Detecting and Characterizing Lateral Phishing at Scale

Grant Ho[†○] *Asaf Cidon*[○Ψ] *Lior Gavish*[○] *Marco Schweighauser*[○]
Vern Paxson[†ϕ] *Stefan Savage*[⋆] *Geoffrey M. Voelker*[⋆] *David Wagner*[†]

[○]*Barracuda Networks* [†]*UC Berkeley* [⋆]*UC San Diego* [Ψ]*Columbia University* [ϕ]*International Computer Science Institute*

Abstract

We present the first large-scale characterization of lateral phishing attacks, based on a dataset of 113 million employee-sent emails from 92 enterprise organizations. In a lateral phishing attack, adversaries leverage a compromised enterprise account to send phishing emails to other users, benefitting from both the implicit trust and the information in the hijacked user's account. We develop a classifier that finds hundreds of real-world lateral phishing emails, while generating under four false positives per every one-million employee-sent emails. Drawing on the attacks we detect, as well as a corpus of user-reported incidents, we quantify the scale of lateral phishing, identify several thematic content and recipient targeting strategies that attackers follow, illuminate two types of sophisticated behaviors that attackers exhibit, and estimate the success rate of these attacks. Collectively, these results expand our mental models of the 'enterprise attacker' and shed light on the current state of enterprise phishing attacks.

1 Introduction

For over a decade, the security community has explored a myriad of defenses against phishing attacks. Yet despite this long line of work, modern-day attackers routinely and successfully use phishing attacks to compromise government systems, political figures, and companies spanning every economic sector. Growing in prominence each year, this genre of attacks has risen to the level of government attention, with the FBI estimating $12.5 billion in financial losses worldwide from 78,617 reported incidents between October 2013 to May 2018 [12], and the US Secretary of Homeland Security declaring that phishing represents "the most devastating attacks by the most sophisticated attackers" [39].

By and large, the high-profile coverage around targeted spearphishing attacks against major entities, such as Google, RSA, and the Democratic National Committee, has captured and shaped our mental models of enterprise phishing attacks [35, 43, 46]. In these newsworthy instances, as well as many of the targeted spearphishing incidents discussed in the academic literature [25, 26, 28], the attacks come from external accounts, created by nation-state adversaries who cleverly craft or spoof the phishing account's name and email address to resemble a known and legitimate user. However, in recent years, work from both industry [7, 24, 36] and academia [6, 18, 32, 41] has pointed to the emergence and growth of *lateral phishing* attacks: a new form of phishing that targets a diverse range of organizations and has already incurred billions of dollars in financial harm [12]. In a lateral phishing attack, an adversary uses a compromised enterprise account to send phishing emails to a new set of recipients. This attack proves particularly insidious because the attacker automatically benefits from the implicit trust in the hijacked account: trust from both human recipients and conventional email protection systems.

Although recent work [10, 15, 18, 19, 41] presents several ideas for detecting lateral phishing, these prior methods either require that organizations possess sophisticated network monitoring infrastructure, or they produce too many false positives for practical usage. Moreover, no prior work has characterized this attack at a large, generalizable scale. For example, one of the most comprehensive related work uses a multi-year dataset from one organization, which only contains two lateral phishing attacks [18]. This state of affairs leaves many important questions unanswered: How should we think about this class of phishing with respect to its scale, sophistication, and success? Do attackers follow thematic strategies, and can these common behaviors fuel new or improved defenses? How are attackers capitalizing on the information within the hijacked accounts, and what does their behavior say about the state and trajectory of enterprise phishing attacks?

In this joint work between academia and Barracuda Networks we take a first step towards answering these open questions and understanding *lateral phishing* at scale. This paper seeks to both explore avenues for practical defenses against this burgeoning threat and develop accurate mental models for the state of these phishing attacks in the wild.

First, we present a new classifier for detecting URL-based lateral phishing emails and evaluate our approach on a dataset of 113 million emails, spanning 92 enterprise organizations. While the dynamic churn and dissimilarity in content across phishing emails proves challenging, our approach can detect 87.3% of attacks in our dataset, while generating less than 4 false positives per every 1,000,000 employee-sent emails.

Second, combining the attacks we detect with a corpus of user-reported lateral phishing attacks, we conduct the first large-scale characterization of lateral phishing in real-world organizations. Our analysis shows that this attack is potent and widespread: dozens of organizations, ranging from ones with fewer than 100 employees to ones with over 1,000 employees, experience lateral phishing attacks within the span

of several months; in total, 14% of a set of randomly sampled organizations experienced at least one lateral phishing incident within a seven-month timespan. Furthermore, we estimate that over 11% of attackers successfully compromise at least one additional employee. Even though our ground truth sources and detector face limitations that restrict their ability to uncover stealthy or narrowly targeted attacks, our results nonetheless illuminate a prominent threat that currently affects many real-world organizations.

Examining the behavior of lateral phishers, we explore and quantify the popularity of four recipient (victim) selection strategies. Although our dataset's attackers target dozens to hundreds of recipients, these recipients often include a subset of users with some relationship to the hijacked account (e.g., fellow employees or recent contacts). Additionally, we develop a categorization for the different levels of content tailoring displayed by our dataset's phishing messages. Our categorization shows that while 7% of attacks deploy targeted messages, most attacks opt for generic content that a phisher could easily reuse across multiple organizations. In particular, we observe that lateral phishers rely predominantly on two common lures: a pretext of a shared document and a fake warning message about a problem with the recipient's account. Despite the popularity of non-targeted content, nearly one-third of our dataset's attackers invest additional time and effort to make their attacks more convincing and/or to evade detection; and, over 80% of attacks occur during the normal working hours of the hijacked account.

Ultimately, this work yields two contributions that expand our understanding of enterprise phishing and potential defenses against it. First, we present a novel detector that achieves an order-of-magnitude better performance than prior work, while operating on a minimal data requirement (only leveraging historical emails). Second, through the first large-scale characterization of lateral phishing, we uncover the scale and success of this emerging class of attacks and shed light on common strategies that lateral phishers employ. Our analysis illuminates a prevalent class of enterprise attackers whose behavior does not fully match the tactics of targeted nation-state attacks or industrial espionage. Nonetheless, these lateral phishers still achieve success in the absence of new defenses, and many of our dataset's attackers do exhibit some signs of sophistication and focused effort.

2 Background

In a *lateral phishing* attack, attackers use a compromised, but legitimate, email account to send a phishing email to their victim(s). The attacker's goals and choice of malicious payload can take a number of different forms, from a malware-infected attachment, to a phishing URL, to a fake payment request. Our work focuses on lateral phishing attacks that employ a malicious URL embedded in the email, which is the most common exploit method identified in our dataset.

Listing 1: An anonymized example of a lateral phishing message that uses the lure of a fake contract document.

```
From: "Alice" <alice@company.com>
To: "Bob" <bob@company.com>
Subject: Company X (New Contract)

New Contract

View Document [this text linked to a phishing website]

Regards,
Alice [signature]
```

Listing 1 shows an anonymized example of a lateral phishing attack from our study. In this attack, the phisher tried to lure the recipient into clicking on a link under the false pretense of a new contract. Additionally, the attacker also tried to make the deception more credible by responding to recipients who inquired about the email's authenticity; and they also actively hid their presence in the compromised user's mailbox by deleting all traces of their phishing email.

Lateral phishing represents a dangerous but understudied attack at the intersection of phishing and account hijacking. Phishing attacks, broadly construed, involve an attacker crafting a deceptive email from any account (compromised or spoofed) to trick their victim into performing some action. Account hijacking, also known as account takeover (ATO) in industry parlance, involves the use of a compromised account for any kind of malicious means (e.g., including spam). While prior work primarily examines each of these attacks at a smaller scale and with respect to personal accounts, our work studies the intersection of both of these at a large scale and from the perspective of enterprise organizations. In doing so, we expand our understanding of important enterprise threats, avenues for defending against them, and strategies used by the attackers who perpetrate them.

2.1 Related Work

Detection: An extensive body of prior literature proposes numerous techniques for detecting traditional phishing attacks [1,3,13,14,44], as well as more sophisticated spearphishing attacks [8, 10, 23, 41, 47]. Hu et al. studied how to use social graph metrics to detect malicious emails sent from compromised accounts [19]. Their approach detects hijacked accounts with false positive rates between 20–40%. Unfortunately, in practice, many organizations handle tens of thousands of employee-sent emails per day, so a false positive rate of 20% would lead to thousands of false alerts each day. IdentityMailer, proposed by Stringhini et al. [41], detects lateral phishing attacks by training behavior models based on timing patterns, metadata, and stylometry for each user. If a new email deviates from an employee's behavioral model,

their system flags it as an attack. While promising, their approach produces false positive rates in the range of 1–10%, which is untenable in practice given the high volume of benign emails and low base rate of phishing. Additionally, their system requires training a behavioral model for each employee, incurring expensive technical debt to operate at scale.

Ho et al. developed methods for detecting lateral spearphishing by applying a novel anomaly detection algorithm on a set of features derived from historical user login data and enterprise network traffic logs [18]. Their approach detects both known and newly discovered attacks, with a false positive rate of 0.004%. However, organizations with less technical expertise often lack the infrastructure to comprehensively capture the enterprise's network traffic, which this prior approach requires. This technical prerequisite begs the question, can we practically detect lateral phishing attacks with a more minimalist dataset: only the enterprise's historical emails? Furthermore, their dataset reflects a single enterprise that experienced only two lateral phishing attacks across a 3.5-year timespan, which prevents them from characterizing the nature of lateral phishing at a general scale.

Characterization: While prior work shows that attackers frequently use phishing to compromise accounts, and that attackers occasionally conduct (lateral) phishing from these hijacked accounts, few efforts have studied the nature of lateral phishing in depth and at scale. Examining a sample of phishing emails, webpages, and compromised accounts from Google data sources, one prior study of account hijacking discovered that attackers often use these accounts to send phishing emails to the account's contacts [6]. However, they concluded that automatically detecting such attacks proves challenging. Onaolapo et al. studied what attackers do with hijacked accounts [32], but they did not examine lateral phishing. Separate from email accounts, a study of compromised Twitter accounts found that infections appear to spread laterally through the social network. However their dataset did not allow direct observation of the lateral attack vector itself [42], nor did it provide insights into the domain of compromised enterprise accounts (given the nature of social media).

