UNIVERSITY OF CALIFORNIA SAN DIEGO

Prioritizing Security Practices via Large-Scale Measurement of User Behavior

A dissertation submitted in partial satisfaction of the
requirements for the degree Doctor of Philosophy

in

Computer Science

by

Ariana Mirian

Committee in charge:

Professor Stefan Savage, Co-Chair
Professor Geoffrey M. Voelker, Co-Chair
Professor Nadia Heninger
Professor Farinaz Koushanafar
Professor Aaron Schulman

2023

The Dissertation of Ariana Mirian is approved, and it is acceptable in quality and form for publication on microfilm and electronically.

University of California San Diego

2023

DEDICATION

To those who never felt like things would get better, or that you could never find somewhere to truly belong: it is out there.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

I would like to start by thanking my co-chairs, Stefan Savage and Geoffrey M. Voelker for their support and encouragement over the years. They were the folks who first showed what it was like to work in a place where I could flourish, and were a very large reason of why I decided to stick around and finish my PhD. The jokes, sarcasm, and general tomfoolery were a most welcome plus, and something I will remember fondly from my time here.

I also would like to share a huge thanks to my committee for their guidance and help over the last few years: Nadia Heninger, Aaron Schulman, Farinaz Koushanafar.

I would also like to thank some really fantastic external collaborators I have been fortunate enough to work with over the years, in no particular order: Adriana Porter Felt, Emily Stark, Chris Thompson, Kurt Thomas, Caitlin Sadowski, and Nik Bhagat. Thank you for further showing that good work can happen with good people.

I also need to thank my internal collaborators, again in no particular order: Joe DeBlasio, Louis DeKoven, Audrey Randall, Gautam Akiwate, Ansel Blume, Lawrence Saul, Aaron Schulman, Ian Foster, Cindy Moore, Taner Halicioglu, Alisha Ukani, Alex Snoeren, and Anil Yelam. A special thanks to Grant Ho, for being my partner in crime these last two years. Thanks to all for proving Stefan and Geoff were not wrong, and this was a great place to work.

To the students I have had the pleasure of working with and who taught me far more than I was ever able to teach them: Nadah Feteih, Jeff Xie, Isabel Suizo, Alisha Ukani, Allison Turner, and the 2019 and 2020 ERSP groups.

To the truly amazing labmates in sysnet: Sunjay Cauligi, Brian Johanesmeyer, Joe DeBlasio, David Kohlbrenner, Arjun Roy, Danny Huang, Louis DeKoven, Nishant Bhaskar, Liz Izhikevich, Nadah Feteih, Stewart Grant, Keegan Ryan, Audrey Randall, Anil Yelam, Alisha Ukani, Alex Liu, and Ben Du.

And also to some outstanding folks in the department who I made a lot of skits with (Jess, Dimo), biked a lot with (Danilo, Steven), and generally just shenaniganed around with (Sunjay, Brian).

A huge thanks to the IT folks who welcomed me with a lot of enthusiasm and endless patience for my questions, and made the last two years of my PhD not feel like a PhD at all: Edward Wade, Mike Corn, Phillip Lopo, Sheena Yarberry, Ferdie Escudero, Elaine Fleming, and James Dotson. Working with you folks really solidified my desire to keep doing practical work, and was literally career changing.

To the amazing CSE staff, for making sure this department kept running even when we tried our best to make sure it did not.

To some fantastic non-research collaborators: Tierra Terrell, Sorin Lerner, Ailie Fraser, Christine Alvarado, Patrick Mallon, Veronica Abreu, Margaret Ramaker, and so many others. We did a lot of work, and really changed this department for the better, even if no one remembers who we are.

To the folks at Michigan, who taught me how to wear a thick skin.

To the folks in my grief group, who showed me it was ok to not always wear a thick skin.

To my close friends, there are not enough words. Thanks for getting me through this: Ailie, Chris, Cat, Ashley, Doug, Nessa, Aaron, Megan, Danilo, Steven, Janet, Sarah, Mitchell, Caity, Zakir, Charlotte, and Ethan.

To the many others that I am forgetting by name in this moment: thank you.

To my partner, Rob Kaufman, who was there for so much and was a bedrock of support and laughter even when I did not feel like laughing.

And finally to my parents, Mirfarhad and Rodabeh Mirian, for encouraging me to dream bigger.

Chapter 2, in part, is a reprint of the material as it appears in the Proceedings of the International Measurement Conference 2019. Louis F. DeKoven, Audrey Randall, Ariana Mirian, Gautam Akiwate, Ansel Blume, Lawrence K. Saul, Aaron Schulman, Geoffrey M. Voelker, and Stefan Savage. The dissertation author was a collaborator and contributor to this paper.

Chapter 3, in full, is currently being prepared for submission for publication of material. Ariana Mirian, Grant Ho, Stefan Savage, Geoffrey M. Voelker. The dissertation author was the

primary investigator and author of this material.

Chapter 4, in full, is a reprint of the material as it appears in The World Wide Web Conference 2019. Ariana Mirian, Joe DeBlasio, Stefan Savage, Geoffrey M. Voelker, and Kurt Thomas. The dissertation author was the primary investigator and author of this material.

VITA

2012–2016 Bachelor of Science in Engineering, University of Michigan

2016–2019 Master of Science, University of California San Diego

2019–2023 Doctor of Philosophy, University of California San Diego

PUBLICATIONS

"Please Act Now: An Empirical Analysis of Enterprise-wide Mandatory Password Updates." Ariana Mirian, Grant Ho, Stefan Savage, Geoffrey M. Voelker.

"In the Line of Fire: Risks of DPI-triggered Data Collection." Ariana Mirian, Alisha Ukani, Ian Foster, Gautam Akiwate, Taner Halicioglu, Cindy Moore, Alex C. Snoren, Geoffrey M. Voelker, Stefan Savage.

"On Knowing a Hawk from a Handsaw: The Challenges of Passive Device Classification." Ariana Mirian, Audrey Randall, Aaron Schulman, Stefan Savage, Geoffrey M. Voelker.

"Locked-In During Lock-Down: Undergraduate Life on the Internet in a Pandemic." Alisha Ukani, Ariana Mirian, and Alex C. Snoeren. In Proceedings of the 21st ACM Internet Measurement Conference (IMC '21). Association for Computing Machinery, New York, NY, USA, 2021.

"Who's Got Your Mail? Characterizing Mail Service Provider Usage." Enze Liu, Gautam Akiwate, Mattijs Jonker, Ariana Mirian, Stefan Savage, and Geoffrey M. Voelker. 2021. Who's got your mail? characterizing mail service provider usage. In Proceedings of the 21st ACM Internet Measurement Conference (IMC '21). Association for Computing Machinery, New York, NY, USA, 2021

"CoResident Evil: Covert Communication In The Cloud With Lambdas." Anil Yelam, Shibani Subbareddy, Keerthana Ganesan, Stefan Savage, and Ariana Mirian. In Proceedings of the Web Conference 2021 (WWW '21). Association for Computing Machinery, New York, NY, USA, 2021.

"Measuring Security Practices and How They Impact Security." Louis F. DeKoven, Audrey Randall, Ariana Mirian, Gautam Akiwate, Ansel Blume, Lawrence K. Saul, Aaron Schulman, Geoffrey M. Voelker, and Stefan Savage. In Proceedings of the Internet Measurement Conference, Amsterdam, Netherlands, 2019.

"Hack for Hire: Exploring the Emerging Market for Account Hijacking." Ariana Mirian, Joe DeBlasio, Stefan Savage, Geoffrey M. Voelker, and Kurt Thomas. In The World Wide Web

Conference (WWW '19). Association for Computing Machinery, New York, NY, USA, 2019.

"Web Feature Deprecation: A Case Study for Chrome." Ariana Mirian, Nik Bhagat, Caitlin Sadowski, Adrienne Porter Felt, Stefan Savage and Geoffrey M. Voelker. 2019 IEEE/ACM 41st International Conference on Software Engineering: Software Engineering in Practice (ICSE-SEIP), Montreal, QC, Canada, 2019.

"HTTPS Adoption in the Longtail." Ariana Mirian, Christopher Thompson, Stefan Savage, Geoffrey M. Voelker, Adrienne Porter Felt. Google Tech Report, 2018.

ABSTRACT OF THE DISSERTATION

Prioritizing Security Practices via Large-Scale Measurement of User Behavior

by

Ariana Mirian

Doctor of Philosophy in Computer Science

University of California San Diego, 2023

Professor Stefan Savage, Co-Chair
Professor Geoffrey M. Voelker, Co-Chair

Security is an ever growing concern for daily Internet users, especially since many facets of a user's daily interactions (banking, commerce, workplace) are now accessed via the Internet. Fortunately, recent technical advancements such as encrypted web browsing, email spam filtering, and login two factor authentication have increased the accessibility and practicality of security for users. However, studies show that the majority of exploited attacks take advantage of the human in the loop. Technology and humans are required to work in harmony for security to be effective. As a result, it is crucial that we understand the extent to which users follow best practices, and that we evaluate whether their behaviors in fact help prevent adverse security

outcomes. In this dissertation, I argue that large-scale empirical measurement is a practical and effective technique to answer these questions as the basis for prioritizing security practices, and I support this argument with three different projects. First, I use network traffic data and measurement methods to quantify user behavior "best practices" and how they relate to an outcome (in this case, compromise). Next I examine how communication about a security policy change can affect an organization by analyzing large-scale organizational data. Finally, I quantify attacker behavior in the Hack for Hire market by hiring and monitoring attackers, which provides insight into which defenses to prioritize for better protecting users from these types of attacks. By empirically understanding and prioritizing effective security practices, we can further improve security for users.

Chapter 1

Introduction

Security is an ever growing concern for daily Internet users, especially since the world is becoming more interconnected. Banking, workplaces, and commerce all have online presences, and a user must understand how to navigate security best practices while also achieving whatever their end goal may be. This is not an easy situation for users.

However, a promising facet of security is that much of the technical work has become more stable in recent years. While technical advancements continue, achievements such as the use of HTTPS for web browsing, spam filtering on email and other communications, and two factor authentication for login have made security far more accessible and practical for users of all backgrounds and capabilities.

As a result, many of the difficulties that plague online security today are not technical issues, but the interplay between technology and the human. Secure technology can only affect change so far if the human in the loop, and the difficulties humans experience using various technologies, is not accounted for.

Indeed, empirical data shows that many attacks rely on exploiting the human, not the technology itself. For example, Verizons yearly data breach report aggregates data across thousands of incidents and organizations to provide a comprehensive list of attack vectors. In 2022, 82% of these attacks were exploited by taking advantage of the human in the loop. While the ubiquity of HTTPS on websites today may represent a significant technological deployment

success, encryption does not help protect a user if a user is tricked into downloading a piece of malware.

Thus, for security to be effective, both the technical and human element must be secure themselves. However, this model assumes that the user will operate under best practices to keep themselves safe. While previous studies have examined user best practices from expert or crowdsourced points of view, two important questions remain unanswered: at large scale, do users follow these best practices, and do these practices have any sort of empirical effect on outcome?

In this dissertation, I argue that large-scale empirical measurement is a practical and effective technique to answer these questions. While the security field has some notion of best practices (e.g., update frequently, have long passwords, use 2FA), it is unclear how effective they may be at protecting the user, and which practices should be prioritized. While it would be ideal for users to follow all best practices, the reality is that time and energy is limited, so understanding which to prioritize would be a more effective practice.

I use large-scale measurement to quantify user behaviors and prioritize security processes in three different projects. First, I use network traffic data to quantify user behavior best practices and how it relates to an outcome (in this case, compromise). I find that in some cases, best practice does in fact line up with outcome, but not to a strong effect, requiring us to reassess whether focusing on these best practices is sufficient.

Next I examine how communication about a security policy change can affect an organization. I find that certain communications are more effective at convincing users to execute the change, which is useful for future organizational efforts that aim to change user behavior en masse. Understanding the most effective change allows an organization to prioritize using that method over others to better its security.

Finally, I quantify attacker behavior in the Hack for Hire market, which is a commodity market that sells email hacking services for $100 - $400 USD. Empirically measuring this market and its attributes provides insight into which defenses to prioritize for better protecting users

from these types of attacks.

Security will continue to remain an important facet of user lives, and using large-scale empirical measurement, we can better prioritize user and organizational time and efforts that matter in increasing security.

This dissertation is structured as follows. In Chapter 2 I discuss a measurement study that empirically quantifies and relates end user behavior to device compromise. In Chapter 3 I describe a study that examines effective communication mechanisms for a security policy change from the perspective of a large orgnaization. Chapter 4 examines a commodity market for breaking into email accounts, which provides insights into defenses against these attacks. Finally, Chapter 5 summarizes this disseration.

Chapter 2, in part, is a reprint of the material as it appears in the Proceedings of the International Measurement Conference 2019. Louis F. DeKoven, Audrey Randall, Ariana Mirian, Gautam Akiwate, Ansel Blume, Lawrence K. Saul, Aaron Schulman, Geoffrey M. Voelker, and Stefan Savage. The dissertation author was a collaborator and contributor to this paper.

Chapter 3, in full, is currently being prepared for submission for publication of material. Ariana Mirian, Grant Ho, Stefan Savage, Geoffrey M. Voelker. The dissertation author was the primary investigator and author of this material.

Chapter 4, in full, is a reprint of the material as it appears in The World Wide Web Conference 2019. Ariana Mirian, Joe DeBlasio, Stefan Savage, Geoffrey M. Voelker, and Kurt Thomas. The dissertation author was the primary investigator and author of this material.

Chapter 2

Measuring Security Practices and How They Impact Security

We start by measuring end user behavior itself. By examining end user behavior, we can empirically quantify the extent to which users follow "best security practices" and also how those practices relate to security outcomes, if at all. In this chapter, I explore end user behaviors of UCSD students who reside in the dormitories via a network vantage point, and how these behaviors relate to device compromise. This analysis allows us to better understand whether certain behaviors should be encouraged or discouraged to improve the security of user devices.

2.1 Overview

Ensuring effective computer security is widely understood to require a combination of both appropriate technological measures and prudent human behaviors; e.g., , rapid installation of security updates to patch vulnerabilities or the use of password managers to ensure login credentials are distinct and random. Implicit in this status quo is the recognition that security is not an intrinsic property of today's systems, but is a byproduct of making appropriate choices — choices about what security products to employ, choices about how to manage system software, and choices about how to engage (or not) with third-party services on the Internet. Indeed, the codifying of good security choices, commonly referred to as security policy or "best practice", has been a part of our lives as long as security has been a concern.

However, establishing the value provided by these security practices is underexamined at best. First, we have limited empirical data about which security advice is adopted in practice. Users have a plethora of advice to choose from, highlighted by Reeder et al.'s recent study of expert security advice, whose title — "152 Simple Steps to Stay Safe Online" — underscores both the irony and the variability in such security lore [80]. Clearly few users are likely to follow all such dicta, but if user behavior is indeed key to security, it is important to know which practices are widely followed and which have only limited uptake.

A second, more subtle issue concerns the efficacy of security practices when followed: Do they work? Here the evidence is scant. Even practices widely agreed upon by Reeder's experts, such as keeping software patched, are not justified beyond a rhetorical argument. In fact, virtually all of the most established security best practices — including "use antivirus software", "use HTTPS/TLS", "update your software regularly", "use a password manager", and so on — have attained this status without empirical evidence quantifying their impact on security outcomes. Summarizing this state of affairs, Herley writes, "[Security] advice is complex and growing, but the benefit is largely speculative or moot", which he argues leads rational users to reject security advice [44].

Our existing models of security all rely on end users to follow a range of best practices. However, we neither understand the extent to which they are following this advice, nor do we have good information about how much this behavior ultimately impacts their future security.

This chapter seeks to make progress on both issues — the prevalence of popular security practices and their relationship to security outcomes — via longitudinal empirical measurement of a large population of computer devices. In particular, we monitor the online behavior of 15,291 independently administered desktop/laptop computers and identify per-device security *behaviors*: is the software patched, how quickly their software is patched, as well as concrete security *outcomes* (i.e., , whether a particular machine becomes compromised). In the course of this work, we describe three primary contributions:

1. Large-scale passive feature collection. Our results are based on large-scale measurement using passive monitoring. In doing so, we develop and test a large dictionary of classification rules to indirectly infer software state on monitored machines (e.g., , that a machine is running antivirus of a particular brand, or if its operating system has been updated). In addition, to ensure that features are consistently associated with particular devices, we describe techniques for addressing a range of aliasing challenges due to DHCP and to DNS caching.

2. Outcome-based analysis. We use a combination of operational security logs and network intrusion detection alerts to identify the subset of machines in our data set that are truly compromised. This outcome data allows us to examine the impact of adopted security practices in terms of individual security outcomes and with respect to concrete time periods surrounding the likely time of compromise.

3. Prevalence and impact of security practices. For our user population, we establish the prevalence of a range of popular security practices as well as how these behaviors relate to security outcomes. We specifically explore the hypotheses that a range of existing "best practices" are negatively correlated with host compromise or that "bad practices" are positively correlated. We consider both behaviors that could *directly* lead to compromise and those which may *indirectly* reflect a user's attentiveness to security hygiene.

Finally, while we find a number of behaviors that are positively correlated with host compromise, few "best practices" exhibit the negative correlations that would support their value in improving end user security.

2.2 Methodology

Our measurement methodology uses passive network traffic monitoring to infer the security and behavioral practices of devices within a university residential network. This

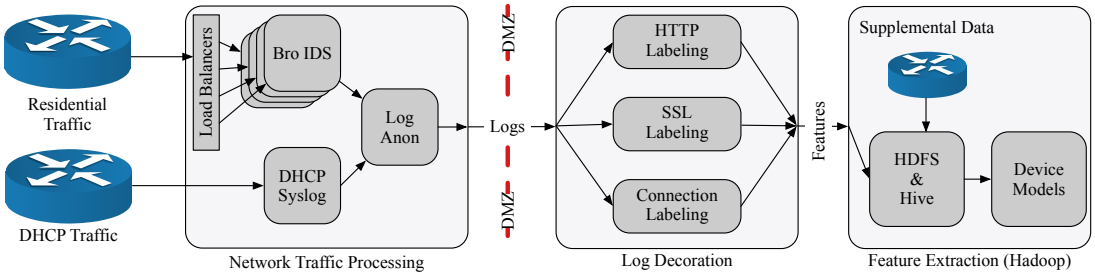


Figure 2.1. System architecture overview. Network traffic is first processed into logs and its addresses anonymized. The next stage replays the network traffic logs to extract further information and label each connection with (also anonymized) MAC address information. The decorated logs are then stored in Hive where they are labeled with security incidents, security practice features, and behavioral features. Lastly, device models are created for analysis.

approach has numerous advantages, including scalability (we are able to collect data from tens of thousands of devices) and granular analysis (we can frequently infer when a device updates a particular application and to what version). However, it also introduces liabilities (a focus on a particular population) and risks (in particular to privacy). In this section we first focus on the technical aspects of our data collection methodology and then discuss some of its attendant challenges and limitations.

2.2.1 Network Traffic Processing

The first stage of our system takes as input 4–6 Gbps of raw bi-directional network traffic from the campus residential network, and outputs logs of processed network events at the rate of millions of records per second. As part of this stage, campus IP addresses are anonymized and, to track the contemporaneous mapping of IP addresses to device MAC addresses, this stage also collects and compatibly anonymizes contemporaneous DHCP syslog traffic.

Residential Network Traffic

As shown in the Network Traffic Processing stage of Figure 2.1, our server receives network traffic mirrored from a campus Arista switch using two 10G fiber optic links. In addition to load balancing, the switch filters out high-volume traffic from popular content distribution networks (CDNs) (e.g., , Netflix, YouTube, Akamai, etc.), resulting in a load of 4–6 Gbps of

traffic on our server.

To minimize loss while processing traffic, we experimented with a number of network processing configurations before settling on the following. We use the PF_RING ZC (Zero Copy) framework [69] to move traffic from the network card directly into user-level ring buffers, bypassing the kernel. We then use the `zbalance_ipc` application from PF_RING ZC to locally perform 4-tuple load balancing across many virtual network interfaces. Instances of the Bro (now Zeek) IDS [73] then read from each virtual network interface, consuming and processing the network traffic into a custom log format. This configuration results in an average daily loss of 0.5% of received packets throughout our six-month measurement period.

While IDS are typically used for detecting threats and anomalous network behavior, we use Bro to convert network traffic into logs since it is extensible, discards raw network traffic as soon as a connection is closed (or after a timeout), and is able to parse numerous network protocols [117]. We also customize the Bro output logs to record only information needed to identify security practice and behavioral features.

In particular, we use the `HTTP`, `SSL`, `DNS`, and `Connection` protocol analyzers. The `HTTP` analyzer provides a summary of HTTP traffic on the network, including components such as the HOST and URI fields. The `SSL` analyzer extracts the SNI field from TLS connections. SNI is an extension of the TLS protocol enabled by most modern browsers, and allows a client to indicate the hostname it is contacting at the start of an encrypted connection. The SNI field is particularly useful for inferring the destination of connections that otherwise are encrypted. The `DNS` analyzer provides a summary of DNS requests and responses. Lastly, the `Connection` analyzer summarizes information about TCP, UDP, and ICMP connections.

Every thirty minutes Bro rotates the previous logs through an address anonymization filter that encrypts campus IP addresses. At this stage of processing, the logs contain IP addresses and not MAC addresses since DHCP traffic is not propagated to our network vantage point. After being so anonymized, the logs are rotated across the DMZ to another server for further processing (Section 2.2.2).

DHCP Traffic

The server also runs a syslog collector that receives forwarded DHCP traffic from the residential network's DHCP servers. DHCP dynamically provides an IP address to a device joining the network. The IP address is leased to the device (by MAC address) for a specified duration, typically 15 minutes. Since we need to track a device's security and behavioral practices for long time periods, we utilize this IP-to-MAC mapping in later processing.

Similar to the Bro IDS logs, every thirty minutes we process the previous DHCP traffic into a (MAC address, IP address, starting time, lease duration) tuple. Then, the entire IP address and identifying lower 24-bits of the MAC address are encrypted using a similar address anonymization filter. The anonymized DHCP logs are then rotated across the DMZ to the Log Decoration server.

2.2.2 Log Decoration

The second stage takes as input these intermediate network event and DHCP logs, and processes them further to produce a single stream of network events associated with (anonymized) device MAC addresses and domain names.

Associating Flows to Devices. Our goal is to model device behavior based upon network activity over long time spans. While we identify unique devices based upon their MAC address, the network events that we collect have dynamically assigned IP addresses. As a result, we must also track dynamic IP address assignments to map IP-based network events to specific device MAC addresses.

We use a Redis key-value store [76] to build a DHCP cache by replaying campus DHCP logs. We use the DHCP cache to assign a MAC address to the inbound and outbound IP of each connection. We consider an IP-to-MAC mapping valid if a connection takes place during the time when the IP address was allocated and the lease is still valid. In the event that there is not a valid mapping (e.g., , the IP address is a non-university IP, or a the device uses a static IP), we do

not assign a MAC address to the IP.

Associating Flows to Domains. When using network activity to model device behavior, it useful to know the domain name associated with the end points devices are communicating with (e.g., categorizing the type of web site being visited). We also extract the registered domain and TLD from each fully qualified domain name using the Public Suffix List [68]. Again, since the network events we observe use IP addresses, we must map IP addresses to domain names. And since the mapping of DNS names to IP addresses also changes over time, we also dynamically track DNS resolutions as observed in the network so that we can map network events to the domain names involved.

Due to our network vantage point (at the campus edge), the DNS traffic our collection server observes generally has the source IP address of our local DNS resolver, and *not* the IP address of the host which will subsequently make a connection to the resolved IP.[1] This constraint limits our ability to use the DNS mapping alone to infer a connection's domain name. Therefore, one of the steps in this stage is to build a local DNS cache by replaying the logs in chronological order and labeling the domain name of observed connections where it is not already provided (i.e., , excluding HTTP and SNI-labeled connections).

We use another Redis key-value store to build a DNS cache by replaying DNS traffic. The cache tracks the mappings of each IP address to domain name at the time the IP address was observed. We consider a mapping to be valid as long as it has not expired — the log time falls between the time at which the DNS request was observed plus the response TTL — and there is one registered domain name mapped to the IP address.

When sites use virtual hosting, it is possible that an IP address has multiple domain names associated with it. In this case, we first check if the registered domain names match (e.g., , bar.bar.com and car.bar.com share a registered domain of bar.com). If the registered domains match, we label the connection using the longest suffix substring match (e.g., , ar.bar.com) and set a flag indicating that the fully qualified domain name has been truncated. In the case where

[1] The primary exceptions are devices configured to use remote DNS resolvers.

there is more than one registered domain with a valid mapping to the IP address, we do not use the mapping to label connections until enough of the conflicting mappings expire such that they share a registered domain, or there is only one mapping.

