A Large-Scale Analysis of Attacker Activity in Compromised Enterprise Accounts

Neil Shah
Grant Ho
Marco Schweighauser
M.H. Afifi
Asaf Cidon
David A. Wagner

Electrical Engineering and Computer Sciences
University of California at Berkeley

Technical Report No. UCB/EECS-2020-80
http://www2.eecs.berkeley.edu/Pubs/TechRpts/2020/EECS-2020-80.html

May 28, 2020

Copyright © 2020, by the author(s).
All rights reserved.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission.

A Large-Scale Analysis of Attacker Activity in Compromised Enterprise Accounts

By

Neil Shah

A thesis submitted in partial satisfaction of the

requirements for the degree of

Master of Science

in

Electrical Engineering and Computer Science

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor David Wagner, Chair

Spring 2020

A Large-Scale Analysis of Attacker Activity in Compromised Enterprise Accounts

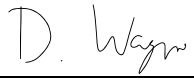
by Neil Shah

Research Project

Submitted to the Department of Electrical Engineering and Computer Sciences, University of California at Berkeley, in partial satisfaction of the requirements for the degree of **Master of Science, Plan II**.

Approval for the Report and Comprehensive Examination:

Committee:



Professor David Wagner
Research Advisor

May 24, 2020

(Date)

* * * * * * *

Professor Raluca Ada Popa
Second Reader

May 15, 2020

(Date)

Acknowledgements

I would like to thank Professor David Wagner for his continued support in my pursuit of this Masters Degree and constant encouragement of my research endeavors. I would also like to thank Grant Ho for his mentorship and for his countless hours of availability in pushing me to explore all avenues of analysis and think outside the box. Thanks to Asaf from Columbia University and Marco and Afifi from Barracuda Networks for their constant collaboration throughout this project, as well as Professor Raluca Ada Popa for her valuable feedback as a reader for my thesis. Last but not least, I want to thank my family and friends for their everlasting words of encouragement and support throughout my masters journey - none of this could have been done without their love and guidance.

To my wonderful, loving family who have always been there and supported me no matter what.

A Large-Scale Analysis of Attacker Activity in Compromised Enterprise Accounts

Neil Shah
UC Berkeley, Barracuda Networks

Grant Ho
UC Berkeley, Barracuda Networks

Marco Schweighauser
Barracuda Networks

M. H. Afifi
Barracuda Networks

Asaf Cidon
Columbia University

David Wagner
UC Berkeley

Abstract

We present the first large-scale characterization of attacker activity in compromised enterprise accounts based on our dataset of 989 enterprise accounts spanning 120 real-world enterprise organizations. Given the wealth of confidential and sensitive information that enterprises have access to, malicious access to enterprise accounts can incur major damage. We develop a novel forensic technique for distinguishing between attacker activity and benign activity in compromised enterprise accounts that yields few false positives and enables us to perform fine-grained analysis. Applying our forensic methods to these accounts, we quantify the length of time attackers spend in enterprise accounts, surface clues about the economy of enterprise accounts, explore a potential attack vector of compromise, and identify what these accounts are used for by attackers. We find that attackers dwell a long time in accounts and there appears to be a specialized market for these accounts in which one set of attackers compromise the accounts and another set of attackers utilize the accounts, possibly for extracting monetary value. Taken together, our findings illuminate differences in how attackers exploit enterprise accounts compared to personal accounts and inform organizations of new defense strategies that can address the state of threats today.

1 Introduction

With the advent of cloud computing, many organizations have changed the way they store and operate on their data. Rather than storing their documents and hosting applications on internally managed, on-premise servers, companies frequently choose from a plethora of third-party cloud applications to perform these same business functions. For example, services such as Microsoft Office 365 and Google Drive have shifted traditionally local document, presentation, and spreadsheet processing to cloud-backed web applications; likewise, commonplace business operations ranging from sales negotiations to time sheet tracking to customer support interactions also have a number of third party web-applications that organizations can choose from. This transition to a cloud-centric world has created a growing reliance on the security of enterprise cloud accounts and their credentials. An attacker who compromises a legitimate employee account now has access to a wealth of internal business information and functionality, in addition to the employee's sensitive enterprise emails.

As a result of the growing value of cloud accounts and the difficulty of compromising an enterprise's internal network, attackers have increasingly shifted to compromising enterprise cloud accounts through attacks such as phishing. For example, several government agencies have issued advisories and reports warning that phishing represents "the most devastating attacks by the most sophisticated attackers" and detailing the billions of dollars in financial harmed caused by enterprise phishing and account compromise [18, 31]. Not limited to financial gain, attackers have also compromised enterprise cloud accounts for personal and political motives, such as in the 2016 US presidential election, when nation-state adversaries dumped a host of internal emails from high-profile figures involved with Hillary Clinton's presidential campaign and the Democratic National Committee [37].

Given the growing importance of online accounts and credentials, a large body of existing work has focused on building mechanisms to defend against attacks through better credential hygiene, detecting phishing attacks, and stronger user authentication [14, 15, 20, 22, 23, 33, 36]. Despite these advances, account hijacking (compromise and malicious use of cloud accounts) remains a widespread and costly problem [8]. Moreover, this existing body of work often focuses on cloud accounts through the specific lens of preventing an account compromise by detecting credential phishing; in particular, most detection work crafts features specific to emails, with the goal of preventing an account from having its credentials phished. However, no detection method is perfect, and there can be many sources of compromise beyond just phishing.

There remain a number of valuable technical questions related to the activity of the attacker *after* they have successfully compromised an account. For example, how can an organiza-

tion comprehensively identify all of the data that an attacker accessed throughout the lifetime of the account compromise? And how much value does post-initial-compromise detection provide (e.g., by developing compromise detection signals based off account and application access behaviors by a user)? Although a number of prior works have characterized what attackers do with an account post-compromise [13,30,35], existing work focuses heavily on compromised personal email accounts. While these insights are useful, it remains unclear how well they generalize to compromised enterprise accounts and whether attacks on enterprise accounts have different characteristics. Unlike personal accounts, enterprise accounts often have access to a wealth of additional sensitive data and functionality through their organization's suite of cloud-based work applications. In addition, an attacker who compromises one enterprise account can use the identities of the compromised account to launch additional attacks on other users, potentially gaining access to other accounts within that enterprise.

Thus, a number of important questions remain unanswered. For example, what is the end-to-end lifecycle of a compromised account: how are these accounts compromised, how long do attackers retain access to a hijacked account, and how useful is detection and mitigation after an account has been initially compromised? What is the economy of enterprise accounts and what modes of attackers compromise these accounts? In particular, what kinds of cloud data and functionality do attackers exploit, and does this vary based on the organization where compromise occurred?

In this joint work between academia and Barracuda Networks, we attempt to close the loop on these unanswered questions and conduct a large-scale analysis of attacker activity within compromised enterprise accounts. First, we present a technique for distinguishing between attacker and benign activity in compromised enterprise accounts, enabling more comprehensive incident forensics. Evaluating our approach on a random sample of enterprise accounts, we find that our forensic technique yields a false positive rate of 11% and a precision of 94%.

Second, we conduct a large-scale analysis of attacker activity in real-world enterprises. We find that although several types of attackers appear to compromise enterprise accounts, in over one-third of the hijacked accounts in our dataset, the attacker dwells in the account for more than one week. This suggests that delayed non-real-time detection can still be beneficial by mitigating an attack even after a compromise has occurred. Furthermore, our analysis suggests the majority of these long-duration compromises reflect an increasingly specialized market of account compromise, where one set of attackers focuses on compromising enterprise accounts and subsequently sells account access to another set of attackers who focus on utilizing the hijacked account. This existence of a specialized underground economy also suggests that the second set of attackers likely use the compromised accounts for monetary purposes. Correlating the compromised accounts in our dataset against a commercial data breach service, we find that 20% of our dataset's compromised accounts appear in at least one online password data breach, which suggests that one potential vector for compromise involves exploiting credential reuse across an employee's personal and enterprise accounts.

Finally, examining the kinds of data and applications that attackers access via these enterprise accounts, we find that most attackers in our dataset do not access many applications outside of email, which suggests that either many enterprise cloud accounts do not have access to interesting data and functionality outside of email, or that attackers have yet to adapt to and exploit these additional sources of information.

Overall, this work yields two main contributions that expand our knowledge and ability to remediate compromised enterprise accounts. First, we present a forensics method for distinguishing between attacker and benign activity in compromised enterprise accounts. Second, we illuminate a number of surprising characteristics of enterprise attacks. We show that in many cases, account compromise remains focused on "traditional" malicious email activity, such as sending phishing or spam from compromised accounts. We also demonstrate signs of a growing sophistication with attackers beginning to specialize and differentiate between those focused on the initial compromise and those focused on utilizing compromised accounts, likely for extracting monetary value. The combination of these two contributions improves our understanding of the nature of enterprise account compromise and informs future directions for remediation and defenses.

2 Background

From an attacker's point of view, it is lucrative to gain access to enterprise email accounts for several reasons. The accounts themselves may have access to sensitive and valuable information through email and cloud-based applications. In addition, the compromised account may be used to impersonate the identity of the employee and their organization in order to gain access to other employee accounts or other related organizations.

The space of enterprise email account security has been relatively understudied, yet is extremely important given the value of the data that can be accessed through these accounts. In general, prior work has been primarily focused on characterizing attacker behavior in compromised personal accounts and has involved relatively small-scale observation of honeypot accounts. Also, much prior work has focused on phishing detection, but doesn't look beyond that to post-compromise analysis. Our work studies and characterizes attacker activity post-compromise at scale in real world enterprises, expanding our understanding of the threats enterprise organizations face and the types of defense strategies needed in practice.

2.1 Related Work

Forensics. There has been an extensive amount of literature proposing various techniques from machine learning and anomaly detection for detecting phishing attacks in personal and enterprise accounts on a smaller scale [9,11,19,24] and on a large scale [14,22,23,33]. In addition, a limited amount of prior work exists on detecting compromised accounts [16,26] through the use of honeypot accounts and personal accounts on social networks. These works focus on building detectors for phishing emails or detecting compromised accounts in real time, but don't study enterprise accounts post compromise.