Open Questions and Challenges: Prior work makes clear that account compromise poses a significant and widespread problem. This literature also presents promising defenses for enterprises that have sophisticated monitoring in place. Yet despite these advances, several key questions remain unresolved. Do organizations without comprehensive monitoring and technical expertise have a practical way to defend against lateral phishing attacks? What common strategies and tradecraft do lateral phishers employ? How are lateral phishers capitalizing on their control of legitimate accounts, and what does their tactical sophistication say about the state of enterprise phishing? This paper takes a step towards answering these open questions by presenting a new detection strategy and a large-scale characterization of lateral phishing attacks.

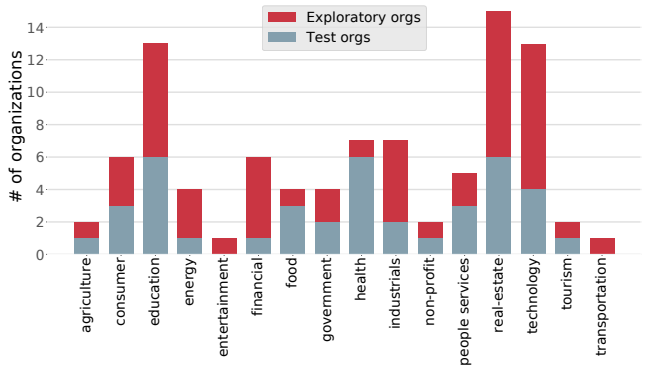

Figure 1: Breakdown of the economic sectors across our dataset's 52 exploratory organizations versus the 40 test organizations.

2.2 Ethics

In this work, our team, consisting of researchers from academia and a large security company, developed detection techniques using a dataset of historical emails and reported incidents from 92 organizations who are active customers of Barracuda Networks. These organizations granted Barracuda permission to access their Office 365 employee mailboxes for the purpose of researching and developing defenses against lateral phishing. Per Barracuda's policies, all fetched emails are stored encrypted, and customers have the option of revoking access to their data at any time.

Due to the sensitivity of the data, only authorized employees at Barracuda were allowed to access the data (under standard, strict access control policies). No personally identifying information or sensitive data was shared with any non-employee of Barracuda. Our project also received legal approval from Barracuda, who had permission from their customers to analyze and operate on the data.

Once Barracuda deployed a set of lateral phishing detectors to production, any detected attacks were reported to customers in real time to prevent financial loss and harm.

3 Data

Our dataset consists of employee-sent emails from 92 English-language organizations; 23 organizations came from randomly sampling enterprises that had reports of lateral phishing, and 69 were randomly sampled from all organizations. Across these enterprises, 25 organizations have 100 or fewer user accounts, 34 have between 101–1000 accounts, and 33 have over 1000 accounts. Real-estate, technology, and education constitute the three most common industries in our dataset, with 15, 13, and 13 enterprises respectively; Figures 1 and 2 show the distribution of the economic sectors and sizes of our dataset's organizations, broken down by *exploratory organizations* versus *test organizations* (§ 3.2).

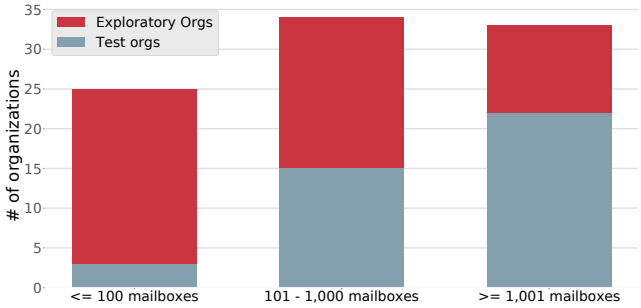

Figure 2: Breakdown of the organization sizes across our dataset's 52 exploratory organizations versus the 40 test organizations.

3.1 Schema

The organizations in our dataset use Office 365 as their email provider. At a high level, each email object contains: a unique Office 365 identifier; the email's metadata (SMTP header information), which describes properties such as the email's sent timestamp, recipients, purported sender, and subject; and the email's *body*, the contents of the email message in full HTML formatting. Office 365's documentation describes the full schema of each email object [29]. Additionally, for each organization, we have a set of *verified domains*: domains which the organization has declared that it owns.

3.2 Dataset Size

Our dataset consists of 113,083,695 unique, employee-sent emails. To ensure our detection techniques generalized (Section 5.1), we split our data into a training dataset of emails from 52 'exploratory organizations' during April–June 2018, and a test dataset covering July–October 2018 from 92 organizations. Our test dataset consists of emails from the 52 exploratory organizations (but from a later, disjoint time period than our training dataset), plus data from an additional, held-out set of 40 'test organizations'. We selected the 40 test organizations via a random sample that we performed prior to analyzing any data. Our training dataset has 25,670,264 emails, and our test dataset has 87,413,431 emails. Both sets of organizations cover a diverse range of industries and sizes as shown in Figures 1 and 2. The exploratory organizations span a total of 89,267 user mailboxes that sent or received email, and the test organizations have 138,752 mailboxes (based on the data from October 2018).[1]

3.3 Ground truth

Our set of lateral phishing emails comes from two sources: (1) attack emails reported to Barracuda by an organization's security administrators, as well as attacks reported by users to their organization or directly to Barracuda, and (2) emails

[1]The number of mailboxes is an upper bound on the number of employees due to the use of mailing lists and aliases.

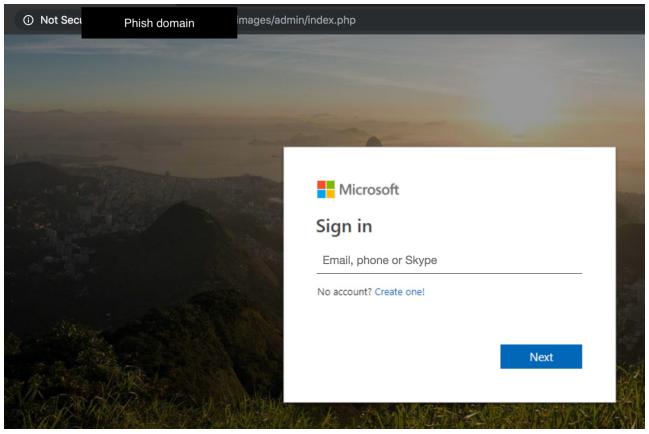

Figure 3: An anonymized screenshot of the web page that a phishing URL in a lateral phishing email led to.

flagged by our detector (§4), which we manually reviewed and labeled before including.

At a high-level, to manually label an email as phishing, or not, we examined its message content, Office 365 metadata, and Internet Message Headers [33] to determine whether the email contained phishing content, and whether the email came from a compromised account (versus an external account, which we do not treat as lateral phishing). For example, if the Office 365 metadata showed that a copy of the email resided in the employee's *Sent Items* folder, or if its headers showed that the email passed the corresponding SPF or DKIM [9] checks, then we considered the email to be lateral phishing. Appendix §A.1 describes our labeling procedure in detail.

Additionally, for a small sample of URLs in these lateral phishing emails, employees at Barracuda accessed the phishing URL in a VM-contained browser to better understand the end goals of the attack. To minimize potential harm and side effects, these employees only visited phishing URLs which contained no unique identifiers (i.e., no random strings or user/organization information in the URL path). To handle any phishing URLs that resided on URL-shortening domains, we used one of Barracuda's URL-expansion APIs that their production services already apply to email URLs, and only visited suspected phishing links that expanded to a non-side-effect URL. Most phishing URLs we explored led to a SafeBrowsing interstitial webpage, likely reflecting our use of historical emails, rather than what users would have encountered contemporaneously. However, more recent malicious URLs consistently led to credential phishing websites designed to look like a legitimate Office 365 login page (the email service provider used by our study's organizations); Figure 3 shows an anonymized example of one phishing website.

In total, our dataset contains 1,902 lateral phishing emails (unique by subject, sender, and sent-time), sent by 154 hijacked employee accounts from 33 organizations. 1,694 of these emails were reported by users, with the remainder found solely by our detector (§ 4); our detector also finds many of the

user-reported attacks as well (§ 5). Among the user-reported attacks, 40 emails (from 12 hijacked accounts) contained a fake wire transfer or malicious attachment, while the remaining 1,862 emails used a malicious URL.

We focus our detection strategy on URL-based phishing, given the prevalence of this attack vector. This focus means that our analysis and detection techniques do not reflect the full space of lateral phishing attacks. Despite this limitation, our dataset's attacks span dozens of organizations, enabling us to study a prevalent class of enterprise phishing that poses an important threat in its own right.

4 Detecting Lateral Phishing

Adopting the *lateral attacker* threat model defined by Ho et al. [18], we focus on phishing emails sent by a compromised employee account, where the attack embeds a malicious URL as the exploit (e.g., leading the user to a phishing webpage).

We explored three strategies for detecting lateral phishing attacks, but ultimately found that one of the strategies detected nearly all of the attacks identified by all three approaches. At a high level, the two less fruitful strategies detected attacks by looking for emails that contained (1) a rare URL and (2) a message whose text seemed likely to be used for phishing (e.g., similar text to a known phishing attack). Because our primary detection strategy detected all-but-two of the attacks found by the other strategies, while finding over ten times as many attacks, we defer discussion of the two less successful approaches to our extended technical report [17]; below, we focus on exploring the more effective strategy in detail. In our evaluation, we include the two additional attacks found by the alternative approaches as false negatives for our detector.

Overview: We examined the user-reported lateral phishing incidents in our training dataset (April–June 2018) to identify widespread themes and behaviors that we could leverage in our detector. Grouping this set of attacks by the hijacked account (*ATO*) that sent them, we found that 95% of these ATOs sent phishing emails to 25 or more distinct recipients.[2] This prevalent behavior, along with additional feature ideas inspired by the lure-exploit detection framework [18], provide the basis for our detection strategy. In the remainder of this section, we describe the features our detector uses, the intuition behind these features, and our detector's machine learning procedure for classifying emails.