User Agent. We parse HTTP user agent strings using the open-source `ua-parser` library. From the user agent string we extract browser, OS, and device information when present.

2.2.3 Feature Extraction

In the final stage of our system we store the log events in a Hive database [4] and process them to extract a wide variety of software and network activity features associated with the devices and their activity as seen on our network. The last critical feature is device outcomes: knowing when a device has become compromised. We derive device outcomes from a log of alerts from a campus IDS appliance, and also store that information in our database.

Software Features

To identify features describing application use on devices, we crafted custom network traffic signatures to identify application use (e.g., a particular browser) as well as various kinds of application behavior (e.g., a software update).

To create our network signatures we use virtual machines instrumented with Wireshark [98]. We then manually exercise various applications and monitor the machine's network behavior to derive a unique signature for each application. Fortunately most applications associated with security risk frequently reveal their presence when checking for updates. In total, we develop network signatures for 68 different applications, including OS. For a subset of applications, we are also able to detect the application's version. Knowing application versions allows us to compare how fine-grained recommended security practices (i.e., updating regularly) correlates with device compromise.

Operating System. We created six signatures to identify the OS running on devices. Since regular OS updating is a popular recommended security practice, we also created signatures

to detect OS updates. While Windows and Mac OS operating system updates are downloaded over a CDN that is removed from the network traffic before reaching our system (Section 2.2.1), we can use OS version information from the host header and User-Agent string provided in HTTP traffic to infer that updates have taken place.

Detecting Security Incidents

While previous work has relied on the use of blacklists or Google Safe Browsing to identify devices that expose users to potential risk, we are able to identify compromised devices with high confidence as a result of post-infection behavior, typically in the form of CNC communication [12, 86]. To identify compromised devices (i.e., , ones with a security incident) we use alerts generated by a campus network appliance running the Suricata IDS [96]. The campus security system uses deep packet inspection with an industry-standard malware rule set to flag devices exhibiting post-compromise behavior [75].

The IDS rules also detect network activity that might occur before a device becomes compromised (e.g., , possible phishing attempts, exploit kit landing pages, etc.). Since we focus on compromised devices, we reduce the rules we consider to ones that explicitly detect post-infection behavior. False positives are likely with any real-world signature-based intrusion detection system. To minimize the frequency of false positives, we manually remove rules that are frequently triggered, but do not indicate that a device has been compromised.

2.3 Ethical Considerations

Having described our measurement methodology in considerable detail, we now consider the risks it presents – both to the privacy of network users and to the validity of conclusions drawn from these measurements.

Protecting user privacy. Foremost among the risks associated with the passive measurement approach is privacy. Even with the prevalence of encrypted connections (e.g., via TLS), processing raw network data is highly sensitive. From an ethical standpoint, the potential

benefits of our research must be weighed against potential harms from any privacy violations. In engaging with this question — and developing controls for privacy risk — we involved a broad range of independent campus entities including our institutional review board (IRB), the campus-wide cybersecurity governance committee and our network operations and cybersecurity staff. Together, these organizations provided necessary approvals, direction and guidance in how to best structure our experiment, and strong support for the goals of our research. The campus security group has been particularly interested in using our measurements to gain insight into the security risks of devices operating on their network.[2]

Operationally, we address privacy issues via minimization, anonymization and careful control over data. First, as soon as each connection has been processed, we discard the raw content and log only metadata from the connection (e.g., a feature indicating that device X is updating antivirus product Y). Thus, the vast majority of data is never stored. Next, for those features we do collect, we anonymize the campus IP and the last 24-bits of each MAC address, using a keyed format-preserving encryption scheme [6].[3] Thus, we cannot easily determine the identity of which machine generated a given feature and, as a matter of policy, we do not engage in any queries to attempt to make such determinations via re-identification. Finally, we use a combination of physical and network security controls to restrict access to both monitoring capabilities and feature data (this is to help foreclose the possibility that any outside party, not bound by our policies, is unable to access the data or our collection infrastructure). Thus, the server processing raw network streams is located in a secure campus machine room with restricted physical access, only accepts communications from a small static set of dedicated campus machines and requires multi-factor authentication for any logins. Moreover, its activity is itself logged and monitored for any anomalous accesses. We use similar mechanisms to protect the processed and anonymized feature data, although these servers are located in our

[2]Indeed, during the course of our work we have been able to report a variety of unexpected and suspicious activity to campus for further action.

[3]Thus, the IP address `192.168.0.1` may be replaced with `205.4.32.501` and the MAC address `00:26:18:a5:38:24` may become `00:26:18:b5:fe:ba`. We do not anonymize the organizationally unique identifier (OUI) to allow us to derive the network device manufacturer.

local machine room. The feature data set is only accessible to members of our group, subject to IRB and our agreements with campus, and will not (and cannot) be shared further.

Limitations of our approach. In addition to privacy risk, it is important to document the implicit limitations of our study arising from its focus on a residential campus population — primarily undergraduates, as well as the use of a particular IDS and rule set to detect security incidents [96, 75].

It is entirely possible that the behavioral modes of this population, particularly with respect to security, are distinct from older, less affluent or more professional cohorts. This population bias is also likely to impact time-of-day effects, as well as the kinds of hardware and software used. Additionally, the security incidents we consider rely on the Suricata IDS, commercial network traffic signatures, and security-related network usage requirements of our university environment (e.g., , residential students are nominally required to have antivirus software installed on their devices before connecting). It is entirely possible that these incident detection biases also influence the behaviors and software applications that correlate with device compromise. Thus, were our same methodology employed in other kinds of networks, serving other populations, using different security incident detection techniques, it is possible that the results may differ. For this reason, we hope to see our measurements replicated in other environments.

2.4 Recommended Practices

There are a variety of security practices widely recommended by experts to help users become safer online. Prior work has explored some of these practices in terms of users being exposed to risky web sites [12, 86]. Since our data includes actual security outcomes, we start our evaluation by exploring the correlation of various security practices to actual device compromises in our user population: operating system choice and keeping software up to date.

	Operating System	*Incidents*	*Total Devices*
	Windows	538 (7.0%)	7,668
	Mac OS	140 (1.9%)	7,339
	ChromeOS	1 (0.5%)	205
	Linux Variant	3 (3.8%)	79

Figure 2.2. Device operating system classification after removing Internet-of-Things and mobile devices, including the total number of devices with each operating system and the number with a security incident.

2.4.1 Operating System

Different operating systems have different security reputations, so it is not surprising that experts have recommendations of the form "Use an uncommon OS" [80]. Part of the underlying reasoning is that attackers will spend their efforts targeting devices with most common systems, so using an uncommon operating system makes that device less of a target.

In terms of device compromise, as with previous work and experience, such advice holds for our user population as well. Using the OS classification method described above, Table 2.2 shows the number of devices using major operating systems and the number of each that were compromised during our measurement period. Most devices use Windows and Mac OS, split nearly equally between the two. The baseline compromise rate among devices is 4.5%, but Windows devices are 3.9× more likely to be compromised than Mac OS devices. The Chrome OS population is small, but only one such device was compromised.

Of course, modulo dual-booting or using virtual machines, this kind of advice is only actionable to users when choosing a device to use, and is no help once a user is already using a system.

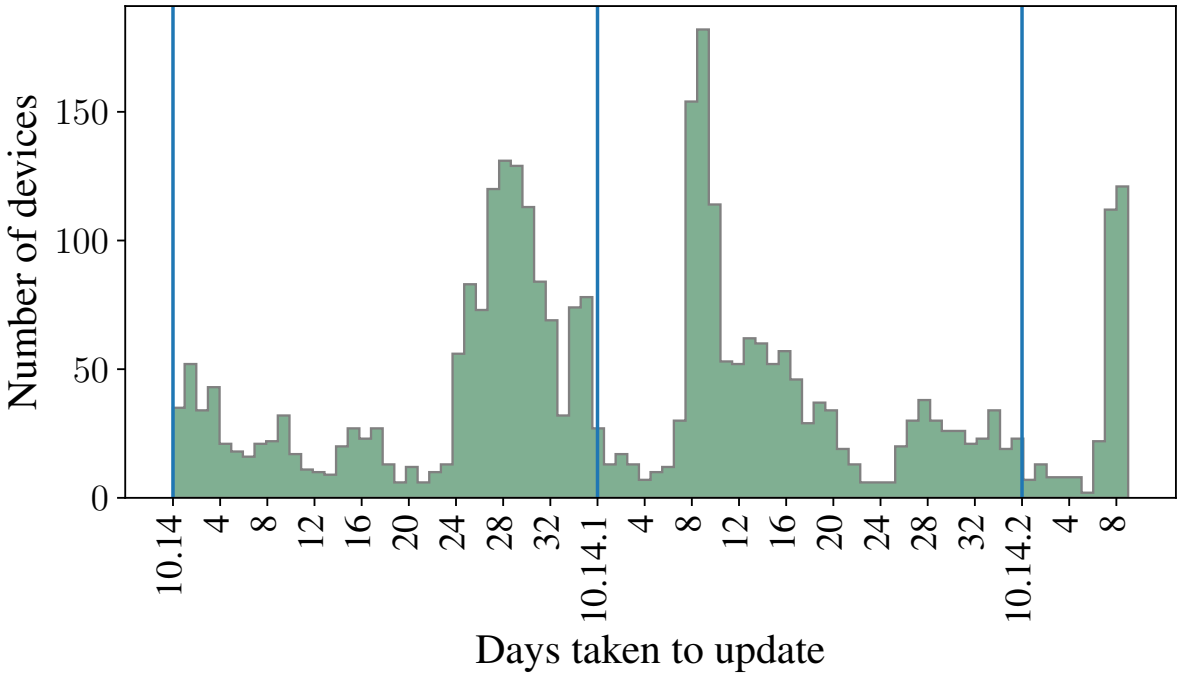


Figure 2.3. Number of days a Mac OS X device takes to update to a specific version. The version number on the x-axis denotes the day that the specified version update was published.

2.4.2 Update Software

Among hundreds of security experts surveyed, by far the most popular advice is to "Keep systems and software up to date" [80]. In this part we explore the operating system, browser, and Flash update characteristics of the devices in our population, and how they correlate with device compromise.

Operating System

Mac OS. We start by analyzing the update behavior of devices running Mac OS. Our system labels each HTTP connection of a device with the type of operating system and its current version number, both extracted from the User Agent string. However, if a device leaves the network and returns with an updated version number in the UA string, then we cannot accurately tell when the device was updated. Thus, we only utilize devices that are absent for less than four days to bound the error on update times. We see 7,268 (47.5%) devices that identify as Mac

according to the User Agent string. Of these devices, we see at least one update for 2,113 (29.1% of all Mac OS devices). Figure 2.3 shows the update pattern of these Mac OS devices over time, anchored around the three OS updates released by Apple during our measurement period. In general, Mac OS users are relatively slow to update, anecdotally because of the interruptions and risks Mac OS updates entail.

Of these devices, 57 (2.7%) of them were compromised. Compromised devices have a mean and median update rate of 16.21 and 14 days, respectively, while their clean counterparts have a mean and median update rate of 17.96 and 16 days. However, this difference is not statistically significant according to the Mann-Whitney U test ($p = 0.13$).[4]

Web Browser

Updating the browser may be as important as updating the operating system. Browsers are also large, complex pieces of software used on a daily basis and, as with most software, these large programs have vulnerabilities. Updating is viewed as such an important process that Chrome and Firefox employ auto-updating by default [103, 28], with UI features to encourage timely updating.

As such, we explore the relationship between compromised and clean devices and browser updating behaviors. Similar to the Mac OS devices, we are able to detect the current browser version number from the User Agent string of a device. Since browser vendors publish the dates when they make updates available,[5] we can check whether the browser on a device is out of date each time we see the device on the network. Across the measurement period, we then calculate how quickly devices update. Also similarly to the Mac OS analysis, we exclude devices that are absent from the network for more than three days.

Moreover, we only analyze the dominant browser for each device. Many devices have User Agent strings naming different browsers. While users may use different browsers for

[4]The Mann-Whitney U test is a non-parametric statistical test that can be used to determine if two independent samples are selected from populations with the same distribution. The null hypothesis for a Mann-Whitney U test is that the populations are selected from the same distribution.

[5]During our measurement period each popular browser had at least three major updates.

Table 2.1. Number of days between when an update is published and when devices update. Compromised devices update faster than their clean counterparts.

Browser	Mean, Median, # (Cmp)	Mean, Median, # (Cln)
Chrome	14.4, 15.0 (421)	15.4, 15.0 (7883)
Firefox	5.64, 3.00 (24)	9.65, 5.00 (424)

different use cases, we identify a dominant browser to remove the noise from user applications that spoof a browser in their User Agent string. Thus, we determine which browser connects to the largest number of distinct registered domains from a device and label the device with that dominant browser. We choose unique registered domains as our metric over number of HTTP connections because there are web sites and applications that "spam" the network, making the device appear to use one browser dominantly when the natural user behavior is actually coming from a different browser.

We analyzed updates for devices that dominantly use Chrome, Edge, Firefox, and Safari. Of the total devices, 10,831 (70.8%) devices use Chrome, 719 (4.7%) devices use Edge, 561 (3.7%) devices use Firefox, and 2993 (19.6%) devices use Safari. However, only 8,304 (76.7%) of the Chrome devices, 132 (18.4%) of the Edge devices, 448 (80.0%) of the Firefox devices, and 1592 (53.2%) of the Safari devices are on the network continuously (absent for less than three days).

Table 2.1 shows the browsers with statistically significant differences in update time between clean and compromised devices (Mann Whitney U: Chrome $p = 4.2 \times 10^{-4}$ and Firefox $p = 0.03$).

Clean devices appear to spend more time out of date than their compromised counterparts. Examining this phenomenon in more detail, we compare the update behavior of compromised devices before and after their compromise date. We focus on devices using Chrome that have two updates spanning the compromise event (other browsers do not have a sufficiently large sample size). Figure 2.4 shows the distribution of times devices were out of date with respect to when a browser update was released for updates before and after the device was compromised. The shift

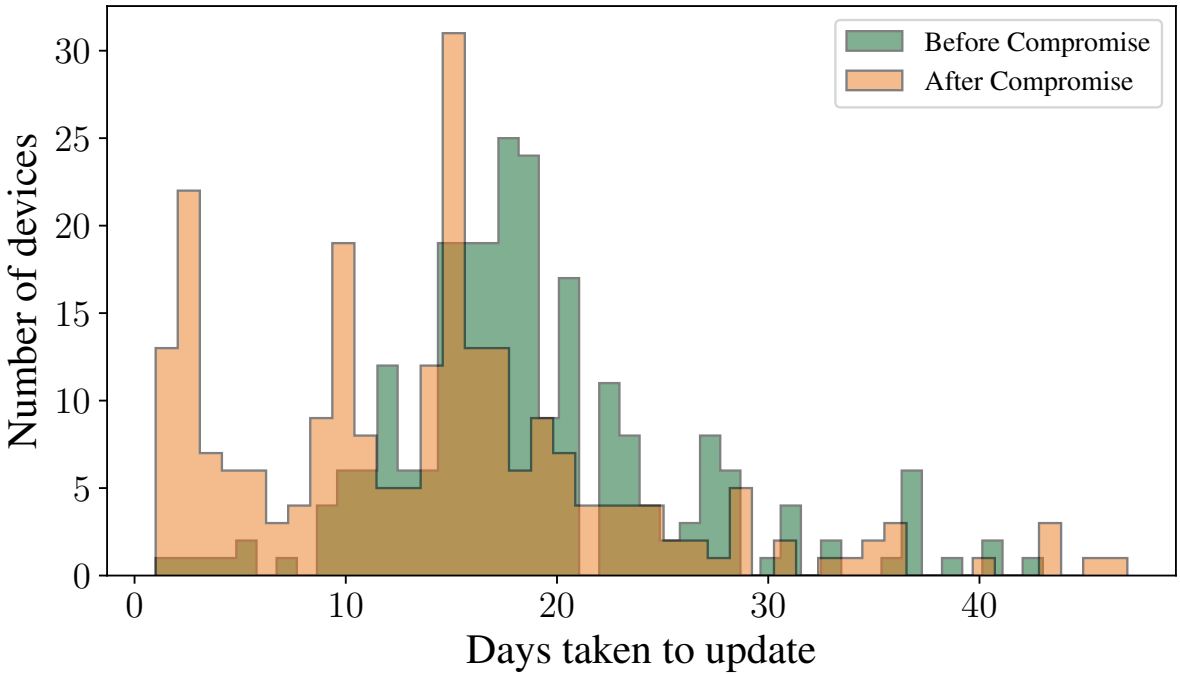


Figure 2.4. Distribution of days a device takes to update Chrome before compromise and after compromise.

in distributions illustrates that devices update faster after compromise. In more detail, devices that use Chrome have a before-compromise mean update rate of 18.9 days (18.0 median days) and an after-compromise mean update rate of 14.2 days (15.0 days median). This difference is significant, with $p = 4.8 \times 10^{-12}$ using the Wilcoxon signed-rank test.[6]

2.5 Related Work

This study follows a large body of prior work that empirically relates user activity to various risk factors, which we highlight in five categories below.

Small scale studies of individuals. In 2008, Carlinet et al. [13] analyzed three-hour long packet traces of ADSL customers (from 200–900 customers) and correlated hosts that experienced at least one Snort IDS alert with other factors. Their study revealed a relationship between those machines raising alerts, and their use of the Windows operating system as well as

[6]The Wilcoxon signed-rank test is a non-parametric paired difference test which indicates if the means of two dependent samples differ. The null hypothesis of the Wilxocon signed-rank test is that the means do not differ.

heavy web browsing habits. Our study is similarly based on passive network data collection, but we operate at a significantly larger scale (in number and diversity of hosts as well as duration) and we also explicitly try to control for a range of confounding factors.

Aggregate studies of user behavior. Others have studied risk factors in aggregate across large organizations. Notably, Yang et al. [59] correlated publicly-declared data breaches and web site hacks with external measurements (e.g., misconfigured DNS or HTTPS certificates). They found that evidence of organizational failures to police security is predictive of attacks. Similarly, recent papers have focused on exploring how differences in deployed defenses (e.g., across ISPs or web sites) relate to the occurrence of particular attacks [108, 97], and Xiao et al. [116] showed that user patterns of security activity can be a predictor of future malware outbreaks in an ISP.

Web access behavior. Other researchers have investigated how a user's web browsing habits reveal risk factors. Levesque et al. [53] monitored web browser activity for 50 users over four months and found that the likelihood of visiting a malware hosting site was correlated with the other kinds of sites a machine visited (e.g., with P2P and gambling sites). Canali et al. [12] replicated this study using antivirus telemetry (100,000 users), and Sharif et al. [86] describe a similar analysis for 20,000 mobile users. Both found that frequent, nighttime, and weekend browsing activity are correlated with security risk.

Software Updates. Another vein of research has correlated poor software update habits with indicators of host compromise. Kahn et al. [48] used passive monitoring of roughly 5,000 hosts to infer software updates and used the Bothunter traffic analysis tool [38] to infer likely infected hosts based on suspicious traffic patterns (e.g., based on outbound scanning). They found a positive correlation between infection indicators and a lack of regular updating practice.

At a larger scale, Bilge et al. [8] used antivirus logs and telemetry from over 600,000 enterprise hosts to retrospectively relate software updates to subsequent infections. They found that devices that do not patch correlate with those that were at some point infected. Finally, Sarabi et al. [83] used a similar data set of 400,000 Windows hosts and found that patching faster provides limited benefit if vulnerabilities are frequently introduced into product code.

Human factors. Finally, there is an extensive literature on the human factors issues involved in relating security advice to users, the extent to which the advice leads to changes in behaviors, and how such effects are driven by both individual self-confidence and cultural norms [31, 84, 111, 77, 78, 79, 112, 113].

2.6 Discussion

The practice of cybersecurity implicitly relies on the assumptions that users act "securely" and that our security advice to them is well-founded. In this chapter, we have sought to ground both assumptions empirically: measuring both the prevalence of key security "best practices" as well as the extent to which these behaviors (and others) relate to eventual security outcomes. We believe that such analysis is critical to making the practice of security a rigorous discipline and not simply an art.

However, achieving the goal of evidence-based security is every bit as formidable as delivering evidence-based healthcare has proven to be. In any complex system, the relationship between behaviors and outcomes can be subtle and ambiguous. For example, our results show that devices using Windows are significantly more likely to be compromised. This is a factual result in our data and is in opposition to the "best practice" that using a common OS will more likely protect a user. However, there are a number of potential explanations for *why* this relationship appears: since the Windows operating system is more widely used, and has been in use for longer than its counterparts, attackers have developed a myriad of attacks in order to target a large number of victims simultaneously.

Thus, while some of our results seem likely to have explanatory power others demand more study and in a broader range of populations. Those results that lack simple explanations are a reflection of the complexity of the task at hand and force us to question which security practices are truly best to prioritize in a user population.

Chapter 2, in part, is a reprint of the material as it appears in the Proceedings of the

International Measurement Conference 2019. Louis F. DeKoven, Audrey Randall, Ariana Mirian, Gautam Akiwate, Ansel Blume, Lawrence K. Saul, Aaron Schulman, Geoffrey M. Voelker, and Stefan Savage. The dissertation author was a collaborator and contributor to this paper.

Chapter 3

Passwords

End user behavior is critical in understanding how to best prioritize security processes. Just as crucial, however, is understanding an organization's efforts in changing user behavior to employe better security practices. In this chapter, I explore UCSD's efforts in motivating it's user population to change the passwords of their campus accounts, and what communication mecahnisms are most effective in prompting organizational security changes.

3.1 Overview

Enterprise-wide mandatory password updates are inevitably fraught affairs. Typically driven by either a change in circumstances (e.g., , evidence of a data breach) or security policy (e.g., , requirements for longer or more complex passwords), such mandates require that all members of an organization update their Single-Sign On (SSO) passwords within a set time period. These dual requirements of completeness and timeliness are particularly challenging given the limited resources of IT service departments. Scale requires that instructions be delivered via mass communication (e.g., , email), yet they must contend with a broad spectrum of understanding, capability, and incentives in the user population. Unsurprisingly, there are few established best practices for how to achieve these goals, and limited empirical data about how to most effectively enact this change at enterprise scale.

This chapter seeks to address this deficit through the empirical analysis of a mandatory

password update event at our institution, one which required almost 10,000 faculty and staff to take independent action. Using data from this experience, we explore how the *operational requirements* of coordinated enterprise-scale password changes — timeliness, completeness, and staff overhead — interact with the behavioral and organizational aspects of the problem that have the potential to create friction.[1] We are guided by concretely motivated questions that, were the answer understood, would directly inform operational practice, such as: How long does it take to effect institution-wide password updates? What impact do notifications have on user compliance? What factors predict efficient password updating behavior and how significant is the staff overhead in managing user problems during the process?

Our work combines detailed records of user notification events, password update logs, and IT help desk reports, to empirically deconstruct the synchronized password update process across our campus population. In doing so our work makes three primary analysis contributions:

1. *Communication Effectiveness*. We demonstrate the effectiveness of repeated email requests in driving timely password update behavior—characterizing how much of the population is responsive to serial pleas over time and what subset is not reached and/or motivated by such efforts. We also analyze the effects of Web-based interstitial login reminders in galvanizing this unresponsive remainder into action.

2. *Completeness hazards*. It is common during such updates to track the fraction of user accounts that have complied with the password update edict. After correcting for inactive accounts (e.g., , for users who have left the institution), we identify the small subset of users who are ultimately unable to meet the password update burden. We show that this set is over-represented in business units whose job function does not require regular computer use.

3. *Quantified IT overhead*. Finally, we explore the costs to IT organizations in supporting

[1] We specifically *do not* focus on issues such as how password policies interact with password strength, for which there is an extensive literature [118, 10, 114, 63, 115, 20, 29, 105, 74, 52].

