Evaluating the Effectiveness of Phishing Reports on Twitter

Sayak Saha Roy
The University of Texas at Arlington
sayak.saharoy@mavs.uta.edu

Unique Karanjit
The University of Texas at Arlington
unique.karanjit@mavs.uta.edu

Shirin Nilizadeh
The University of Texas at Arlington
shirin.nilizadeh@uta.edu

2021 APWG Symposium on Electronic Crime Research (eCrime) | 978-1-6654-8029-1/21/$31.00 ©2021 IEEE | DOI: 10.1109/eCrime54498.2021.9738786

***Abstract*—Phishing attacks are an increasingly potent web-based threat, with nearly 1.5 million such websites being created on a monthly basis. In this work, we present the first study towards identifying phishing attacks through reports shared by security conscious users on Twitter. We evaluated over 16.4k such reports posted by 701 Twitter accounts between June to August 2021, which contained 11.1k unique URLs, and analyzed their effectiveness using various quantitative and qualitative measures. Our findings indicate that not only these reports share a high volume of legitimate phishing URLs, but they also contain more information regarding the phishing websites (which can expedite the process of identifying and removing these threats), when compared to two popular open-source phishing feeds: PhishTank and OpenPhish. We also noticed that the URLs in the Twitter reports had very little overlap with the URLs found on PhishTank and OpenPhish, and also remained active for longer periods of time. However, despite having these attributes, we found that these reports have very low interaction from other users on Twitter, especially from the domains and organizations which were targeted by the reported URLs. Moreover, nearly 31% of these URLs were still active even after a week of them being reported while also being detected by very few anti-phishing tools. This suggests that a large majority of these reports remain undiscovered and underutilized. Thus, this work highlights the utility of phishing reports shared on Twitter, and the benefits of using them as an open source knowledge base for identifying new phishing websites.**

I. Introduction

Phishing websites are web-based social engineering attacks which often impersonate legitimate organizations to trick users into sharing their personal information. Their volume has significantly increased over the past few years [1], [2], encouraging a prolonged effort from the security community towards detecting these attacks, which have manifested in the form of automated tools using sophisticated machine learning [3]–[6], deep-learning [7], [8], rule-based/heuristic based techniques [9]–[11]. The phishing websites found in the wild using these approaches, along with several others (including manual reporting of the attacks) are often consolidated into dedicated knowledgebases known as phishing blocklist feeds. Information from these blocklists (which are frequently updated) is then used by anti-phishing tools, website hosting registrars and commercial organizations to prevent end-users from accessing malicious websites through their systems. However, as noted by Oest et al. [12], phishing blocklists are *reactive* in nature, with a considerable time gap between the appearance of a website in the wild and it subsequently appearing in one of these blocklists. This is further exacerbated by the increase of newer variants of phishing threats which are highly elusive in nature, often evolving to leverage several loopholes and adversarial tactics to circumnavigate automated tool detection [13]–[16] and are able to prevent themselves from appearing in phishing blocklists for long periods of time. Thus, there is a need for identifying outlets outside of traditional blocklisting efforts, which can not only recognize newer and more elusive forms of phishing threats, but also provide more context-aware information about them, to expedite the blocklisting process.

Thus to further this effort, this work investigates phishing reports which are shared on Twitter [17], the popular micro-blogging platform. To the best of our knowledge, our work constitutes the first study on evaluating Twitter as a formidable blocklisting knowledge-base for identifying new phishing websites. Throughout the course of this paper, we concentrate on finding the effectiveness of phishing reports shared on Twitter and compare their characteristics and performance with two other popular open source phishing feeds - PhishTank and OpenPhish. More specifically, this paper: (i) determines the reliability and volume of information shared by these phishing reports, and how they compare against PhishTank and OpenPhish; (ii) being hosted on Twitter, these reports can also be visibly interacted upon by other users on the platform. This paper, thus, evaluates the frequency of these interactions, including those coming from domain registrars (which host the reported phishing websites) and organizations (which the reported phishing URLs have targeted), and examine the impacts of these interactions on the detection/removal of the reported URLs; and (iii) it also determines what happens after users sharing these reports on Twitter, i.e., how long the URLs remain active after getting reported, as well as how quickly anti-phishing tools detect them. Both are factors which can protect the user from inadvertently accessing the threat.