Liu et al. in [26] monitored the dangers of private file leakage in P2P file-sharing networks through the use of honeyfiles containing forged private information. Their work focused more on the use of honeyfiles instead of account credentials and doesn't study compromised accounts outside of P2P.

Egele et al. in [16] developed a system, called COMPA, for detecting compromised personal accounts in social networks. COMPA constructs behavior profiles for each account and evaluates new messages posted by these social networking accounts by comparing features such as time of day and message text to the behavioral profiles. They measured a false positive rate of 4% on a large-scale dataset from Twitter and Facebook. However, their work only studies how to detect compromised personal accounts and doesn't include enterprise accounts.

Overall, none of the works in the literature have performed forensic analysis to understand attacker activity in personal or enterprise accounts post-compromise. We therefore are the first to experiment with forensic methods on enterprise accounts and present a novel technique for distinguishing between attacker and benign activity in these accounts.

Characterization. Although there has been prior work on understanding attacker behavior and patterns within compromised accounts, most of this research has been primarily focused on attacker characteristics within personal accounts on a small scale; few efforts have been examined the nature of attackers in compromised enterprise accounts at large scale. TimeKeeper, proposed by Fairbanks et al. [17], explored the characteristics of the file system in honeypot accounts controlled by attackers. Although their work applied forensic techniques to honeypot accounts post-compromise, they operated at small scale and only characterized attacker behavior in relation to file systems on these accounts. Onaolapo et al. [30] studied attacker behavior in hijacked Gmail accounts post-compromise, but their work did not examine compromised enterprise accounts. In addition, they were only able to monitor certain actions, such as opening an email or creating a draft of an email. Bursztein et al. [13] examine targeted account compromise, but their work focuses on compromised personal accounts and not on enterprise accounts at scale.

Open Questions. Prior work illuminates the fact that the space of understanding enterprise accounts post-compromise is understudied; yet, compromises to these accounts pose large threats to enterprise organizations. There are many key questions that remain unanswered in the space of compromised enterprise accounts: how can we distinguish between attacker activity and benign activity for forensic purposes? What can we infer about the economy of compromised enterprise accounts and its relationship to detection and mitigation post-compromise? What kinds of techniques do attackers employ to compromise accounts? What types of data and functionality are attackers interested in exploiting? Our work investigates these questions at a large-scale.

2.2 Ethics and Justification

This work consisted of a team of researchers from UC Berkeley and Columbia University, as well as from a large security company, Barracuda Networks. The set of organizations included in our dataset are customers of Barracuda Networks and are anonymized and stored encrypted by Barracuda.

In addition, due to the confidential nature of account data, no sensitive data was released to anyone outside of Barracuda. This project also received approval from Barracuda, and strong security controls were implemented to ensure confidentiality and limited scope of access.

3 Data

In order to be able to characterize attacker activity in compromised enterprise accounts, we must first obtain a set of enterprise accounts that have been confirmed to be compromised. In this section, we give an overview of various data sources that were used to address the key questions as laid out above in Section 2. For the remainder of the paper, we will use the terms *compromised enterprise account*, *compromised user*, and *account* interchangeably.

3.1 Schema and Data Sources

Each organization in our dataset is a registered customer of Barracuda Networks through Barracuda Sentinel. Barracuda Sentinel is a fraud detection platform that protects enterprises and their employees from phishing attacks and account takeover (ATOs) [2,14]. All organizations in our dataset utilize Microsoft Office 365 as their email account provider. In collaboration with Barracuda Networks, we have access to various data sources for each compromised user in our dataset. For each user, we have anonymized access to their respective Microsoft Office 365 audit events. These audit events are metadata about a user's access to their respective Office 365 account. Audit events are generated by Office 365 for research purposes and allow organizations to maintain a record of how

their employees are utilizing their accounts. At a high level, an Office 365 audit event includes the following fields:

- `Id` - Unique identifier for an audit event
- `UserId` - Email of user who performed the operation
- `UserAgent` - User agent string of device that performed the operation
- `ClientIp` - IP address of device that performed the operation
- `Operation` - Operation performed by the user
- `ApplicationId` - Id of Office 365 application that was acted upon

The `Operation` field indicates which operation was performed by the user; some examples of operations that are logged include UserLoggedIn, UserLoginFailed, and ResetUserPassword. The `UserAgent` and `ClientIp` fields are populated with non-empty values *only* when the user successfully logs in, i.e., when `Operation` = UserLoggedIn and `LogonError` = None. The `LogonError` field identifies any errors resulting from a user logging into their account. For more details, the Office 365 schema for audit events can be found in [3, 6]. In addition, we use MaxMind [28] to map each IP address of a successful login audit event to a country and a country subdivision. For simplicity, we will refer to a successful login audit event as a *login audit event* or *login event* throughout the remainder of the paper.

Other data sources that are available to us include raw emails sent by users in our dataset, as well as any audit events, emails, and email forwarding rules flagged by Barracuda's machine learning detectors for the compromised users. Each email contains many fields including a unique identifier, the sender of the email, the recipient of the email, the time the email was sent, and an extended properties field which in many cases contains the IP address of the device from which the email was sent. More details on the Office 365 schema for emails can be found in [4].

3.2 Size of Data Sources

Our dataset consists of 989 compromised users between October 1, 2019 – December 27, 2019 across 120 total organizations. 887 of the compromised users were flagged at least once by Barracuda's detector that examines login audit events, while 101 users were flagged at least once by the detector that examines emails sent from an account. 106 users were flagged at least once by the detector that examines changes to the mail forwarding rules associated with an email account. We note that the total number of compromised users flagged by each detector adds up to more than 989 because some account compromises were detected by multiple detectors.

Barracuda Networks then verified with the IT teams at the users' respective organizations that these users' accounts were in fact compromised. To protect the effectiveness of Barracuda's detectors, we are not able to disclose further details on how these detectors work for the remainder of the paper.

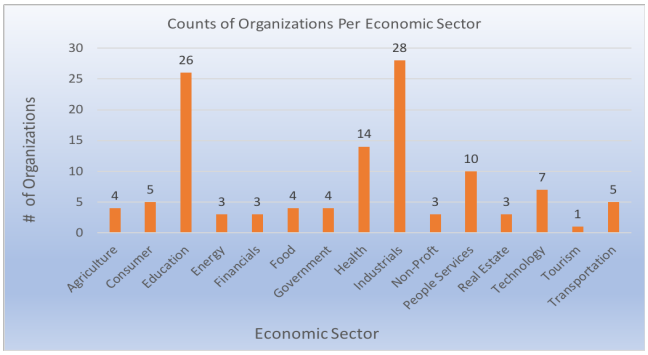


Figure 1: Categorization of the 120 organizations in our dataset across various economic sectors.

Figure 1 shows the distribution of the 120 organizations by economic sector. A majority of these organizations belong to the industrials, education, health, and people services economic sectors, with counts of 28, 26, 14, and 10 respectively. These four sectors represent 65% of the set of organizations in our dataset.

For each of the 989 compromised users, we also collect all audit events over the time period of August 1, 2019 – January 27, 2020. We choose a larger date range for collecting audit events as part of our forensic technique for identifying attacker behavior, which will be discussed in Section 4. For all 989 users, there were a total of 927,822 audit events, 565,165 of which were login events to certain Office 365 applications. Most of our analysis relies primarily on login events because Office 365 records the IP address and user agent string for each of these events.

4 Detecting Attacker Activity

Given a known compromised account, one of our goals is to distinguish between attacker activity and benign activity for forensic purposes. In this section, we develop a set of rules for identifying which of a user's logins are from an attacker vs from the legitimate user. Our rule set is not guaranteed to find every attack, nor does it guarantee robustness against motivated attackers trying to evade detection. However our rule set is still relatively comprehensive and generates few false positives, as shown in Section 5.

4.1 Overview

We use well-known ideas from the field of anomaly-based detection. Anomaly-based detection [10, 32] is the identifi-

cation of a group of data points in a larger set of data that doesn't follow the general pattern of the majority of values in the dataset. Anomaly-based detection is useful in many fields and contexts, from intrusion detection to credit-card fraud detection. At a high level, we aim to design a rule set that will identify the events where the attacker accesses the compromised account. Throughout this section, when describing the components of our rule set, we use the name Bob to refer to a generic compromised user from our dataset. Our rule set first builds a historical profile for Bob that represents his typical IP addresses, location, and user agent strings that he uses to log into his account. Then, we use this profile to classify future login events as either attacker-related or benign. The following subsections describe the components of our rule set in detail.

4.2 Historical User Profile and Features

In order to be able to identify login events as suspicious or not, we need an indication of true user activity in an account. Our idea involves creating a historical user profile using historical login events from a user's account that occurred significantly before the account was first confirmed as compromised. We assume that the historical events are benign and reflect the true user's characteristics. We then extract several features from each login event.

Historical User Profile. The historical user profile for a user Bob characterizes how Bob typically logs into his account. The idea of a behavioral profile has been proposed in many earlier works. Egele et al. [16] create a behavioral profile for each social network user based on a stream of messages the user posts and use it to detect if a user's social network account has been compromised based on examining messages subsequently posted by the user's account. Stringhini et al. [33] create a behavioral profile based on the emails sent by a user in order to determine if an attacker has compromised the user's account and is sending phish or spam.

As mentioned in Section 3, each user's account in our dataset was flagged by at least one of Barracuda's detectors and their respective organization's IT team confirmed the user's account to have been compromised. For each compromised user Bob, we find the earliest time that he was flagged by one of Barracuda's detectors between October 1, 2019 – December 27, 2019; call this time t. Bob could have been compromised earlier than t, but we assume that Bob has definitely been compromised by time t. To create Bob's historical user profile, we first retrieve his historical login events from the time window of 2 months prior to t until 1 month prior to t (a one-month time window). Bob's historical user profile consists of 3 sets of values: the set of country subdivisions that he has been observed to log in from during that time period, the set of countries he has logged in from, and the set of user agents that he has logged in with. For some context, a country subdivision [38] is a portion of a country delimited by a government body. For US-based IP addresses, MaxMind treats each state as a country subdivision, while for IP addresses residing outside the US, a country subdivision corresponds to a province.