Our techniques provide neither an all-encompassing approach to finding every attack, nor guaranteed robustness against motivated adversaries trying to evade detection. However, we show in Section 5 that our approach finds hundreds of lateral phishing emails across dozens of real-world organizations, while incurring a low volume of false positives.

[2]To assess the generalizability of our approach, our evaluation uses a withheld dataset, from a later timeframe and with new organizations (§ 5).

Features: Our detector extracts three sets of features. The first set consists of two features that target the popular behavior we observed earlier: contacting many recipients. Given an email, we first extract the number of unique recipients across the email's `To`, `CC`, and `BCC` headers. Additionally, we compute the Jaccard similarity of this email's recipient set to the closest set of historical recipients seen in any employee-sent email from the preceding month. We refer to this latter (similarity) feature as the email's recipient likelihood score.

The next two sets of features draw upon the lure-exploit phishing framework proposed by Ho et al. [18]. This framework posits that phishing emails contain two necessary components: a 'lure', which convinces the victim to believe the phishing email and perform some action; and an 'exploit': the malicious action the victim should execute. Their work finds that using features that target both of these two components significantly improves a detector's performance.

To characterize whether a new email contains a potential phishing lure, our detector extracts a single, lightweight boolean feature based on the email's text. Specifically, Barracuda provided us with a set of roughly 150 keywords and phrases that frequently occur in phishing attacks. They developed this set of 'phishy' keywords by extracting the link text from several hundred real-world phishing emails (both external and lateral phishing) and selecting the (normalized) text that occurred most frequently among these attacks. Thematically, these suspicious keywords convey a call to action that entices the recipient to click a link. For our 'lure' feature, we extract a boolean value that indicates whether an email contains any of these phishy keywords.

Finally, we complete our detector's feature set by extracting two features that capture whether an email might contain an exploit. Since our work focuses on URL-based attacks, this set of features reflects whether the email contains a potentially dangerous URL.

First, for each email, we extract a *global URL reputation* feature that quantifies the rarest URL an email contains. Given an email, we extract all URLs from the email's body and ignore URLs if they fall under two categories: we exclude all URLs whose domain is listed on the organization's *verified domain* list (§ 3.1), and we also exclude all URLs whose displayed, hyperlinked text exactly matches the URL of the hyperlink's underlying destination. For example, in Listing 1's attack, the displayed text of the phishing hyperlink was "Click Here", which does not match the hyperlink's destination (the phishing site), so our procedure would keep this URL. In contrast, Alice's signature from Listing 1 might contain a link to her personal website, e.g., `www.alice.com`; our procedure would ignore this URL, since the displayed text of `www.alice.com` matches the hyperlink's destination.

This latter filtering criteria makes the assumption that a phishing URL will attempt to obfuscate itself, and will not display the true underlying destination directly to the user. After these filtering steps, we extract a numerical feature by

mapping each remaining URL to its registered domain, and then looking up each domain's ranking on the Cisco Umbrella Top 1 Million sites [20];[3] for any unlisted domain, we assign it a default ranking of 10 million. We treat two special cases differently. For URLs on shortener domains, our detector attempts to recursively resolve the shortlink to its final destination. If this resolution succeeds, we use the global ranking of the final URL's domain; otherwise, we treat the URL as coming from an unranked domain (10 million). For URLs on content hosting sites (e.g., Google Drive or Sharepoint), we have no good way to determine its suspiciousness without fetching the content and analyzing it (an action that has several practical hurdles). As a result, we treat all URLs on content hosting sites as if they reside on unranked domains.

After ranking each URL's domain, we set the email's *global URL reputation* feature to be the worst (highest) domain ranking among its URLs. Intuitively, we expect that phishers will rarely host phishing pages on popular sites, so a higher *global URL reputation* indicates a more suspicious email. In principle a motivated adversary could evade this feature; e.g., if an adversary can compromise one of the organization's verified domains, they can host their phishing URL from this compromised site and avoid an accurate ranking. However, we found no such instances in our set of user-reported lateral phishing. Additionally, since the goal of this paper is to begin exploring practical detection techniques, and develop a large set of lateral phishing incidents for our analysis, this feature suffices for our needs.

In addition to this global reputation metric, we extract a local metric that characterizes the rareness of a URL with respect to the domains of URLs that an organization's employees typically send. Given a set of URLs embedded within an email, we map each URL to its fully-qualified domain name (FQDN) and count the number of days from the preceding month where at least one employee-sent email included a URL on the FQDN. We then take the minimum value across all of an email's URLs; we call this minimum value the *local URL reputation* feature. Intuitively, suspicious URLs will have both a low global reputation and a low local reputation. However, our evaluation (§ 5.2) finds that this *local URL reputation* feature adds little value: URLs with a low *local URL reputation* value almost always have a low *global URL reputation* value, and vice versa.

Classification: To label an email as phishing or not, we trained a Random Forest classifier [45] with the aforementioned features. To train our classifier, we take all user-reported lateral phishing emails in our training dataset, and combine them with a set of likely-benign emails. We generate this set of "benign" emails by randomly sampling a subset of the training window's emails that have not been reported as phishing; we sample 200 of these benign emails for each attack email to form our set of benign emails for training. Following standard machine learning practices, we selected both the hyperparameters for our classifier and the exact downsampling ratio (200:1) using cross-validation on this training data. Appendix A.2 describes our training procedure in more detail.

Once we have a trained classifier, given a new email, our detector extracts its features, feeds the features into this classifier, and outputs the classifier's decision.

[3] We use a list fetched in early March 2018 for our feature extraction, but in practice, one could use a continuously updated list.

5 Evaluation

In this section we evaluate our lateral phishing detector. We first describe our testing methodology, and then show how well the detector performs on millions of emails from over 90 organizations. Overall, our detector has a high detection rate, generates few false positives, and detects many new attacks.

5.1 Methodology

Establishing Generalizability: As described earlier in Section 3.2, we split our dataset into two disjoint segments: a *training dataset* consisting of emails from the 52 exploratory organizations during April–June 2018 and a *test dataset* from 92 enterprises during July–October 2018; in § 5.2, we show that our detector's performance remains the same if our test dataset contains only the emails from the 40 withheld test organizations. Given these two datasets, we first trained our classifier and tuned its hyperparameters via cross validation on our training dataset (Appendix A.2). Next, to compute our evaluation results, we ran our detector on each month of the held-out test dataset. To simulate a classifier in production, we followed standard machine learning practices and used a continuous learning procedure to update our detector each month [38]. Namely, at the end of each month, we aggregated the user-reported and detector-discovered phishing emails from all previous months into a new set of phishing 'training' data; and, we aggregated our original set of randomly sampled benign emails with our detector's false positives from all previous months to form a new benign 'training' dataset. We then trained a new model on this aggregated training dataset and used this updated model to classify the subsequent month's data. However, to ensure that any tuning or knowledge we derived from the training dataset did not bias or overfit our classifier, we did not alter any of the model's hyperparameters or features during our evaluation on the test dataset.

Our evaluation's temporal-split between the training and test datasets, along with the introduction of new data from randomly withheld organizations into the test dataset, follows best practices that recommend this approach over a randomized cross-validation evaluation [2, 31, 34]. A completely randomized evaluation (e.g., cross-validation) risks training on data from the future and testing on the past, which might lead us to overestimate the detector's effectiveness. In contrast,

Metric	Training April – June 2018	Testing July – October 2018
Organizations	52 Exploratory	52 Exploratory + 40 Test
Detected Known Attacks	34	47
Detected New Attacks	28	49
Missed Attacks (FN)	8	14
Detection Rate	88.6%	87.3%
Total Emails	25,670,264	87,413,431
False Positives (FP)	136	316
False Positive Rate	0.00053%	0.00036%
Precision	31.3%	23.3%

Table 1: Evaluation results of our detector. 'Detected Known Attacks' shows the number of incidents that our detector identified, and were also reported by an employee at an organization. 'Detected New Attacks' shows the number of incidents that our detector identified, but were not reported by anyone. 'Missed Attacks (FN)' shows all incidents either reported by a user or found by any of our detection strategies, but our detector marked it as benign (false negative). Of the 22 incidents our detector misses, 12 are attachment-based attacks, a threat model which our detector explicitly does not target but which we include in our FN and Detection Rate results for completeness.

our methodology evaluates our detector with fresh data from a "future" time period and introduces 40 new organizations, neither of which our detector saw during training time; this also reflects how a detector operates in practice.

Alert Metric (Incidents): We have several choices for modeling our detector's alert generation process (i.e., how we count *distinct* attacks). For example, we could evaluate our detector's performance in terms of how many unique emails it correctly labels. Or, we could measure our detector's performance in terms of how many distinct employee accounts it marks as compromised (modeling a detector that generates one alert per account and suppresses the rest). Ultimately, we select a notion commonly used in practice, that of an *incident*, which corresponds to a unique (subject, sender email address) pair. At this granularity, our detector's alert generation model produces a single alert per unique (subject, sender) pair. This metric avoids biased evaluation numbers that overemphasize compromise incidents that generate many identical emails during a single attack. For example, if there are two incidents, one which generates one hundred emails to one recipient each, and another which generates one email to 100 recipients, a detector's performance on the hundred-email incident will dominate the result if we count attacks at the email level.

In total, our training dataset contains 40 lateral phishing incidents from our user-reported ground truth sources, and our test dataset contains 61 user-reported incidents. Our detector finds an additional 77 unreported incidents (row 2 of Table 1).

5.2 Detection Results

Table 1 summarizes the performance metrics for our detector. We use the term *Detection Rate* to refer to the percentage of lateral phishing incidents that our detector finds, divided by all known attack incidents in our dataset (i.e., any user-reported incident and any incident found by any detection technique we tried). For completeness, we include the 12 attachment-based incidents in our False Negative and Detection Rate computations, which our detector obviously misses since we designed it to catch URL-based lateral phishing. Additionally, we also include, as false negatives, 2 training incidents that our less successful detectors identified [17]; these two alternative strategies did not find any new attacks in the test dataset. Thus, the *Detection Rate* reflects a best-effort assessment that potentially overestimates the true positive rate of our detector, since we have an imperfect ground truth that cannot account for narrowly targeted attacks that go unreported by users. *Precision* equals the percent of attack alerts (incidents) produced by our detector divided by the total number of alerts our detector generated (attacks plus false positives).