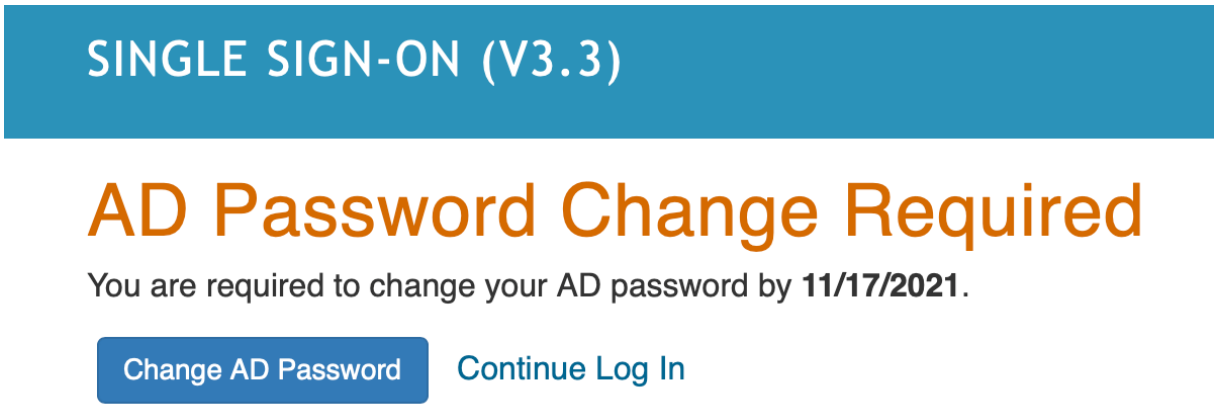


Figure 3.1. Example of the browser intercept message that campus's SSO portal displayed to users who had not updated their passwords by mid October 2021.

universal mandatory password updates, using the number of help desk tickets as a proxy for the IT staff time that must be spent to help shepherd users through the process.

From these results we provide guidelines for reasoning about mandatory password update costs in terms of effectiveness and IT staff effort. We believe this is a pragmatic example of a more general and analytical approach to managing enterprise IT security processes.

3.2 Background

This study describes a natural experiment driven by a security policy directive that required all users at our university to update their campus Active Directory passwords, used for Single-Sign On (SSO) across a range of university IT services (i.e., , including organizational e-mail, calendaring, financial services, etc.). For a variety of reasons, campus faculty and staff were prioritized in this effort and thus our work focuses on the experience of this population.

In the summer of 2021, our campus Information Technology Services (ITS) team enacted a campaign to reach out to affected employees, inform them of this policy, and direct them to online resources for updating their passwords.[2] These resources included two self-service Web portals: one for updating passwords after a valid login and one for (re)setting a password without

[2]We were not involved in the design or implementation of this password change campaign and are simply studying its effects retrospectively.

a valid login (requiring employee specific identification). As well, employees using "managed" Windows or Mac devices were able to update their SSO password locally with a valid login.

Employees *new* passwords were required to be different from their previous password, to be at least 12 characters in length, to not include their username as a substring, and to utilize three of four character classes (uppercase, lowercase, numbers, symbol).[3] Our work does not concern the quality of the resulting passwords, but we document these requirements to the extent that the additional burden may have caused some users to delay or fail to change their password as directed.

The password update campaign consisted of three kinds of actions performed by ITS staff: asynchronous reminder emails, synchronous login intercepts, and actively resetting non-compliant users' passwords to random strings ("password scrambling"). Initially, a campus-wide email was sent to all employees on August 10th notifying them of the upcoming password update requirement. After this initial email, there were two stages of correspondence. The first stage consisted of a set of four email messages (we refer to them as communications) that were sent to disjoint "waves" of users that were staggered in time. Waves were segregated based on the first letter of a user's last name: A–B, C–G, H–N, O–Z. Each subsequent wave increased in size as the ITS team became increasingly confident about their ability to manage technical or user issues that arose.

Each wave received the same set of four email messages, each of which was staggered by one week, as shown in Table 3.1. For example, users in Wave 1 received their initial communication on August 18th, a second on August 25th, a third on September 1st, and a final communication on September 8th. If a user updated their password, they did not receive subsequent communications.[4]

[3]This requirement, as well as a further filter against using "known compromised" passwords provided by a third-party service, were enforced mechanically by rejecting new passwords that did not comply with these requirements.

[4]Because communication lists were constructed by querying the Active Directory (AD) system for password update information, updates were not strictly atomic. Thus, a user who updated their password after the second communication list was constructed, but before it was sent, would still receive the reminder even though they had already updated their password.

Table 3.1. Dates for the email communications sent during each of the four waves.

Wave	Comm #1	Comm #2	Comm #3	Comm #4
Wave 1	2021/08/18	2021/08/25	2021/09/01	2021/09/08
Wave 2	2021/08/25	2021/09/01	2021/09/08	2021/09/15
Wave 3	2021/09/01	2021/09/08	2021/09/15	2021/09/22
Wave 4	2021/09/08	2021/09/15	2021/09/22	2021/09/29

The first three email communications were very similar to each other, and the fourth differed slightly. The first email served as the initial notification, informing users that they needed to update their password, and that their deadline was four weeks from the initial email. The second and third email reiterated the deadline and requirement to update the password. The fourth email ("last wave communication") did NOT mention any deadline, but instead informed users that this was their *final* notification, and that they should "Avoid account access complications and change your AD password now".

The second stage of the campaign started roughly one month after the last communication of Wave 4. During this second stage, users who had not updated their password received an active notification (an "SSO intercept") each time they logged into a campus service. These intercept messages were initially rolled out to a small subset of users and gradually deployed to all users who had not updated their password. As seen in Figure 3.1, the login intercept told users that they were required to update their AD password by a certain date (and provided an inline button that, if clicked, brought them to the password update portal). After the deadline passed, this intercept became modal and would not allow a login without a password update.

Finally, two more email notifications were emailed to users, which we refer to as the "Final" notifications and "Scramble" notifications. These notifications were sent in conjunction with the later stages of the login intercept to further convince users to update their password. The "Final" email communication told users that "Unless changed, your AD password will expire on <Deadline>", while the "Scramble" notifications informed users that "Your AD Account password will be removed on <Deadline> and you will lose access to all AD-accessed university

systems...".

Any users who had not updated their password after receiving these final messages and SSO intercepts had their account password "scrambled" (i.e., , set to a random value) by an ITS administrator. Such users would thus be unable to login to the vast array of campus IT services and would need to trigger the password reset mechanism themselves or with the help of the ITS help desk to obtain a new valid password. We were given the list of users whose passwords had been scrambled as well as the date and time at which this action was taken.

Our university has a number of closely-affiliated but semi-independent organizations, such as separately endowed research institutes and a medical center, which have their own IT infrastructure. A small subset of accounts in our data set reflect "secondary accounts" of users who have a primary appointment at one of these sister organizations, but who happen to have an account in main campus's IT systems as a result of joint initiatives. As we discuss later (§ 3.5), these users might not frequently check or use these secondary campus accounts, since their day-to-day online activities could revolve around an account at their home organization.

3.3 Ethics

Our analysis does not expose any vulnerabilities, nor does it indirectly create harms by virtue of its results. The benefits of our research include better understanding the dynamics around mandatory password policy changes, how to do so more efficiently and, by generalization, improving mass compliance with other changes in security policy. Our analysis is based on secondary use of data already routinely logged by our institution's IT services group and this data is de-identified for our analysis. Further, we only pursue analyses of population aggregates and do not present results about individual users (even de-identified). Our project has been reviewed by our institutional review board (IRB) and considered exempt. Additionally, our work takes place with the full knowledge of our institution's CISO and with the associated IT staff (our work is driven, in part, by helping this organization understand how to better manage their security

communications).

3.4 Methodology

In this section, we discuss our university authentication process, the data sources we used, and the set of accounts we focus on in this study.

3.4.1 Authentication into Campus Services

Our institution uses Active Directory (AD) for basic authentication and Duo for two-factor authentication for all major systems. Thus, users accessing campus services ranging from email to payroll first need to login using an Active Directory username and password, and then authenticate via Duo (typically a phone-based app) to access their service. Our Duo deployment supports a *remembrance window* of seven days which, if configured, reduces the two-factor authentication requirement to once per week *per device*.

3.4.2 Data Sources

We conduct our analysis using four data sources from August 2021 to March 2022: Splunk logs, email correspondence logs, Active Directory metadata, and Help Desk tickets. We explain each of these data sources in further detail.

Splunk Logs. Our institution collects various logs of user activity and stores them in Splunk, a third-party service for capturing, indexing, and querying system log information. For this study we use Splunk-managed event logs from our campus' Active Directory and Duo deployments.

The Active Directory logs contain password update information—notably Windows Event IDs 4724 (account password reset attempt) or 4723 (account password change attempt) paired with 4738 (account changed)—as well as metadata about the password update itself (i.e., , who initiated the change). The event codes and metadata allow us to differentiate password updates into four different semantic categories: a password change by a user via a campus

self-service online password change portal, a password change by a user via the user's campus-administered machine (e.g., , via the Windows Sign-in/Password dialog), a password reset by a user via the password change portal, and an administrative reset (e.g., , help desk, departmental IT support).

The Duo logs contain every Duo authentication success and failure for users on campus. For password updates initiated by the online password update portal, users must already be authenticated via both Active Directory and Duo. For password resets initiated via the portal, no authentication is necessary (although failed authentications appear in the logs if the user attempted to authenticate but forgot their password).

Email Correspondence and Scrambled Accounts. The campus security team notified users about the new password update requirements and deadlines via a series of email messages (§ 3.2). These messages used Emma [22], an e-mail marketing service which incorporates a tracking pixel into messages to identify when each email is delivered, opened, or bounced. This team provided the Emma logs to us, as well as which accounts were ultimately scrambled and when.

Active Directory Metadata. Each user profile in Active Directory has additional metadata, including their Organizational Unit (OU). This metadata indicates user roles and departmental affiliations. We use this profile information to correlate behavior with user demographics in our analysis. We have anonymized OU values unique to our institution where necessary to support blind review.

Help Desk Tickets. Finally, we used aggregate statistics collected from logs of campus Help Desk tickets to help understand the IT staff burdens created by the password update campaign. As discussed in more detail in Section 3.6.1, our data consists of coded tickets (i.e., , tagged as related to password updates) from the period in question that are de-identified and tagged with associated OU membership. This process produced 919 password update related tickets submitted by 762 distinct users.

Table 3.2. Distribution of the different kinds of users in our study.

Category	Number of Users
Single Change Users	7925 (81.3%)
Multiple Change Users	1291 (12.2%)
Nonresponsive Users	528 (5.42%)
Total	9744 (100%)

3.4.3 User Population

Finally, for the purposes of our study, we focus specifically on active users who successfully received the password update correspondences. Concretely, we consider users that satisfy the following two criteria:

1) Users successfully contacted. We only consider users who were successfully contacted by the email notification campaign. We consider users "successfully contacted" if the email tracking service indicates that they received (although not necessarily opened) all messages in the notification campaign until they updated their password. This avoids confounding effects caused by non-human accounts that do not have e-mail accounts or the minority of users who, for one reason or another, have no working email point of contact.

2) Users are active. Like any large organization, ours has user accounts that are accessible but largely inactive (e.g., , alumni "email-for-life" accounts). Since we are interested in the behavior of active users—those for whom password expiration will have a direct impact on their activity—we restrict the account population to accounts that have had at least one successful login authentication (both Active Directory and Duo two-factor) during the password update campaign.

Table 3.2 summarizes the user population we consider in this study. Among 9,744 users, 7,925 (81.3%) of them updated their password exactly once during the password campaign ("single change" users), 1,291 (13.2%) updated their password more than once ("multiple change" users), and 528 (5.42%) did not update their password by the communicated deadline ("nonresponsive" users, whose passwords were scrambled by the IT staff due to their failure

to act in a timely fashion). We note that most users are "single change" users—changing their password once during this campaign—with a smaller percentage deviating from this behavior and incurring additional costs (either on individual users or the IT organization).

3.5 User Responsiveness

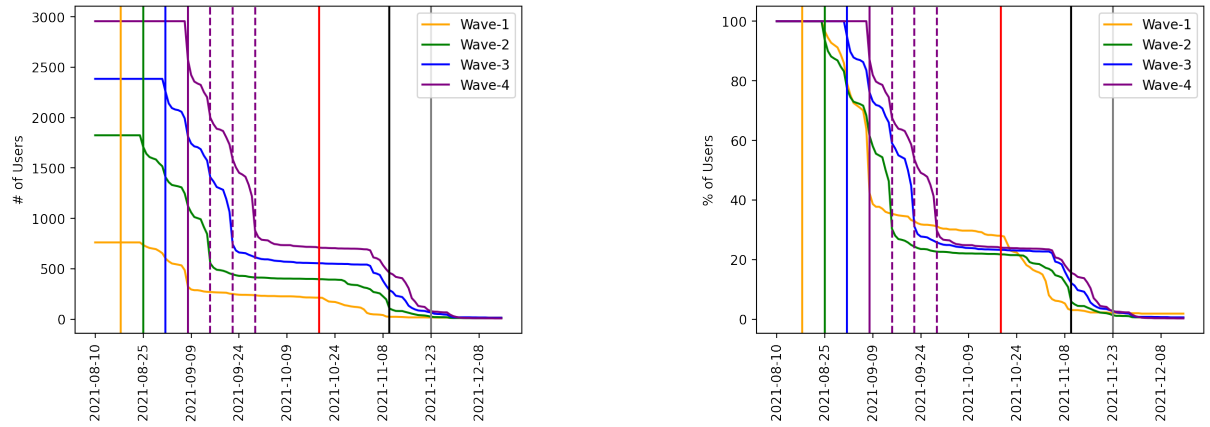
In this section we analyze how the users in our study responded to the password update campaign. In particular, we explore the following research questions:

- RQ1: Were repetitive emails effective in prompting user change?
- RQ2: Were login intercepts effective in prompting user change?
- RQ3: In what ways did multiple change users react differently than single change users?
- RQ4: Which users utilized password reset more than a password change?
- RQ5: Were users who opened email more likely to update their password?
- RQ6: Which organizational units were slower in updating their passwords?

We focus on this set of questions when analyzing user responsiveness to understand which actions are most effective for the organization (RQ1, RQ2, RQ3, RQ5), which update mechanisms are utilized the most and thus should be embraced (RQ4), and why organizations should use different mechanisms for certain employee groups to more effectively promote password updates (RQ6).

3.5.1 Single Change Users

We begin by examining the behavior of single change users. Figure 3.2 shows the password change behavior of these users over time. The left graph (a) shows four curves, each corresponding to one of the communication waves. Each curve shows, on a daily granularity, the remaining number of users in that wave who still need to change their password. The right graph



(a) Number of users in each wave

(b) Percentage of users in each wave

Figure 3.2. (a) The number of single change users *without* a password update in the different waves over time, and (b) the same results showing the percentage of users in each wave. The first four solid lines denote the beginning of the communication series for each wave. The line at October 19, 2021 denotes the start of login intercept, while the line after November 8, 2021 denotes the start of final notifications, and the line at November 23, 2021 for scramble notifications.

(b) shows the same results, but with each wave normalized to the number of users in that wave: the curves show the percentage of users in each wave who have yet to change their password. The solid vertical lines in the graph correspond to the start of various actions taken by campus during the campaign, including when campus sent initial email communications to each wave (first four solid vertical lines), intercepted logins (solid vertical line at October 19, 2021), and final/scrambled notifications (last two solid vertical lines). For a subset of users who had not yet updated their password, the IT staff began scrambling their passwords on November 16, 2021 prior to sending email notifications. Additionally, we note that each wave received four communications, but the communications were staggered by a week and thus are overlapped. We denote the trailing last communications in the final wave with dashed lines. For reference, Table 3.1 shows the dates of each communication in the different waves.

From the timeline in Figure 3.2, we define periods of user activity based on user response to the various notifications. Each wave begins with a "responsive" period that engages with responsive users until seven days after the final communication for a given wave. Each wave then

has an "idle" period between the email notifications and the first use of the login portal intercept. Finally, each wave ends with an "intervention" period engaging with unresponsive users and spanning the login intercepts, expiration warning email communications, and account scrambling. The intervention period is the same for each wave (October 19th until December 15th), but the responsive period and idle period are shifted by each wave start date. For example, the responsive period for Wave 1 is August 18th to September 15th and the idle period is September 15th to October 19th, while for Wave 2 the responsive and idle periods are from August 25th to September 22nd and September 22nd to October 19th, respectively. Combining the waves, 71.0% of these users changed their password during the responsive period, 5.28% changed during the idle period, and 23.7% changed during the intervention period.

RQ1: Repetitive emails are effective in prompting a majority of user updates with diminishing returns. As seen in Figure 3.2 by the stairstep shape of the curves from August 25 to September 29, multiple email communications were necessary and effective for the majority of users. An immediate question for an organization planning to use email notifications is how many iterations to perform. We measure effectiveness of each iteration by quantifying the number of users who initiated a password update within a week of a given communication. For our campus, multiple communications was clearly impactful and the plan of four communications was a good one. The first three communications resulted in a roughly uniform response from users proportion to the size of the wave, roughly 15%.[5] An interesting question is whether a fifth communication would have induced a similar response as the previous four. Given the much smaller response of the fourth communication (around 5% across each wave) and subsequent email notifications, we speculate that a fifth email would only have further diminishing returns and that the campus decision to change how it interacted with the remaining nonresponsive users after the fourth communication was the right one.

Our results suggest that our organization clearly needed to be proactive throughout the

[5] An exception is the first communication of the first wave, which does not appear to have prompted any password updates. Upon investigating, this apparent lack of response was due to a data collection error in that timeframe.

campaign. The initial email communications were effective for roughly three-quarters of users. Subsequently, very few remaining users changed their passwords during the long idle period, even though these users had already received four email reminders. Not until campus activated the login portal intercept reminder and sent the final warnings did the remaining users start to react again.

RQ2: Login intercepts are an effective tool for user updates. While our organization used login intercepts for well over a month (from October 19, 2021 to November 11, 2021), they staggered their use for the different waves. Moreover, towards the end of the campaign they continued displaying login intercepts in addition to sending a final round of email warnings (note that the IT staff sent the first batch of final email warnings on November 9, 2021). To more clearly assess the impact of login intercepts, we examine their impact on just users in the first two waves, users who had the longest exposure and response to just the login intercepts (before the final email warnings were sent). For this time period preceding November 9, 88% of the remaining non-updated users in Wave 1 responded to the portal intercept and successfully updated their password. For Wave 2, 51% of the remaining users updated their password during the login intercept period (note that Wave 2 users had one fewer week in which to respond compared to Wave 1 users). Email notifications are clearly effective for the majority of our population, but require action out of context. The portal intercept, in contrast, happens exactly as the user is in the process of logging in, and was successful in leading users to update their password.

3.5.2 Multiple Change and Nonresponsive Users

Compared to single change users, multiple change users have more than one password update during the campaign, suggesting these users experienced more friction with the password update process.

Of the 1,291 multiple change users, 72.03% have two password changes, 18.20% have three, and 9.76% have more than three. For simplicity, we focus on the 90.23% of multiple change

Table 3.3. Breakdown of the first, second, and third password changes for multiple change users across the different time periods of the campaign.

	Responsive	Idle	Intervention
% First Change	57.25%	8.33%	34.33%
% Second Change	22.23%	12.79%	60.69%
% Third Change	18.30%	10.21%	64.68%

users that have two or three password changes since they capture the bulk of this population. For each password update these users made, Table 3.3 shows which period during the campaign the user made the update.

RQ3: Multiple change users are less responsive to email communications than single change users. However, multiple change users have similar password update attempts as nonresponsive users. Compared with the single change users, the multiple change users are less responsive to the email communications: 71% of single change users update their password during the responsive period, but only 57% of the multiple change users make their initial password update during the period. Correspondingly, more multiple change users (34%) wait until the intervention period than single change users (23%) before making an update.

The majority of the second and third password updates for the multiple change users happen later in the intervention period (60% and 64%, respectively), rather than closely associated with the first password update in the responsive period. We originally suspected that most users who had multiple password updates had issues involving multiple devices. For instance, they might first change their password on their laptop, but then soon after attempt to login via their phone (e.g., which had the older password cached). At that point the most expedient action would be to reset their password via their phone so that they could continue to login. This scenario would lead to multiple password updates in quick succession, but the long duration between the first and subsequent password updates for the multiple change users indicates this explanation does not hold for most of them.

Two other situations could explain the behavior of multiple change users and their delayed

subsequent password updates. The first is that the users were confused because they have multiple accounts on our campus (e.g., , a faculty account on main campus and another account on the health campus). For example, if a user has two accounts, changed the password on their first account, and then received an intercept for their second account, they may not have paid attention to the account targeted in the notification and instead re-initiated a password change on their first account.

We explored this hypothesis by comparing the anonymized legal name attributed to each user account and counting how many single change, multiple change, and nonresponsive users have user accounts with the same legal name. Overall, there are less than 50 instances where two different user accounts have the same legal name, indicating 1) the legal name attribute is not correct, 2) most users do not have multiple accounts, or 3) their additional accounts are hosted on separate IT infrastructure that we do not have access to (see note about various infrastructures in § 3.2).

The second hypothesis is that users became confused about messaging and initiated another password change: they forgot whether they changed their password, were reminded about the password change out of band, and re-initiated a change. Given the granularity of our data, we unfortunately could not explore this hypothesis further.

We finally compare multiple change and nonresponsive user reactions. Among the nonresponsive users 68.62% had two changes, 29.20% had three changes, and the remaining 7.56% had over three changes, a distribution similar to the multiple change users. If we use the number of changes as a proxy for how many issues a user faced (with a higher number of changes approximating more issues), then the nonresponsive users experience no more issues than multiple change users and simply encounter them in a later time period.

3.5.3 Password Update Mechanisms

We next investigate the different mechanisms that users selected to update their passwords, providing insight into time and energy spent on these updates. Recall that users can change or

Table 3.4. Password change mechanisms across the three different user populations. Note that there are two ways to perform a reset, and thus Admin Reset is a subset of the Reset and Both columns.

Category	% Change	% Reset	% Both	% Admin Reset
Single Change	92.72%	7.28%	—	0.52%
Multiple Change	29.20%	13.90%	56.85%	23.86%
Nonresponsive	2.36%	77.12%	19.66%	27.22%

reset their password via a self-service Web portal, via their campus-managed work computer, or by invoking the help of campus administrative staff. To minimize procedural costs, organizations want to maximize the use of the first two methods and minimize the third.

RQ4: Multiple change and nonresponsive users utilize password resets more than single change users. Table 3.4 summarizes the actions taken by the three responsiveness categories of users in the study. Note that there are two ways for a user to execute a reset, and thus "Admin Reset" is a subset of the "Reset" and "Both" columns. Single change users, as desired, overwhelmingly perform their password change on their own: only 0.52% of these users require administrator assistance with updating their password. In contrast, multiple change and nonresponsive users require significant administrative help. Roughly a quarter of each user category (23.86% of multiple change users, 27.22% of nonresponsive users) initiate a password reset with the assistance of an administrator. To further reduce procedural costs, organizations can focus on reducing circumstances that lead to users making multiple changes. Nonresponsive users represent a difficult case since they generally have minimal interaction with campus already (§ 3.5.5).

As a final observation, in addition to the self-service Web portal and IT help desk service, our campus also allows users to change their password via the operating system of their work machine. More than 22% of the single change users updated their password using their work machine, and all of these updates were successful (the users were already logged in). Since this method is both effective and low cost, organizations should continue to support it and encourage its use.

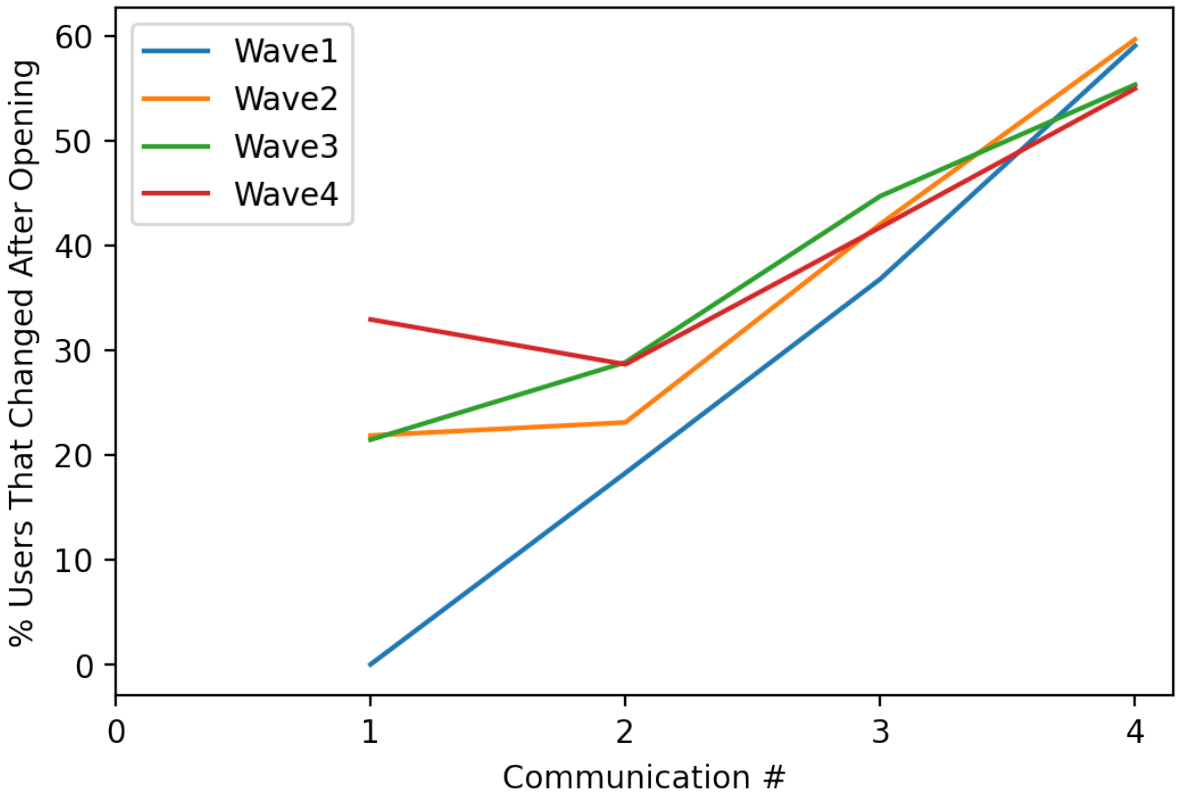


Figure 3.3. The percentage of users across different waves who update their password when opening various communications. Note that the denominator is users who open a given communication.