Features. Our rule set extracts 2 features for each login event based on the information stored in a user's historical user profile. First, for each login event e we wish to classify, we extract a geolocation feature by geolocating the IP address used to log into the account and comparing this location to the historical profile of the account (Bob) that is being logged into:

(a) If e represents a login from a country that was never seen in Bob's historical user profile, then assign e's geolocation feature value a **2** (most suspicious).

(b) Otherwise, if e represent a login from a country subdivision not found in in Bob's historical user profile, then assign e's geolocation feature value a **1** (medium suspicion).

(c) Otherwise, assign e's geolocation feature value a **0** (least suspicious).

We also extract a user agent feature from e that captures the suspiciousness of the user agent of e. All user agents are normalized in a pre-processing step: the version number is removed and only the device and model identifiers are retained, so a user agent string such as `iPhone9C4/1706.56` is normalized to `iPhone9C4`. Thus, `iPhone9C4/1706.56` and `iPhone9C4/1708.57` yield the same normalized user agent. The user agent feature is then defined as follows:

(a) If e's normalized user agent does not match any of the normalized user agents in Bob's historical user profile, then assign e's user agent feature value a **1** (most suspicious).

(b) Otherwise, assign e's user agent feature value a **0** (least suspicious).

For each user Bob, we classify all login events from one month prior to t to t using a historical user profile based on events from two months prior to t to one month prior to t. Then, we classify all events from t to one month after t using a historical user profile based on events from two months prior to t to one month prior to t, and all events from one month prior to t to t that were classified as benign. Thus we update the historical user profile for each user after classifying the first month of login events [7]. Malekian et al. also describes a similar approach [27] where the historical profile is updated to reflect new patterns in user behaviors in e-commerce for the purposes of detecting online user profile-based fraud.

Therefore, the last month of Bob's events are analyzed using an updated historical user profile that incorporates benign activity from his previous month of login events.

4.3 Applying the Rule Set

In this section, we combine the historical user profile and the two features discussed above in Section 4.2 using a set of rules for distinguishing between attacker and benign activity.

Rules. To classify a new login event e for Bob, we first extract its geolocation feature, denoted as **geo**, and its user agent feature, denoted as **ua**. We then apply a set of rules to the values of **geo** and **ua** to obtain a classification of the event e. The pseudocode below outlines the rule set.

```
if geo == 2
    mark e as attacker-related
else if (geo == 1) and (ua == 1)
    mark e as attacker-related
else
    mark e as benign
```

Justification and Assumptions. The geolocation and user agent features quantify the suspiciousness of a new login event in relation to a user's historical profile. We assume that the historical login events for each user do not contain attacker activity. Since we don't know precisely when each user was compromised in our dataset, we conservatively analyze events at least one month prior to when the user was first flagged as compromised by Barracuda's detectors in an effort to limit the amount of poisoning of the historical login events by potential attacker events.

We also assume that it is less common for users to travel to another country than to another state or province within their home country. Although traveling abroad is common in some industries, we assume that most employees travel more frequently to another country subdivision than to a different country. As a result, if a login event contains an IP address mapped to a country that was never seen before in a user's historical login events, the event in question is marked as an attacker event. For travel within the same country, the country subdivision and user agent need to be new for a login event to be marked as an attacker event.

Applying Rule Set to Compromised Users. We apply the same procedure that was done for Bob to each of the 989 compromised users in our dataset. For each user, a historical user profile is created based on their historical login events, features are extracted for login events within a two-month time window of when the user first appeared in at least one of Barracuda's detectors, and the rule set is applied to each user's login events, generating a set of attacker events and benign events.

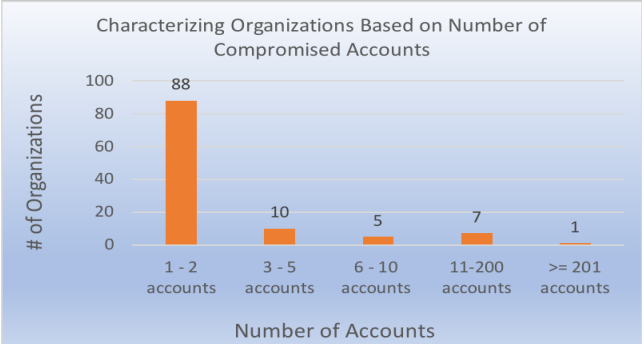


Figure 2: Categorization of the 111 organizations in our dataset based on number of compromised user accounts.

276 of the 989 compromised users didn't have any historical login events due to the fact that these users' enterprises registered with Barracuda as a customer after the start of our study period, and we did not have logs from before then. As a result, our rule set couldn't be applied to these users. Of the remaining 713 users that had historical login events, 653 had at least one attacker event that our rule set classified. These 653 users were found in 111 different organizations. Across the 653 compromised users that had attacker events, our attacker rule set labelled 17,842 audit events as attacker-related. Figure 2 shows the distribution of compromised accounts among organizations. 98 of the 111 organizations (89%) had 1–5 compromised users, 12 organizations had 6–200 compromised users, and 1 organization had over 200 compromised users, precisely 206. Moreover, 68% of the 653 compromised accounts belong to 6 organizations. We do not know what accounts for this skewed distribution, but it is possible that one or a few attackers specifically targeted those 6 organizations. Therefore, to ensure that we obtain many different attackers and our results are not biased by a few attackers that compromise many accounts, we grouped each of the 653 compromised users by organization and month of earliest attack event flagged by our rule set and randomly selected one compromised user from each group. This resulted in a final sample of 159 compromised users across the 111 organizations, which is the dataset we use to evaluate our rule set in Section 5 and obtain our findings in Section 6.

5 Evaluation of Rule Set

In this section, we evaluate our attacker rule set. We first describe how we sampled a set of users and established ground truth labels for login events. We then show how well our rule set performs on the sample of users. Overall, our rule set generates few false positives and offers a promising direction for distinguishing between attacker activity and benign activity at the granularity of login events.

5.1 Evaluation Methodology

As described at the end of Section 4.3, we obtained a set of 159 compromised enterprise accounts that each have attacker events classified by the rule set. To understand how effective our rules are, we randomly sample 20 users and manually evaluate the accuracy of the rule set on these users. For each of the 20 sampled users, we also randomly sample up to 2 sessions labelled by our rule set as attacker-related and 1 session labelled as benign, where we define a *session* to consist of all events with the same (IP address, user agent) pair value; all events within the same session are assigned the same label by our rule set. Across the sample of 20 users, we evaluate a total of 54 sessions, 34 of which are labelled as attacker-related sessions and 20 as benign.

5.2 Establishing Ground Truth

In order to evaluate whether the labels our rule set applies to sessions are correct, we must develop a way to reason about what the ground truth labels for sessions are. Just knowing that a user has been compromised does not give us much information on which particular login events are from the attacker. In this section, we describe four basic indicators that we apply to each of the 54 sessions to help us gain confidence on what the "true" labels are for the sessions when evaluating the rule set. We note that since the four basic indicators discussed in this section are not perfect in terms of determining the true label for sessions, we also perform a more extensive manual analysis for sessions in which the basic indicators label a session as benign and our rule set labels the session as attacker-related. Due to our conservative approach, we aim to limit the false positives of our rule set and thus obtain a more refined label through manual analysis when the basic indicators are not comprehensive enough. Throughout the remainder of this section, we refer to a compromised user out of our sample of 20 as Bob and one of Bob's sessions as s.

Phishing Email Indicator. For some session s, we retrieve all emails sent from Bob's account within ± 5 hours from a login event in s. The time window of ± 5 hours serves as a heuristic for relating the email to the login event it is close in time to, as users may not send email immediately after logging into their accounts. In addition, there are sometimes delays as to when Office 365 creates login events in the data buckets for retrieval by Barracuda.

Once all emails are retrieved that are close in time to s, we first determine if any of the emails were flagged as phishing by Barracuda; if so, then we assign the phishing email indicator for s a value of 1. If none of the emails were flagged, we then iterate over all emails and manually label them as phishing or not, using properties of the email header and body. Our method for manually labeling emails as phishing is similar to approaches taken in previous work [14, 22], in which we first analyze the subject of an email in the context of the sender's organization and the number and domains of recipients of the email. For example, an email of the form "Bob has shared a file with you" sent to many recipients across many types of domains is very likely to be phishing. For emails in which the subject is not suspicious and the number of recipients is small, we look through the bodies of the emails along with any links present to determine if the domains of the links are unrelated to the sender's organization. For many of the emails that we looked at, these steps were sufficient to determine if emails were related to phishing or not. We assign the phishing email indicator for s a value of 1 if there was at least one phishing email we labelled; else, we assign the phishing email indicator for s a value of 0 if there were no flagged emails by Barracuda and no manually labelled phishing emails.

Inbox Rules Detection Indicator. We retrieve all inbox rules detections that are ± 5 hours from a login event during Bob's session s. An inbox rule detection indicates that a suspicious rule was created in a user's account, such as emails being forwarded to the trash or to an external account. The inbox rules detection indicator is assigned a value of 1 for session s if at least one inbox rules detection exists close in time to s.

Interarrival Time Indicator. The *interarrival time* between 2 login events e_1 and e_2 is the absolute value of the difference in timestamps between e_1 and e_2. The general idea for including this interarrival time indicator is for detecting if the absolute time difference between login events for the same user with two different locations is shorter than the expected travel time between the two locations. We first obtain the country subdivision that is most common in Bob's historical user profile (i.e. the country subdivision that is associated with the most number of Bob's historical login events). For simplicity, we call this country subdivision Bob's home territory. Then, for each of Bob's login events e during session s, we compute the interarrival time between e and the closest login event in time to e that contains Bob's home territory. Among all events during Bob's session s, we take the smallest interarrival time and if that value is smaller than the expected travel time between Bob's home territory and the location mapped to session s, we mark the interarrival time indicator for s with a value of 1.