Training and Tuning: On the training dataset, our detector correctly identified 62 out of 70 lateral phishing incidents (88.6%), while generating a total of 62 false positives (on 25.7 million employee-sent emails).

Our PySpark Random Forest classifier exposes a built-in estimate of each feature's relative importance [40], where each feature receives a score between 0.0–1.0 and the sum of all the scores adds up to 1.0. Based on these feature weights, our model places the most emphasis on the *global URL reputation* feature, giving it a weight of 0.42, and the email's 'number of recipients' feature (0.34). In contrast, our model essentially ignores our *local URL reputation*, assigning it a score of 0.01, likely because most globally rare domains tend to also be locally rare. Of the remaining features, the recipient likelihood feature has a weight of 0.17 and the 'phishy' keyword feature has a weight of 0.06.

Test Dataset: Our detector correctly identified 96 lateral phishing incidents out of the 110 test incidents (87.3%) across our ground truth dataset. Additionally, our detector discovered 49 incidents that, according to our ground truth, were not reported by a user as phishing. With respect to its cost, our detector generated 312 total false positives across the entire test dataset (a false positive rate of less than 0.00035%, assuming that emails not identified as an attack by our ground truth are benign). Across our test dataset, 82 out of the 92 organizations accumulated 10 or fewer false positives across the entire four month window, with 44 organizations encountering zero false positives across this timespan. In contrast, only three organizations had more than 40 total false positives across all four months (encountering 44, 66, and 83 false positives, respectively). Our detector achieves similar results if we evaluate on just the data from our 40 withheld test organizations, with a Detection Rate of 91.0%, a precision of 23.1%, and a false positive rate of 0.00038%.

Bias and Evasion: We base our evaluation numbers on the best ground truth we have: a combination of all user-reported

lateral phishing incidents (including some attacks outside our threat model), and all incidents discovered by any detection technique we tried (which includes two approaches orthogonal to our detector's strategy). This ground truth suffers from a bias towards phishing emails that contact many potential victims, and attacks that users can more easily recognize. Additionally, since our detector focuses on URL-based exploits, our dataset of attacks likely underestimates the prevalence of non-URL-based phishing attacks, which come solely from user-reported instances in our dataset. As a result, our work does not capture the full space of lateral phishing attacks, such as ones where the attacker targets a narrow, select set of victims with stealthily deceptive content. Rather, given that our detector identifies many known and unreported attacks, while generating only a few false positives per month, we provide a starting point for practical detection that future work can extend. Moreover, even if our detector does not capture every possible attack, the fact that the attacks in our dataset span dozens of different organizations, across a multi-month timeframe, allows us to illuminate a class of understudied attacks that many enterprises currently face.

Aside from obtaining more comprehensive ground truth, more work is needed to explore defenses against potential evasion attacks. Attackers could attempt to evade our detector by targeting different features we draw upon, such as the composition or number of recipients they target. Against many of these evasion attacks, future work could leverage additional features and data, such as the actions a user takes within an email account (e.g., reconnaissance actions, such as unusual searches, that indicate an attacker mining the account for targeted recipients to attack) or information from the user's account log-on (e.g., the detector proposed by Ho et al. used an account's login IP address [18] to detect lateral phishing). At the same time, future work should study which evasion attacks remain economically feasible for attackers to conduct. For example, an attacker could choose to only target a small number of users in the hopes of evading our detector; but even if this evasion succeeded, the conversion rate of fooling a recipient might be so low that the attack ultimately fails to compromise an economically viable number of victims. Indeed, as we explore in the following section (§ 6), the attackers captured in our dataset already engage in a range of different behaviors, including a few forms of sophisticated, manual effort to increase the success of their attacks.

6 Characterizing Lateral Phishing

In this section, we conduct an analysis of real-world lateral phishing using all known attacks across our entire dataset (both training and test). During the seven month timespan, a total of 33 organizations experienced lateral phishing attacks, with the majority of these compromised organizations experiencing multiple incidents. Examining the thematic message content and recipient targeting strategies of the attacks, our analysis suggests that most lateral phishers in our dataset do not actively mine a hijacked account's emails to craft personalized spearphishing attacks. Rather, these attackers operate in an opportunistic fashion and rely on commonplace phishing content. This finding suggests that the space of enterprise phishing has expanded beyond its historical association with sophisticated APTs and nation-state adversaries.

Scale and Success	
# distinct phishing emails	1,902
# incidents	180
# ATOs	154
# organizations w/ 1+ incident	33
# phishing recipients	101,276
% successful ATOs	11%
# employee recip (average) for compromise	542

Table 2: Summary of the scale and success of the lateral phishing attacks in our dataset (§ 6.1).

At the same time, these attacks nonetheless succeed, and a significant fraction of attackers do exhibit some signs of sophistication and attention to detail. As an estimate of the success of lateral phishing attacks, at least 11% of our dataset's attackers successfully compromise at least one other employee account. In terms of more refined tactics, 31% of lateral phishers invest some manual effort in evading detection or increasing their attack's success rate. Additionally, over 80% of the attacks in our dataset occur during the normal working hours of the hijacked account. Taken together, our results suggest that lateral phishing attacks pose a prevalent enterprise threat that still has room to grow in sophistication.

In addition to exploring attacks at the incident granularity (as done in § 5), this section also explores attacks at the granularity of a lateral phisher (hijacked account) when studying different attacker behaviors. As described in Section 2, industry practitioners often refer to such hijacked accounts as ATOs, and throughout this section, we use the terms *hijacked account*, *lateral phisher*, and *ATO* synonymously.

6.1 Scale and Success of Lateral Phishing

Scale: Our dataset contains 1,902 distinct lateral phishing emails sent by 154 hijacked accounts.[4] A total of 33 organizations in our dataset experience at least one lateral phishing incident: 23 of these organizations came from sampling the set of enterprises with known lateral phishing incidents (§ 3), while the remaining 10 came from the 69 organizations we sampled from the general population. Assuming our random sample reflects the broader population of enterprises, over 14% of organizations experience at least one lateral phishing incident within a 7 month timespan. Furthermore, based

[4]Distinct emails are defined by having a fully unique tuple of (sender, subject, timestamp, and recipients).

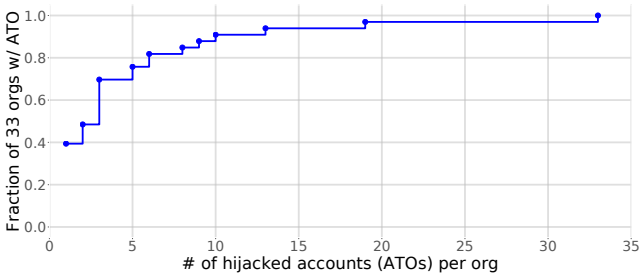

Figure 4: Fraction of organizations with x hijacked accounts that sent at least one lateral phishing email. 13 organizations had only 1 ATO; the remaining 20 saw lateral phishing from 2+ ATOs (§ 6.1).

on Figure 4, over 60% of the compromised organizations in our dataset experienced lateral phishing attacks from at least two hijacked employee accounts. Given that our set of attacks likely contains false negatives (thus underestimating the prevalence of attacks), these numbers illustrate that lateral phishing attacks are widespread across enterprise organizations.

Successful Attacks: Given our dataset, we do not definitively know whether an attack succeeded. However, we conservatively (under)estimate the success rate of lateral phishing using the methodology below. Based on this procedure, we estimate that at least 11% of lateral phishers successfully compromise at least one new enterprise account.

Let Alice and Bob represent two different ATOs at the same organization, where P_A and P_B represent one of Alice's and Bob's phishing emails respectively, and $Reply_B$ represents a reply from Bob to a lateral phishing email he received from Alice. Intuitively, our methodology concludes that Alice successfully compromised Bob if (1) Bob received a phishing email from Alice, (2) shortly after receiving Alice's phish, Bob then subsequently sent his own phish, and (3) we have strong evidence that the two employees' phishing emails are related (reflected in criteria 3 and 4 below).

Formally, we say that P_A succeeded in compromising Bob's account if *all* of the following conditions are true:

1. Bob was a recipient of P_A
2. After receiving P_A, Bob subsequently sent his own lateral phishing emails (P_B)
3. Either of the following two conditions are met:
   (a) P_B and P_A used similar phishing content: if the two attacks used identical subjects or if both of the phishing URLs they used belonged to the same fully-qualified domain
   (b) Bob sent a reply ($Reply_B$) to P_A, where his reply suggests he fell for Alice's attack and where Bob sent $Reply_B$ prior to his own attack (P_B)
4. Either of the following two conditions are met:
   (a) P_B was sent within two days after Bob received P_A
   (b) P_B and P_A used identical phishing messages or their phishing URLs' paths followed nearly identical structures (e.g., '`http://X.com/z/office365/index.html`' vs. '`http://Y.com/z/office365/index.html`')

Unpacking the final criteria (#4), in the first case (4.a), we settled on a two-day interarrival threshold based on prior literature [21, 22], which suggests that 50% of users respond to an email within 2 days and roughly 75% of users who click on a spam email do so within 2 days. Assuming that phishing follows similar time constants for how long it takes a recipient to take action, 2 days represented a conservative threshold to establish a link between P_A and P_B. At the same time, both prior works show there exists a long tail of users who take weeks to read and act on an email. The second part (4.b) attempts to address this long tail by raising the similarity requirements between Alice and Bob's attacks before concluding that former caused the latter. For successful attackers labeled by heuristic 4.b, the longest observed time gap between P_A and P_B is 17 days, which falls within a plausible timescale based on the aforementioned literature.

From this methodology, we conclude that 17 ATOs successfully compromised at least 23 future ATOs. While our procedure might erroneously identify cases where an attacker has concurrently compromised both Alice and Bob (rather than compromising Bob's account via Alice's), the first two criteria (requiring Bob to be a recent recipient of Alice's phishing email) help reduce this error. Our procedure likely underestimates the general success rate of lateral phishing attacks, since it does not identify successful attacks where the attacker does not subsequently use Bob's account to send phishing emails, nor does it account for false negatives in our dataset or attacks outside of our visibility (e.g., compromise of recipients at external organizations).

