND/OR BE PROVIDED PARTICIPANT IDENTIFYING INFORMATION (E.G., NAME, CONTACT INFORMATION, VIDEO/AUDIO RECORDINGS)?

☐ **No,** go to question 7
☒ **Yes,** answer questions within table

IF YES
Describe if/how the study will utilize study codes: We will use hash algorithms, format preserving encryption (FPE), and prefix preserving encrypt (PPE) to generate study codes that correspond to individual persons' identifying information.
If applicable, where will the key [i.e., linked code and identifying information document (for instance, John Doe = study ID 001)] **be stored and who will have access?** **The key (mapping between study codes and identifying information) will not be stored. The specific hash, PPE, or FPE algorithm will not be masked. For PPE and FPE a secret password will only be known to specific co-investigators (AKA "honest brokers") specified by the principal investigator.** *Note: the key should be stored separately from subjects' completed data documents and accessibility should be limited.*

7. HOW WILL DATA BE STORED TO ENSURE SECURITY (E.G., PASSWORD PROTECTED COMPUTERS, ENCRYPTION) AND LIMITED ACCESS?

Data will be stored on Virginia Tech owned, password protected computers. Data will only be explicitly accessible by co-investigators who have executed a non-disclosure agreement (NDA) with the IT Security Office and completed initial IRB training.

8. WHO WILL HAVE ACCESS TO STUDY DATA?

Co-investigators who have executed a non-disclousre agreement (NDA) with the IT Security Office and completed initial IRB training. Co-investigators and personel from the IT Security Office will have access to the study data.

9. DESCRIBE THE PLANS FOR RETAINING OR DESTROYING STUDY DATA:

Data will be retained for no more than 24 months. The start of the time period will be the creation date of the event recorded in observational data. In general, we expect most data to be expired and destroyed within 6 months of collection. Specific data subsets may be retained for longer periods (more than 24 months). In this case the research protocol will be updated to identify the data retained. Data will be destroyed by electronically removing stored data files.

10. FROM WHERE DOES THE EXISTING DATA ORIGINATE?

Existing data originates from Virginia Techs operational network. As people and information systems interact with network mid-points (servers and network equipment) they produce event data. Some of this data is electronically aggregated in a centralized Log Aggregation & Analysis (LAA) system that is operated by Virginia Tech's Communications Network Services (CNS) and Network Infrastructure & Services (NIS). IT Security Office and IT Security lab personel will query data subsets from the LAA system and then apply de-identification procdures.

11. PROVIDE A DETAILED DESCRIPTION OF THE EXISTING DATA:

The existing observational data contains the following directly identifying data elements:
- user names (the user's VT PID or Hokies user name)
- user e-mail addresses

The identifying data elements will be de-identified by IT Security Office or IT Security Lab personel before the data is made availble for research.

Other data elments (that are not identifying) include:
- End-point device Media Access Codes (MACs) which are associated with users devices when they connect to network resources
- Internet Protocol (IP) numbers which are associated with the MAC of a device and a network session
- Informational, warning, and error messages about the network resources and users interaction with the network resources

12. IS THE SOURCE OF THE DATA PUBLIC?

☒ **No,** go to question 13
☐ **Yes,** you are finished with this application

13. WILL ANY INDIVIDUAL ASSOCIATED WITH THIS PROJECT (INTERNAL OR EXTERNAL) HAVE ACCESS TO OR BE PROVIDED WITH EXISTING DATA CONTAINING INFORMATION WHICH WOULD ENABLE THE IDENTIFICATION OF SUBJECTS:

- **Directly** (e.g., by name, phone number, address, email address, social security number, student ID number), or
- **Indirectly through study codes** even if the researcher or research team does not have access to the master list linking study codes to identifiable information such as name, student ID number, etc

or

- **Indirectly through the use of information that could reasonably be used in combination to identify an individual** (e.g., demographics)

☒ **No,** collected/analyzed data will be completely de-identified
☐ **Yes,**

If yes,

Research will not qualify for exempt review; therefore, if feasible, written consent must be obtained from individuals whose data will be collected / analyzed, unless this requirement is waived by the IRB.

Will written/signed or verbal consent be obtained from participants prior to the analysis of collected data?
-select one-

This research protocol represents a contract between all research personnel associated with the project, the University, and federal government; therefore, must be followed accordingly and kept current.

Proposed modifications must be approved by the IRB prior to implementation except where necessary to eliminate apparent immediate hazards to the human subjects.

Do not begin human subjects activities until you receive an IRB approval letter via email.

It is the Principal Investigator's responsibility to ensure all members of the research team who collect or handle human subjects data have completed human subjects protection training prior to handling or collecting the data.

----------END----------

VirginiaTech

Office of Research Compliance
Institutional Review Board
North End Center, Suite 4120, Virginia Tech
300 Turner Street NW
Blacksburg, Virginia 24061
540/231-4606 Fax 540/231-0959
email irb@vt.edu
website http://www.irb.vt.edu

MEMORANDUM

DATE: November 29, 2017

TO: David Richard Raymond, Mark Edward DeYoung, Alex Hsu, Philip D Kobezak, Amanda Teh, [redacted] Zachary Burch, [redacted]

FROM: Virginia Tech Institutional Review Board (FWA00000572, expires January 29, 2021)

PROTOCOL TITLE: IT Security Office & Lab Data Analytics

IRB NUMBER: **17-375**

Effective November 28, 2017, the Virginia Tech Institution Review Board (IRB) approved the Amendment request for the above-mentioned research protocol.

This approval provides permission to begin the human subject activities outlined in the IRB-approved protocol and supporting documents.

Plans to deviate from the approved protocol and/or supporting documents must be submitted to the IRB as an amendment request and approved by the IRB prior to the implementation of any changes, regardless of how minor, except where necessary to eliminate apparent immediate hazards to the subjects. Report within 5 business days to the IRB any injuries or other unanticipated or adverse events involving risks or harms to human research subjects or others.

All investigators (listed above) are required to comply with the researcher requirements outlined at:

http://www.irb.vt.edu/pages/responsibilities.htm

(Please review responsibilities before the commencement of your research.)

PROTOCOL INFORMATION:

Approved As:	**Expedited, under 45 CFR 46.110 category(ies) 5**
Protocol Approval Date:	**May 30, 2017**
Protocol Expiration Date:	**May 29, 2018**
Continuing Review Due Date*:	**May 15, 2018**

*Date a Continuing Review application is due to the IRB office if human subject activities covered under this protocol, including data analysis, are to continue beyond the Protocol Expiration Date.

FEDERALLY FUNDED RESEARCH REQUIREMENTS:

Per federal regulations, 45 CFR 46.103(f), the IRB is required to compare all federally funded grant proposals/work statements to the IRB protocol(s) which cover the human research activities included in the proposal / work statement before funds are released. Note that this requirement does not apply to Exempt and Interim IRB protocols, or grants for which VT is not the primary awardee.

The table on the following page indicates whether grant proposals are related to this IRB protocol, and which of the listed proposals, if any, have been compared to this IRB protocol, if required.