An example of a set of login events (anonymized for privacy) is shown in Table 1. If IL, US is Bob's home territory and we are evaluating one of Bob's sessions tied to 27, JP, the interarrival time for a login event tied to 27, JP (last login event in the table) would be about 46 minutes. However, the expected travel time between Illinois and Japan is about 13 hours. Therefore, the session tied to 27, JP would be suspicious and would be marked with a value of 1 by the

Country Subdivision	Timestamp
MO, US	2019-11-29 08:00:05
IL, US	2019-11-29 15:06:45
IL, US	2019-11-29 21:14:32
27, JP	2019-11-29 22:00:07

Table 1: Illustration of an example of a set of login events, with country subdivision in the first column and timestamp in the second column.

indicator. Note that in applying the indicator to some session *s*, we use Bob's home territory for computing interarrival times for login events during *s* to reduce the amount of manual analysis needed to be done in evaluating the rule set (we had to manually look up the expected travel time between all sessions and the respective home territories for our random sample of 20 users). To make this indicator more general, for each event during *s*, we could calculate the smallest interarrival time between the event and any country subdivision within Bob's historical user profile. However, using the home territory for each user was sufficient for the evaluation.

Tor Exit Node Indicator. If the IP address for *s* is a Tor exit node, then we assign the Tor exit node indicator for session *s* as a 1.

Applying the Basic Indicators and Refinement. For each of the 54 sessions across our random sample of 20 users that we evaluate our rule set on, we apply the four basic indicators described above. If at least one indicator labels a session with a value of **1**, then we say that it is an attacker-related session; otherwise, we label it a benign session.

Out of the 54 sessions, there were seven that were labeled as benign by the basic indicators and attacker by our rule set. To ensure that we obtained the highest confidence ground truth label for these seven sessions (possible false positives), we performed manual analysis to obtain a more refined label. Each of the seven sessions involved a different compromised user. From their respective historical user profiles, four users primarily use US-based IP addresses and the remaining three primarily use IP addresses based in countries outside the US. Through our analysis that we present below for each of the seven sessions, we find that five of these sessions should be labelled as attacker-related by ground truth. For simplicity in our analysis, we will refer to each of the 7 sessions as session *x*, where *x* is a number between 1–7.

For session 1, the IP address mapped to a country that had never been seen before in the historical user profile. In addition, there was only one login event from this new country and it occurs implausibly soon after a benign login event where it is impossible to travel between the two locations within the interarrival time. The interarrival time indicator didn't flag for session 1 because interarrival times were only calculated with respect to user's home territories and not any location seen in a historical user profile. As a result, this session is truly an attack. In both session 2 and session 3, the user agent string matches that of the second randomly sampled session in which our rule set correctly labelled as attacker (verified via ground truth). As a result, we declare session 2 and session 3 as attacker-related sessions.

The analysis for session 4 and session 5 is very similar. Both sessions map to new countries that have never been seen before in their respective users' login events from August 1, 2019 - January 27, 2020. In addition, both sessions involve user agents that are totally different from what their respective users use during their benign sessions (historical login events + benign sessions labelled by rule set). In session 4, there are a total of 6 login events over a time period of 3.5 weeks from the new country. Halfway through the time period, there is an interspersed login mapped to the user's home territory (most common country subdivision in historical user profile). Then, 2 weeks later, there is a final login from the new country. The only way that session 4 could be benign is if the user decided to travel back-and-forth between the new country and their home territory over the 3.5 week window; based on the fact that there was one interspersed login from the user's home territory, the user would need to make 3 total back-and-forth trips between the home territory and new country over the 3.5 week period, which seems unlikely. Also, since the country has never been seen before throughout any of the user's previous login events, we declare this an attacker-related session. Similarly, in session 5, at least 4 back-and-forth trips between the home territory and the new country would be required over a 2 week period. As a result, session 5 is attacker-related.

Sessions 6 and 7 are likely benign sessions, as captured by the basic indicators above. For session 6, login events during the session do not happen close in time to other attacker sessions that our rule set correctly classifies for the user. In addition, the user agent doesn't stand out as suspicious and is a standard Firefox user agent string similar to the form "Mozilla/5.0 (Windows NT 10.0; Win64; x64) Gecko/20100101 Firefox/70.0". Even though the country mapped to session 6 is a new country that has never been seen before in the historical user profile, the user's organization has offices in this country. As a result, we believe this is a benign session. For session 7, the associated country subdivision has never seen before in the user's historical profile, but the country appears in all of the user's historical login events. This session was flagged by our rule set because the user agent had never been seen before in the historical user profile, as the the device that the user typically uses was of an older model. An example is if the user's historical login events frequently contain the user agent string `iPhone9C4`

Metric	
Compromised Users	20
Sessions	54
False Positives (FP)	2
False Positive Rate	11%
Precision	94%
False Negatives (FN)	9
False Negative Rate	22%
Recall	78%

Table 2: Evaluation results of our rule set. 'False Positives (FP)' shows the number of sessions that the rule set labels as attack but ground truth labels as benign. 'False Negatives (FN)' shows the number of sessions that the rule set labels as benign but ground truth labels as attack.

and this session in question contains the user agent string `iPhone10C2`. However, during the two-month window of login events that we applied our rule set on, the typical user updated their device before we see any session 7 events, but this update wasn't reflected in the historical user profile until after running the rule set on session 7. We also continue to see the use of session 7's user agent after the two-month window in login events tied to the user's home territory. As a result, we label session 7 as benign.

Therefore, through our manual analysis, we obtain more refined labels for the seven sessions mentioned above and find that five of the sessions are attacker-related and two are benign.

5.3 Evaluation Results

Tables 2 and 3 summarize the performance metrics of our rule set and display the confusion matrix for the 54 sessions we evaluate the rule set on across the set of 20 randomly sampled users. As mentioned above in Section 5.1, our rule set labelled 34 of the sessions as attacker and 20 of the sessions as benign. Based on the ground truth analysis discussed above, each session also has a "true" label. Our rule set generates 2 false positive sessions (FP) and a false positive rate of 11%. *Precision* is defined as the number of sessions that our rule set correctly marks as attacker divided by the total number of sessions that our rule set marks as attacker (true attacker sessions plus false positives). The precision for our rule set is 94%. We base our evaluation numbers on ground truth labels that we assign to sessions and we acknowledge that these are not perfect. However, due to the extensive manual analysis we perform to obtain more refined labels after applying the four basic indicators as discussed in Section 5.2, our source of ground truth is relatively comprehensive.

Our rule set also generates 9 false negative sessions (FN) and a false negative rate of 22%. This seems to suggest that attackers show some level of sophistication in trying to evade

		True Label		
		Attacker	Benign	Total
Rule Set Label	Attacker	32	2	34
	Benign	9	11	20
	Total	41	13	54

Table 3: Confusion Matrix for 54 sessions across 20 users.

detection (i.e. accessing user accounts with locations that blend or match with the user's typical login locations).

6 Characterizing Attacker Behavior

In this section, we conduct an analysis of attacker behavior across our dataset of 159 compromised users belonging to a total of 111 organizations. Examining the duration of compromises in enterprise accounts, our analysis suggests that a substantial number of accounts (51%) are compromised for 1 or more days and 38% of accounts are compromised for 1 or more weeks. In addition, out of 11 enterprise accounts that had at least one email flagged by Barracuda's email detector, 7 had a 3-day gap between their first attacker event and first sent phishing email. These findings together suggest that the space of enterprise account security has much room for improvement and that even taking steps post compromise to mitigate the threat of attack can prevent a lot of damage to enterprise accounts.

At the same time, we find that the economy of compromised enterprise accounts is impacted by many modes of attackers operating in tandem. We uncover at least two modes of attackers that compromise enterprise accounts and offer the first-large scale analysis of the prevalence of these attacker modes. We estimate that 50% of enterprise accounts are compromised by one attacker who also utilizes the account themselves. However, we also uncover a specialized economy of enterprise accounts in which 31% of accounts are compromised by one attacker and utilized by a different attacker that mines for information. We also hypothesize that the second set of attackers use the accounts to extract monetary value from the accounts. As a result, a more mature and specialized economy around compromised accounts is emerging, where some attackers specialize in compromising accounts, while others specialize in extracting information and potentially monetary value from accounts.

In trying to understand how enterprise accounts are compromised, we find that 20% of the accounts were found in data breaches of online company databases, indicating a potential source of compromise for these accounts if password reuse was omnipresent among these accounts. We also find that attackers compromising enterprise accounts primarily use the accounts for accessing email-related Office 365 applications; 78% of the enterprise accounts only accessed email applications through their attacker events. As a result, either

many enterprise cloud accounts don't have access to much interesting data outside of email or attackers have yet to adapt to and exploit additional cloud based applications. In addition, since email seems to be a common attack vector within enterprise accounts, the nature of the compromised enterprise does not seem to influence the ways in which an attacker utilizes an enterprise account. As measured by our study, it is likely not beneficial for enterprises to develop features for defenses that measure accesses to many cloud-based applications, as these are rarely accessed by attackers. However, we acknowledge that things could change in the future with the emerging popularity of other applications that may store sensitive information or be used for communication within the Office 365 ecosystem, like Microsoft Teams or OneDrive.

Taken together, our findings expand upon the limited research in the space of compromised enterprise accounts and uncover similarities and differences between compromised enterprise and personal accounts.

6.1 Duration of Account Compromises and Damage Prevention

In this section, we estimate the duration of compromise for enterprise accounts. We also illustrate that detectors don't necessarily need to react in real time; in fact, those that operate over long periods of time can still prevent attackers from inflicting a lot of damage in enterprise accounts.