3.5.4 User Interactions with Communications

Recall from Section 4.2 that the communications sent to users included an email tracking mechanism that records whether and when users receive or open the email message. We use these analytics to examine the relationship between user password update behavior and their interaction with the email communications.

RQ5: Users who open email are more likely to act, and are more likely to act quicker than their counterparts. Users in our organization who opened the email communications are strongly correlated with users who successfully update their password: 83.85% of single change users open at least one of the email communications, 79.16% of multiple change users do the same, but only 38.83% of nonresponsive users open any of the communications.

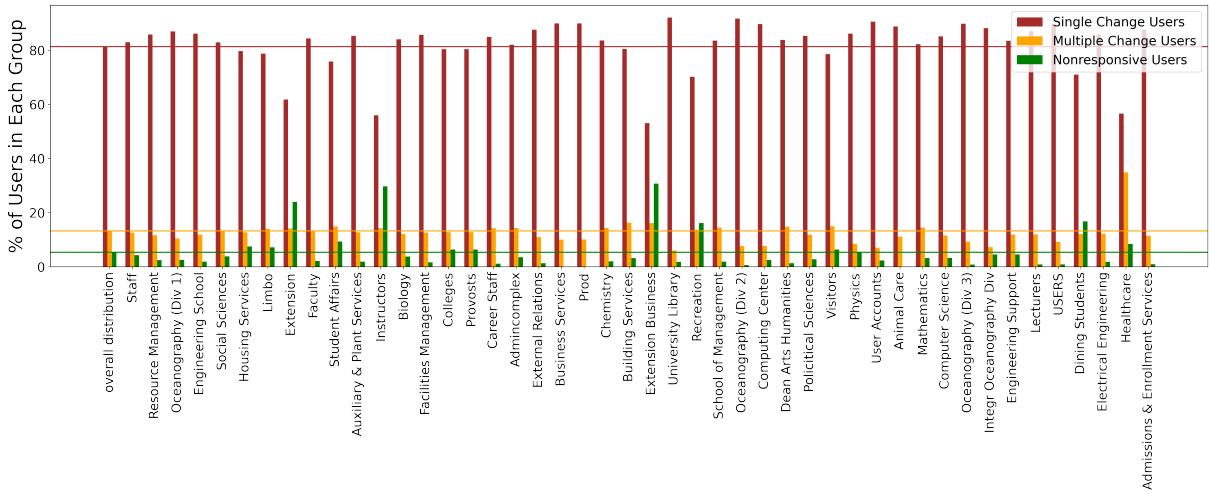


Figure 3.4. Most popular OU distributions across single change, multiple change, and nonresponsive users.

These results are another indication that the nonresponsive users have far less engagement with the university than the rest of the user population.

Moreover, about 71% and 60% of single change and multiple change users, respectively, reacted in the responsive period of the campaign, which is less than 10% lower than the percentage of users who opened an email. This may be indicative that users who opened their email are more likely to react, and users who do not open email need other interventions sooner.

The email campaign consisted of four waves of communications, and users stopped receiving communications once they updated their password. We use the timestamps of user password updates to infer which of the four communications they responded to, and how they interacted with their last communication. For example, if a user received a communication on Day 1, opened it on Day 3, and then changed their password after Day 3, we consider them responsive to opening the email message. However, if the user changed their password on Day 1 or 2, we consider their responsiveness due to receiving the email.

Confirming expectations, users are more likely to update their password after opening the email communication than just receiving it. Figure 3.3 shows the percentage of users who update after opening their communication. The graph has four curves corresponding to the waves of users, and each curve shows the percentage of users in that wave who have updated

their password after *opening* an email communication.[6] By the end of the communications all waves have a 50–60% response rate. In contrast, less than 10% of users across all waves and communications update their password after only *receiving* the email.

Users who open their last email communication also respond much more quickly than users who do not. More than half of the users who open the communication update their password within 24 hours of opening it, whereas more than half of the users who update just based on receiving their last communication take multiple days to update. In short, users who open emails are more responsive in aggregate and also are faster to update than users who only receive the email.

Recall that the communications to the different waves of users are staggered in time: users in Wave 4 receive their first communication three weeks after users in Wave 1. When looking at various metrics, we noticed that users were increasingly more responsive in later waves. For instance, between Wave 1 and Wave 4: users took less time to update their password after receiving their last email communication (median time decreasing from 5 to 3 hours); slightly more users updated within a day (increasing from 60% to 66%) and slightly more users opened at least one email message (increasing from 84% to 87%).

The trends are slight, but we speculate that to the extent there is an effect, it could be due in part to out of band communication about the password change (e.g., mentioning in conversations among co-workers and friends). The longer the campaign lasts, the more opportunity for such an out-of-band mechanism to contribute. However, the benefits over time are modest at best and seem an ancillary benefit to a long campaign.

3.5.5 User Role

Next, we explore how the password update behavior of users correlates with their role on campus. Recall that the account profiles for the users on our campus specify the Organizational

[6]Note that these results are different from Figure 3.2, which includes all user responsiveness regardless of email open status.

Units (OUs) that the user is associated with.

For the 50 largest OUs on campus, Figure 3.4 shows the percentage of single change, multiple change, and nonresponsive users in each OU who update their password. For reference, the "Overall Distribution" bars show the percentage of total users for each user group. Note that users can have multiple OU labels and we count the users in all OUs that they are associated with.

Using these values, we construct three distributions (one for each user responsiveness category) and calculate the Z-Score of each OU, which characterizes how far the value deviates from the mean. For each user responsiveness group (single change, multiple change, and nonresponsive), we identify OUs that are either above or below 1.96 standard deviations from the mean as outliers.[7] Among these outlier OUs, users in Extension, Instructors, and Extension Business OUs are over-represented in nonresponsive users, and under-represented in single change users. These OUs are interesting because they correspond to users who can perform their daily jobs without needing to interact with campus accounts or systems as often as other roles.

Focusing on only single change users, we again examine the 50 largest OUs, but this time across the three time periods (Responsive, Idle, Intervention). Specifically, we investigate if single change users in the intervention period differ from those in other periods. Using Z-Scores to identify outliers, we see that users in the Building Services, Recreation, and Dining Services OUs are over-represented in the intervention time period. Once again, a common thread among many of these OUs is that they correspond to users more on the periphery of the campus: users who may not need to interact with main campus systems regularly.

RQ6: Users in peripheral organizations take longer than their counterparts. Both of these findings reinforce the point that users who take longer or have difficulty updating their password are correlated with roles that have less online interaction with campus systems. In this light, it is not surprising that email notifications are less effective or that such users

[7]Examining data that is above or below 1.96 standard deviations is considered common practice when using Z-Scores.

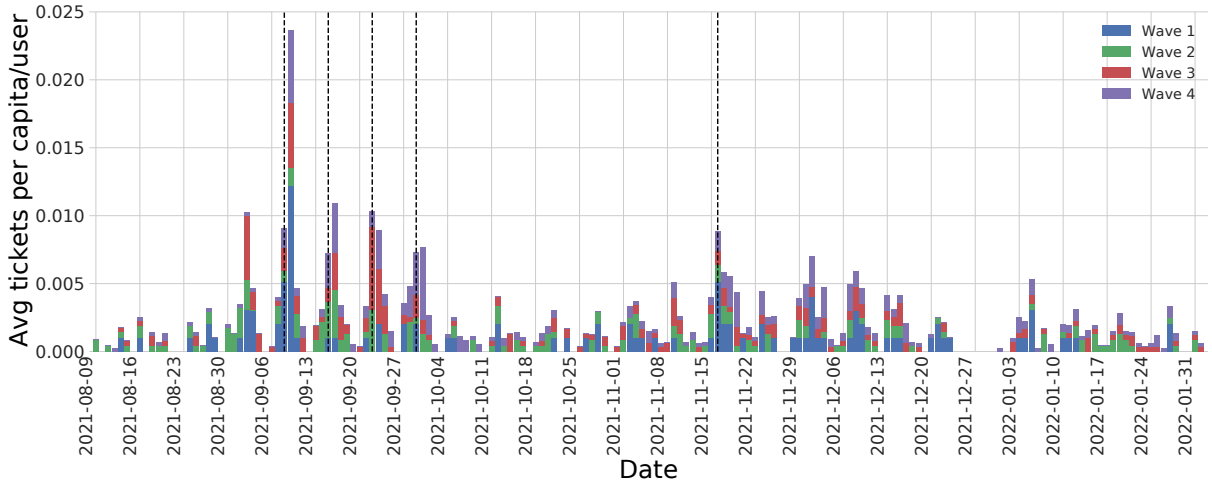


Figure 3.5. Ticket volume per day, normalized (divided) by the total number of users in each wave notification group.

have more difficulty performing the update (e.g., , resulting in a disproportionate number of password scrambles). While this finding may seem obvious, it is still important to understand from an organizational standpoint, as this may change how the organization approaches these departments in future campaigns. In particular, to improve the responsiveness of these kinds of users, organizations may want to target these users differently: e.g., , targeting such users earlier, or forgoing email reminders and using login intercepts from the start, or even using a different notification mechanism such as text messages.

3.6 Help Ticket Workload

Although password update initiatives can improve the security of an organization, these efforts generate extra work for users and the IT staff at the organization, particularly when issues arise during the password update process. To better understand these associated costs, we analyzed changes in the volume of help desk tickets regarding password and account changes during the password update time period. In particular, we examine the following research questions surrounding the costs of enterprise password update campaigns:

- RQ7: Did the password campaign increase the number of help desk tickets?

- RQ8: What were the costs of different enterprise actions in terms of help desk ticket load?
- RQ9: Do users in different departments produce heavier help desk ticket loads?

3.6.1 Help Desk Ticket Data

Our university uses ServiceNow, a centralized ticketing service, to manage all help tickets and requests generated by users. Users can submit help tickets via a standard web portal or by emailing specific help aliases; additionally, users can call specific campus phone numbers to speak with support staff, who then manually create a ticket on behalf of the user during the assistance process. To identify help desk tickets related to the password update process, we created aggregate statistics from the ticket database related to the password update roll-out. Concretely, tickets that met the following criteria are involved in the analysis:

1. The ticket was assigned to the "Service Desk" team (which handles all password and account related issues).
2. The ticket's customer was a user from the 9,744 users in the population we investigate (§ 4.2).
3. The ticket was created between August 9, 2021 and February 1, 2022 (i.e., , between the start of the password reset notifications and approximately one month after the final password reset notification).
4. The ticket satisfied the following keyword requirements: the ticket contained at least one word from each of two lists — ["password", "account"] and ["lock","reset", "change", "update", "sign in"] — and it also did *not* contain any "false positive" words identified based on manual sampling (e.g., , "compromise", "new", etc.).

In total, this search yielded 919 help desk tickets filed by 762 distinct users. For the remainder of this section, we refer to these 762 users as "ticket-filing users" and any password-update related ticket they file simply as a "help ticket". Over 85% of these users (653) filed only

Table 3.5. Percentage of users with password help tickets one year apart.

	Password Update Campaign		Prior Year	
All Waves	7.82%	(762 / 9,744)	2.21%	(215 / 9,744)
Wave 1	7.94%	(78 / 983)	2.24%	(22 / 983)
Wave 2	7.66%	(174 / 2,272)	2.60%	(59 / 2,272)
Wave 3	8.04%	(237 / 2,948)	2.37%	(70 / 2,948)
Wave 4	7.71%	(273 / 3,541)	1.81%	(64 / 3,541)

one help ticket during the update time frame, and 12% of users (93) submitted exactly 2 tickets. Among the remaining 3% of users (16), the maximum number of tickets filed by any single user was 12 tickets, and upon manual inspection it appeared that this user is an IT staff member who created help desk tickets on behalf of users who called the support hotline.

3.6.2 Changes to Help Ticket Volume

Using the volume and timing of tickets, we investigated how much additional work our institution's IT staff encounters as a result of initiating an enterprise-wide password update.

RQ7: Password updates increase the overall ticket volume by a factor of 3–4×. Table 3.5 shows the percentage of ticket-filing users during the password update time period (second column) and the percentage of ticket-filing users from this same exact population during the same time frame one year prior to the password change campaign (third column). We observe a 3–4× increase in the proportion of ticket-filing users during the password update time period (7.5–8%) when compared to the same set of users during the same time period in the prior year (1.8–2.6%). The proportion of users who submit tickets, and the relative increase over the preceding year, remains consistent across all wave groups.

RQ8: Actions lead to different ticket volumes. As described in Section 3.2, over the course of the password update roll-out, campus IT staff employed multiple types of actions to encourage users to update their password.

Figure 3.5 displays the total volume of tickets that users from different wave groups submitted during each day, where the daily volume is normalized (divided) by the total number

Table 3.6. Proportion of single change, multiple change, and nonresponsive users who file exactly one ticket and multiple tickets.

User Responsiveness	% w/ 1 Ticket	% w/ 2+ Tickets
Single Change Users	5.6%	0.6%
Multiple Change Users	18.6%	3.8%
Nonresponsive Users	14.6%	2.3%

of users in each respective wave group. Figure 3.6 shows the cumulative fraction of tickets submitted by all users over time. As marked by vertical dashed lines in both figures, two types of actions led to noticeable increases in the volume of tickets.

First, we see large spikes in the proportion of users who submit tickets after each of the first four dashed lines; these dates correspond to when the IT staff sent their fourth ("last") communication email to users in each of the waves. These notifications stated that users must immediately change their passwords to "avoid account access complications". As we observed in Figure 3.2, this set of email messages galvanized a significant fraction of users into updating their password, which likely accounts for the increase in help ticket volumes immediately following these email notifications.

The last dashed line in Figure 3.5 corresponds to the date (Nov 16) when the IT staff began to automatically scramble the passwords of any user who had not yet updated their password. Unsurprisingly, this intervention led to a significant increase in the proportion of users who filed help desk tickets. Among the 528 nonresponsive users, 77 users filed password help tickets (14.6%); in contrast, only 7.6% (700) of the 9,216 single change and multiple change users without a password scramble submitted a help desk ticket. Furthermore, of the 77 nonresponsive users, only 8 users submitted a ticket prior to having their password reset by the IT team, which suggests that the vast majority of these users filed tickets as a result of the IT team's actions.

In contrast to these two actions, from October 19, 2021 to November 15, 2021, the IT staff configured the university's SSO login portal to display a browser interstitial message after every

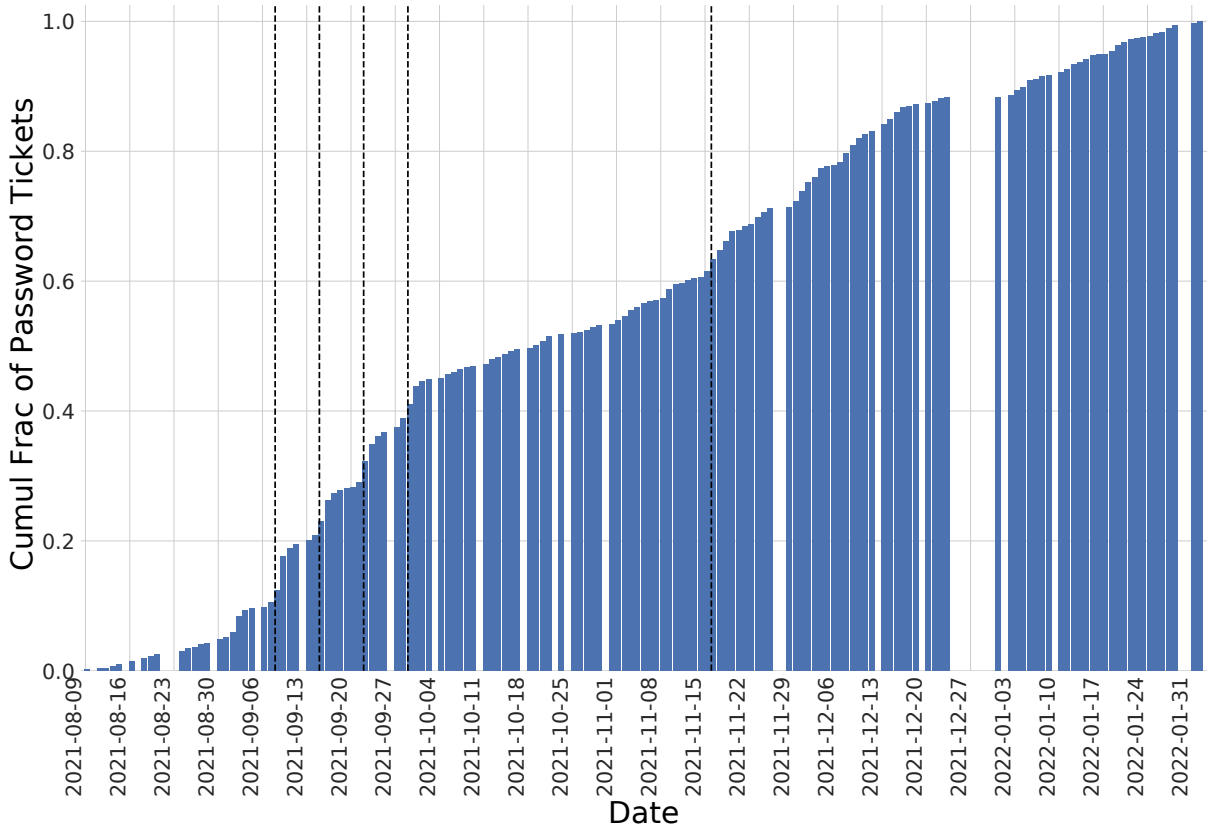


Figure 3.6. Cumulative fraction of password-update help tickets over time.

successful login to users who had not updated their password; from November 9 to November 16, the IT staff also sent out an additional email notification and the SSO portal continued to produce browser interstitial pop-ups. As seen in Figures 3.5 and 3.6, this institutional action generated noticeably fewer tickets than both the earlier email message notifications and the password scrambling: only 8% of tickets were submitted between October 19 and November 9 (the period where only active action was login intercept).

We hypothesize that the SSO intercepts created a lower ticket volume because they presented a more concise message and direct, in-situ path to updating a user's password. Namely, whereas the email notifications contained a detailed description of the update roll-out and list of instructions for users to complete, the SSO intercept message displayed a short message with a link for the user to immediately update their password (as shown in Figure 3.1). Furthermore, users are more likely to successfully update their password independently because they only saw the SSO intercept message after successfully authenticating with their old password (which they then can use to change their password).

3.6.3 Help Ticket User Demographics

Update Responsiveness and Help Ticket Volume. We next explore whether the single change users' apparent efficiency at successfully resetting their password correlated with needing less help from IT staff members. As seen in Table 3.6, single change users in fact submit 3–6× fewer tickets than users in other categories: only 5.6% (445 / 7,925) single change users submit one help ticket, compared to 18.6% (240 / 1,291) multiple change users and 14.6% (77 / 528) nonresponsive users.

RQ9: Help Tickets volume are non-uniform by Organizational Unit. We also investigated whether a user's specific department (a proxy for job role and technical familiarity) correlated with the likelihood of them requesting help. As discussed earlier in Section 4.2, our institution uses Active Directory to manage information about users and their accounts, and each user has an associated set of Organizational Unit (OU) affiliations (e.g., , Computer Science

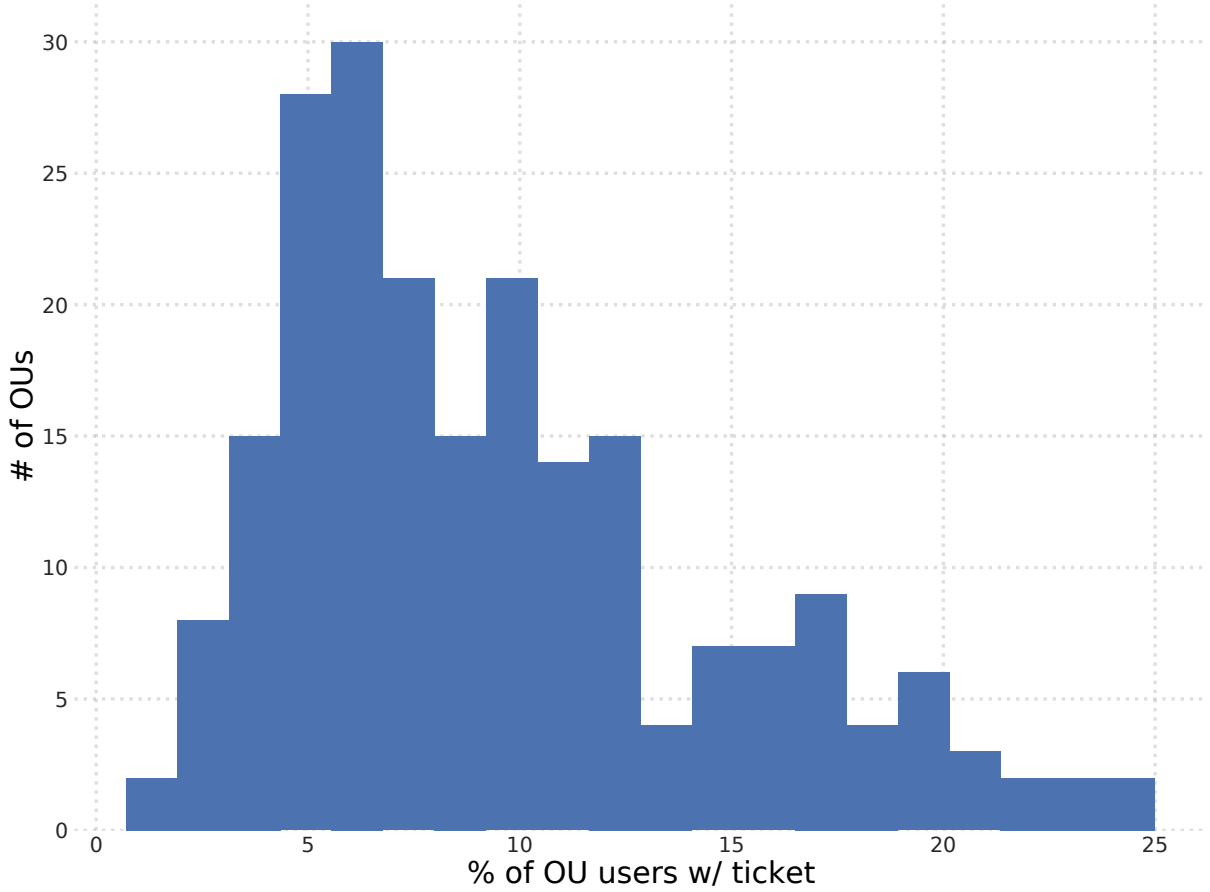


Figure 3.7. Percentage of users in each OU (proxy for department/role) who filed a password-related help desk ticket during the password update period.

department, Staff Tech Support, etc.). Users can and often do have multiple OU affiliations as a result of being affiliated with multiple departments or groups on campus. For our analysis, a member of campus's IT staff computed the set of OUs that each user in our dataset belonged to.

The 762 users who filed a password-reset related ticket span a total of 314 distinct OUs. Figure 3.7 shows the proportion of each OU's users that filed a password-related help ticket: for each OU, this proportion equals the number of ticket-filing users affiliated with the OU divided by the total number of users in our data set who had an affiliation with the OU (i.e., , if any of a user's OU affiliations match, then we count them as part of the OU). As seen by the right-skewed distribution, users in several OUs submit help tickets at over twice the rate as the median OU (8.19%).