We collected and analyzed more than 16k tweets which contained 11k unique URLs, over the period of 21st June to 17th August 2021. Using a combination of automated and manual investigations, we repeatedly tracked several properties that were extracted from these phishing reports which included checking the activity of these URLs, anti-phishing tool detection, information shared by these posts (such as relevant hashtags, images, etc.), true positive rate (percentage of URLs which were legitimate phishing attacks), interactions with

978-1-6654-8029-1/21/$31.00 ©2021 IEEE

other users (likes, comments and replies). We also compared the relevant statistics with two other phishing feeds- PhishTank and OpenPhish.

In Section III, we underline how we collected the phishing reports from Twitter, and also illustrate the distribution of the phishing websites across several hosting domains. In Section III-C, we discuss how phishing attacks covered by these reports distributed Drive-by downloads, a family of threats which are rarely covered by PhishTank and OpenPhish. In Section IV we evaluate the information shared by these phishing reports (IV-A) such as website screenshots, IP address, name of targeted registrar/ organization, labelling of threats, etc., and compare them with data shared by PhishTank (IV-A1) and OpenPhish (IV-B), which we later also summarize in **Table I**. We also report on the reliability and validity of these Twitter phishing reports in Section IV-C. Unlike PhishTank and OpenPhish, where user interaction with the feed is limited, phishing reports shared on Twitter has a huge scope of being received by the digital community. Thus, in section V, we find the volume of interactions (favourites/ retweets/ comments) that these reports get on Twitter, and whether interaction from the targeted domain/organization has an impact on how quickly these reported websites are removed. We also qualitatively explore how these interactions look like (Fig 6) and determine the technological proficiency of users who typically interact with these reports (V-B). Finally, in Section VI, we determine how long these URLs stay active after being reported, and how the rate (and pace) of removal compares with URLs which are posted on PhishTank and OpenPhish (Section IV). We also check for the coverage of the phishing URLs by anti-phishing engines VI-C. Our main findings can be summarized as below:

1) Twitter is a viable candidate for being utilized as a knowledge-base for phishing reports. Over the course of three months, users consistently shared over 16.4k such reports which covered 11.1k unique URLs hosted over 203 unique registrars, and targeted 146 different organizations. Unlike PhishTank and OpenPhish, these accounts also reported URLs distributing Drive-by downloads (7%).
2) The majority of the phishing reports shared on Twitter contained more information compared to PhishTank and OpenPhish which can help domain registrars and anti-phishing tools to expedite the process of identifying these attacks.
3) The majority of the phishing reports posted on Twitter were found to be accurate, with the URLs shared in these reports having a high true positive rate (87%), with only one account contributing to the majority (11%) of the false positives in our dataset.
4) These reports receive very low engagement, with only 13.8% of the posts receiving at least one comment. The domain registrars and organizations which the reported URLs targeted (referred to as *targets* henceforth) contributed to only 4% of these comments. However, when these entities did respond to the reports, the URLs became inactive more quickly when compared to the URLs which did not receive such interaction. Moreover, only 10.2% of the targets follow at least one such Phishing report account, indicating that they are either not aware of these reporting accounts, or do not consider them as a credible source.
5) About 31% of the URLs, which were reported on Twitter, remained active even after a week of their first appearance in our dataset. Moreover, anti-phishing tools consistently had lower detection rates for these URLs when compared to URLs which showed up on PhishTank and OpenPhish.

Thus, our evaluation indicates that phishing reports shared on Twitter are a reliable and efficient source for conveying information regarding new phishing websites. Using these reports can help domain registrars and anti-phishing tools in expediting the process of identifying these threats. However, the present scenario indicates that the targeted entities seldom interact with these reports. and thus, our contribution through this work aims to raise awareness towards the existence of this information-rich phishing knowledge-base on Twitter and hopefully accelerate its integration with prevalent traditional blocklisting efforts. Additionally, based on the volume of information shared by these reports, it proves to be a valuable resource for researchers in building detailed ground truth datasets with less effort and more efficiency compared to other open phishing feeds like PhishTank and OpenPhish. We explore our findings in broader details from the proceeding section on-wards.