Duration of Account Compromises. Given our dataset of 159 compromised users and their respective login events, we can not definitively determine how long an attacker compromised the account for, but we can form a reasonable estimation of the duration of compromise using the attacker events our rule set classified over each user's two-month window surrounding the earliest time they were flagged by Barracuda's detectors. For each user, we computed the difference in seconds of the time of the earliest attacker login event and the time of the latest attacker login. The distribution of time differences is shown in Figure 3 across all 159 compromised enterprise accounts. From the figure, we notice that accounts tend to be compromised for long periods of time. Assuming our random sample of organizations reflects the larger population of organizations, almost 51% of all enterprise accounts (81 out of 159) are compromised for at least 1 day and 37% of all enterprise accounts are compromised for at least 1 week. As a result, while it's important to detect attacks in real-time, detectors of compromised accounts at organizations do not necessarily need to operate in this fashion to provide useful defense; namely, it is equally important to build detectors that operate over longer time periods to gather more detection signals for forensics and remediation of accounts post-compromise. This second point in particular will be illustrated below through an analysis of phishing in our set of enterprise accounts.

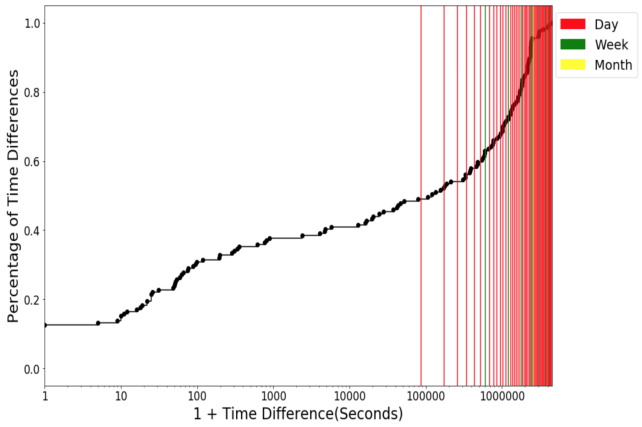


Figure 3: CDF of the distribution of differences in time in seconds between the first and last attacker login for each of the 159 compromised enterprise accounts. Red lines represent days, green lines represent weeks, and yellow lines represent months. **Note**: The x-axis has been log-scaled and each user's time difference was increased by 1 second as part of the log scaling.

Damage Prevention. To illustrate the importance of early intervention of compromised enterprise accounts, we analyzed accounts that sent at least one email flagged by Barracuda's email detector during their respective two-month time window in which login events were classified. There were 11 out of the 159 compromised enterprise accounts that sent emails flagged by Barracuda during their respective two-month windows.

Figure 4 shows the distribution of differences in time in seconds between first phishing email and first attacker login event for each of the 11 compromised enterprise accounts. We can see that 4 out of 11 compromised accounts (37%) had less than 1 day between the first phishing email and first attacker login event. The remaining 7 compromised accounts (73%) had over 3 days of time difference. Although a small sample size, this finding suggests that attackers who aim to send phish vary their approach upon first accessing a compromised account; some send phishing emails almost immediately, while others wait for some time to pass.

For our sample of 11 compromised accounts that sent phishing email, a detector that can react within 3 days of initial attacker access can prevent damage from phishing for the majority of enterprise accounts. However, how much damage is actually prevented by detector intervention? For the 7 compromised accounts that had over 3 days of time difference between first phishing email sent and first attacker login event, we found that in 6 of the 7 compromised enterprise accounts (all from different organizations), the first phishing email was

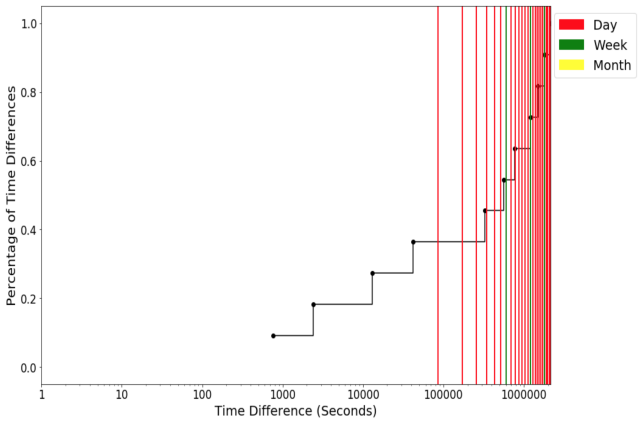


Figure 4: CDF of the distribution of differences in time in seconds between first phishing email and first attacker login event for each of the 11 compromised enterprise accounts that sent at least one phishing email during their attack windows. **Note**: The x-axis has been log-scaled.

part of a large "burst" of emails sent in 25 minutes or less, where each email in the burst had the same subject. 4 of these accounts had 400 or more emails as part of the "burst", while the remaining 2 accounts had fewer than 100 emails as part of the "burst". The 6 accounts with "bursts" sent emails to various numbers of accounts within and outside their respective organizations and 3 of these accounts had at least one email where BCC was used to send phishing emails.

This finding, along with the duration of compromises, illustrates the importance of detectors to react as soon as possible to attacker activity, though not necessarily in real-time. For accounts with phishing email, since the majority of attackers wait for some time before sending their first phishing email, detectors at organizations have a time window of around 3 days to intervene and react to potential attacker activity in enterprise accounts before more damage is done through phishing emails. From our analysis, most of the damage can be done in a short period of time through the mass amounts of phishing emails that are sent to various numbers of recipients.

6.2 Economy of Compromised Enterprise Accounts

In this section, we explore the economy of compromised enterprise accounts and the different modes of attackers that operate in this space. We estimate that in 50% of our sample of enterprise accounts, a single attacker conducts both the compromise and utilization of the account. We also uncover a specialized underground economy of compromised enterprise accounts where in 31% of our enterprise accounts, one attacker conducts the compromise and another likely buys the compromised account and utilizes it to extract information. In addition, we find that within this specialized economy, in the majority of compromised enterprise accounts, the second set of attackers that gain access to the account inflict more damage than the first set of attackers that perform the compromise, leading to an even greater importance of early detection and mitigation. For the remaining 19% of compromised enterprise accounts, it is unclear as to what mode of attacker these accounts face (i.e. one of the 2 modes we have discovered or another mode not yet explored).

First Mode of Attackers. Revisiting our findings from Section 6.1, we found that 81 out of 159 enterprise accounts (51%) are compromised for at least 1 day and 59 out of 159 enterprise accounts (37%) are compromised for at least 1 week. This finding suggests that there are largely two main segments of compromised enterprise accounts; those that are compromised for less than a day and the remaining that appear to be compromised for a day or more. Given this preliminary result, we aim to investigate the relationship between compromise duration and the economy and existence of various modes of attackers operating in the enterprise account space.

We start by trying to investigate whether enterprise accounts are generally accessed regularly by attackers or in isolated pockets of time during the compromise lifecycle. Using a similar idea to our `interarrival time ground truth indicator` from Section 5.2, for each of the 159 compromised enterprise accounts, we compute the interarrival time between every pair of successive attack events. We then take the max interarrival time for each user, which represents the longest time gap between any two successive attacker accesses within an account. From Figure 5, which shows a CDF

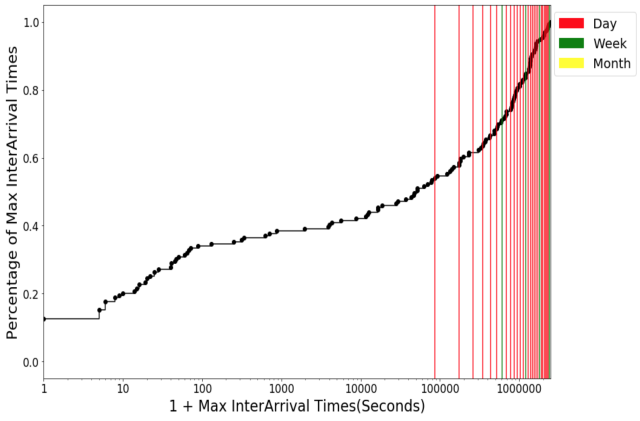


Figure 5: CDF of the max attacker interarrival times in seconds for all 159 enterprise accounts. **Note**: The x-axis has been log-scaled and each account's max interarrival time was increased by 1 second as part of the log scaling.

of the max attacker interarrival times in seconds for all 159

compromised enterprise accounts, we can see that at around the 1 day mark (first red line from the left), the inflection and trend of the CDF starts to change. In 53% of compromised enterprise accounts, the largest time gap between successive attacker accesses is less than 1 day, while the remaining 47% of compromised enterprise accounts (74 out of 159) have 1 or more days as their longest time gap.

Assuming our random sample of 159 enterprise accounts generalizes to a broader range of enterprises, we believe that accounts with small max attacker interarrival times and compromise duration, both less than 1 day, comprise one segment of the economy of compromised enterprise accounts. In our dataset, 78 out of the 159 enterprise accounts (50%) fall into this category. Due to the small time gaps between successive attacker events and relatively small compromise duration, these 78 accounts are likely compromised by a single set of attackers that both perform the compromise and use the accounts for a short period of time. We also note that there were 7 enterprise accounts that had small max attacker interarrival times (< 1 day), but longer duration of compromise of over a day. It is unclear if the same mode of attackers compromise these accounts or if a different mode presents itself.

Second Mode of Attackers. As we saw above from the CDF of max attacker interarrival times in Figure 5, 53% of enterprise accounts (74 out of 159) experienced a maximum of 1 or more days between successive attacker events. One possible explanation of the large time gap is that the initial set of attackers that compromised these accounts sell them to another set of attackers; hence, the time gaps represent the time needed for the transaction to complete. To accrue evidence for this claim, we compared attacker events before and after the max attacker interarrival time in these 74 accounts on the basis of geolocation, user agent, and internet service providers (ISPs). The goal was to determine if there was a discernible difference in these 3 properties of attacker events before and after the time gap, which could uncover the presence of two sets of attackers operating in these accounts.

To quantify the similarity of two sets of values before and after the max interarrival time for each of the 74 compromised enterprise accounts, we use the *Jaccard Index*, also know as the *Jaccard Similarity Coefficient*. The Jaccard Similarity Coefficient is a method for quantifying how similar 2 sets of data A and B are by relating the number of elements in the set intersection of A and B to the number of elements in the set union of A and B, as shown below.