6.2 Recipient Targeting

In this section, we estimate the conversion rate of our dataset's lateral phishing attacks, and discuss four recipient targeting strategies that reflect the behavior of most attackers in our dataset.

Recipient Volume and Estimated Conversation Rate: Cumulatively, the lateral phishers in our dataset contact 101,276 unique recipients, where 41,740 belong to the same organization as the ATO. As shown in Figure 5, more than 94% of the attackers send their phishing emails to over 100 recipients; with respect to the general population of all lateral phishers, this percentage likely overestimates the prevalence of high "recipient-volume" attackers, since our detector draws on recipient-related features.

Targeting hundreds of people gives attackers a larger pool of potential victims, but it also incurs a risk that a recipient will detect and flag the attack either to their security team or

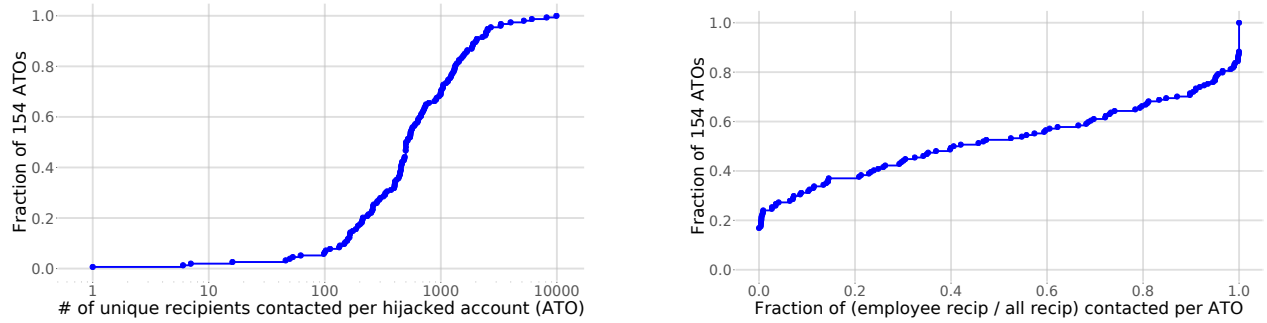

Figure 5: The left CDF shows the distribution of the total number of phishing recipients per ATO. The right CDF shows the fraction of ATOs where $x\%$ of their total recipient set consists of fellow employees.

their fellow recipients (e.g., via Reply-All). To isolate their victims and minimize the ability for fellow recipients to warn each other, we found that attackers frequently contact their recipients via a mass `BCC` or through many individual emails. Aside from this containment strategy, we also estimate that our dataset's lateral phishing attacks have a difficult time fooling an individual employee, and thus might require targeting many recipients to hijack a new account. Earlier in Section 6.1, we found that 17 ATOs successfully compromised 23 new accounts. Looking at the number of accounts they successfully hijacked divided by the number of fellow employees they targeted, the median conversation rate for our attackers was one newly hijacked account per 542 fellow employees; the attacker with the best conversation rate contacted an average of 26 employees per successful compromise. We caution that our method for determining whether an attack succeeded (§ 6.1) does not cover all cases, so our conversation rate might also underestimate the success of these attacks in practice. But if our estimated conversion rate accurately approximates the true rate, it would explain why these attackers contact so many recipients, despite the increased risk of detection.

Recipient Targeting Strategies: Anecdotally, we know that some lateral phishers select their set of victims by leveraging information in the hijacked account to target familiar users; for example, sending their attack to a subset of the account's "Contact Book". Unfortunately our dataset does not include information about any reconnaissance actions that an attacker performed to select their phishing recipients (e.g., explicitly searching through a user's contact book or recent recipients).

Instead, we empirically explore the recipient sets across our dataset's attackers to identify plausible strategies for how these attackers might have chosen their set of victims. Four recipient targeting strategies, summarized in Table 3 (explained below), reflect the behavior of all but six attackers in our dataset. To help assess whether a recipient and the ATO share a meaningful relationship, we compute each ATO's *recent contacts*: the set of all email addresses whom the ATO sent at least one email to in the 30 days preceding the ATO's phishing emails. While some attackers (28.6%) specifically target many of an account's recent contacts, the majority of lateral phishers appear more interested in either contacting many arbitrary recipients or sending phishing emails to a large fraction of the hijacked account's organization.

Recipient Targeting Strategy	# ATOs
Account-agnostic	63
Organization-wide	39
Lateral-organization	2
Targeted-recipient	44
Inconclusive	6

Table 3: Summary of recipient targeting strategies per ATO (§ 6.2).

Account-agnostic Attackers: Starting with the least-targeted behavior, 63 ATOs in our dataset sent their attacks to a wide range of recipients, most of whom do not appear closely related to the hijacked account. We call this group *Account-agnostic attackers*, and identify them using two heuristics.

First, we categorize an attacker as Account-agnostic if less than 1% of the recipients belong to the same organization as the ATO, and further exploration of their recipients does not reveal a strong connection with the account. Examining the right-hand graph in Figure 5, 37 ATOs target recipient sets where less than 1% of the recipients belong to the same organization as the ATO. To rule out the possibility that these attackers' recipients are nonetheless related to the account, we computed the fraction of recipients who appeared in each ATO's recent contacts; for all of the 37 possible Account-agnostic ATOs, less than 17% of their attack's total recipients appeared in their recent contacts. Among these 37 candidate Account-agnostic ATOs, 33 of them contact recipients at 10 or more organizations (unique recipient email domains), 2 of them exclusively target either Gmail or Hotmail accounts, and the remaining 2 ATOs are best described as Lateral-organization attackers (below).[5] Excluding the 2 Lateral-organization attackers, the 35 ATOs identified by this

[5] Our extended technical report provides the distribution of recipient domains contacted by all ATOs [17].

first criteria sent their attacks to predominantly external recipients, belonging to either many different organizations or exclusively to personal email hosting services (e.g., Gmail and Hotmail), and only a small percentage of these recipients appeared in the ATO's recent contacts; as such, we label these 35 attackers as Account-agnostic.

Second, we expand our search for Account-agnostic attackers by searching for attackers where less than 50% of the ATO's total recipients also belong to the ATO's organization, and where the ATO contacts recipients at many different organizations; specifically, where the ATO's phishing recipients belonged to over twice as many unique domains as all of the email addresses in ATO's recent contacts. This search identified 63 ATOs. To filter out attackers in this set who may have drawn on the hijacked account's recent contacts, we exclude any ATO where over 17% of their attack's total recipients also appeared in the ATO's recent contacts (17% was the maximum percentage among ATOs from the first Account-agnostic heuristic). After applying this last condition, our second heuristic identifies 54 Account-agnostic attackers.

Combining and deduplicating the ATOs from both criteria results in a total of 63 Account-agnostic attackers (40.9%): lateral phishers who predominantly target recipients without close relationships to the hijacked account or its organization.

Lateral-organization Attackers: During our exploration of potential Account-agnostic ATOs, we uncovered 2 attackers whom we label under a different category: *Lateral-organization attackers*. In both these cases, less than 1% of the attacker's recipients belonged to the same organization as the ATO, but each attacker's recipients did belong to organizations within the same industry as the ATO's organization. This thematic characteristic among the recipients suggests a deliberate strategy to spread across organizations within the targeted industries, so accordingly, we categorize them as Lateral-organization attackers.

Organization-wide Attackers: Office 365 provides a "Groups" feature that lists the different groups that an account belongs to [30]. For some enterprises, this feature enumerates most, if not all, employees at the organization. Thus, lateral phishers who wish to cast a wide phishing net might adopt a simple strategy of sending their attack to everyone at the organization. We call these ATOs *Organization-wide attackers* and identify them through two ways.

First, we search for any attackers where at least half of their phishing recipients belong to the ATO's organization, and where at least 50% of the organization's employees received the phishing email (i.e., the majority of a phisher's victims were employees and the attacker targeted a majority of the enterprise); this search yielded a total of 16 ATOs. We estimate the list of an organization's employees by building a set of all employee email addresses who sent or received email from

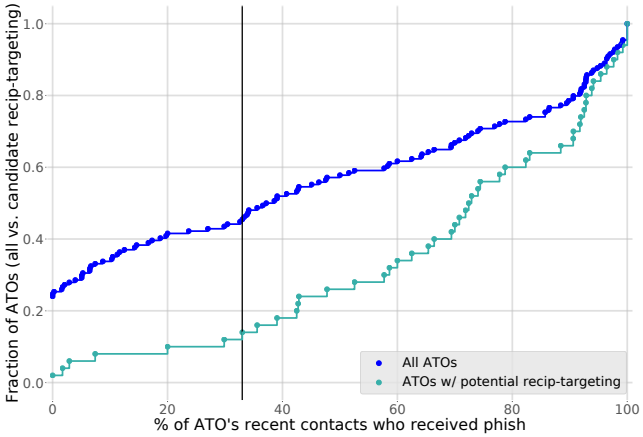

Figure 6: CDF: the x-axis displays what % of the ATO's recent contacts received a lateral phishing email (§ 6.2). The bottom teal graph filters the ATOs to *exclude* any ATO identified as Account-agnostic, Lateral-organization, and Organization-wide attackers; at the vertical black line, 88% of these filtered ATOs send phishing emails to at least $x = 33\%$ of addresses from their recent contacts.

anyone during the entire month of the phishing incident.[6] For all of these 16 ATOs, less than 11% of the recipients they target also appear in their recent contacts. Coupled with the fact that each of these ATOs contacts over 1,300 recipients, their behavior suggests that their initial goal focuses on phishing as many of the enterprise's recipients as possible, rather than targeting users particularly close to the hijacked account. Accordingly, we categorize them as Organization-wide attackers.