II. Background and Related work

Phishing website detection: Based on recent measurements, nearly 1.4 million phishing websites are created every month [18]. As highlighted by Vayansky and Kumar [19], unlike malware threats, the success of a phishing attacks is largely dependent on human interaction factors. However, several qualitative and quantitative studies have found that end-users are not proficient at identifying phishing websites [20], [21]. Additionally, phishing attacks often evolve to tailor themselves based on socio-economic disasters such as the COVID-19 pandemic [22], [23] and also leverage several adversarial tactics to circumnavigate automated tool detection [13], [15], [16]. Thus, domain hosting services, antiphishing tools and web-browsers often rely on one or more phishing feeds/blocklists, which are knowledge-bases containing frequently updated lists of new phishing URLs. These URLs are either manually annotated by security conscious individuals or discovered using crawlers which leverage several automated approaches, such as machine learning and deep learning based models [3], [4] and heuristic or rule-based implementations [9], [11]. Our work in this paper focuses on one such knowledge-base distributed across Twitter [17], the micro-blogging platform, to determine its efficiency and how it compares with two popular and open-source phishing feeds - PhishTank [24] and Openphish [25].

Effectiveness of Phishing feeds: These are specialized feeds dedicated towards keeping track of new phishing threats which are distributed across the web. These feeds can be either closed (proprietary) or open-source in nature. In this work we focus on comparing phishing reports shared on Twitter with two feeds belonging to the open source category - PhishTank (PT) and OpenPhish (OP). Despite the utility of these open-source phishing feeds, academic research on them is limited. Even then, previous work has highlighted several pitfalls that these phishing feeds suffer from. For example, Sheng et al. [26] noted how these feeds had a very low efficiency at identifying newer threats at hour zero(the time when phishing threats are at their most potent state) and continued to have a low coverage several hours after that. Bell et al. [27] notes that Phishtank and OpenPhish have very few URLs overlapping, suggesting that using them collectively can help in covering a larger volume of these threats. In this work we determine the volume of URLs that are reported exclusively by phishing reports posted on Twitter, and the need for using it as an anti-phishing knowledge-base. Moreover, Moore et al. [28] points out that a significant number of URLs posted on PhishTank, one of the most popular community driven phishing feed, are false positives which might be contributed by malicious actors to deliberately poison the feed and make it an unreliable resource . Several of these phishing feeds also report very minimal information about the threats that they report, which was noted by Oest et al. [12], who suggested using an *evidence-based* phishing reporting feed containing additional artifacts such as screenshots can expedite the process of detecting and removing the threats. Considering that these feeds are being used exhaustively by several web browsers [29], [30], as well as anti-phishing tools and organizations [31], these shortcomings can inadvertently impact the protection that is offered to end-users. Thus, in the process of critically evaluating the effectiveness of phishing reports shared on Twitter, we also find the scope for discussing the shortcomings of PhishTank and OpenPhish towards their reliability and coverage of phishing threats.

III. Data collection and characterisation

To automate the process of collecting tweets containing phishing reports, we initially looked for 500 random tweets which contain such reports and qualitatively analysed them to identified the common keywords and features that they contained. We found that the majority of such tweets report the URLs in an obfuscated format, usually replacing 'http'/'https' with case insensitive variants of 'hxxp/hxxps.' This strategy is popularly known as *'URL defanging'* [32], and is used to prevent users from accidentally visiting the malicious link. However, in some cases, other parts of the URL are also defanged, for example, *http://abc[.].com*, but they were usually accompanied with the hashtags *#phishing* and *#scam*. Thus, to populated our dataset, we collected tweets using the Twitter API [33] using the keywords *'hxxp'* and *'hxxps'* as well as the hashtags *#phishing* and *#scam*. We then used a regular expression which reverses the defanging by replacing the obfuscating characters from the URL to make them usable for our experiments.

We thus utilized this data collection approach to collect new phishing reports every 30 mins from the period of June 21st to August 17th 2021, getting 16,486 tweets, which contained 11,139 unique URLs, posted by 701 unique reporters in the process. During each 30 min period several companion processes were also run, including tracking whether the URL was active, checking if the URL was present in the phishing feeds provided by PhishTank and Openphish, and also tracking how many anti-phishing engines detected the URL by using VirusTotal [34]. VirusTotal [35] is an online URL scanning tool which scans URLs using 80 different anti-phishing engines, and returns an aggregated total of the engines which detected the URL as malicious. It is used frequently by researchers to create a ground-truth of malicious URLs [36]–[39]. These companion processes enabled us to get a full picture of how domain registrars, OpenPhish, PhishTank and antiphishing engines reacted to the URLs which were shared by these reports.