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|}$$

In general, the smaller the Jaccard similarity coefficient for A and B, the less similar A and B are to one another. It has been widely used in many fields [12, 21, 29, 39] such as keyword similarity matching in search engines, test case selection for industrial software systems, and in secure multi-party computation.

For each of the 74 compromised enterprise accounts, we gather a set of country subdivisions mapped to attacker events before the max attacker interarrival time and a set of country subdivisions mapped to attacker events after. Similarly, we gather two sets of user agents and two sets of ISPs in the same manner. For each account, we compute 3 Jaccard similarity coefficients for geolocation, user agent, and ISP respectively. In Figure 6, we can see that generally, most of the

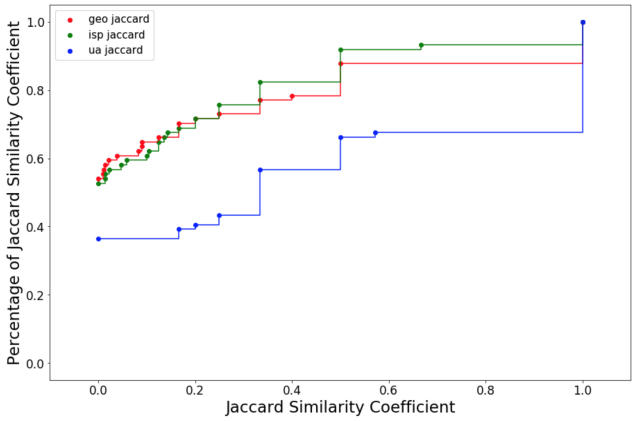


Figure 6: CDF of Jaccard Similarity Coefficients for Geolocation, User Agent, and ISP usage for 74 compromised enterprise accounts.

enterprise accounts have low Jaccard similarity coefficients for geolocation and ISP; one reason the user agent curve follows a different pattern is because of the normalization we performed for user agents, where we treat user agent strings with different device versions as the "same" underlying user agent. 50 of the enterprise accounts (around 70% of 74) had Jaccard similarity coefficients of 0.3 or less for geolocation and ISP indicating that the set of country subdivisions and set of ISPs before and after the large time gaps in these accounts were substantially different.

One can argue that the low geolocation Jaccard similarity coefficients might be a result of attackers using unstable anonymized IP proxies or even Tor. For each of the 74 accounts, we computed the number of unique hours and number of unique country subdivisions seen across all attack events after the account's respective max attacker interarrival time. For each account, we calculated the following stability ratio of the form

$$\text{stability} = \frac{\text{number of unique country subdivisions}}{\text{number of unique login hours}}.$$

If attackers are using unstable proxy services or Tor, we would expect this ratio to be large for many of the enterprise accounts. As we can see from Figure 7, which shows a CDF of

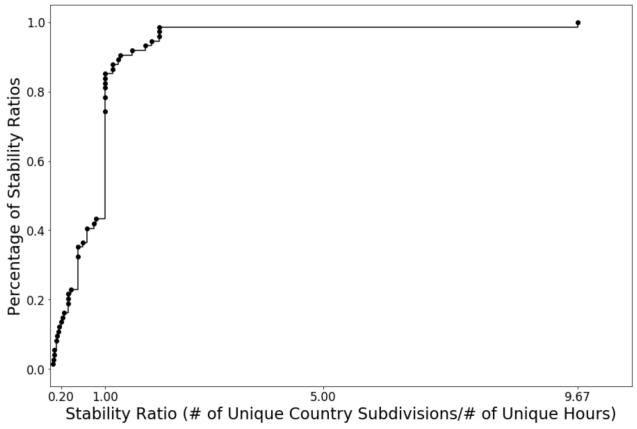


Figure 7: CDF of the Stability Ratios for each of the 74 Compromised Enterprise Accounts

the stability ratios for each of the 74 enterprise accounts, 85% of the accounts have stability ratios of at most 1 and 45% of the accounts have stability ratios less than 1. After looking into the enterprise account that had a stability ratio of 9.67, it was obvious that the attacker was using a specialized proxy service that generated a different IP address upon each login. In general, if attackers are using proxy services to obtain IP addresses, these services seem to be fairly stable and as a result, geolocation seems to be a viable way to distinguish between different attackers.

Given the large time gaps between successive attacker events and low Jaccard similarity coefficients, we believe that 50 of the 159 enterprise accounts (31%) comprise a specialized economy in the space of enterprise accounts, where one set of attackers compromise the accounts and sells the account credentials to another set of attackers that utilize the accounts for extracting information. The presence of this second mode of attackers implies that a more mature and specialized economy around compromised accounts has or is emerging, where attackers are starting to specialize in their roles (a skill set to compromise accounts vs. a skill set to extract value out of the compromised account's data and functionality).

In terms of understanding the potential damage inflicted by the two sets of attackers, we developed an *application access rate* metric that measures the number of Office 365 applications accessed by attack events divided by the number of unique hours the attack events span over a certain time period. For each of the 50 accounts, we computed the *application access rate* before and after the max attacker interarrival time and in 30 of the 50 accounts (60%), we find that the *application access rate* after the max attacker interarrival time is larger than that before the max attacker interarrival time. As a result, this further shows the importance of early mitigation in compromised enterprise accounts and that detectors that don't run in real-time should be designed to monitor continuous activity in order to prevent future damage after an account is sold.

Given that we have found the presence of a specialized underground economy for enterprise accounts that involves credential selling between attackers, as well as evidence of the second set of attackers inflicting more damage (higher application access rates) than the first set of attackers that perform the compromise, we believe that the second set of attackers utilizes the accounts for extracting monetary value. As a result, monetary information is likely obtained through sending phishing emails or looking through various applications within the accounts. We note that this is more of an observation, but it is guided by our evidence of the presence of a specialized underground economy of enterprise accounts and a second mode of attackers.

Overall, in this section, we were able to identify modes of attackers for 81% of enterprise accounts; however, for the remaining 19%, future work would involve determining if these accounts uncover a different mode of attackers or follow similar patterns to what we have discussed above.

6.3 How Enterprise Accounts Are Compromised

There are many ways in which enterprise accounts are compromised [1]. Some common methods include phishing, lateral phishing [22], password reuse, and the compromise of web-based databases. In this section, we analyze how enterprise accounts are compromised from the perspective of data breaches.

Data Breaches. We obtained data from a company, whom we will keep anonymous for security purposes, that mines the criminal underground and dark web for compromised credentials involved in breaches of online company databases. From our dataset of 159 compromised enterprise accounts, 31 of the accounts (20%) were involved in data breaches.

Economic Sector	Total
Consumer	1
Education	11
Food	1
Government	1
Health	2
Industrials	4
Technology	1
Tourism	1
Grand Total	23

Table 4: Table representing the number of organizations within each economic sector had at least one of its employee accounts found in the data breach.

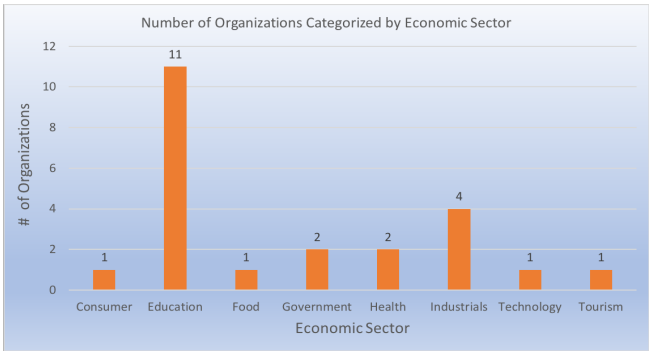


Figure 8: Bar chart of number of organizations within each economic sector that had at least one of its employee accounts found in a data breach.

Users of these accounts likely used their enterprise email address to create personal accounts on websites and when the websites' databases were breached, their associated personal account credentials were leaked. As a result, if these users reused credentials across their personal and enterprise accounts, their corresponding enterprise account was also likely compromised through the same data breach.

Figure 8 and Table 4 display economic sectors and the number of organizations within those economic sectors that had at least one of their accounts involved in a data breach. The 31 enterprise accounts belong to 21% of the organizations in our dataset (23 out of 111 organizations). We can see these 23 organizations span 8 of the 15 economic sectors that were shown in Figure 1. Although data breaches and credential leaks do not seem to discriminate against economic sectors, the education and industrials sectors seem to be hit the hardest in our dataset; there were 11 education organizations that had at least one compromised enterprise account found in a data breach and similarly, 4 industrials organizations.

From our findings, educational accounts, such as those belonging to `.edu` organizations, are the most common accounts involved in data breaches and credential leaks. In many cases, users of these academic accounts tend to also create personal accounts on study websites and password reuse is common; as a result, if the databases backing the websites are breached, then the original academic accounts are also subject to compromise. There has been previous research in the field of analyzing the lure of compromising academic accounts, such as the work done by Zhang et al. [25]. Zhang et al. note that academic accounts often offer free and unrestrained access to information due to less stringent security restrictions on these accounts. In addition, given that universities and schools are dormant for periods of time during the year and that upon graduation, users rarely access their educational accounts, attackers can go unnoticed for certain amounts of time in these accounts.

The findings in this section offer an insight into how enterprise accounts can be compromised. We saw that 21% of enterprise accounts were found in a data breach of online company databases; although we don't know for sure if these enterprise accounts were compromised as a result of the data breach, we nevertheless show that data breaches are fairly common among enterprise accounts and credential reuse with personal accounts can cause a lot of damage. As a result, enterprises should frequently remind their employees of the dangers of credential reuse among their accounts to avoid additional compromises of their accounts.

6.4 Uses of Compromised Enterprise Accounts

In this section, we aim to understand if there are certain operations attackers perform once inside an enterprise account. We find that attackers rarely change account passwords and never grant OAuth to cloud applications. In addition, within the Office 365 ecosystem, we find that attackers are not very interested in many cloud applications outside of email; 78% of the enterprise accounts only accessed email applications through attack events.