Table 3.7. Top 12 OUs with the highest proportion of active users (undergoing a password reset) who filed a password-reset related ticket.

OU	% of Notified OU Users
Teaching & Learning Commons	25.0%
Sponsored	24.4%
Academic Affairs	23.5%
Counseling	23.1%
IT Services	22.5%
Emeritus	21.7%
Provost (Div 1)	21.2%
Emeriti	20.6%
Provost (Div 2)	20.6%
Employment Community Outreach	20.0%
Copy Center	20.0%
Provost (Div 3)	20.0%

Table 3.7 shows the OUs with the highest proportion of users who submitted a help ticket (again, with some modifications to the OU names to blind our organization). Among these OUs, we note that the tickets submitted by users in IT Services correspond to members of the IT staff submitting tickets on behalf of users who contacted help / support out-of-band (e.g., , via a phone call to the help desk). A total of 301 OUs (not shown in Figure 3.7) had 0 affiliated users who submitted a password-reset related ticket; of these, only 31 OUs (10.3%) have more than 10 users and span a variety of different parts of campus with no clear thematic grouping (e.g., , they cover a variety of different academic departments and groups, such as postdocs, the campus registrar's OU, and OUs for technical institutes co-located and affiliated with campus).

Similarly, the OUs with the highest proportion of users who submit tickets lack easily discernible patterns: on one hand, this set of OUs contains both users with looser present-day affiliations to campus (e.g., , emeritus) as well as groups involved in day-to-day campus interactions (e.g., , Academic Affairs, the Copy Center, and staff in various Provosts' offices). Based on this heterogeneous mix across both high and low ticket-filing OUs, it appears that other underlying factors (beyond a user's department affiliation or working ground) may be more predictive in determining whether users will need help during the update process (e.g., , technical

aptitude and familiarity).

3.7 Related Work

Password research is a rich subfield with influential work dating back to the 1970s [66]. However, much of this work can be categorized into three primary sub-areas:

Password Guessability and Cracking: Many studies have explored ways to quantify password strength by developing and applying online and offline attacks to guess (crack) users passwords [118, 10, 114, 63, 115, 20, 29, 47, 107, 64, 57]. Guiding users with UIs to create stronger passwords less susceptible to cracking has also been examined [105], as has mnemonic cracking [52]. As a result of some of this work, researchers have recommended using length and/or mnemonic devices to reduce password guessability [74, 52].

User mental models and Password policies: Another important area of work focuses on understanding user's mental models about password policies and their impact on security [43, 41, 50, 89, 90, 87, 104, 106]. For example, studies have found that users report rarely changing passwords unless asked [45]. Some studies have further shown that users are generally proactive in changing their passwords when a deadline is provided, while others find that they postpone as long as possible [72, 91, 5]. Much of the work in this sub-area has been used to motivate password expiration policy updates [88, 30, 120, 15, 85, 40]. Notably, Florencio et al. summarizes and synthesizes much of the relevant work in a SoK-style paper aimed towards conveying best practices of password policies for system administrators [30].

Empirical analyses: Finally, a number of papers have conducted empirical measurement studies to understand various aspects of the password lifecycle [5]. For example, studies have documented that users do not change their passwords proactively after a notification of a password breach, and when they do, the updated password is similar to the old one [7]. Moreover, others have examined password update metrics at scale, and found strength meters were effective in prompting users to produce higher entropy passwords [105]. There have also been a number

of studies examining and quantifying password guessability and password reset policies of university populations at scale [63, 72]. Most recently, a new system has been proposed to study logging attempts at a university in real-time and securely [9].

Comparison to our Work: Despite this well-established body of research, relatively little work has focused on the operational needs of enterprise administrators who must, despite individual behavioral priorities, ensure that disparate users comply with password policies. Moreover, prior efforts in this space have largely focused on small-scale or single factor studies. Our research addresses this lack via a large-scale empirical analysis of the enterprise password update process. Critically, we validate, at scale, key aspects of the interaction between user behavior and the password update process. This includes a range of basic factors such as the prevailing use of both self-administered update mechanisms [45], online password reset [72], and the prevalence of ticket submission [91, 72] among others. Moreover, our dataset captures all of these aspects together, allowing a fuller understanding of how these issues play holistically during a mandatory update campaign. Most importantly, we have been able to analyze the effectiveness of the most commonly deployed operational treatments for driving mass updates (repeated e-mail notifications and interstitial Web intercepts) as well as the overhead incurred on help desk resources when using these treatments. Together, these findings provide an empirical basis for establishing best practices (including more aggressive, and hence more timely, efforts employed in the campaign we studied).

Finally, while prior work has discussed the importance of a central management system (like Active Directory) in easing the user experience of policy changes [30, 72], our work shows that a range of user difficulties persist, some likely reflecting individual priorities and capabilities but others reflecting anticipatable differences in organizational use of IT services.

Other Security Communications Outside the password literature, two other lines of research offer related contributions. The first is around security communication—how the content and modality of security information plays a role in how it is acted upon [17]. This includes studies on the efficacy of both user interface elements such as phishing toolbars [18]

and browser TLS security indicators [2, 25, 26, 81, 27] as well as email-based vulnerability notifications [55, 95, 56, 14, 39, 60]. That said, we are unaware of experimental research concerning the structure or timing of password update communications.

The other key line of research is around the user overhead and adoption issues around new security technologies, notably two-factor authentication (2FA) [32, 16, 82, 1]. Notably, both Abbot et al. and Reynolds et al. characterize the support costs for incorporating two-factor authentication into their respective universities by analyzing related help desk tickets (similar to the kinds of analysis we explore in our work on password changes).

3.8 Discussion

Organizational password updates are strenuous because they put the needs of the organization (potentially increased security) at odds with the user population (time, energy, and technical knowledge). For many users, security is a secondary or tertiary concern, and as the complexity of an organization increases, so do potential pitfalls for a user. Moreover, large organizations are routinely supported by modestly-staffed IT departments whose success is predicated on extensive use of automation. Large-scale mandatory password changes have the potential, when combined with confused or unmotivated users, to upend this calculus and overwhelm IT staff with large numbers of support requests. However, in spite of these issues and the common nature of this situation, we lack an empirical basis for establishing the best methods for prompting user updates or for anticipating potential support costs.

In this work, we empirically deconstructed a password update campaign at our academic institution aimed at employees and staff. From this analysis the key observations are that:

1. Single change users, who comply to email change requests in a timely manner, only comprise 80% of our user population. This is a vast majority, but the remaining 20% is a significant population that requires more focused outreach.

2. Email communications are effective, but have diminishing returns. We observed that the

first three reminder emails prompted 15% of users to change their passwords, but the fourth only elicited a response from an additional 5% of users. A significant aspect of this issue appears to revolve around users who are less likely to engage with email. Indeed, opening email remains a strong predictor of whether a user will update their password for all reminders. Alternative mechanisms, or advance scheduling, may be appropriate for users whose roles require less email engagement.

3. A proactive stance is needed. This campaign had an idle period of about a month between the initial email communication and the login intercept and final/scramble notifications. We observe very little user action during this idle period, suggesting that reminders create a short attention window for this task and users are not "waiting" to change their password later.

4. Help Desk costs are non-uniform. A small fraction of users induce most of the IT support burden. Indeed, almost 40% of help desk tickets are driven by nonresponsive users (whose passwords were scrambled) in spite of the fact that they comprise less than 3% of the user population. Conversely, the IT support costs during the email campaign are quite modest compared to the size of the population. Of particular note, the login intercept action does not appear to trigger significant numbers of help desk tickets, in spite of the selection bias in our data that prevents the most responsive users from being exposed to login intercepts.

Based on these overall results, we believe that login intercepts show promise as the most expedient and least costly mechanism for prompting password updates. We observe that these intercepts are effective at prompting password updates for users who are unresponsive to the email campaign, the incur little cost in terms of IT support and, finally, they locate a password change request in the midst of an authentication action—a context in which the user is already prepared to enter their password—and in so doing removes the cognitive load of reading and understanding documentation and deciding how and when to schedule a future password change. While the intercept capability must be built and implemented, the cost afterwards appears to be

quite low, with large returns in user efficacy and IT staff efficiency.

Finally, in this study, our organization valued cost in tandem with efficacy, and thus implemented a longer password update campaign that was cognizant of the unknown burden that might be placed on IT staff. However, every organization has different constraints and incentives that define their operational logistics. For some organizations, in some situations, expediency trumps all—however, for others cost may be a larger factor. We believe that this study provides an initial step in uncovering the hidden factors of large scale organizational updates, and that other organizations can use these results, even if their constraints differ, to design and implement password change campaigns suited to their needs.

Chapter 3, in full, is currently being prepared for submission for publication of material. Ariana Mirian, Grant Ho, Stefan Savage, Geoffrey M. Voelker. The dissertation author was the primary investigator and author of this material.

Chapter 4

Hack for Hire: Exploring the Emerging Market for Account Hijacking

The last perspective that is useful to understand empirically is that of the attacker. In this chapter, I characterize the commodity market for "Hack for Hire" services which compromise email accounts for payment. The results of this study provide insight into attacker behavior that can directly improve defenses to protect users against this specific category of attacks.

4.1 Overview

It has long been understood that email accounts are the cornerstone upon which much of online identity is built. They implicitly provide a root of trust when registering for new services and serve as the backstop when the passwords for those services must be reset. As such, the theft of email credentials can have an outsized impact—exposing their owners to fraud across a panoply of online accounts.

Unsurprisingly, attackers have developed (and sell) a broad range of techniques for compromising email credentials, including exploiting password reuse, access token theft, password reset fraud and phishing among others. While most of these attacks have a low success rate, when applied automatically and at scale, they can be quite effective in harvesting thousands if not millions of accounts [99]. In turn, email providers now deploy a broad range of defenses to address such threats—including challenge questions to protect password reset actions, mail

scanning to filter out clear phishing lures, and two-factor authentication mechanisms to protect accounts against password theft [35, 34, 33]. Indeed, while few would claim that email account theft is a solved problem, modern defenses have dramatically increased the costs incurred by attackers and thus reduce the scale of such attacks.

However, while these defenses have been particularly valuable against large-scale attacks, targeted attacks remain a more potent problem. Whereas attackers operating at scale expect to extract small amounts of value from each of a large number of accounts, targeted attackers expect to extract large amounts of value from a small number of accounts. This shift in economics in turn drives an entirely different set of operational dynamics. Since targeted attackers focus on specific email accounts, they can curate their attacks accordingly to be uniquely effective against those individuals. Moreover, since such attackers are unconcerned with scale, they can afford to be far nimbler in adapting to and evading the defenses used by a particular target. Indeed, targeted email attacks—including via spear-phishing and malware—have been implicated in a wide variety of high-profile data breaches against government, industry, NGOs and universities alike [46, 36, 109, 42].

While such targeted attacks are typically regarded as the domain of sophisticated adversaries with significant resources (e.g., state actors, or well-organized criminal groups with specific domain knowledge), it is unclear whether that still remains the case. There is a long history of new attack components being developed as vertically integrated capabilities within individual groups and then evolving into commoditized retail service offerings over time (e.g., malware authoring and distribution, bulk account registration, AV testing, etc. [99]). This transition to commoditization is commonly driven by both a broad demand for a given capability and the ability for specialists to reduce the costs in offering it at scale.

In this chapter, we present the first characterization of the *retail* email account hacking market. We identified dozens of underground "hack for hire" services offered online (with prices ranging from $100 to $500 per account) that purport to provide targeted attacks to all buyers on a retail basis. Using unique online buyer personas, we engaged directly with 27 such account

hacking service providers and tasked them with compromising victim accounts of our choosing. These victims in turn were "honey pot" Gmail accounts, operated in coordination with Google, and allowed us to record key interactions with the victim as well as with other fabricated aspects of their online persona that we created (e.g., business web servers, email addresses of friends or partner). Along with longitudinal pricing data, our study provides a broad picture of how such services operate—both in their interactions with buyers and the mechanisms they use (and do not use) to compromise victims.

We confirm that such hack for hire services predominantly rely on social engineering via targeted phishing email messages, though one service attempted to deploy a remote access trojan. The attackers customized their phishing lures to incorporate details of our fabricated business entities and associates, which they acquired either by scraping our victim persona's website or by requesting the details during negotiations with our buyer persona. We also found evidence of re-usable email templates that spoofed sources of authority (Google, government agencies, banks) to create a sense of urgency and to engage victims. To bypass two-factor authentication, the most sophisticated attackers redirected our victim personas to a spoofed Google login page that harvested both passwords as well as SMS codes, checking the validity of both in real time. However, we found that two-factor authentication still proved an obstacle: attackers doubled their price upon learning an account had 2FA enabled. Increasing protections also appear to present a deterrent, with prices for Gmail accounts at one service steadily increasing from $125 in 2017 to $400 today.

As a whole, however, we find that the commercialized account hijacking ecosystem is far from mature. Just five of the services we contacted delivered on their promise to attack our victim personas. The others declined, saying they could not cover Gmail, or were outright scams. We frequently encountered poor customer service, slow responses, and inaccurate advertisements for pricing. Further, the current techniques for bypassing 2FA can be mitigated with the adoption of U2F security keys. We surmise from our findings, including evidence about the volume of real targets, that the commercial account hijacking market remains quite small and niche. With

prices commonly in excess of $300, it does not yet threaten to make targeted attacks a mass market threat.

4.2 Methodology

In this section we describe our methodology for creating realistic, but synthetic, victims to use as targets, the infrastructure we used to monitor attacker activity, and the services we engaged with to hack into our victim email accounts. We also discuss the associated legal and ethical issues and how we addressed them in our work.

4.2.1 Victims

We created a unique victim persona to serve as the target of each negotiation with a hack for hire service. We never re-used victim personas among services, allowing us to attribute any attacks deployed against the persona back to the service we hired. In creating victim personas, we spent considerable effort to achieve three goals:

1. *Victim verisimilitude*. We created synthetic victims that appeared sufficiently real that the hacking services we hired would treat them no differently from other accounts that they are typically hired to hack into.

2. *Account non-attributability*. We took explicit steps to prevent attackers from learning our identities while we engaged with them as buyers, when they interacted with us as victims, and even if they successfully gained access to a victim email account.

3. *Range of attacker options*. We did not know a priori what methods the hacking services would use to gain access to victim email accounts. Since there are many possibilities, including brute-force password attacks, phishing attacks on the victim, and malware-based attacks on the victim's computers, we created a sufficiently rich online presence to give attackers the opportunity to employ a variety of different approaches.

The remainder of this section details the steps we took to achieve these goals when creating fictitious victims, the monitoring infrastructure we used to capture interactions with our fake personas, and the selection of "hack for hire" services we engaged with.

Victim Identities.

Each victim profile consisted of an email address, a strong randomly-generated password, and a name. While each of our victims 'lived' in the United States, in most cases we chose popular first and last names for them in the native language of the hacking service, such as "Natasha Belkin" when hiring a Russian-language service.[1] The email address for the victim was always a Gmail address related to the victim name to further reinforce that the email account was related to the victim (e.g., `natasha.r.belkin@gmail.com`). We loaded each email account with a subset of messages from the Enron email corpus to give the impression that the email accounts were in use [23]. We changed names and domains in the Enron messages to match those of our victim and the victim's web site domain (described below), and also changed the dates of the email messages to be in this year.

Each victim Gmail account used SMS-based 2-Factor Authentication (2FA) linked to a unique phone number.[2] As Gmail encourages users to enable some form of 2FA, and SMS-based 2FA is the most utilized form, configuring the accounts accordingly enabled us to explore whether SMS-based 2FA was an obstacle for retail attackers who advertise on underground markets [3] (in short, yes, as discussed in detail in Section 4.4.4).

Online Presence.

For each victim, we created a unique web site to enhance the fidelity of their online identity. These sites also provided an opportunity for attackers to attempt to compromise the web server as a component of targeting the associated victim (server attacks did not take place). Each victim's web site represented either a fictitious small business, a non-governmental

[1] These example profile details are from a profile that we created, but in the end did not need to use in the study.

[2] These phone numbers, acquired via prepaid SIM cards for AT&T's cellular service, were also non-attributable and included numbers in a range of California area codes.

organization (NGO), or a blog. The sites included content appropriate for its purported function, but also explicitly provided contact information (name and email address) of the victim and their associates (described shortly). We hosted each site on its own server (hosted via third-party service providers unaffiliated with our group) named via a unique domain name. We purchased these domain names at auction to ensure that each had an established registration history (at least one year old) and the registration was privacy-protected to prevent post-sale attribution to us (privacy protection is a common practice; one recent study showed that 20% of .com domains are registered in this fashion [58]). The sites were configured to allow third-party crawling, and we validated that their content had been incorporated into popular search engine indexes before we contracted for any hacking services. Finally, we also established a passive Facebook profile for each victim in roughly the style of Cristofaro et al. [19]. These profiles were marked 'private' except for the "About Me" section, which contained a link to the victim's web site.[3]

Associate Identity.

In addition to the victim identity, we also created a unique identity of an associate to the victim such as a spouse or co-worker. The goal with creating an associate was to determine whether the hacking services would impersonate the associate when attacking the victim (and some did, as detailed in Section 4.4.2) or whether they would use the associate email account as a stepping stone for compromising the victim email account (they did not). Similar to victim names, we chose common first and last names in the native language of the hacking service. Each victim's web site also listed the name and a Gmail address of the associate so that attackers could readily discover the associate's identity and email address if they tried (interestingly, most did not try as discussed in Section 4.4.2). Finally, if the victim owned their company, we also included a company email address on the site (only one attack used the company email address in a phishing lure).

[3]None of the service providers we contracted with appeared to take advantage of the Facebook profile, either by visiting the victim's web site via this link or communicating with the victim via their Facebook page.

Buyer Identity.

We interacted anonymously with each hack for hire service using a unique buyer persona. When hiring the same service more than once for different victims, we used distinct buyer personas so that each interaction started from scratch and was completely independent. In this role, we solely interacted with the hacking services via email (exclusively using Gmail), translating our messages into the native languages of the service when necessary.

Many hacking services requested additional information about the victim from our buyers, such as names of associates, to be able to complete the contract. Since we made this information available on the victim web sites, we resisted any additional requests for information to see if the services would make the effort to discover this information themselves, or if services would be unable to complete the contract without it (Section 4.4.1).

4.2.2 Monitoring Infrastructure

Email Monitoring.

For each Gmail account, we monitored activity on the account by using a modified version of a custom Apps Script shared by Onaolapo et al. [71]. This script logged any activity that occurs within the account, such as sending or deleting email messages, changing account settings, and so on (Section 4.4.6 details what attackers did after gaining access to accounts). The script then uploaded all logged activity to a service running in Google's public cloud service (Google App Engine) as another level-of-indirection to hide our infrastructure from potential exposure to attackers. Since the script runs from within the Gmail account, it is possible in principle for an attacker to discover the script and learn where the script is reporting activity to, though only after a successful attack. We found no evidence that our scripts were detected.

Login Monitoring.

In addition to monitoring activity from within the accounts, the accounts were also monitored for login activity by Google's system-wide logging mechanisms. Google's monitoring, shared with us, reported on login attempts and whether they were successful, when attackers

Table 4.1. We contacted 27 hacking services attempting to hire them to hack 34 different victim Gmail accounts. We communicated with the services in the language in which they advertised, translating when necessary. The prices were advertised in their native currency, and we normalized them to USD for ease of comparison. (Yes†: for first-time customers.)

Service	Price	Lang	Prepay	Payment	Respond	Attack
A.1	$229	RU	50%	Qiwi	Yes	Yes
A.2	$229	RU	50%	Qiwi	Yes	Yes
A.3	$458	RU	50%	Qiwi	Yes	Yes
B.1	$380	RU	No	Webmoney, Yandex	Yes	Yes
B.2	$380	RU	No	Webmoney, Yandex	Yes	Yes
C.1	$91	RU	No	Bitcoin	Yes	Yes
C.2	$91	RU	No	–	Yes	Yes
D.1	$76	RU	No	–	Yes	Yes
E.1	$122	RU	No	–	Yes	Yes
E.2	$122	RU	No	–	Yes	No
D.2	$76	RU	No	–	Yes	No
F	$91	RU	No	–	Yes	No
G	$91	RU	No	–	Yes	No
H.1	$152	RU	No	Webmoney	Yes	No
H.2	$152	RU	No	Webmoney	Yes	No
J	–	EN	–	–	Yes	No
K	$200–300	EN	Yes	Bitcoin	Yes	No
L	$152	RU	No	–	Yes	No
M	$84	RU	No	–	Yes	No
N	$69	RU	No	Webmoney, Yandex	Yes	No
O	–	RU	No	Webmoney, Yandex	Yes	No
P	$305	RU	No	–	Yes	No
Q	$46	RU	Yes†	–	Yes	No
R	$100	EN	No	–	No	No
S	$400–500	EN	50%	–	No	No
T	$95 or 113	EN	No	Bitcoin, Credit Card	No	No
U	$98	RU	No	Webmoney	No	No
V	$152	RU	No	Webmoney, Yandex, Qiwi	No	No
W	$152	RU	No	–	No	No
X	$152	RU	No	Webmoney, Yandex	No	No
Y	$23 – $46	RU	No	–	No	No
Z	$61	RU	No	–	No	No
AA	$46	RU	No	–	Yes	No
BB	–	CN	–	–	No	No

were presented with a 2FA challenge, and whether they were able to successfully respond to the challenge (Section 4.4.4). These monitoring logs also include the infrastructure and devices used to make login attempts, which Google used to identify other Gmail accounts attacked by these services (Section 4.5.1).

Phone Monitoring.

As described earlier, each victim account was associated with a unique cell number (used only for this purpose) which was configured in Gmail to be the contact number for SMS-based 2FA. To capture attacks against these phone numbers or notifications from Google (e.g., for 2FA challenges or notification of account resets) we logged each SMS message or phone call received.

Web Site Monitoring.

To monitor activity on the web sites associated with the victims, we recorded HTTP access logs (which included timestamp, client IP, user agent, referrer information, and path requested). For completeness, we also recorded full packet traces of all incoming traffic to the target server machines in case there was evidence of attacker activity outside of HTTP (e.g., attempts to compromise the site via SSH). Overall, we found no evidence of attackers targeting our web sites.

4.2.3 Hacking Services

Recruitment.

We identified hacking services through several mechanisms: browsing popular underground forums, searching for hacking services using Google search, and contacting the abuse teams of several large Internet companies. We looked for services that specifically advertised the ability to hack into Gmail accounts. While we preferred services that explicitly promised the passwords of targeted accounts, we also engaged with services that could instead provide an archive of the victim's account. Figure 4.1 shows an example service advertisement (one we did not purchase from).

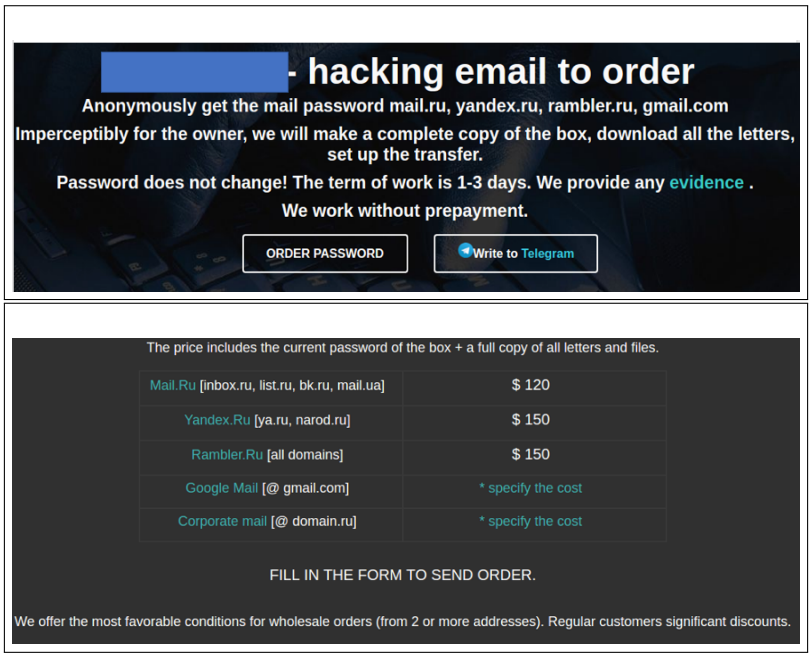


Figure 4.1. An online advertisement for Gmail hacking services. We remove any identifiable information and translate the page from Russian to English.