Additionally, to evaluate the amount of information shared by the phishing reports, as well as their efficiency (i.e., if the URLs shared were actually phishing websites), we also collected the screenshots of both the tweets as well as the website that they had reported. To study how other users on Twitter engaged with these reports, we collected a snapshot of all interactions towards these tweets at end of every day. The data which was collected in this scenario included the comments posted on these tweets, as well as the user ids of the individuals who liked and retweeted them.

A. Distribution of websites across registrars and targeted organizations

We used WHOIS [40] to determine the hosting records of the 11,139 unique URLs found in the phishing reports, and found that they were distributed across 203 unique domains. Moreover, 5% (n=631) of the URLs consisted of an adversarial threat category highlighted in work by Saha Roy et al. [15]. These URLs leverage the use of popular free web-hosting domains (which are often white-listed by anti-phishing vendors and phishing feeds alike) to host phishing websites, and in turn remain active for a longer time compared to traditional phishing threats, while also evading detection by several anti-phishing engines. Overall, around 52% of the phishing reports used hashtags to refer to the names of the domains or organization targeted by the URLs. Hashtags are widely used to define a shared context for specific events or topics [41], and we assume that the reporters use them to: (a) inform the domain registrar service and organization that the reported website is phishing and should be investigated, and (b) inform other users about where the website is being hosted and/or which brand or organization it is targeting, We explore the other informational attributes shared by these reports (such as screenshots of the URL, threat category, location, etc.), and how they compare with PhishTank and Openphish in Section IV. Unlike other phishing feeds, the reports on Twitter

can be interacted upon by other users in the Twitter community, including accounts maintained by the domain registrars and organizations which are targeted by the reported URL. Thus, in Section V, we explore the responsiveness of these aforementioned parties towards the post, and how it affects the activity of the respective phishing URLs. Figure 1 illustrates the distribution of the URLs across different registrars and drive-by download categories (n=11,139). We find that large amounts of these reported URLs are hosted across popular domain registrars such as GoDaddy, Namecheap, Namesilo, Public domain registry etc. This indicates that these posts are not focused on a particular registrar/ group of registrar, but cover URLs from several sources. Similarly, these reports also cover URLs which host a wide range of file-based threats ranging from Trojan horses, infected PDFs and malicious APKs. We explore the distribution of these threats in Section III-C.

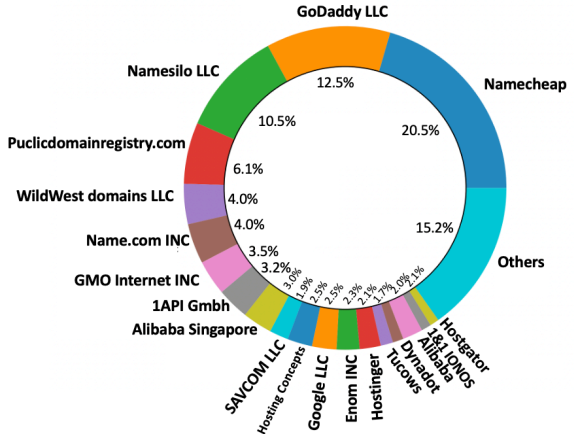

Fig. 1. Distribution of URLs across different registrars

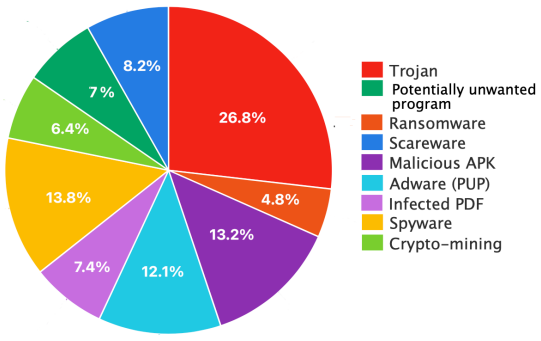

Fig. 2. Distribution of Drive-by download URLs found in the phishing reports *(n=829)*

B. Distribution of phishing reporters

Phishing report tweets in our dataset were posted by 701 accounts. Interestingly, one account posted more than 48.2% of URLs in our dataset (n=7,946 tweets), 25 accounts posted more than 100 such tweets, and 21 accounts posting more than 50 tweets. Due to only one user contributing to such a large portion of the tweets, we report our findings by both considering and not considering this one user (whom we refer to as *top reporter* henceforth) separately. Also, 65% of the users in our dataset shared only one tweet. Infact, the distribution of the posts contributed by these accounts is heavily skewed towards some particular users as illustrated in Figure 3. But, our goal is to not concentrate on any one user, but instead investigate the content shared by all these accounts as a form of distributed knowledge-base and determine the reliability of information provided by these reports and if it can benefit the identification of new phishing threats.