Other Operations Performed During Attacker Window. As we discussed in Section 3, every audit event has an `Operation` field that specifies the action that was taken through the audit event. The operations we are most interested in learning if attackers perform are ones that affect a user's ability to access their account; namely, operations such as "Change user password" and "Add OAuth". The operation "Change user password" enables the user to change the password to their account, while the "Add OAuth" operation enables a user to grant applications access to certain data within their account. Since our rule set only classifies successful login audit events due to the non-empty IP and user agent fields, we gather all "Change user password" and "Add OAuth" audit events that are close in time to each account's attack events.

We find that only 2 out of 159 compromised enterprise accounts (2%) had at least one "Change password for user" operation performed close in time to attacker activity. Looking deeper into the 2 accounts, we see the presence of more attacker activity after the change password operations were performed, indicating that these operations were performed by the attacker themselves. None of the 159 accounts had a single "Add OAuth" operation performed during the time period of attacker activity. Taken together, these findings suggest that attackers are not interested in changing a user's password or adding OAuth to a user's account, as this might reveal to the user that their account has been compromised and limit the amount of time the attacker can operate in the account. As a result, a "Change password" event or "Add OAuth" event are likely not good features for detectors, as they are rarely found performed by an attacker.

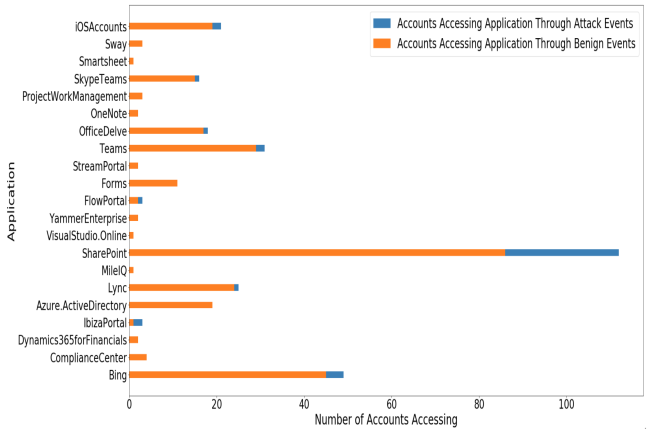


Figure 9: Bar chart comparing number of accounts accessing each of the 21 non-email applications via only attacker-labelled events and number of accounts accessing non-email applications via only benign successful login audit events from August 1, 2019 – January 27, 2020.

Unusual Applicaton Accesses by Attackers. A common finding among previous works is that many attackers tend to access email applications within accounts. Given that we have information on specific applications that are accessed through attack events, we aim to understand if there are specific Office 365 applications outside of frequently accessed email applications, such as Microsoft Exchange and Microsoft Outlook, that attackers access but the true users of the accounts don't access.

Across all audit events for the 159 enterprise accounts, there were a total of 21 non email-related Office 365 applications that were accessed by at least one account. For each of the 21 non-email applications, we determined the number of accounts that only accessed the application through their attack events and the number of accounts that only accessed the application through their benign events. The results for each of the 21 non-email applications are shown in the stacked bar chart in Figure 9. Surprisingly, other than Ibiza Portal, none of the remaining 20 applications had the characteristic of more accounts accessing it only through attack events than number of accounts accessing it through benign events. For Ibiza Portal, there were 3 accounts that accessed it only through attack events, while only one account accessed it solely through benign events. Ibiza Portal, or Microsoft Azure portal, [5] is an application that allows users to build and monitor their enterprise's web and cloud applications in a simple, unified place. Microsoft Azure Portal might allow an attacker to view confidential data within an enterprise's applications, but retrieving that data may take longer compared to other file-sharing applications, such as Microsoft SharePoint and Microsoft Forms. In addition, Microsoft Azure Portal is very rarely accessed by true users of enterprise accounts (only one account ever accessed Microsoft Azure Portal during their benign events). Overall, based on our dataset of compromised enterprise accounts, it does not appear that attackers are accessing unusual cloud-based applications that typical users don't access within the Office 365 ecosystem. Therefore, in the current state, building features for detectors around atypical accesses to cloud-based applications may not aid much in detecting attacker activity post-compromise. In future work, we hope to explore additional cloud-based applications outside of Office 365 to determine attacker interest.

Applications that Attackers Commonly Use. In the previous subsection, we saw that attackers who compromise cloud enterprise accounts don't seem to be accessing any exotic cloud applications that are unusual for enterprises to access. Therefore, in this section, we aim to understand the types of cloud applications that attackers exploit in enterprise accounts, regardless of how common the application is for enterprises to use.

Most attackers favor email-related applications. We found that in 98% of compromised enterprise accounts (156 out of 159), attackers accessed at least one email-related Office 365 application. Much of the previous work in understanding compromised personal accounts found that attackers tended to go through user's inboxes and send phishing emails; we now see that at scale, attackers seem to be exhibiting similar behavior in enterprise accounts. We also found that in 78% of compromised enterprise accounts (124 out of 159), attackers only accessed email-related Office 365 applications. We speculate that this may be because examining a user's inbox is sufficient for attackers who want to learn more about the user and the enterprise the account belongs to. Attackers

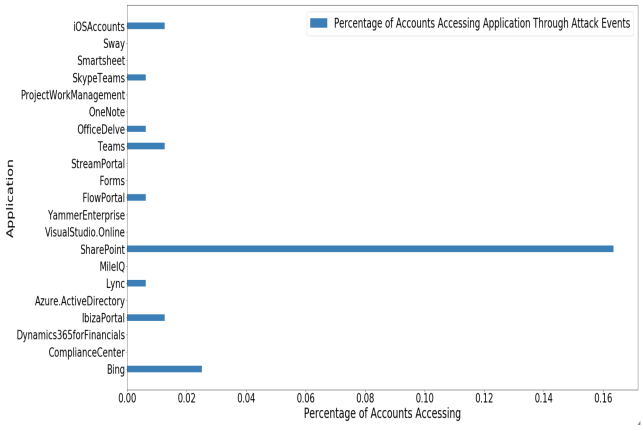


Figure 10: Bar chart showing the percentage of enterprise accounts that access each non-email related Office 365 application through attack events.

accessed non-email-related Office 365 applications in only 22% of compromised enterprise accounts (34 out of 159). In Figure 10, we can see that of the non-email related applications, Microsoft Sharepoint has the highest percentage of accounts that access it through attack events (17%), with Bing as the second highest percentage at 3%. Microsoft Sharepoint is a document management and storage application within Office 365 that allows users to share documents amongst one another. As a result, an attacker can use Microsoft Sharepoint as a way to gain access to confidential documents pertaining to the enterprise in a quick and easy way. Bing is a search engine developed by Microsoft so it is likely not a useful tool for an attacker.

Given the wide range of Office 365 cloud applications accessible by attackers, such as Office Delve and Microsoft Forms, and the abundance of documents and files shared among enterprises, it is surprising that attackers don't access these applications more often. Other than Microsoft Sharepoint, attackers of enterprise accounts still favor email-related applications, such as Microsoft Outlook, which offer a quick and convenient way for an attacker to gain access to contact lists and learn about any confidential and financial information tied to the employee and or enterprise. Taking all our findings in this section together, email appears to be by far the most common way of exploitation by attackers of enterprise accounts.

6.5 Compromised Enterprise Accounts and Personal Accounts

In this section, we compare characteristics of attacker activity within compromised enterprise accounts and personal accounts, showing that although the two spaces share some overlap, enterprises in particular face a different set of threats that inform new types of defense strategies.

Compromise Duration. There has been an extensive amount of prior work aimed at understanding duration of compromises within personal accounts, but none have studied this characteristic in compromised enterprise accounts. Thomas et al. [34] studied criminal account hijacking through the analysis of 40 million tweets a day over a ten-month period originating from personal accounts on Twitter. They found that 60% of compromises of Twitter accounts lasted a single day and 90% of compromises lasted fewer than 5 days. However, in our work with compromised enterprise accounts, we find that 38% of accounts are still compromised after 1 or more weeks. As a result, this suggests that compared to personal accounts, the state of enterprise account security actually has a lot of room to grow in terms of improving detection.

Economy of Compromised Accounts. Onaolapo et al [30] studied the motives of attackers accessing a 100 stolen Gmail accounts based on where the account credentials were leaked. They devised a taxonomy of attacker activity accessing the Gmail accounts, noting the presence of four attacker types (curious, gold diggers, spammers, and hijackers). They also observed that spammers tend to send phishing emails in bursts for a certain period of time. In our work, we found similar behavior for some enterprise accounts; attackers in 6 of the 11 compromised enterprise accounts that had at least one phishing email flagged by Barracuda sent their first phishing email as part of a large burst of emails sent in 25 minutes or less.

Onaolapo et al.'s findings are based on personal honeypot accounts leaked to paste sites, underground forums, and information-stealing malware. In our work, we analyzed the economy of compromised enterprise accounts at scale where the source of compromise was not definitively known. As a result, we illuminated at least 2 distinct modes of attackers that operate in the space of compromised enterprise accounts and devised techniques for distinguishing between the two. One mode of attackers in which a single attacker compromises and utilizes the account for a short period of time comprises 51% of the compromised enterprise accounts in our dataset. We also uncover a second mode of attackers: for 31% of accounts in our dataset, one set of attackers compromise the accounts and another set of attackers utilize these purchased accounts to extract information. This second set of attackers are similar in motive to the gold diggers defined by Onaolapo et al.

We also find that the second set of attackers inflict more damage to the account than the first set of attackers, which shows that in the space of enterprise accounts, detectors need not react in real-time; rather, given the long duration of compromises in enterprise accounts and presence of a specialized economy of attackers, detectors should be designed with more signals in mind in order to prevent future damage after an account is sold.

Uses of Compromised Accounts. Much of the prior work in the space of enterprise and personal accounts have studied attacker activity from the perspective of email applications and phishing. For example, Ho et al. [22] conducted the first large-scale detection of lateral phishing attacks in enterprise accounts. With the continued transition of data to cloud applications, part of our work aimed to understand if attackers are interested in different applications outside of email in enterprise accounts. What we find is that even in enterprise accounts, most attackers do not access many applications outside of email (78% of accounts access only email-related applications), which suggests that either many enterprise cloud accounts may not have access to interesting data outside of email or that attackers have yet to exploit these additional sources of information in enterprise accounts. Overall, based on our findings, attackers seem to be primarily interested in email applications in both personal and enterprise accounts.