Our second heuristic looks for attackers whose recipient set consists nearly entirely of fellow employees, but where the majority of the organization does not necessarily receive a phishing email. Revisiting Figure 5, 36 candidate Organization-wide ATOs sent over 95% of their phishing emails to fellow employee recipients. However, we again need to exclude and account for ATOs who leverage their hijacked account's recent contacts. From the first Organization-wide heuristic discussed previously, we saw that less than 11% of the recipients of that heuristic's Organization-wide attackers came from the ATO's recent contacts. Using this value as a final threshold for this second candidate set of Organization-wide attackers, we identify 29 Organization-wide attackers where over 95% of their recipients belong to the ATO's organization but less than 11% of the recipients came from the ATO's recent contacts; a combination that suggests the attacker seeks primarily to compromise other employees, but who do not necessarily have a personal connection with the hijacked account.

Aggregating and deduplicating the two sets of lateral phishers from above produces a total of 39 Organization-wide attackers (25.3%), who take advantage of the information in a hijacked account to target many fellow employees.

[6] This collection likely overestimates the actual set of employees because of service addresses, mailing list aliases, and personnel churn.

Targeted-recipient Attackers: For the remaining, uncategorized 50 ATOs, we cannot conclusively determine the attackers' recipient targeting strategies because our dataset does not provide us with the full set of information and actions available to the attacker. Nonetheless, Figure 6 presents some evidence that 44 of these remaining attackers do draw upon the hijacked account's prior relationships. Specifically, 44 attackers sent their attacks to at least 33% of the addresses in the ATO's recent contacts.[7] Since these ATOs sent attacks to at least 1 out of every 3 of the ATO's recently contacted recipients, these attackers appear interested in targeting a substantial fraction of users with known ties to the hijacked account. As such, we label these 44 ATOs as Targeted-recipient attackers.

6.3 Message Content: Tailoring and Themes

Since lateral phishers control a legitimate employee account, these attackers could easily mine recent emails to craft personalized spearphishing messages. To understand how much attackers do leverage their privileged access in their phishing attacks, this section characterizes the level of tailoring we see among lateral phishing messages. Overall, only 7% of our dataset's incidents contain targeted content within their messages. Across the phishing emails that used non-targeted content, the attackers in our dataset relied on two predominant narratives (deceptive pretexts) to lure their victim into performing a malicious action. The combination of these two results suggests that, for the present moment, these attackers (across dozens of organizations) see more value in opportunistically phishing as many recipients as possible, rather than investing time to mine the hijacked accounts for personalized spearphishing fodder.

Content Tailoring: When analyzing the phishing messages in our dataset, we found that two dimensions aptly characterized the different levels of content tailoring and customization. The first dimension, "Topic tailoring", describes how personalized the topic or main idea of the email is to the victim or organization. The second dimension, "Name tailoring", describes how specifically the attacker addresses the victim (e.g., "Dear user" vs. "Dear Bob"). For each of these two dimensions, we enumerate three different levels of tailoring and provide an anonymized message snippet below; we use Bob to refer to one of the attack's recipients and FooCorp for the company that Bob works at.

1. Topic tailoring: the uniqueness and relevancy of the message's topic to the victim or organization:
   (a) Generic phishing topic: an unspecific message that could be sent to any user ("You have a new shared document available.")
   (b) Broadly enterprise related topic: a message that appears targeted to enterprise environments, but one that would also make sense if the attacker used it at many other organizations ("Updated work schedule. Please distribute to your teams.")
   (c) Targeted topic: a message where the topic clearly relies on specific details about the recipient or organization ("Please see the attached announcement about FooCorp's 25th year anniversary.", where FooCorp has existed for exactly 25 years.)
2. Name tailoring: whether the phishing message specifically uses the recipient or organization's name:
   (a) Non-personalized naming: the attack does not mention the organization or recipient by name ("Dear user, we have detected an error in your mailbox settings...")
   (b) Organization specifically named: the attack mentions just the organization, but not the recipient ("New secure email message from FooCorp...")
   (c) Recipient specifically named: the attack specifically uses the victim's name in the email ("Bob, please review the attached purchase order...")

	Generic	Enterprise	Targeted
No naming	90	35	9
Organization named	23	16	4
Recipient named	0	3	0

Table 4: Distribution of the number of *incidents* per message tailoring category (§ 6.3). The columns correspond to how unique and specific the message's topic pertains to the victim or organization. The rows correspond to whether the phishing email explicitly names the recipient or organization.

Taken together, this taxonomy divides phishing content into nine different classes of tailoring; Table 4 shows how many of our dataset's 180 incidents fall into each category. From this categorization, two interesting observations emerge. First, only 3 incidents (1.7%) actually address their recipients by name. Since most ATOs (94%) in our dataset email at least 100 recipients, attackers would need to leverage some form of automation to both send hundreds of individual emails and customize the naming in each one. Based on our results, it appears these attackers did not view that as a worthwhile investment. For example, they might fear that sending many individual emails might trigger an anti-spam or anti-phishing mechanism, which we observed in the case of one ATO who attempted to send hundreds of individual emails. Second,

[7] When examining and applying thresholds for the Account-agnostic and Organization-wide Attackers, we used a slightly different fraction: how many of the ATO's phishing recipients also appeared in their recent contacts? Here, we seek to capture attackers who make a specific effort to target a considerable number of familiar recipients. Accordingly, we look at the fraction of the ATO's recent contacts that received phishing emails, where the denominator reflects the number of users in the ATO's recent contacts, rather than the ATO's total number of phishing recipients.

Word	# Incidents	Word	# Incidents
document	89	sent	44
view	76	review	43
attach	56	share	37
click	55	account	36
sign	50	access	34

Table 5: Top 10 most common words across all 180 lateral phishing incidents.

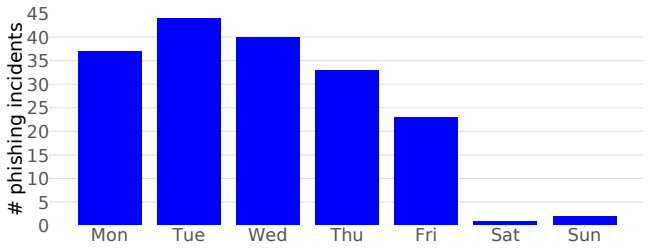

Figure 7: Number of lateral phishing incidents per day of week.

looking at the last column of Table 4, only 13 incidents (7%) use targeted content in their messages. The overwhelming majority (92.7%) of incidents opt for more generic messages that an attacker could deploy at a large number of organizations with minimal changes (e.g., by only changing the name of the victim organization).

While our attack dataset captures a limited view of all lateral phishing attacks, it nonetheless reflects all known lateral phishing incidents across 33 organizations over a 7-month timeframe. Thus, despite the data's limitations, our results show that a substantial fraction of lateral phishers do not fully draw upon their compromised account's resources (i.e., historical emails) to craft personalized spearphishing messages. This finding suggests these attackers act more like an opportunistic cybercriminal, rather than an indomitable APT or nation-state. However, given the arms-race and evolutionary nature of security, these lateral phishers could in the future increase the sophistication and potency of their attacks by drawing upon the account's prior emails to craft more targeted content.

Thematic Content (Lures): When labeling each phishing incident with a level of tailoring, we noticed that the phishing messages in our dataset overwhelmingly relied on one of two deceptive pretexts (lures): (1) an alarming message that asserts some problem with the recipient's account (and urges them to follow a link to remediate the issue); and (2) a message that notifies the recipient of a new / updated / shared document. For the latter 'document' lure, the nature and specificity of the document varied with the level of content tailoring. For example, whereas an attack with generic topic tailoring will just mention a vague document, attacks that use enterprise-related tailoring will switch the terminology to an invoice, purchase order, or some other generic but work-related document.

To characterize this behavior further, we computed the most frequently occurring words across our dataset's phishing messages. First, we selected one phishing email per incident, to prevent incidents with many identical emails from biasing (inflating) the popularity of their lures. Next, we normalized the text of each email: we removed auto-generated text (e.g., user signatures), lowercased all words, removed punctuation, and discarded all non-common English words; all of these can be done with open source libraries such as Talon [27] and NLTK [5]. Finally, we built a set of all words that occurred in any phishing email across our incidents and counted how many incidents each word appeared in.

Interestingly, our dataset's phishing messages draw on a relatively small pool of words: there are just 444 distinct, common English words across the texts of every phishing message in our dataset (i.e., every phishing email's text consists of an arrangement from this set of 444 words). In contrast, a random sample of 1,000 emails from our dataset contained a total of 2,516 distinct words, and only 176 of these emails consisted entirely of words from the phishing term set.

Beyond this small set of total words across lateral phishing emails, all but one incident contained at least one of the top 20 words, illustrating the reliance on the two major lures we identified. Our extended technical report shows the occurrence distribution of each word [17]. Focusing on just the top ten words and the number incidents that use them (Table 5), the dominance of these two thematic lures becomes apparent. Words indicative of the "shared document" lure, such as 'document', 'view', 'attach', and 'review', each occur in over 23% of incidents, with the most popular (document) occurring in nearly half of all incidents. Similarly, we also see many words from the account-related lure in the top ten: 'access', 'sign' (from 'sign on'), and 'account'.

Overall, while our dataset contains several instances of targeted phishing messages, the majority of the lateral phishing emails we observe rely on more mundane lures that an attacker can reuse across multiple organizations with little effort. The fact that we see this behavior recur across dozens of different organizations suggests either the emergence of a new, yet accessible, form of enterprise phishing, or an evolution in the way "ordinary" cybercriminals execute phishing attacks (moving from external accounts that use clever spoofing to compromised, yet legitimate accounts).

6.4 Temporal Aspects of Lateral Phishing

Because attackers might not live or operate in the same geographic region as the hijacked account, prior work has suggested using features that capture unusual timing properties inherent in phishing emails [11, 15, 41]. Contrary to this intuition, in our dataset most lateral phishing attacks occur at "normal" times of the day and week. First, for 98% of lateral phishing incidents, the attacker sent the phishing email dur-

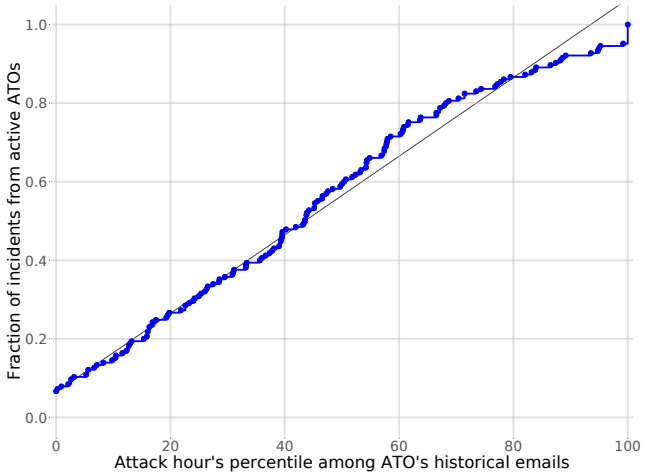

Figure 8: CDF of the fraction of incidents from active ATOs where the time (hour) of day fell within the x'th percentile of the hours at which the ATO's benign emails in the preceding 30 days were sent. Active ATOs are hijacked accounts that sent at least 1 non-phishing email within the 30 days preceding their lateral phishing email.

ing a weekday. Additionally, the majority of attackers in our dataset send their phishing emails during the true account's normal working hours.