When hiring these services, we followed their instructions for how to contact them. Typically, interactions with the services consisted of a negotiation period, focused on a discussion of what they would provide, their price, and a method of payment. The majority of the services were non-English speaking. In these cases, we used a native speaker as a translator when needed. We always asked whether they could obtain the password of the account in question as the objective, and always offered to pay in Bitcoin. If the sellers did not want to use Bitcoin, we used online conversion services to convert into their desired currency (the minority of cases). Interestingly, only a handful of services advertised Bitcoin as a possible payment vector, though many services were generally receptive towards using Bitcoin when we mentioned it.

Table 4.1 summarizes the characteristics of all services that we contacted, which we anonymize so that our work does not advertise merchants or serve as a performance benchmark. In total, we reached out to 27 different services and attempted to hire them to hack 34 unique victim Gmail accounts. When a service successfully hacked into an account, we later hired them again (via another unique buyer persona) with a different victim to see if their methods changed

over time (we denote different purchases from the same service by appending a number after the letter used to name the service).

Service reliability.

Of the twenty-seven services engaged, ten refused to respond to our inquiries. Another twelve responded to our initial request, but the interactions did not lead to any attempt on the victim account. Of these twelve, nine refused up front to take the contract for various reasons, such as claiming that they no longer hacked Gmail accounts contrary to their contemporary advertisements. The remaining three appear to be pure scams (i.e., they were happy to take payment, but did not perform any service in return). One service provided a web-based interface for entering the target email address, which triggered an obviously fake progress bar followed by a request for payment.[4] Another service advertised payment on delivery, but after our initial inquiry, explained that they required full prepayment for first-time customers. After payment, they responded saying that they had attempted to get into the account but could not bypass the 2FA SMS code without further payment. They suggested that they could break into the mobile carrier, intercept the SMS code, and thus break into the Gmail account. We paid them, and, after following up a few times, heard nothing further from them. During this entire exchange, we did not see a single login attempt on the victim's Gmail account from the hacking service. The third site similarly required pre-payment and performed no actions that we could discern.

Finally, five of the services made clear attempts (some successful, some unsuccessful) to hack into eleven victim accounts. We focus on these services going forwards.

Pricing.

The cost for hiring the hacking services often varied significantly between the advertised price and the final amount we paid. Table 4.2 shows a breakdown of the price differences during engagement with the hacking services we successfully hired. The table shows the service, the purported price for that service from their online advertisement, the initially agreed upon price

[4]We did not pay them since we would learn nothing more by paying.

Table 4.2. The changes in negotiated prices when advertised, when initially hired, and when finally successful at hacking into victim Gmail accounts. All prices were originally in rubles, but are converted to USD for easier comparison.

Service	Advertised	Discussed	Final
A.1	$230	$230	$307
A.2	$230	$230 - $307	Failed
A.3	$460	$460	$460
B.1	$383	$383	Failed
B.2	$383	$383	$383
C.1	$92	$102	$100
C.2	$92	–	Failed
D.1	$77	$184	Failed
D.2	$77	$184	Failed
E.1	$123	$383 - $690	$383
E.2	$123	$690	Failed

for their services, and then any price increase that may have incurred during the attack period. When services failed to hack into the account, they did not request payment. Several factors influenced the changes in prices, in particular the use of 2FA on the accounts (Section 4.7).

As a rule, we always paid the services, even when they requested additional money, and even when we strongly suspected that they might not be able to deliver when they asked for payment up front.[5] Our goal was to ultimately discover what each service would actually do when paid.

4.3 Legal and Ethical Issues

Any methodology involving direct engagement with criminal entities is potentially fraught with sensitivities, both legal and ethical. We discuss both here and how we addressed them.

There are two legal issues at hand in this study: unauthorized access and the terms of service for account creation and use. Obtaining unauthorized access to third-party email accounts is unlawful activity in most countries and in the United States is covered under 18 USC 1030,

[5]The one exception to this rule is the aforementioned service whose automated web site immediately told us they had hacked the site when all evidence was to the contrary.

Table 4.3. Overview of attack scenarios per service. Lure emails include impersonating an associate (A), bank (B), Google (G), government (V), or a stranger (S). In the event a service indicated they could not succeed without additional information, we indicate what details they requested. In one case (marked *), this was only for the second attempt.

Service	Method	Lure	Inbox or Spam	Promised goods	Requested	Success
A.1	Phishing	A, G, S	Inbox	Archive	–	Y
A.2	Phishing	A, G, S	Inbox	Archive	Victim/associate name, phone number	N
A.3	Phishing	A, G, S	Inbox	Archive	Victim/associate name, phone number	Y
B.1	Phishing	B	Inbox, Spam	Password	–	N
B.2	Phishing	A, G, V	Inbox, Spam	Password	Victim name, associate name/email, phone number*	Y
C.1	Phishing	G	Inbox	Password	–	Y
C.2	Phishing	G	Inbox, Spam	Password	–	N
D.1	Malware	V	Spam	Password	Victim name and occupation	N
E.1	Phishing	G, V	Inbox, Spam	Password	–	Y

the Computer Fraud and Abuse Act (CFAA). Contracting for such services, as we did in this study, could constitute aiding and abetting or conspiracy if the access was, in fact, unauthorized. However, in this study, the email accounts in question are directly under our control (i.e., we registered them), and since we are acting in coordination with the account provider (Google), our involvement in any accesses was explicitly authorized. The other potential legal issue is that this research could violate Google's terms of service in a number of ways (e.g., creating fake Gmail accounts). We addressed this issue by performing our study with Google's explicit permission (including a written agreement). Both our institution's general counsel and Google's legal staff were appraised of the study, its goals, and the methods employed before the research began.

This study is not considered human subjects research by our Institutional Review Board because, among other factors, it focuses on measuring organizational behaviors and not those of individuals. Nevertheless, outside traditional human subjects protections, there are other ethical considerations that informed our approach. First, by strictly using fictitious victims, associates and web sites, we minimized the risk to any real person resulting from the account hacking

contracted for in this study. Second, to avoid indirect harms resulting from implicitly advertising for such services (at least the effective ones), we made the choice to anonymize the names of each service. Finally, to minimize our financial contributions to a potentially criminal ecosystem, we limited the number of purchases to those needed to establish that a service "worked" and, if so, that its modus operandi was consistent over time.

4.4 Hack for Hire Playbook

Our study characterizes the operational methods that hack for hire services employ when making a credible attempt to hijack our victim personas. We limit our analysis exclusively to the five services where the attackers made a detectable attempt to gain access to our victim account. We note that the ultimate "success" of these attacks is partially dependent on our experimental protocol: in some cases, we supplied 2FA SMS codes to phishing attacks or installed a provided executable, while in other cases, we avoided such actions to see if the attackers would adapt.

4.4.1 Attacks Overview

We present a high-level breakdown of each hack for hire service's playbook in Table 4.3. Four of the five services relied on phishing, while just one relied on malware. In all cases, attacks began with an email message to our victim persona's Gmail address. We never observed brute force login attempts, communication with a victim's Facebook account, or communication to our associate personas of any kind.[6] On average, attackers would send roughly 10 email messages over the course of 1 to 25 days—effectively a persistent attack until success. All of the services but one were able to bypass Gmail spam filtering (though to varying degrees of success) until at least one of their messages appeared in our victim's inbox. However, this outcome is expected: since these are targeted attackers with more focused motivation, they have strong incentives to adapt to phishing and spam defenses to ensure that their messages arrive in the victim's

[6] In practice, a victim's password may be exposed in a third-party data breach. Our use of synthetic identities prevents this as a potential attack vector.

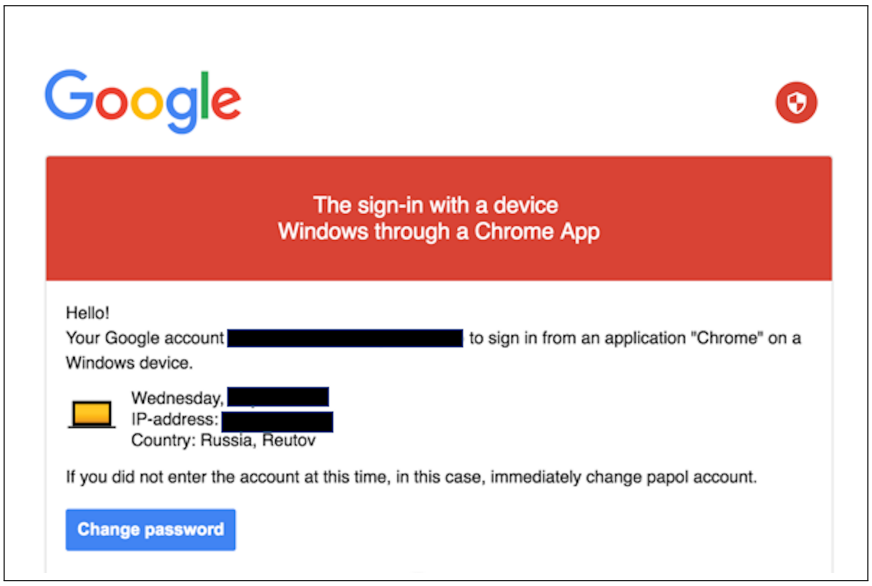


Figure 4.2. An example Google lure mimicking a real warning that Gmail will send to users. Identifying information removed and translated to English.

inbox. For example, attackers can create honeypot accounts of their own to test and modify their techniques, thereby ensuring a higher success rate; unlike their high-volume counterparts, targeted attackers only produce a modest number of examples and thus may pass "under the radar" of defenses designed to recognize and adapt to new large-scale attacks.

4.4.2 Email Lures

Each email message contained a lure impersonating a trusted associate or other source of authority to coerce prospective victims into clicking on a link. We observed five types of lures: those impersonating an associate persona, a stranger, a bank, Google, or a government authority. The associate lures tempted the user to click on an "image" for the victim's associate (using the personal connection as a sense of safety), while the Google, bank, and government lures conveyed a sense of urgency to induce a user to click on the link. Figure 4.2 shows a sample Google lure that mimics a real warning used by Google about new device sign-ins. Such lures highlight the challenge of distinguishing authentic communication from service providers, whereby attackers repurpose potentially common experiences to deceive victims into taking an unsafe action.

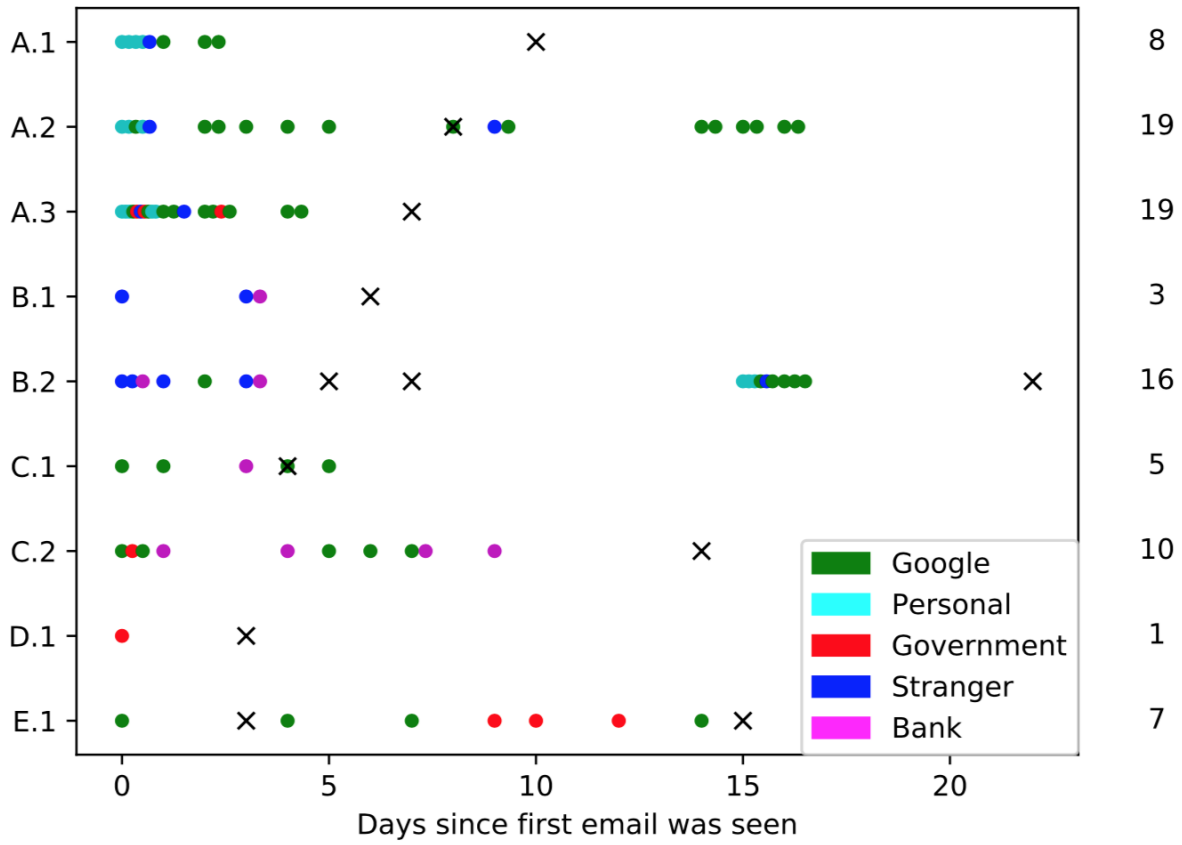


Figure 4.3. Different types of lures used by services that attempted to access a victim account. An 'X' marks when we clicked on a link in a message sent to a victim. Numbers on the right denote the total number of emails sent by a service.

Attackers cycled through multiple lures over time in an apparent attempt to find any message that would entice a victim into clicking on a link. Figure 4.3 shows the elapsed time since attackers sent their first email message to our victim account, the type of lure they used for each message, and when we clicked on the lure acting as a victim (potentially halting further attempts). Each row corresponds to one attack on a victim, and the *x*-axis counts the number of days since the service sent their first message to the victim. The numbers on the right *y*-axis show the number of messages sent by the service to the victim. The most popular lure mimicked Google, followed by associates and then lures from strangers.

Of the five services, two relied on personalized messages when communicating with four victim personas. In three of these cases, the service asked for additional details upfront

Table 4.4. For services that attempted to hack a victim account, we show whether Google was used in the phishing URL, whether the phishing page used HTTPS, and the number of redirects to the phishing page. We include separate rows for the services that sent multiple messages (services B and E).

Service	'google' in URL?	HTTPS	# redirects to phishing page
A.1	Yes	Yes	2
A.2	Yes	Yes	2
A.3	Yes	Yes	2
B.1	Yes	No	1
B.2.1	Yes	No	1
B.2.2	Yes	No	1
B.2.3	Yes	Yes	2
C.1	No	No	0
C.2	NA	NA	NA
D.1	NA	NA	NA
E.1.1	Yes	Yes	1
E.1.2	Yes	Yes	2

about the victim persona during negotiation. Only service A.1 was able to construct personal lures without requesting assistance from the buyer, finding the details from the victim persona's website. The extent of personalization was limited, though, consisting either of mimicking the victim persona's company or their associate's personal email address. No additional branding was lifted from our web sites.

4.4.3 Phishing Landing Pages

All services but one relied on phishing as their attack vector. Once we clicked on the links sent to the victim personas, we were redirected to a spoofed Google login page that requested the credentials from the victim. Table 4.4 lists the different attack attempts and the degree to which attackers tried to spoof a Google domain, use HTTPS, or mask URLs from a crawler via multiple redirects. All services but one used "combo" domain name squatting [49] with the keyword 'google' in the URL, presumably to trick the victim into thinking that the URL was a real Google subdomain. Services A.2 and B.2 used the same fully qualified domain name for the phishing landing page, suggesting that they share a business relationship (i.e., they may both be

value-added resellers for the same phishing page service). Long-lived, reused domains suggest that they are valuable and perhaps relatively costly to acquire.

All but one service tried to obscure the URL to their phishing page with at least one layer of redirection. (The exception was the link in the phishing message from C.2, which redirected to an error page on a Russian hosting service indicating that the page had been taken down.) The redirection URLs seemed to be one-time use URLs, since we were not able to visit them after the attack executed and did not see repeat redirection URLs in any of the attacks. One-time use URLs are attractive for attackers because they can greatly complicate investigating attacks after the fact or sharing attack information among organizations.

Figure 4.4 shows an example page flow used by one hacking service. We always entered the Gmail credentials of the victim to see how the hacking attempt would progress. After collecting the password, all but one of the hacking services would redirect to a new screen which asked for the 2FA code that the victim had just received on their phone from Google.

Six of the nine hacking attempts captured the password from the phishing page and then immediately tried to use it to login to the victim's account (as verified with our Gmail access logging). Due to the similar behavior and speed at which these logins occurred, we believe that most of these services used an automated tool, similar to Evilginx [24], for this step.

Moreover, three of five of these attacks captured the necessary information in one session visiting the phishing pages. This sophistication suggests that attackers can readily adapt any additional information requested by Google as a secondary factor. Since our study, Google launched additional protections at login to prevent automated access attempts [94]. However, hardware security keys remain the best protection mechanism against phishing for users.

4.4.4 Live Adaptation

Services B.2 and E.1 exhibited phishing attacks that adapted over time to overcome obstacles. These services, once realizing that the account used 2FA, sent new phishing email messages with a different structure than the ones they sent previously. Service E.1, for example,

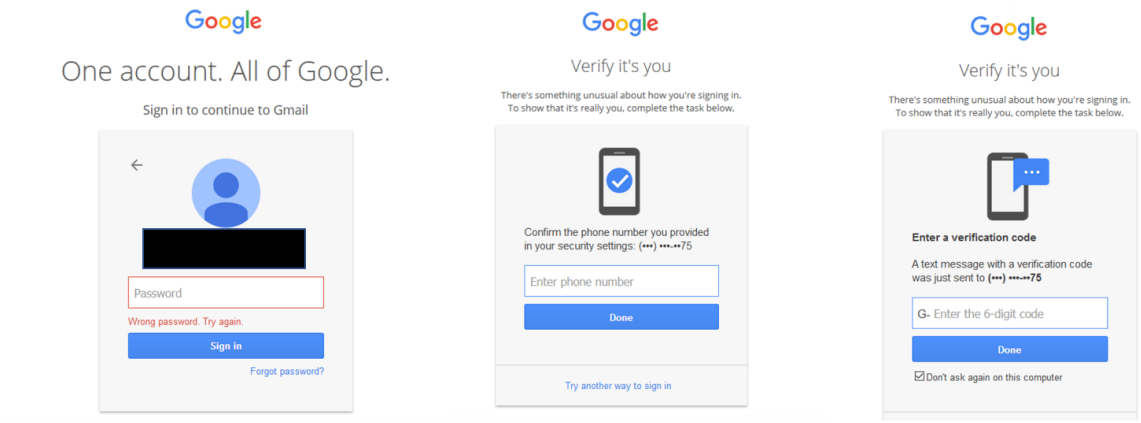


Figure 4.4. A service phishing flow, with identifiable information redacted. The flow is purposefully designed to mimic Gmail to trick the user into trusting the site.

initially used a phishing attack that only captured the Gmail password. When the service attempted to login, they were blocked by the 2FA prompt. The service then contacted our buyer persona asking for the victim's phone number. The victim's email account subsequently received more phishing messages in their inbox. Clicking on the link in the phishing messages led to a page that requested the 2FA code that was sent to the victim's phone. When we entered the 2FA code into the phishing page, the service was able to successfully login. This behavior indicates live testing of password validity, as the attackers were able to determine if the account had 2FA.

Service B.2 was similar to service E.1, but when they were blocked by the 2FA challenge they switched to phishing messages that looked exactly like the messages from service A. Upon collecting the password and the 2FA code that was sent to the phone number for the victim, the service was able to login.

4.4.5 Malware Attachments

Service D was the only service that attempted to hijack our victim account using malware. The attacker in this case sent just one email message to our victim persona—flagged as spam—that contained a link to a `rar` archive download (Gmail forbids executable attachments). The archive contained a sole executable file. We unpacked and ran the executable in an isolated

environment, but to no effect. According to VirusTotal [110], it is a variant of TeamViewer (a commercial tool for remote system access) which would have enabled the attacker to hijack any existing web browsing sessions.

After no further visible activity, the service eventually contacted our buyer persona to say that they could not gain access to our victim account. We decided to hire them again via a different contract (and different buyer and victim personas) to see if the seller would adapt to Gmail's defenses. However, we observed no email messages from the attacker the second time around, even in our spam folder. The seller eventually responded stating that they could not gain access to our second persona's account. While this malware vector proved unsuccessful, the presence of remote access tools poses a significant risk for adaptation, as session hijacking would enable an attacker to bypass any form of two-factor authentication.

4.4.6 Post Compromise

For those services that did obtain our victims' credentials and 2FA codes, the attackers proceeded to sign in to each account and immediately removed all Google email notifications (both from the inbox and then trash) related to a new device sign-in. None changed the account password. We also observed that services A, B, and E removed the 2FA authentication and the recovery number from our victim accounts as well. Presumably they took these steps to regain access to the account at a later time without having to phish an SMS code again, but we did not see any service log back into the accounts after their initial login. However, these changes to the account settings could alert a real victim that their account had been hijacked, a discovery which the attackers are willing to risk.

Once accessed, all but one of the services abused a portability feature in Google services (Takeout) to download our victim account's email content and then provided this parcel to our buyer persona. One advantage of this approach is that it acquires the contracted deliverable in one step, thus removing risks associated with subsequent credentials changes, improvements in defenses, or buyer repudiation. Only service C avoided logging into our victim account and

only provided the buyer persona with a password.[7] These findings highlight an emerging risk with data portability and regulations around streamlining access to user data. While intended for users, such capabilities also increase the ease with which a single account hijacking incident can expose all of a user's data to attackers. Since our study, Google has added additional step-up verification on sensitive account actions.

4.5 Real Victims & Market Activity

Based on our findings from the hack for hire process, we returned to the forums of the most successful attackers to understand their pricing for other services and how they attract buyers. Additionally, we present an estimate of the number of real victims affected by these services based on login traces from Google. Our findings suggest that the hack for hire market is quite niche, with few merchants providing hijacking capabilities beyond a handful of providers.

4.5.1 Victims Over Time

Of the 27 initial services we contacted, only three—services A, E, and B—could successfully login to our honeypot accounts. Google examined metadata associated with each login attempt and found that all three services rely on an *identical* automation process for determining password validity, bypassing any security check such as producing an SMS challenge, and downloading our honey account's email history. Whereas the email messages from the services had varied senders and delivery paths for each contracted campaign, this automation infrastructure remained stable despite eight months between our successive purchases. This stability in turn allowed Google to develop a signature allowing the retrospective analysis of all such login attempts from the three services in aggregate.

Over a seven-month period from March 16 to October 15, 2018, Google identified 372 accounts targeted by services A, B, and E. Figure 4.5 shows a weekly breakdown of activity. On

[7]The service demanded additional payment to defeat the 2FA, which we paid, at which point they stopped responding to our requests.

an average week, these services attacked 13 targets, peaking at 35 distinct accounts per week. We caution these estimates are likely only lower bounds on compromise *attempts* as we cannot observe users who received a phishing URL, but did not click it (or otherwise did not enter their password on the landing page). Despite these limitations, the volume of activity from these hack for hire services is quite limited when compared to off-the-shelf phishing kits which impact over 12 million users a year [101]. Thus, we surmise that the targeted account hacking market is likely small when compared to other hacking markets, e.g., for malware distribution [37]. While the damage from these commercialized hacking services may be more potent, they are only attractive to attackers with particular needs.

Apart from the volume of these attacks, we also examine the sophistication involved. As part of its authentication process, Google may trigger a "challenge" for sign-in attempts from previously unseen devices or network addresses [65]. All of the hack for hire attempts triggered this detection. In 68% of cases, the attacker was forced to solve an SMS challenge, while in 19% of cases the attacker only had to supply a victim's phone number. The remaining 13% involved a scattering of other secondary forms of authentication. This layered authentication approach provides better security when compared to passwords alone, with attackers only correctly producing a valid SMS code for 34% of accounts and a valid phone number in 52% of cases. These rates take into consideration repeated attacks: Google observed that attackers would attempt to access each account a median of seven times before they either succeeded or abandoned their efforts. As such, even though these attacks may be targeted, Google's existing account protections can still slow and sometimes stop attackers from gaining access to victim accounts.