7 Summary

In this work, we presented the first large-scale characterization of attacker activity in compromised enterprise accounts using a dataset of 159 compromised accounts from 111 enterprise organizations. We also developed and evaluated a novel forensics technique for distinguishing between attacker activity and benign activity in compromised enterprise accounts, yielding few false positives and enabling us to perform comprehensive incident forensics. Through a thorough analysis of compromised enterprise organizations in our dataset, we discovered many important findings that contribute to the understanding of the threats enterprises face everyday in relation to employee accounts and how to better defend against compromises in the future. Enterprise accounts tend to be compromised for long periods of time (51% of our accounts were compromised for at least one day). We also brought to light the economy of compromised enterprise accounts and the various modes of attackers that operate in this space. We find that a majority of enterprise accounts are compromised and utilized by a single set of attackers, but there also exists a specialized market of account compromises in which one set of specialized attackers compromise enterprise accounts and another set of attackers utilize the accounts. Using a commercial data breach service, we also note that 20% of our dataset's enterprise accounts appeared in at least one online password data breach, which suggests that credential reuse across an employee's personal and enterprise accounts can be a potential attack vector for compromise. Finally, we find that most attackers in our dataset do not access many applications outside of email, which suggests that attackers have yet to explore the wide-range of information within cloud applications. Overall, our work provides the first large-scale characterization of attacker behavior in compromised enterprise accounts and opens the door for better defenses and further research in the avenues of attacks these accounts face.

References

[1] 10 ways companies get hacked. https://www.cnbc.com/2012/07/06/10-Ways-Companies-Get-Hacked.html. Accessed: 2020-04-27.

[2] Barracuda sentinel. https://barracuda.com/products/sentinel. Accessed: 2020-03-29.

[3] Detailed properties in the office 365 audit log. https://docs.microsoft.com/en-us/microsoft-365/compliance/detailed-properties-in-the-office-365-audit-log?view=o365-worldwide. Accessed: 2020-03-29.

[4] Message resource type. https://docs.microsoft.com/en-us/graph/api/resources/message?view=graph-rest-1.0. Accessed: 2020-03-30.

[5] Microsoft azure portal. https://azure.microsoft.com/en-us/features/azure-portal/. Accessed: 2020-04-30.

[6] Office 365 management activity api schema. https://docs.microsoft.com/en-us/office/office-365-management-api/office-365-management-activity-api-schema#common-schema. Accessed: 2020-03-29.

[7] Retraining models on new data. https://docs.aws.amazon.com/machine-learning/latest/dg/retraining-models-on-new-data.html. Accessed: 2020-03-28.

[8] 2019 internet crime report released, Feb 2020.

[9] Saeed Abu-Nimeh, Dario Nappa, Xinlei Wang, and Suku Nair. A comparison of machine learning techniques for phishing detection. In *Proceedings of the Anti-Phishing Working Groups 2nd Annual ECrime Researchers Summit*, eCrime '07, page 60–69, New York, NY, USA, 2007. Association for Computing Machinery.

[10] Auth0. Protect your users with anomaly detection. https://auth0.com/learn/anomaly-detection/. Accessed: 2020-03-27.

[11] André Bergholz, Jeong Ho Chang, Gerhard Paass, Frank Reichartz, and Siehyun Strobel. Improved phishing detection using model-based features. In *CEAS*, 2008.

[12] Carlo Blundo, Emiliano De Cristofaro, and Paolo Gasti. Espresso: Efficient privacy-preserving evaluation of sample set similarity. In Roberto Di Pietro, Javier Herranz, Ernesto Damiani, and Radu State, editors, *Data Privacy Management and Autonomous Spontaneous Security*, pages 89–103, Berlin, Heidelberg, 2013. Springer Berlin Heidelberg.

[13] 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 *Proc. of 14th ACM IMC*, 2014.

[14] Asaf Cidon, Lior Gavish, Itay Bleier, Nadia Korshun, Marco Schweighauser, and Alexey Tsitkin. High precision detection of business email compromise. In *28th USENIX Security Symposium (USENIX Security 19)*, pages 1291–1307, Santa Clara, CA, August 2019. USENIX Association.

[15] Periwinkle Doerfler, Kurt Thomas, Maija Marincenko, Juri Ranieri, Yu Jiang, Angelika Moscicki, and Damon McCoy. Evaluating login challenges as adefense against account takeover. In *The World Wide Web Conference*, WWW '19, 2019.

[16] Manuel Egele, Gianluca Stringhini, Christopher Kruegel, and Giovanni Vigna. Compa: Detecting compromised accounts on social networks. In Proc. of 20th ISOC NDSS, 2013.

[17] K. D. Fairbanks, C. P. Lee, Y. H. Xia, and H. L. Owen. Timekeeper: A metadata archiving method for honeypot forensics. In *2007 IEEE SMC Information Assurance and Security Workshop*, pages 114–118, 2007.

[18] FBI. Business E-Mail Compromise The 12 Billion Dollar Scam, Jul 2018. https://www.ic3.gov/media/2018/180712.aspx.

[19] Ian Fette, Norman Sadeh, and Anthony Tomasic. Learning to detect phishing emails. In *Proceedings of the 16th International Conference on World Wide Web*, WWW '07, page 649–656, New York, NY, USA, 2007. Association for Computing Machinery.

[20] Hugo Gascon, Steffen Ullrich, Benjamin Stritter, and Konrad Rieck. Reading Between the Lines: Content-Agnostic Detection of Spear-Phishing Emails. In *Proc. of 21st Springer RAID*, 2018.

[21] H. Hemmati and L. Briand. An industrial investigation of similarity measures for model-based test case selection. In *2010 IEEE 21st International Symposium on Software Reliability Engineering*, pages 141–150, 2010.

[22] Grant Ho, Asaf Cidon, Lior Gavish, Marco Schweighauser, Vern Paxson, Stefan Savage, Geoffrey M Voelker, and David Wagner. Detecting and characterizing lateral phishing at scale. In *28th {USENIX} Security Symposium ({USENIX} Security 19)*, 2019.

[23] Grant Ho, Aashish Sharma, Mobin Javed, Vern Paxson, and David Wagner. Detecting Credential Spearphishing Attacks in Enterprise Settings. In *Proc. of 26th USENIX Security*, 2017.

[24] Xuan Hu, Banghuai Li, Yang Zhang, Changling Zhou, and Hao Ma. Detecting compromised email accounts from the perspective of graph topology. In *Proceedings of the 11th International Conference on Future Internet Technologies*, CFI '16, page 76–82, New York, NY, USA, 2016. Association for Computing Machinery.

[25] Jing Zhang, R. Berthier, W. Rhee, M. Bailey, P. Pal, F. Jahanian, and W. H. Sanders. Safeguarding academic accounts and resources with the university credential abuse auditing system. In *IEEE/IFIP International Conference on Dependable Systems and Networks (DSN 2012)*, pages 1–8, 2012.

[26] Bingshuang Liu, Zhaoyang Liu, Jianyu Zhang, Tao Wei, and Wei Zou. How many eyes are spying on your shared folders? In *Proceedings of the 2012 ACM Workshop on Privacy in the Electronic Society*, WPES '12, page 109–116, New York, NY, USA, 2012. Association for Computing Machinery.

[27] D. Malekian and M. R. Hashemi. An adaptive profile based fraud detection framework for handling concept drift. In *2013 10th International ISC Conference on Information Security and Cryptology (ISCISC)*, pages 1–6, 2013.

[28] MaxMind. Maxmind database website. https://www.maxmind.com/en/home.

[29] Suphakit Niwattanakul, Jatsada Singthongchai, Ekkachai Naenudorn, and Supachanun Wanapu. Using of jaccard coefficient for keywords similarity. 03 2013.

[30] 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 Internet Measurement Conference*, IMC '16, page 65–79, New York, NY, USA, 2016. Association for Computing Machinery.

[31] Jeff John Roberts. Homeland Security Chief Cites Phishing as Top Hacking Threat. http://fortune.com/2016/11/20/jeh-johnson-phishing/, Nov 2016.

[32] William Robertson, Giovanni Vigna, Christopher Krügel, and Richard Kemmerer. Using generalization and characterization techniques in the anomaly-based detection of web attacks. 01 2006. In Proc. of NDSS, 2006.

[33] Gianluca Stringhini and Olivier Thonnard. That Ain't You: Blocking Spearphishing Through Behavioral Modelling. In *Proc. of 12th Springer DIMVA*, 2015.

[34] Kurt Thomas, Frank Li, Chris Grier, and Vern Paxson. Consequences of connectivity: Characterizing account hijacking on twitter. In *Proceedings of the 2014 ACM SIGSAC Conference on Computer and Communications Security*, CCS '14, page 489–500, New York, NY, USA, 2014. Association for Computing Machinery.

[35] 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 SIGSAC Conference on Computer and Communications Security*, CCS '17, page 1421–1434, New York, NY, USA, 2017. Association for Computing Machinery.

[36] Kurt Thomas, Jennifer Pullman, Kevin Yeo, Ananth Raghunathan, Patrick Gage Kelley, Luca Invernizzi, Borbala Benko, Tadek Pietraszek, Sarvar Patel, Dan Boneh, et al. Protecting accounts from credential stuffing with password breach alerting. In *28th {USENIX} Security Symposium ({USENIX} Security 19)*, 2019.

[37] Lisa Vaas. How hackers broke into John Podesta, DNC Gmail accounts. https://nakedsecurity.sophos.com/2016/10/25/how-hackers-broke-into-john-podesta-dnc-gmail-account Oct 2016.

[38] Wikepedia. Administrative division. https://en.wikipedia.org/wiki/Administrative_division. Accessed: 2020-03-27.

[39] C. Wu and B. Wang. Extracting topics based on word2vec and improved jaccard similarity coefficient. In *2017 IEEE Second International Conference on Data Science in Cyberspace (DSC)*, pages 389–397, 2017.