Day of the Week: From Figure 7, all but three lateral phishing incidents occurred during a work day (Monday–Friday). This pattern suggests that attackers send their phishing emails on the same days when employees typically send their benign emails, and that the day of the week will provide an ineffective or weak detection signal. Moreover, 67% of incidents occur in the first half of the week (Mon–Wed), indicating that the lateral phishers in our dataset do not follow the folklore strategy where attackers favor launching their attacks on Friday (hoping to capitalize on reduced security team operations over the coming weekend) [37].

Time (Hour) of Day: In addition to operating during the usual work week, most attackers tend to send their lateral phishing emails during the typical working hours of their hijacked accounts. To assess the (ab)normality of an attack's sent-time, for each ATO, we gathered all of the emails that the account sent in the 30 days prior to their first lateral phishing email. We then mapped the sent-time of each of these historical (and presumably benign) emails to the hour-of-day on a 24 hour scale, thus forming a distribution of the typical hour-of-day in which each hijacked account usually sent their emails. Finally, for each lateral phishing incident, we computed the percentile for the phishing email's hour-of-day relative to the hour-of-day distribution for the ATO's historical emails. For example, phishing incidents with a percentile of 0 or 100 were sent at an earlier or later hour-of-day than any email that the true account's owner sent in the preceding 30 days.

Across all lateral phishing incidents sent by an active ATO, Figure 8 shows what hour-of-day percentile the phishing incident's first email occurred at, relative to the hijacked account's historical emails. Out of the 180 incidents, 15 incidents were sent by an "inactive" (quiescent) ATO that sent zero emails across all 30 days preceding their lateral phishing emails; Figure 8 excludes these incidents. Of the remaining 165 incidents sent by an active ATO, 18 incidents fall completely outside of the hijacked account's historical operating hours, which suggests that a feature looking for emails sent at atypical times for a user could help detect these attacks. However, for the remaining 147 incidents, the phishing emails' hour-of-day evenly cover the full percentile range. As shown in Figure 8, the percentile distribution of phishing hours closely resembles the CDF of a uniformly random distribution (a straight $y = x$ line); i.e., the phishing email's hour-of-day appears to be randomly drawn from the true account's historical hour-of-day distribution. This result indicates that for the majority of incidents in our dataset (147 out of 180), the time of day when the ATO sent the attack will not provide a significant signal, since their sent-times mirror the timing distribution of the true user's historical email activity.

Thus, based on the attacks in our dataset, we find that two weak timing-related features exist: searching for quiescent accounts that suddenly begin to send suspicious emails (15 incidents), and searching for suspicious emails sent completely outside of an account's historically active time window (18 incidents). Beyond these two features and the small fraction of phishing attacks they reflect, neither the day of the week nor the time of day provide significant signals for detection.

6.5 Attacker Sophistication

Since most of our dataset's lateral phishers do not mine the hijacked account's mailbox to craft targeted messages, one might naturally conclude that these attackers are lazy or unsophisticated. However, in this subsection, we identify two kinds of sophisticated behavior that required some investment of additional time and manual effort: attackers who continually engage with their attack's recipients in an effort to increase the attack's success rate, and attackers who actively "clean up" traces of their phishing activity in an attempt to hide their presence from the account's legitimate owner. In contrast to the small number of attackers who invested time in crafting tailored phishing messages to a personalized set of recipients, nearly one-third (31%) of attackers engage in at least one of these two sophisticated behaviors.

Interaction with potential victims: Upon receiving a phishing message, some recipients naturally question the email's validity and send a reply asking for more information or assurances. While a lazy attacker might ignore these recipients' replies, 27 ATOs across 15 organizations actively engaged with these potential victims by sending follow-up messages assuring the victim of the phishing email's legitimacy. For example, at one organization, an attacker consistently sent brief follow-up messages such as "Yes I sent it to you" or "Yes,

have you checked it yet?". In other cases, attackers replied with significantly more elaborate ruses: e.g., "Hi [Bob], its a document about [X]. It's safe to open. You can view it by logging in with your email address and password."

To find instances where a phisher actively followed-up with their attack's potential victims, we gathered all of the messages in every lateral phishing email thread and checked to see if the attacker ever received and responded to a recipient's reply (inquiry).[8] In total, we found that 107 ATOs received at least one reply from a recipient. Of these reply-receiving attackers, 27 ATOs (25%) sent a deceptive follow-up response to one or more of their recipients' inquiries.

Stealthiness: Separate from interacting with their potential victims, attackers might expend manual effort to hide their presence from the account's true owner by removing any traces of their phishing emails, particularly since lateral phishers appear to operate during the hijacked account's normal working hours (§ 6.4). To estimate the number of these ATOs, we searched for whether any of the following emails ended up in the hijacked account's Trash folder, and were deleted within 30 seconds of being sent or received: any phishing emails, replies to phishing emails, or follow-up emails sent by the attacker. The 30 second threshold distinguishes stealthy behavior from deletion resulting from remediation of the compromised account. In total, 30 attackers across 16 organizations engage in this kind of evasive clean-up behavior.

Of the 27 ATOs who interactively responded to inquiries about their attack, only 9 also exhibited this stealthy clean-up behavior. Thus, counting the number of attackers across both sets, 48 ATOs engaged in at least one of these behaviors.

The sizeable fraction of attackers who engage in a sophisticated behavior creates a more complex picture of the attacks in our dataset. Given that these attackers do invest dedicated and (often) manual effort in enhancing the success of their attacks, why do so many of them (over 90% in our dataset) use non-targeted phishing content and target dozens to hundreds of recipients? One plausible reason for this generic behavior is that the simple methods they currently use work well enough under their economic model: investing additional time to develop more tailored phishing emails just does not provide enough economic value. Another reason might be that growth of lateral phishing attacks reflects an evolution in the space of phishing, where previously "simple" external phishers have moved to sending their attacks via lateral phishing because attacks from (spoofed) external accounts have become too difficult, due to user awareness and/or better technical mitigations against external phishing. Ultimately, based on our work's dataset, we cannot soundly answer why so many lateral phishers employ simple attacks, and leave it as an interesting question for future work to explore.

[8] Office 365 includes a *ConversationID* field, and all emails in the same thread (the original email and all replies) get assigned the same ConversationID value.

7 Summary

In this work we presented the first large-scale characterization of lateral phishing attacks across more than 100 million employee-sent emails from 92 enterprise organizations. We also developed and evaluated a new detector that found many known lateral phishing attacks, as well as dozens of unreported attacks, while generating a low volume of false positives. Through a detailed analysis of the attacks in our dataset, we uncovered a number of important findings that inform our mental models of the threats enterprises face, and illuminate directions for future defenses. Our work showed that 14% of our randomly sampled organizations, ranging from small to large, experienced lateral phishing attacks within a seven-month time period, and that attackers succeeded in compromising new accounts at least 11% of the time. We uncovered and quantified several thematic recipient targeting strategies and deceptive content narratives; while some attackers engage in targeted attacks, most follow strategies that employ non-personalized phishing attacks that can be readily used across different organizations. Despite this apparent lack of sophistication in tailoring and targeting their attacks, 31% of our dataset's lateral phishers engaged in some form of sophisticated behavior designed to increase their success rate or mask their presence from the hijacked account's true owner. Additionally, over 80% of attacks occurred during a typical working day and hour, relative to the legitimate account's historical emailing behavior; this suggests that these attackers either reside within a similar timezone as the accounts they hijack or make a concerted effort to operate during their victim's normal hours. Ultimately, our work provides the first large-scale insights into an emerging, widespread form of enterprise phishing attacks, and illuminates techniques and future ideas for defending against this potent threat.

Acknowledgements

We thank Itay Bleier, the anonymous reviewers, and our shepherd Gianluca Stringhini for their valuable feedback. This work was supported in part by the Hewlett Foundation through the Center for Long-Term Cybersecurity, NSF grants CNS-1237265 and CNS-1705050, an NSF GRFP Fellowship, the Irwin Mark and Joan Klein Jacobs Chair in Information and Computer Science (UCSD), by generous gifts from Google and Facebook, a Facebook Fellowship, and operational support from the UCSD Center for Networked Systems.

References

[1] Saeed Abu-Nimeh, Dario Nappa, Xinlei Wang, and Suku Nair. A Comparison of Machine Learning Techniques for Phishing Detection. In *Proc. of 2nd ACM eCrime*, 2007.

[2] Kevin Allix, Tegawendé F Bissyandé, Jacques Klein, and Yves Le Traon. Are Your Training Datasets Yet Relevant? In *Proc. of 7th Springer ESSoS*, 2015.

[3] Andre Bergholz, Jeong Ho Chang, Gerhard Paaß, Frank Reichartz, and Siehyun Strobel. Improved Phishing Detection using Model-Based Features. In *Proc. of 5th CEAS*, 2008.

[4] James Bergstra and Yoshua Bengio. Random search for hyper-parameter optimization. *JMLR*, 13(Feb), 2012.

[5] Steven Bird, Edward Loper, and Ewan Klein. Natural Language Toolkit. https://www.nltk.org/, 2019.

[6] 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.

[7] Asaf Cidon. Threat Spotlight: Office 365 Account Takeover — the New "Insider Threat". https://blog.barracuda.com/2017/08/30/threat-spotlight-office-365-account-compromise-the-new-insider-threat/, Aug 2017.