4.5.2 Alternate Services and Pricing

While our investigation focused on Google—due in large part to our ethical constraints and abiding by Terms of Service requirements—the hack for hire services we engaged with also purport to break into multiple mail providers (Yahoo, Mail.ru, Yandex), social networks

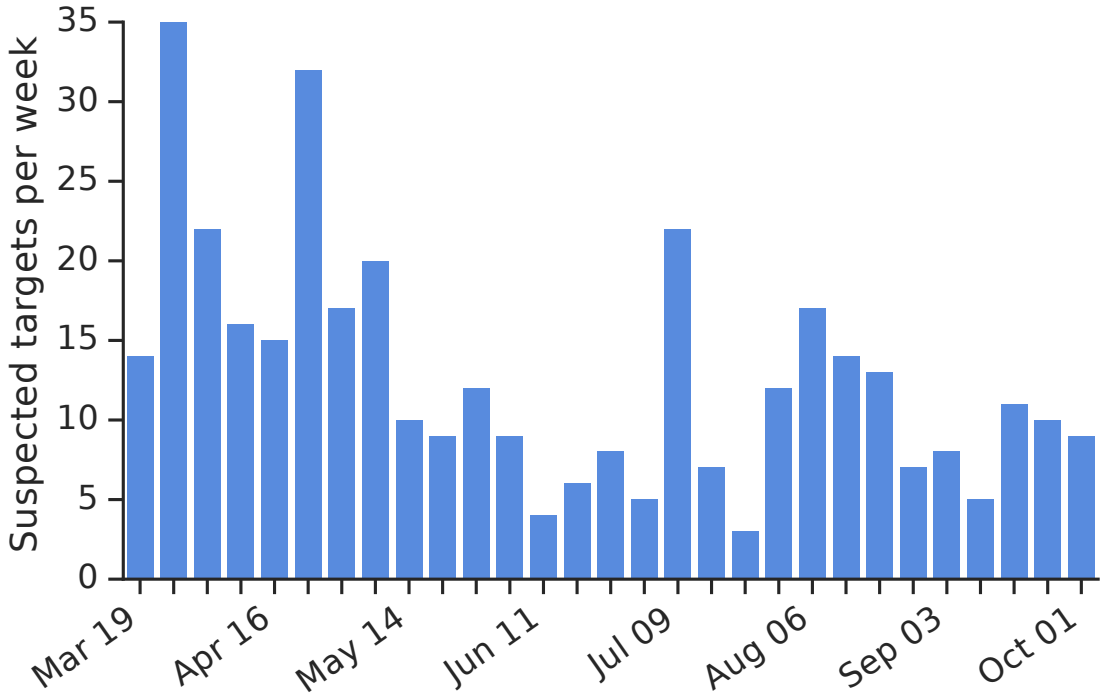


Figure 4.5. Weekly target accounts retroactively associated with hack for hire services.

(Facebook, Instagram), and messaging apps (WhatsApp, ICQ, Viber). To provide a price comparison between offerings, in preparation for our study we performed a weekly crawl of the forum page or dedicated web site advertising each service starting in January 1, 2017. However, as detailed previously in Section 4.4, only a fraction of the services are authentic, and just three—services A, B, and C—had online prices that matched (or were close) to the final price we paid. We treat these as trusted sources of pricing information. We also include services E and D, but note their prices were higher than advertised. We exclude all other services as they failed to attack any of our victim personas.

We present a breakdown of pricing information as of October 10, 2018 in Table 4.5 for the five services that executed an attempt to access the accounts. Across all five services, Russian mail provider hacking (i.e., Mail.ru, Rambler and Yandex) was the cheapest, while other mail providers such as Gmail and Yahoo were more expensive. The cost of hacking a social media account falls in the middle of these two extremes.

Some services increased their prices over time. For services B and C, prices on the

Table 4.5. Purported prices to access various accounts, based on an October 10, 2018 snapshot. All prices USD, converted from rubles. An '*' indicates the service's advertised price was lower than the final payout requested.

Target	Service A	Service B	Service C	Service D*	Service E*
Mail.ru	$77	$77	$62	$54	$77
Rambler	$152	$108	$77	$77	$108
Yandex	$106	$108	$77	$77	$108
Gmail	$384	$385	$92	$77	Negotiable
Yahoo	$384	$231	$92	–	–
Facebook	$306	–	–	–	–
Instagram	$306	–	–	–	$231

forums they advertise have been stable since we first began our monitoring. Only service A provided dynamic pricing, with rates increasing as shown in Figure 4.6. Since 2017, Gmail prices have steadily increased from $123 to $384, briefly peaking at $461 in February 2018. The advertised rates for targeting Yahoo accounts has largely tracked this same rate, while Facebook and Instagram were initially priced higher before settling at $307. We hypothesize that the price differences between services and the change in prices for a service over time are likely driven by both operational and economic factors. Thus, prices will naturally increase as the market for a specific service shrinks (reducing the ability to amortize sunk costs on back-end infrastructure for evading platform defenses) and also as specific services introduce more, or more effective, protection mechanisms that need to be bypassed (increasing the transactional cost for each hacking attempt).

4.5.3 Advertising & Other Buyers

As a final measure, we examined the forum advertisements each service used to attract buyers. Here, we limit our analysis to the five successful hack for hire services. Across seven underground forums, we identified two types of advertisements—pinned posts and banner ads—which require paying forum operators. Services A, B, and E, the three services that were able to bypass two-factor authentication, all had pinned posts on forums where this option was available. Only service A paid for banner advertisements on all of these forums. Together, this suggests

that the services are profitable enough to continue advertising via multiple outlets. Additionally, these three services had verified accounts, indicating that a forum moderator had vetted the service stated. Further, services A, B, D, and E all stated they could work with a "guarantor", an escrow service proxying for payment between service and buyer to avoid fraud risks. Generally, feedback on the forum was positive, though we caution this may be biased due to the ability to delete posts and the difficulty in distinguishing between legitimate customers and virtual "shills". We avoid using forum posts as a count of purchases as most negotiation activity occurs via private messaging.

In addition to this qualitative search, we received an email advertisement from one of the services for upcoming changes to the service, which was sent to 44 other buyers as well (exposing their clientele's email addresses). The message was an announcement that the service now had a Telegram channel that was available (with a link to the channel), and to join the channel to keep up to date with relevant news. The only response to that initial email message was another customer exclaiming their excitement for this new development. Of the 44 email addresses that were leaked, 23 were accounts with `mail.ru` or `yandex.ru`, 9 were Gmail addresses, and the rest were various other providers, like Tutanota, Protonmail, or iCloud. We were unable to find these buyers online, which indicates that they did not engage in forum postings, or used a burner (one-purpose use) email address. However, the concentration of Russian mail providers suggests that interest in the market may largely be geographically limited, potentially due to language barriers or culturally-biased demands for account hacking.

4.6 Related Work

Phishing is a well studied, yet continuing concern in the security community. Sheng et al. studied the demographic of people who are susceptible to phishing attacks, and found that users 18–25 years of age are most likely to click on phishing messages [92]. Egleman et al. studied the effectiveness of phishing warnings and found they can be successful in preventing

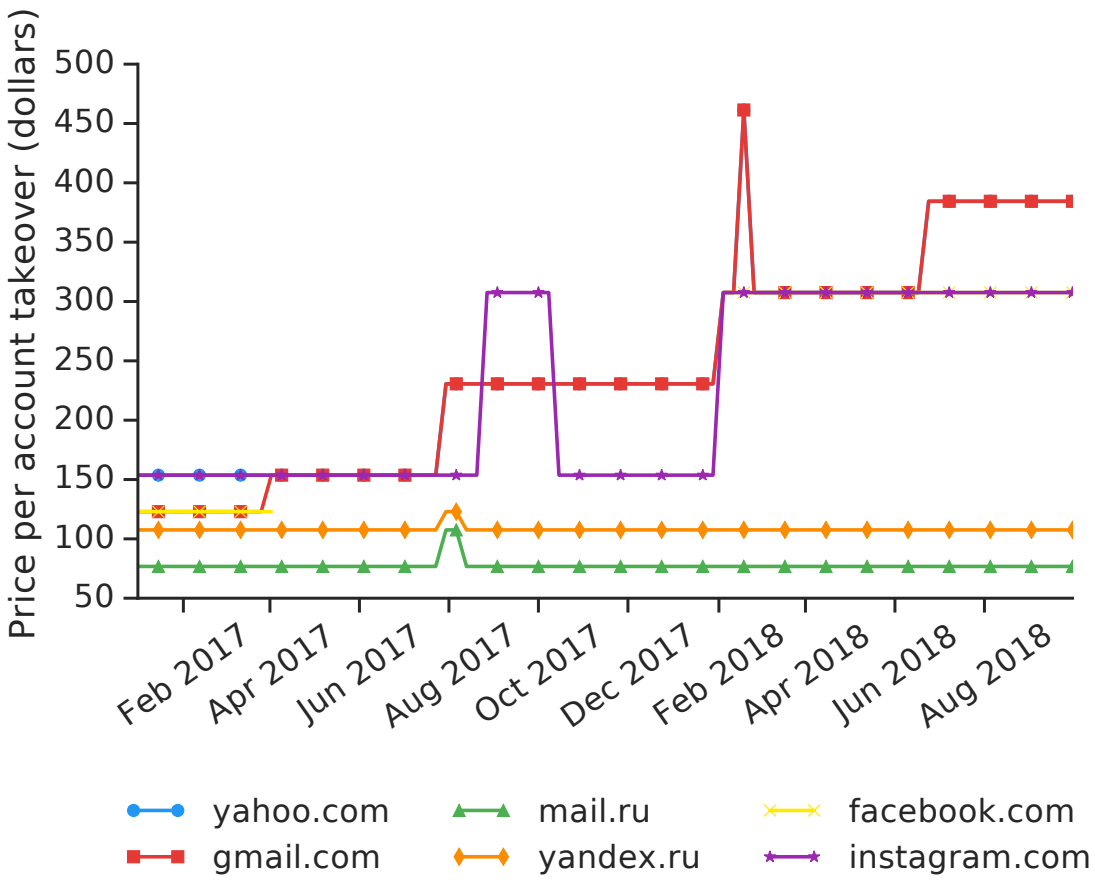


Figure 4.6. Monthly price that Service A charges across email and social network account providers. Over two years, the price per Gmail account increased from $123 to $384.

account hijacking [21]. Following this mode of thought, there are a variety of studies on effective anti-phishing training as well as the creation of a content-based approach to detect phishing sites [119, 93, 51]; in all of these studies, the percentage of users' susceptible to phishing emails dropped. Similarly, Zhang et al. evaluted anti-phishing tools, and found that many of them are not effective on new URLs and have exploits of their own. Oest et al. also studied the phishing ecosystem via an analysis of phishing kits, and developed a URL-based scheme to detect phishing URLs [70].

Account hijacking threats represent a spectrum that ranges from financially motivated, large-scale attacks to highly-targeted incidents motivated by political, personal, or financial incentives. Thomas et al. identified billions of credentials stolen via data breaches and millions

of credentials stolen by phishing kits and keyloggers, with phishing posing the largest hijacking risk [101]. Once an account was accessed, hijackers searched for financial records or used the account as a stepping stone to control connected online identities [71, 11]. While techniques such as risk-aware authentication [65] or two-factor authentication help protect against unsophisticated bulk attacks, the hack for hire outfits we studied were more dedicated, with attackers stealing SMS two-factor codes as part of their phishing pages to bypass the additional layers of security. Security keys would prevent this attack vector.

At the other end of the spectrum, Marczak et al. investigated government actors targeting political dissidents [62]. The hijackers in these cases relied on exploits or social engineering to have victims install commercial or off-the-shelf spyware to enable long-term monitoring of the victim's activities. Email was a common delivery mechanism, where attackers customized their lures to the NGOs where employees worked or to the human rights topics they were involved with [54, 42]. Given the risks involved here, researchers have focused on how to improve the security posture of at-risk users [61]. Compared to our work, we found more generalized lures that can work for any target (e.g., your account is running out of storage space or there was a security incident), while phishing was the most popular technique for gaining one-off access to a victim's account. Pressure on the hack for hire playbook, or wider-scale adoption of security keys, may cause them to move towards malware and thus mirror government attackers.

4.7 Discussion and Conclusion

When starting this study, we had very little knowledge of what to expect in terms of attacker methods, behaviors, and ability. At a high level, we find that the commercial account hijacking ecosystem is far from mature. When such attackers are successful, they can be potentially devastating to individuals. Yet, as an overall market it is not poised to cause widespread harm.

Retail account hijacking is a niche market. Many aspects of engaging with account

hijackers strongly indicate that these services are a fledgling market:

1. Most telling is that only five of the 27 services we contacted were willing to take our business, a third never responded to repeated requests as buyers, and some were outright fraudulent.

2. Services have inconsistent and poor customer service. For example, three of the services charged significantly higher prices than their advertised price, and two services changed their initial prices while they were executing the hack. Moreover, customer service is slow and inconsistent in their communication with the buyer, sometimes taking more than a day to respond.

3. Attackers showed little initiative. Most attacks made no effort to gather information independently about their victims. Of the nine attempts, only services A.1 and A.2 discovered additional information about the victim on their web sites, such as the name of their associate. The others, including different contracts within service A, would not attempt hacking the account without explicitly requesting additional information from the buyer.

In contrast, studies on markets for CAPTCHA solving [67], Twitter spam [102], and Google phone verified accounts [100] show that those services are quick to respond, and stable in their services and pricing. This differentiation between other underground service offerings and the retail hacking market suggests that account hacking may not be the main focus of these attackers, and may simply be a "side hustle" — a method to gain opportunistic income in addition to other activities they are more fully engaged in.

Services predominantly mount social engineering attacks using targeted phishing email messages. All but one of the nine attacks used targeted email phishing to hack into our Gmail accounts. The attackers customized their phishing messages using details that we made available

about the businesses and associates of our fictitious victims. To prompt engagement with a victim, the phishing messages created a sense of urgency by spoofing sources of authority.

These methods are a subset of those used in other targeted attack ecosystems. In particular, in *addition* to targeted phishing (frequently much more tailored than any attacks mounted by the services we studied), government-targeted attackers use malware and long-term monitoring of victim behavior to gain access to the account, requiring much more overhead than phishing alone [62]. Indeed, although these two classes of attackers are superficially similar in focusing on individual users, they are distinct in most other respects including the nature of the populations they target, their resource investment per target, their goals upon compromising an account, and a far greater requirement for covert operations.

Two-factor authentication creates friction. Even though phishing can still be successful with 2FA enabled, our results demonstrate that 2FA adds friction to attacks. Various services said that they could not hack into the account without the victim's phone number, had to adapt to 2FA challenges by sending new phishing messages to bypass them, and one renegotiated their price (from $307 to $690) when they discovered that the account had 2FA protection. Based on these results, we recommend major providers encourage or require their user base to use a 2FA physical token

Minimal service differentiation. Even with a variety of services advertising in the account hijacking market, they have remarkably little differentiation in their methods and infrastructure. Four services sent very similar re-usable phishing email messages to their respective victims, and all services that successfully hacked our accounts used identical automation tools for determining password validity, bypassing security checks, and downloading victim data.

Gmail as a vantage point. Overall, our study indicates that the attack space against Gmail is quite limited. Since we focused on hiring services to hack solely into Gmail accounts, it is possible that the landscape of the commercialized hacking market would look much different when deployed against native email services such as `mail.ru` or `yandex.ru`.

Chapter 4, in full, is a reprint of the material as it appears in The World Wide Web

Conference 2019. Ariana Mirian, Joe DeBlasio, Stefan Savage, Geoffrey M. Voelker, and Kurt Thomas. The dissertation author was the primary investigator and author of this material.

Chapter 5

Conclusion

Security is an aspect that touches many users lives, and while important, it is infeasible for a user to execute best security practices constantly. However, a user might not *need* to employ all best practices in order to remain safe on the Internet. As such, in this disseration, I argued that using large-scale measurement would allow for better prioritization of security practices.

I first explored a large-scale analysis of end user behavior from the perspective of a network tap at UCSD's campus, and related canonical security "best practices" to compromise. I found that many relations are counterintuitive to popular beliefs. While this does not mean a user should stop executing best practices, it does suggest that these best practices are not what a user should prioritize, given limited time and energy.

Next, I examined how we can prioritize security processes from an organizational perspective. I determined which of UCSD's organizational efforts to change user security behavior were most effective in changing user behaviors, which helps the organization better prioritize how to communicate similar future efforts.

Finally, I investigated this question from the attacker perspective. Using measurement, I quantified the commodity market that provides hacking services for hire, which subsequently supplied insight into defenses that would best protect against these types of attacks.

Using these three projects, I exemplify how empirical measurement of large-scale behaviors is an effective tool in prioritizing security practices from many different perspectives. This

tool is one that can be used moving forward to continue to gain knowledge and insight into how best to prioritize security pratices for all.

Bibliography

[1] Jacob Abbott and Sameer Patil. How Mandatory Second Factor Affects the Authentication User Experience. In *Proceedings of the ACM SIGCHI Conference on Human Factors in Computing Systems*, pages 1–13, 2020.

[2] Devdatta Akhawe and Adrienne Porter Felt. Alice in Warningland: A Large-Scale Field Study of Browser Security Warning Effectiveness. In *Proceedings of the 22nd USENIX Security Symposium*, pages 257–272, Washington, D.C., August 2013.

[3] Olabode Anise and Kyle Lady. State of the Auth: Experiences and Perceptions of Multi-Factor Authentication. *Duo Security*, https://duo.com/assets/ebooks/state-of-the-auth.pdf, November 2017. Accessed: 2018-10-22.

[4] Apache Software Foundation. Apache Hive Website. https://hive.apache.org/, 2019.

[5] Ingolf Becker, Simon Parkin, and M. Angela Sasse. The Rewards and Costs of Stronger Passwords in a University: Linking Password Lifetime to Strength. In *Proceedings of the USENIX Security Symposium*, pages 239–253, August 2018.

[6] Mihir Bellare and Phillip Rogaway. The FFX Mode of Operation for Format-Preserving Encryption. *Manuscript (standards proposal) submitted to NIST*, January 2010.

[7] Sruti Bhagavatula, Lujo Bauer, and Apu Kapadia. (How) Do People Change Their Passwords After a Breach? In *Workshop on Technology and Consumer Protection*. IEEE, 2020.

[8] Leyla Bilge, Yufei Han, and Matteo Dell'Amico. RiskTeller: Predicting the Risk of Cyber Incidents. In *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security (CCS)*, Dallas, Texas, USA, November 2017.

[9] Marina Sanusi Bohuk, Mazharul Islam, Suleman Ahmad, Michael Swift, Thomas Ristenpart, and Rahul Chatterjee. Gossamer: Securely measuring password-based logins. In *31st USENIX Security Symposium (USENIX Security 22)*, pages 1867–1884, Boston, MA, August 2022. USENIX Association.

[10] Joseph Bonneau. The Science of Guessing: Analyzing an Anonymized Corpus of 70 Million Passwords. In *Proceedings of the IEEE Symposium on Security and Privacy*, pages 538–552, 2012.

[11] Elie Bursztein, Borbala Benko, Daniel Margolis, Tadek Pietraszek, Andy Archer, Allan Aquino, Andreas Pitsillidis, and Stefan Savage. Handcrafted Fraud and Extortion: Manual Account Hijacking in the Wild. In *Proceedings of the 2014 ACM Internet Measurement Conference (IMC)*, Vancouver, BC, Canada, November 2014.

[12] Davide Canali, Leyla Bilge, and Davide Balzarotti. On the effectiveness of risk prediction based on users browsing behavior. In *Proceedings of the 9th ACM Symposium on Information, Computer and Communications Security (CCS)*, Kyoto, Japan, June 2014.

[13] Yannick Carlinet, Ludovic Mé, Hervé Debar, and Yvon Gourhant. Analysis of Computer Infection Risk Factors Based on Customer Network Usage. In *2008 Second International Conference on Emerging Security Information, Systems and Technologies*, Cap Esterel, France, August 2008.

[14] Orun etin, Carlos Gan, Lisette Altena, Samaneh Tajalizadehkhoob, and Michel van Eeten. Tell Me You Fixed It: Evaluating Vulnerability Notifications via Quarantine Networks. In *Procedings of the 4th IEEE European Symposium on Security and Privacy (EuroS&P)*, pages 326–339, June 2019.

[15] Sonia Chiasson and P. C. Oorschot. Quantifying the Security Advantage of Password Expiration Policies. *Des. Codes Cryptography*, 77(23):401–408, December 2015.

[16] Jessica Colnago, Summer Devlin, Maggie Oates, Chelse Swoopes, Lujo Bauer, Lorrie Cranor, and Nicolas Christin. "It's Not Actually That Horrible": Exploring Adoption of Two-Factor Authentication at a University. In *Proceedings of the ACM SIGCHI Conference on Human Factors in Computing Systems*, pages 1–11, 2018.

[17] Lorrie Faith Cranor. A Framework for Reasoning about the Human in the Loop. In *Proceedings of the Conference on Usability, Psychology, and Security*, 2008.

[18] Lorrie Faith Cranor, Serge Egelman, Jason I. Hong, and Yue Zhang. Phinding phish: An evaluation of anti-phishing toolbars. In *NDSS*, 2007.

[19] Emilano De Cristofaro, Arik Friedman, Guillaume Jourjon, Mohamed Ali Kaafar, and M. Zubair Shafiq. Paying for Likes? Understanding Facebook Like Fraud Using Honeypots. In *Proceedings of the 2014 ACM Internet Measurement Conference (IMC)*, Vancouver, BC, Canada, November 2014.

[20] Matteo Dell'Amico, Pietro Michiardi, and Yves Roudier. Password Strength: An Empirical Analysis. In *Proceedings of IEEE INFOCOM*, pages 983–991, 2010.

[21] Serge Egelman, Lorrie Faith Cranor, and Jason Hong. You've been warned: An empirical study of the effectiveness of web browser phishing warnings. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '08, pages 1065–1074, New York, NY, USA, 2008. ACM.

[22] Emma. Email analytics. https://myemma.com/email-marketing-features/email-analytics/.

[23] Enron Email Dataset. https://www.cs.cmu.edu/~enron/. Accessed: 2018-11-03.

[24] Evilginx — Advanced Phishing with Two-factor Authentication Bypass. https://breakdev.org/evilginx-advanced-phishing-with-two-factor-authentication- bypass/. Accessed: 2018-10-22.

[25] Adrienne Porter Felt, Alex Ainslie, Robert W. Reeder, Sunny Consolvo, Somas Thyagaraja, Alan Bettes, Helen Harris, and Jeff Grimes. Improving SSL Warnings: Comprehension and Adherence. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, CHI '15, 2015.

[26] Adrienne Porter Felt, Robert W. Reeder, Alex Ainslie, Helen Harris, Max Walker, Christopher Thompson, Mustafa Emre Acer, Elisabeth Morant, and Sunny Consolvo. Rethinking Connection Security Indicators. In *Twelfth Symposium on Usable Privacy and Security (SOUPS 2016)*, pages 1–14, Denver, CO, 2016. USENIX Association.

[27] Adrienne Porter Felt, Robert W. Reeder, Hazim Almuhimedi, and Sunny Consolvo. Experimenting At Scale With Google Chrome's SSL Warning. In *ACM CHI Conference on Human Factors in Computing Systems*, 2014.

[28] Firefox. How to stop Firefox from making automatic connections. https://support.mozilla.org/en-US/kb/how-stop-firefox-making-automatic-connections, 2019.

[29] Dinei Florencio and Cormac Herley. A large-scale study of web password habits. In *Proceedings of the 16th International Conference on World Wide Web*, WWW '07, page 657666, New York, NY, USA, 2007. Association for Computing Machinery.

[30] Dinei Florêncio, Cormac Herley, and Paul C. Van Oorschot. An Administrator's Guide to Internet Password Research. In *Proceedings of the USENIX Conference on Large Installation System Administration (LISA)*, pages 35–52. USENIX Association, 2014.

[31] Alain Forget, Sarah Pearman, Jeremy Thomas, Alessandro Acquisti, Nicolas Christin, Lorrie Faith Cranor, Serge Egelman, Marian Harbach, and Rahul Telang. Do or Do Not, There Is No Try: User Engagement May Not Improve Security Outcomes. In *Procedings of the 12th Symposium on Usable Privacy and Security (SOUPS)*, Denver, CO, USA, June 2016.

[32] Maximilian Golla, Grant Ho, Marika Lohmus, Monica Pulluri, and Elissa M. Redmiles. Driving 2FA Adoption at Scale: Optimizing Two-Factor Authentication Notification Design Patterns. In *Proceedings of the USENIX Security Symposium*, pages 109–126, August 2021.

[33] Google. Add 2-Step Verification. https://support.google.com/a/answer/175197. Accessed: 2018-10-22.

[34] Google. Guard Against Targeted Attacks. https://support.google.com/a/answer/9010419. Accessed: 2018-10-22.

[35] Google. Verify a user's identity with a login challenge. https://support.google.com/a/answer/6002699. Accessed: 2018-10-22.