[8] Asaf Cidon, Lior Gavish, Itay Bleier, Nadia Korshun, Marco Schweighauser, and Alexey Tsitkin. High Precision Detection of Business Email Compromise. In *Proc. of 28th Usenix Security*, 2019.

[9] DomainKeys Identified Mail. https://en.wikipedia.org/wiki/DomainKeys_Identified_Mail. Accessed: 2018-11-01.

[10] Sevtap Duman, Kubra Kalkan-Cakmakci, Manuel Egele, William Robertson, and Engin Kirda. EmailProfiler: Spearphishing Filtering with Header and Stylometric Features of Emails. In *Proc. of 40th IEEE COMPSAC*, 2016.

[11] Manuel Egele, Gianluca Stringhini, Christopher Kruegel, and Giovanni Vigna. COMPA: Detecting Compromised Accounts on Social Networks. In *Proc. of 20th ISOC NDSS*, 2013.

[12] FBI. BUSINESS E-MAIL COMPROMISE THE 12 BILLION DOLLAR SCAM, Jul 2018. https://www.ic3.gov/media/2018/180712.aspx.

[13] Ian Fette, Norman Sadeh, and Anthony Tomasic. Learning to Detect Phishing Emails. In *Proc. of 16th ACM WWW*, 2007.

[14] Sujata Garera, Niels Provos, Monica Chew, and Aviel D Rubin. A Framework for Detection and Measurement of Phishing Attacks. In *Proc. of 5th ACM WORM*, 2007.

[15] 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.

[16] Google. Classification: ROC and AUC. https://developers.google.com/machine-learning/crash-course/classification/roc-and-auc, 2019.

[17] 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 (Extended Report). In *arxiv*, 2019.

[18] 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.

[19] Xuan Hu, Banghuai Li, Yang Zhang, Changling Zhou, and Hao Ma. Detecting Compromised Email Accounts from the Perspective of Graph Topology. In *Proc. of 11th ACM CFI*, 2016.

[20] Dan Hubbard. Cisco Umbrella 1 Million. https://umbrella.cisco.com/blog/2016/12/14/cisco-umbrella-1-million/, Dec 2016.

[21] Chris Kanich, Christian Kreibich, Kirill Levchenko, Brandon Enright, Geoffrey M Voelker, Vern Paxson, and Stefan Savage. Spamalytics: An Empirical Analysis of Spam Marketing Conversion. In *Proc. of 15th ACM CCS*, 2008.

[22] Thomas Karagiannis and Milan Vojnovic. Email information flow in large-scale enterprises. Technical report, Microsoft Research, 2008.

[23] Mahmoud Khonji, Youssef Iraqi, and Andrew Jones. Mitigation of spear phishing attacks: A content-based authorship identification framework. In *Proc. of 6th IEEE ICITST*, 2011.

[24] FT Labs. A sobering day. https://labs.ft.com/2013/05/a-sobering-day/?mhq5j=e6, May 2013.

[25] Stevens Le Blond, Cédric Gilbert, Utkarsh Upadhyay, Manuel Gomez Rodriguez, and David Choffnes. A Broad View of the Ecosystem of Socially Engineered Exploit Documents. In *Proc. of 24th ISOC NDSS*, 2017.

[26] Stevens Le Blond, Adina Uritesc, Cédric Gilbert, Zheng Leong Chua, Prateek Saxena, and Engin Kirda. A Look at Targeted Attacks Through the Lense of an NGO. In *Proc. of 23rd USENIX Security*, 2014.

[27] Mailgun Team. Talon. https://github.com/mailgun/talon, 2018.

[28] William R Marczak, John Scott-Railton, Morgan Marquis-Boire, and Vern Paxson. When Governments Hack Opponents: A Look at Actors and Technology. In *Proc. of 23rd USENIX Security*, 2014.

[29] Microsoft Graph: message resource type. https://developer.microsoft.com/en-us/graph/docs/api-reference/v1.0/resources/message. Accessed: 2018-11-01.

[30] Microsoft. People overview - Outlook Web App. https://support.office.com/en-us/article/people-overview-outlook-web-app-5fe173cf-e620-4f62-9bf6-da5041f651bf. Accessed: 2018-11-01.

[31] Brad Miller, Alex Kantchelian, Michael Carl Tschantz, Sadia Afroz, Rekha Bachwani, Riyaz Faizullabhoy, Ling Huang, Vaishaal Shankar, Tony Wu, George Yiu, et al. Reviewer Integration and Performance Measurement for Malware Detection. In *Proc. of 13th Springer DIMVA*, 2016.

[32] Jeremiah Onaolapo, Enrico Mariconti, and Gianluca Stringhini. What Happens After You Are Pwnd: Understanding the Use of Leaked Webmail Credentials in the Wild. In *Proc. of 16th ACM IMC*, 2016.

[33] J. Palme. Common Internet Message Headers. https://tools.ietf.org/html/rfc2076.

[34] Feargus Pendlebury, Fabio Pierazzi, Roberto Jordaney, Johannes Kinder, and Lorenzo Cavallaro. Tesseract: Eliminating experimental bias in malware classification across space and time. In *Proc. of 28th Usenix Security*, 2019.

[35] Kevin Poulsen. Google disrupts chinese spear-phishing attack on senior u.s. officials. https://www.wired.com/2011/06/gmail-hack/, Jul 2011.

[36] Steve Ragan. Office 365 phishing attacks create a sustained insider nightmare for it. https://www.csoonline.com/article/3225469/office-365-phishing-attacks-create-a-sustained-insider-nightmare-for-it.html, Sep 2017.

[37] Fahmida Y. Rashid. Don't like Mondays? Neither do attackers. https://www.csoonline.com/article/3199997/don-t-like-mondays-neither-do-attackers.html, Aug 2017.

[38] Retraining models on new data. https://docs.aws.amazon.com/machine-learning/latest/dg/retraining-models-on-new-data.html, 2019.

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

[40] Apache Spark. PySpark DecisionTreeClassificationModel v2.1.0. http://spark.apache.org/docs/2.1.0/api/python/pyspark.ml.html?highlight=featureimportance#pyspark.ml.classification.DecisionTreeClassificationModel.featureImportances.

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

[42] Kurt Thomas, Frank Li, Chris Grier, and Vern Paxson. Consequences of Connectivity: Characterizing Account Hijacking on Twitter. In *Proc. of 21st ACM CCS*, 2014.

[43] 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-accounts/, Oct 2016.

[44] Colin Whittaker, Brian Ryner, and Marria Nazif. Large-Scale Automatic Classification of Phishing Pages. In *Proc. of 17th ISOC NDSS*, 2010.

[45] Wikipedia. Random forest. https://en.wikipedia.org/wiki/Random_forest, 2019.

[46] Kim Zetter. Researchers uncover rsa phishing attack, hiding in plain sight. https://www.wired.com/2011/08/how-rsa-got-hacked/, Aug 2011.

[47] Mengchen Zhao, Bo An, and Christopher Kiekintveld. Optimizing Personalized Email Filtering Thresholds to Mitigate Sequential Spear Phishing Attacks. In *Proc. of 13th AAAI*, 2016.

A Detector Implementation and Evaluation Details

A.1 Labeling Phishing Emails

Labeling an email as phishing or benign: When manually labeling an email, we started by examining five pieces of information whether the email was a reported phishing incident, the message content, the suspicious URL flagged and if its domain made sense in context, the email's recipients, and the sender. With the exception of a few incidents, we could easily identify a phishing email from the above steps. For example: an email about a "shared Office 365 document" sent to hundreds of unrelated recipients and whose document link pointed to a bit.ly shortened [non-Microsoft] domain; or an

email describing an "account security problem" sent by a non-IT employee, where the "account reset" URL pointed to an unrelated domain. For the more difficult cases, we analyzed all replies and forwards in the email chain, and labeled the email as phishing if it either received multiple replies / forwards that expressed alarm or suspicious, or if the hijacked account eventually sent a reply saying that they did not send the phishing email. Finally, as described in Section 3.3 we visited the non-side-effect, suspicious URLs from a sample of the labeled phishing emails; All of the URLs we visited led to either an interstitial warning page (e.g.,, Google SafeBrowsing), or a spoofed log-on page. For the emails flagged by our detector, but which appeared benign based on examining all the above information, we conservatively labeled them as false positives. In many cases, false positives were readily apparent; e.g., emails where the "suspicious URL" flagged by our detector occurred in the sender's signature and linked to their personal website.

Training exercises vs. actual phishing emails: In addition to distinguishing between a false positive and an attack, we checked to ensure that our lateral phishing incidents represented actual attacks, and not training exercises. First, based on the lateral phishing emails' headers, we verified that all of the sending accounts were legitimate enterprise accounts. Second, all but five of the attack accounts sent one or more unrelated-to-phishing emails in the preceding month. These two points gave us confidence that the phishing emails came from existing, legitimate accounts, and thus represented actual attacks; i.e., training exercises will not hijack an existing account, due to the potential reputational harm this could incur (and enterprise security teams we've previously engaged with do not do this). Furthermore, none of our dataset's lateral phishing incidents are training exercises known to Barracuda, and none of the lateral phishing URLs used domains of known security companies.

A.2 Model Tuning and Hyperparameters

Most machine learning models, including Random Forest, require the user to set various (hyper)parameters that govern the model's training process. To determine the optimal set of hyperparameters for our classifier, we followed machine learning best practices by conducting a three-fold cross-validation grid search over all combinations of the hyperparameters listed below [4].

1. Number of trees: 50–500, in steps of 50 (i.e., 50, 100, 150, . . . , 450, 500)
2. Maximum tree depth: 10–100, in steps of 10
3. Minimum leaf size: 1, 2, 4, 8
4. Downsampling ratio of (benign / attack) emails: 10, 50, 100, 200

Because our training dataset contained only a few dozen incidents, we used three folds to ensure that each fold in the cross-validation contained several attack instances. Our experiments used a Random Forest model with 64 trees, a maximum depth of 8, a minimum leaf size of 4 elements, and a downsampling of 200 benign emails per 1 attack email, since this configuration produced the the highest AUC score [16]. But we note that many of the hyperparameter combinations yielded similar results.