[36] Garrett M. Graff. DOJ Indicts 9 Iranians For Brazen Cyberattacks Against 144 US Universities. *Wired*, https://www.wired.com/story/iran-cyberattacks-us- universities-indictment/. Accessed: 2018-10-22.

[37] Chris Grier, Lucas Ballard, Juan Caballero, Neha Chachra añd Christian J. Dietrich, Kirill Levchenko, Panayiotis Mavrommatis, Da mon McCoy, Antonio Nappa, Andreas Pitsillidis, Niels Provos, M. Zub air Rafique, Moheeb Abu Rajab, Christian Rossow, Kurt Thomas, Vern Paxson, Stefan Savage, and Geoffrey M. Voelker. Manufacturing Compromise: The Emergence of Exploit-as-a-Service. In *Proceedings of the ACM Conference on Computer and Communications Security*, Raleigh, NC, October 2012.

[38] Guofei Gu, Phillip Porras, Vinod Yegneswaran, Martin Fong, and Wenke Lee. BotHunter: Detecting Malware Infection Through IDS-driven Dialog Correlation. In *Proceedings of 16th USENIX Security Symposium on USENIX Security Symposium*, Boston, MA, USA, August 2007.

[39] Carlos H. Ganan and Michel Eeten. Make Notifications Great Again: Learning How to Notify in the Age of Large-Scale Vulnerability Scanning. In *Proceedings of the Workshop on the Economics of Information Security (WEIS)*, pages 1–15, June 2017.

[40] Hana Habib, Jessica Colnago, William Melicher, Blase Ur, Sean Segreti, Lujo Bauer, Nicolas Christin, and Lorrie Cranor. Password creation in the presence of blacklists. 02 2017.

[41] Hana Habib, Pardis Emami-Naeini, Summer Devlin, Maggie Oates, Chelse Swoopes, Lujo Bauer, Nicolas Christin, and Lorrie Faith Cranor. User Behaviors and Attitudes under Password Expiration Policies. In *Proceedings of the USENIX Conference on Usable Privacy and Security (SOUPS)*, pages 13–30, 2018.

[42] Seth Hardy, Masashi Crete-Nishihata, Katharine Kleemola, Adam Senft, Byron Sonne, Greg Wiseman, Phillipa Gill, and Ronald J Deibert. Targeted Threat Index: Characterizing and Quantifying Politically-Motivated Targeted Malware. In *Proceedings of the 23rd USENIX Security Symposium*, San Diego, CA, USA, August 2014.

[43] Delbert Hart. Attitudes and practices of students towards password security. 23(5):169174, may 2008.

[44] Cormac Herley. So Long, and No Thanks for the Externalities: The Rational Rejection of Security Advice by Users. In *Proceedings of the 2009 Workshop on New Security Paradigms Workshop*, Oxford, United Kingdom, September 2009.

[45] Philip G. Inglesant and M. Angela Sasse. The True Cost of Unusable Password Policies: Password Use in the Wild. In *Proceedings of the ACM SIGCHI Conference on Human Factors in Computing Systems*, pages 383–392, 2010.

[46] Ian Karambelas. Spear Phishing: The Secret Weapon Behind the Worst Cyber Attacks. *Cloudmark*, https://blog.cloudmark.com/2016/01/13/spear-phishing-secret-weapon-in-worst-cyber-attacks/, January 2016. Accessed: 2018-10-22.

[47] Patrick Gage Kelley, Saranga Komanduri, Michelle L. Mazurek, Richard Shay, Timothy Vidas, Lujo Bauer, Nicolas Christin, Lorrie Faith Cranor, and Julio Lopez. Guess Again (and Again and Again): Measuring Password Strength by Simulating Password-Cracking Algorithms. In *Proceedings of the IEEE Symposium on Security and Privacy*, pages 523–537, 2012.

[48] Moazzam Khan, Zehui Bi, and John A. Copeland. Software updates as a security metric: Passive identification of update trends and effect on machine infection. In *Proceedings of IEEE Military Communications Conference (MILCOM)*, Orlando, Florida, USA, October 2012.

[49] Panagiotis Kintis, Najmeh Miramirkhani, Charles Lever, Yizheng Chen, Roza Romero-Gomez, Nikolaos Pitropakis, Nick Nikiforakis, and Manos Antonakakis. Hiding in Plain Sight: A Longitudinal Study of Combosquatting Abuse. In *Proceedings of the 2017 ACM Conference on Computer and Communications Security (CCS)*, Dallas, TX, USA, October 2017.

[50] Saranga Komanduri, Richard Shay, Patrick Gage Kelley, Michelle L. Mazurek, Lujo Bauer, Nicolas Christin, Lorrie Faith Cranor, and Serge Egelman. Of passwords and people: Measuring the effect of password-composition policies. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '11, page 25952604, New York, NY, USA, 2011. Association for Computing Machinery.

[51] Ponnurangam Kumaraguru, Yong Rhee, Alessandro Acquisti, Lorrie Faith Cranor, Jason Hong, and Elizabeth Nunge. Protecting People from Phishing: The Design and Evaluation of an Embedded Training Email System. In *Proceedings of the 2007 Conference on Human Factors in Computing Systems (CHI)*, pages 905–914, San Jose, CA, USA, April 2007.

[52] Cynthia Kuo, Sasha Romanosky, and Lorrie Faith Cranor. Human selection of mnemonic phrase-based passwords. In *Proceedings of the Second Symposium on Usable Privacy and Security*, SOUPS '06, page 6778, New York, NY, USA, 2006. Association for Computing Machinery.

[53] Fanny Lalonde Lévesque, Jude Nsiempba, José M. Fernandez, Sonia Chiasson, and Anil Somayaji. A Clinical Study of Risk Factors Related to Malware Infections. In *Proceedings of the 2018 ACM SIGSAC Conference on Computer and Communications Security (CCS)*, Berlin, Germany, November 2013.

[54] Stevens Le Blond, Adina Uritesc, Cédric Gilbert, Zheng Leong Chua, Prateek Saxena, and Engin Kirda. A Look at Targeted Attacks Through the Lense of an NGO. In *Proceedings of the 23rd USENIX Security Symposium*, San Diego, CA, USA, August 2014.

[55] Frank Li, Zakir Durumeric, Jakub Czyz, Mohammad Karami, Michael Bailey, Damon McCoy, Stefan Savage, and Vern Paxson. You've Got Vulnerability: Exploring Effective Vulnerability Notifications. In *25th USENIX Security Symposium (USENIX Security 16)*, 2016.

[56] Frank Li, Grant Ho, Eric Kuan, Yuan Niu, Lucas Ballard, Kurt Thomas, Elie Bursztein, and Vern Paxson. Remedying Web Hijacking: Notification Effectiveness and Webmaster Comprehension. In *International World Wide Web Conference*, 2016.

[57] Enze Liu, Amanda Nakanishi, Maximilian Golla, David Cash, and Blase Ur. Reasoning analytically about password-cracking software. In *2019 IEEE Symposium on Security and Privacy (SP)*, pages 380–397, 2019.

[58] Suqi Liu, Ian Foster, Stefan Savage, Geoffrey M. Voelker, and Lawrence K. Saul. Who is .com? Learning to Parse WHOIS Records. In *Proceedings of the 2015 ACM Internet Measurement Conference (IMC)*, Tokyo, Japan, October 2015.

[59] Yang Liu, Armin Sarabi, Jing Zhang, Parinaz Naghizadeh, Manish Karir, Michael Bailey, and Mingyan Liu. Cloudy with a Chance of Breach: Forecasting Cyber Security Incidents. In *Proceedings of the 24th USENIX Conference on Security Symposium*, Washington, DC, USA, August 2015.

[60] Max Maaß, Henning Pridöhl, Dominik Herrmann, and Matthias Hollick. Best Practices for Notification Studies for Security and Privacy Issues on the Internet. In *Proceedings of The 16th International Conference on Availability, Reliability and Security (ARES)*, August 2021.

[61] William R Marczak and Vern Paxson. Social Engineering Attacks on Government Opponents: Target Perspectives. In *Proceedings of the 17th Privacy Enhancing Technologies Symposium (PETS)*, Minneapolis, MN, USA, July 2017.

[62] William R Marczak, John Scott-Railton, Morgan Marquis-Boire, and Vern Paxson. When Governments Hack Opponents: A Look at Actors and Technology. In *Proceedings of the 23rd USENIX Security Symposium*, San Diego, CA, USA, August 2014.

[63] Michelle L. Mazurek, Saranga Komanduri, Timothy Vidas, Lujo Bauer, Nicolas Christin, Lorrie Faith Cranor, Patrick Gage Kelley, Richard Shay, and Blase Ur. Measuring password guessability for an entire university. In *Proceedings of the 2013 ACM SIGSAC Conference on Computer & Communications Security*, CCS '13, page 173186, New York, NY, USA, 2013. Association for Computing Machinery.

[64] William Melicher, Blase Ur, Sean M. Segreti, Saranga Komanduri, Lujo Bauer, Nicolas Christin, and Lorrie Faith Cranor. Fast, lean, and accurate: Modeling password guessability using neural networks. In *25th USENIX Security Symposium (USENIX Security 16)*, pages 175–191, Austin, TX, August 2016. USENIX Association.

[65] Grzegorz Milka. Anatomy of Account Takeover. *Enigma*, https://www.usenix.org/node/208154, January 2018.

[66] Robert Morris and Ken Thompson. Password security: A case history. *Commun. ACM*, page 594597, nov 1979.

[67] Marti Motoyama, Kirill Levchenko, Chris Kanich, Damon McCoy, Geoffrey M. Voelker, and Stefan Savage. Re: CAPTCHAs: Understanding CAPTCHA-solving Services in an Economic Context. In *Proceedings of the 19th USENIX Security Symposium*, Washington, DC, USA, August 2010.

[68] Mozilla Foundation. Public Suffix List Website. https://publicsuffix.org/, 2019.

[69] ntop. PF_RING ZC (Zero Copy) Website. https://www.ntop.org/products/packet-capture/pf_ring/pf_ring-zc-zero-copy/, 2018.

[70] Adam Oest, Yeganeh Safei, Adam Doupé, Gail-Joon Ahn, Brad Wardman, and Gary Warner. Inside a Phisher's Mind: Understanding the Anti-phishing Ecosystem Through Phishing Kit Analysis. In *Proceedings of the 2018 APWG Symposium on Electronic Crime Research (eCrime)*, San Diego, CA, USA, September 2018.

[71] Jeremiah Onaolapo, Enrico Mariconti, and Gianluca Stringhini. What Happens After You Are Pwnd: Understanding the Use of Leaked Webmail Credentials in the Wild. In *Proceedings of the 2016 ACM Internet Measurement Conference (IMC)*, Santa Monica, CA, USA, November 2016.

[72] Simon Parkin, Samy Driss, Kat Krol, and M. Angela Sasse. Assessing the User Experience of Password Reset Policies in a University. In *Proceedings of the International Conference on the Technology and Practice of Passwords (PASSWORDS)*, pages 21–38, 2015.

[73] Vern Paxson. Bro: a System for Detecting Network Intruders in Real-Time. *Computer Networks*, 31(23-24):2435–2463, 1999.

[74] Robert Proctor, Mei-Ching Lien, Kim-Phuong Vu, E Schultz, and Gavriel Salvendy. Improving computer security for authentication of users: Influence of proactive password restrictions. *Behavior research methods, instruments, & computers : a journal of the Psychonomic Society, Inc*, 34:163–9, 06 2002.

[75] ProofPoint. ET Pro Ruleset. https://www.proofpoint.com/us/threat-insight/et-pro-ruleset, 2019.

[76] Redislabs. Redis Website. https://redis.io/, 2019.

[77] Elissa M. Redmiles, Sean Kross, and Michelle L. Mazurek. How I Learned to Be Secure: A Census-Representative Survey of Security Advice Sources and Behavior. In *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*, Vienna, Austria, October 2016.

[78] Elissa M. Redmiles, Sean Kross, and Michelle L. Mazurek. Where is the Digital Divide?: A Survey of Security, Privacy, and Socioeconomics. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, Denver, Colorado, USA, May 2017.

[79] Elissa M. Redmiles, Sean Kross, and Michelle L. Mazurek. How Well Do My Results Generalize? Comparing Security and Privacy Survey Results from MTurk, Web, and Telephone Samples. In *Proceedings of the 2019 IEEE Symposium on Security and Privacy*, San Fransisco, CA, USA, May 2019.

[80] Robert Reeder, Iulia Ion, and Sunny Consolvo. 152 Simple Steps to Stay Safe Online: Security Advice for Non-tech-savvy Users. *IEEE Security and Privacy*, 15(5):55–64, June 2017.

[81] Robert W. Reeder, Adrienne Porter Felt, Sunny Consolvo, Nathan Malkin, Christopher Thompson, and Serge Egelman. An Experience Sampling Study of User Reactions to Browser Warnings in the Field. In *Proceedings of the 2018 Conference on Human Factors in Computing Systems*, CHI '18, April 2018.

[82] Joshua Reynolds, Nikita Samarin, Joseph Barnes, Taylor Judd, Joshua Mason, Michael Bailey, and Serge Egelman. Empirical Measurement of Systemic 2FA Usability. In *Proceedings of the USENIX Security Symposium*, 2020.

[83] Armin Sarabi, Ziyun Zhu, Chaowei Xiao, Mingyan Liu, and Tudor Dumitras. Patch Me If You Can: A Study on the Effects of Individual User Behavior on the End-Host Vulnerability State. In *Proceedings of the 18th Passive and Active Measurement PAM*, Sydney, Australia, March 2017.

[84] Yukiko Sawaya, Mahmood Sharif, Nicolas Christin, Ayumu Kubota, Akihiro Nakarai, and Akira Yamada. Self-Confidence Trumps Knowledge: A Cross-Cultural Study of Security Behavior. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, Denver, Colorado, USA, May 2017.

[85] Sean M. Segreti, William Melicher, Saranga Komanduri, Darya Melicher, Richard Shay, Blase Ur, Lujo Bauer, Nicolas Christin, Lorrie Faith Cranor, and Michelle L. Mazurek. Diversify to survive: Making passwords stronger with adaptive policies. In *Thirteenth Symposium on Usable Privacy and Security (SOUPS 2017)*, pages 1–12, Santa Clara, CA, July 2017. USENIX Association.

[86] Mahmood Sharif, Jumpei Urakawa, Nicolas Christin, Ayumu Kubota, and Akira Yamada. Predicting Impending Exposure to Malicious Content from User Behavior. In *Proceedings of the 2018 ACM SIGSAC Conference on Computer and Communications Security (CCS)*, Toronto, Canada, October 2018.

[87] Richard Shay, Lujo Bauer, Nicolas Christin, Lorrie Faith Cranor, Alain Forget, Saranga Komanduri, Michelle L. Mazurek, William Melicher, Sean M. Segreti, and Blase Ur. A spoonful of sugar? the impact of guidance and feedback on password-creation behavior. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing*

Systems, CHI '15, page 29032912, New York, NY, USA, 2015. Association for Computing Machinery.

[88] Richard Shay, Abhilasha Bhargav-Spantzel, and Elisa Bertino. Password policy simulation and analysis. In *DIM '07*, 2007.

[89] Richard Shay, Patrick Gage Kelley, Saranga Komanduri, Michelle L. Mazurek, Blase Ur, Timothy Vidas, Lujo Bauer, Nicolas Christin, and Lorrie Faith Cranor. Correct horse battery staple: Exploring the usability of system-assigned passphrases. In *Proceedings of the Eighth Symposium on Usable Privacy and Security*, SOUPS '12, New York, NY, USA, 2012. Association for Computing Machinery.

[90] Richard Shay, Saranga Komanduri, Adam L. Durity, Phillip (Seyoung) Huh, Michelle L. Mazurek, Sean M. Segreti, Blase Ur, Lujo Bauer, Nicolas Christin, and Lorrie Faith Cranor. Can long passwords be secure and usable? In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '14, page 29272936, New York, NY, USA, 2014. Association for Computing Machinery.

[91] Richard Shay, Saranga Komanduri, Patrick Gage Kelley, Pedro Giovanni Leon, Michelle L. Mazurek, Lujo Bauer, Nicolas Christin, and Lorrie Faith Cranor. Encountering stronger password requirements: User attitudes and behaviors. In *Proceedings of the Sixth Symposium on Usable Privacy and Security*, SOUPS '10, New York, NY, USA, 2010. Association for Computing Machinery.

[92] Steve Sheng, Mandy Holbrook, Ponnurangam Kumaraguru, Lorrie Faith Cranor, and Julie Downs. Who Falls for Phish?: A Demographic Analysis of Phishing Susceptibility and Effectiveness of Interventions. In *Proceedings of the 2019 Conference on Human Factors in Computing Systems (CHI)*, pages 373–382, Atlanta, GA, USA, April 2010.

[93] Steve Sheng, Bryant Magnien, Ponnurangam Kumaraguru, Alessandro Acquisti, Lorrie Faith Cranor, Jason Hong, and Elizabeth Nunge. Anti-Phishing Phil: The Design and Evaluation of a Game That Teaches People Not to Fall for Phish. In *Proceedings of the 3rd Symposium on Usable Privacy and Security (SOUPS)*, pages 88–99, July 2007.

[94] Jonathan Skelker. Announcing some security treats to protect you from attackers' tricks. https://security.googleblog.com/2018/10/announcing-some-security-treats- to.html, October 2018.

[95] Ben Stock, Giancarlo Pellegrino, Frank Li, Michael Backes, and Christian Rossow. Didn't You Hear Me? — Towards More Successful Web Vulnerability Notifications. In *Proceedings of the Network and Distributed Systems Security (NDSS) Symposium*, January 2018.

[96] Suricata. Suricata IDS Website. https://suricata-ids.org/, 2019.

[97] Samaneh Tajalizadehkhoob, Tom Van Goethem, Maciej Korczyński, Arman Noroozian, Rainer Böhme, Tyler Moore, Wouter Joosen, and Michel van Eeten. Herding Vulnerable

Cats: A Statistical Approach to Disentangle Joint Responsibility for Web Security in Shared Hosting. In *Proceedings of the ACM Symposium on Information, Computer and Communications Security (CCS)*, Dallas, TX, USA, November 2017.

[98] Wireshark The Wireshark Team. Wireshark Website. https://www.wireshark.org/, 2019.

[99] Kurt Thomas, Danny Yuxing Huang, David Wang, Elie Bursztein, Chris Grier, Tom Holt, Christopher Kruegel, Damon McCoy, Stefan Savage, and Giovanni Vigna. Framing Dependencies Introduced by Underground Commoditization. In *Proceedings of the 2015 Workshop on the Economics of Information Security (WEIS)*, Delft, The Netherlands, June 2015.

[100] Kurt Thomas, Dmytro Iatskiv, Elie Bursztein, Tadek Pietraszek, Chris Grier, and Damon McCoy. Dialing Back Abuse on Phone Verified Accounts. In *Proceedings of the 2014 ACM Conference on Computer and Communications Security (CCS)*, pages 465–476, Scottsdale, AZ, USA, November 2014.

[101] Kurt Thomas, Frank Li, Ali Zand, Jacob Barrett, Juri Ranieri, Luca Invernizzi, Yarik Markov, Oxana Comanescu, Vijay Eranti, Angelika Moscicki, Daniel Margolis, Vern Paxson, and Elie Bursztein. Data Breaches, Phishing, or Malware?: Understanding the Risks of Stolen Credentials. In *Proceedings of the 2017 ACM Conference on Computer and Communications Security (CCS)*, Dallas, TX, USA, October 2017.

[102] Kurt Thomas, Damon McCoy, Chris Grier, Alek Kolcz, and Vern Paxson. Trafficking Fraudulent Accounts: The Role of the Underground Market in Twitter Spam and Abuse. In *Proceedings of the 22nd USENIX Security Symposium*, Washington, DC, USA, August 2013.

[103] Update Google Chrome. Update Google Chrome. https://support.google.com/chrome/answer/95414?co=GENIE.Platform%3DDesktop&hl=en, 2019.

[104] Blase Ur, Jonathan Bees, Sean M. Segreti, Lujo Bauer, Nicolas Christin, and Lorrie Faith Cranor. Do users' perceptions of password security match reality? In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, CHI '16, page 37483760, New York, NY, USA, 2016. Association for Computing Machinery.

[105] Blase Ur, Patrick Gage Kelley, Saranga Komanduri, Joel Lee, Michael Maass, Michelle L. Mazurek, Timothy Passaro, Richard Shay, Timothy Vidas, Lujo Bauer, Nicolas Christin, and Lorrie Faith Cranor. How Does Your Password Measure up? The Effect of Strength Meters on Password Creation. In *Proceedings of the USENIX Security Symposium*, Security'12, 2012.

[106] Blase Ur, Fumiko Noma, Jonathan Bees, Sean M. Segreti, Richard Shay, Lujo Bauer, Nicolas Christin, and Lorrie Faith Cranor. "i added '!' at the end to make it secure": Observing password creation in the lab. In *Eleventh Symposium On Usable Privacy and Security (SOUPS 2015)*, pages 123–140, Ottawa, July 2015. USENIX Association.

[107] Blase Ur, Sean M. Segreti, Lujo Bauer, Nicolas Christin, Lorrie Faith Cranor, Saranga Komanduri, Darya Kurilova, Michelle L. Mazurek, William Melicher, and Richard Shay. Measuring Real-World accuracies and biases in modeling password guessability. In *24th USENIX Security Symposium (USENIX Security 15)*, pages 463–481, Washington, D.C., August 2015. USENIX Association.

[108] Tom van Goethem, Ping Chen, Nick Nikiforakis, Lieven Desmet, and Wouter Joosen. Large-Scale Security Analysis of the Web: Challenges and Findings. In *Proceedings of the International Conference on Trust and Trustworth Computing*, Heraklion, Crete, Greece, July 2014.

[109] Verizon. 2018 Data Beach Investigations Report. https://www.verizonenterprise.com/resources/reports/rp_DBIR_2018_Report_en_xg.pdf. Accessed: 2018-10-22.

[110] Virus Total. https://www.virustotal.com/#/home/upload. Accessed: 2018-10-22.

[111] Francesco Vitale, Joanna McGrenere, Aurélien Tabard, Michel Beaudouin-Lafon, and Wendy E. Mackay. High Costs and Small Benefits: A Field Study of How Users Experience Operating System Upgrades. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, Denver, Colorado, USA, May 2017.

[112] Rick Wash. Folk Models of Home Computer Security. In *Proceedings of the Sixth Symposium on Usable Privacy and Security*, Redmond, Washington, USA, July 2010.

[113] Rick Wash and Emilee Rader. Too Much Knowledge? Security Beliefs and Protective Behaviors Among United States Internet Users. In *Proceedings of the Eleventh USENIX Conference on Usable Privacy and Security*, Ottawa, Canada, July 2015.

[114] Matt Weir, Sudhir Aggarwal, Michael Collins, and Henry Stern. Testing Metrics for Password Creation Policies by Attacking Large Sets of Revealed Passwords. In *Proceedings of the ACM Conference on Computer and Communications Security*, page 162175, 2010.

[115] Matt Weir, Sudhir Aggarwal, Breno de Medeiros, and Bill Glodek. Password Cracking Using Probabilistic Context-Free Grammars. In *Proceedings of the IEEE Symposium on Security and Privacy*, pages 391–405, 2009.

[116] Chaowei Xiao, Armin Sarabi, Yang Liu, Bo Li, Mingyan Liu, and Tudor Dumitras. From Patching Delays to Infection Symptoms: Using Risk Profiles for an Early Discovery of Vulnerabilities Exploited in the Wild. In *Procedings of the 27th USENIX Security Symposium (USENIX Security)*, Baltimore, MD, USA, August 2018.

[117] Zeek. Zeek Protocol Analyzers Website. https://docs.zeek.org/en/stable/script-reference/proto-analyzers.html, 2019.

[118] Yinqian Zhang, Fabian Monrose, and Michael K. Reiter. The Security of Modern Password Expiration: An Algorithmic Framework and Empirical Analysis. In *Proceedings of the ACM Conference on Computer and Communications Security*, pages 176–186, 2010.

[119] Yue Zhang, Jason I. Hong, and Lorrie F. Cranor. CANTINA: A Content-based Approach to Detecting Phishing Web Sites. In *Proceedings of the 16th International Conference on World Wide Web (WWW)*, pages 639–648, May 2007.

[120] Leah Zhang-Kennedy, Sonia Chiasson, and Paul C. van Oorschot. Revisiting Password Rules: Facilitating Human Management of Passwords. In *Proceedings of the APWG Symposium on Electronic Crime Research (eCrime)*, pages 81–90, 2016.