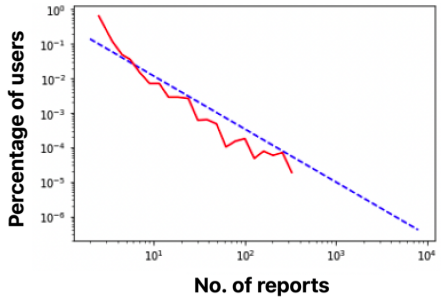

Fig. 3. Distribution of the volume of reports shared by the users.

C. Distribution of drive-by download websites

While phishing URLs leverage social engineering techniques using various persuasion tactics such as authority and distraction to deceive the users into sharing their private information [42], websites distributing drive by downloads contain an additional attack vector which can further steal sensitive information by installing malicious files or applications in the user's system and exploiting critical vulnerabilities [43]. We found 829 unique URLs shared over 902 phishing reports which distributed drive-by downloads. We monitored file downloads which were triggered automatically by visiting the URLs in our dataset, and the downloaded files were then scanned using VirusTotal, and were labelled as *Drive-by download* only if those files were detected by at least two different engines (a threshold considered as a standard for labelling malicious files in both the industry and research communities alike [44]). We then distributed these files equally between our team of four security researchers, who executed each of them in a secure VM environment, and based on their characteristics, each file was assigned a label indicative of the threat family that they belonged to. We adhered to the threat family labels mentioned in Cisco's Cyber-security Trend report [45], but also added two more categories which were distinctly present in our dataset: Malicious APK (Android based malware), and Infected PDFs. Figure 2 breaks down the distribution of the malicious files across 9 different threat families. We find that 26.8% of these drive-by downloads distributed trojan horses [46], malicious files disguised as

legitimate software. A good portion (13.2%) of the URLs distributed Android based malware [47], which ranged from apps which attempted to send premium text messages, showing intrusive advertisements, trying to gain access to system resources etc. We also find cryto-mining malware files, or *crypto-jacking* attacks (6.4%), which tend to use large amount of system resources to illicitly mine crypto-currency for the attackers gain [48]. 13.8% of files also consisted of Spyware attacks, ranging from Browser hijackers [49] to device keyloggers. Also present were Scarewares and Ransomwares which restrict/deny access to system resources and/or personal data and then encourage the victims into sharing their private information to regain access. Thus, it is evident that the Drive-by download URLs shared by the phishing reports on Twitter cover a large array of threat families. In Section IV, we further look into the coverage of drive-by download URLs by PhishTank and OpenPhish.

IV. Comparison with other phishing feeds

In this section, we determine the characteristics and volume of information shared by the phishing reports posted on Twitter, and also compare those attributes with the URLs which are found on PhishTank and OpenPhish in Sections IV-A1 and IV-B respectively. We further extend the comparison in Section IV-D by looking at what portion of URLs overlap between the phishing reports on Twitter with those on PhishTank and OpenPhish. Finally, in IV-C, we use sophisticated machine learning tools as well as qualitative analysis to determine what portion of URLs shared by the reporters on Twitter are legitimate(true-positives).

A. Content shared by phishing reports

Using regular expressions, as well as extracting the hashtags from these tweets, we were able to analyze the content presented by these reports. Overall, we found that phishing reports posted on Twitter shared much more information than just the URL of the offending website. These included the IP address (31%), hosting registrar (52%), targeted organization (47%), the threat category of the URL (for example phishing, scam or malware - 36%) as well as a full image (23.5%) of the phishing website. Figure 4 provides examples for two such reports and highlights the information shared by them. Without considering the *top reporter*, these statistics increased considerably, with 44% of the phishing reports sharing IP addresses, 61% and 53% sharing hosting registrar and domain targets respectively, 28% sharing full images of the websites and 42% sharing the threat category of the website. This indicates that the tweets shared by the *top poster* often contain less information compared to other reporting accounts. We now consider each of the features (IP address, hosting registrar, targeted organization, etc.) that we have identified from these reports and determine if the other phishing feeds- PhishTank and OpenPhish provide similar information:

1) PhishTank: PhishTank allows any individual to add URLs to their feed, which can then be verified by other users on the website. It does not provide any information about

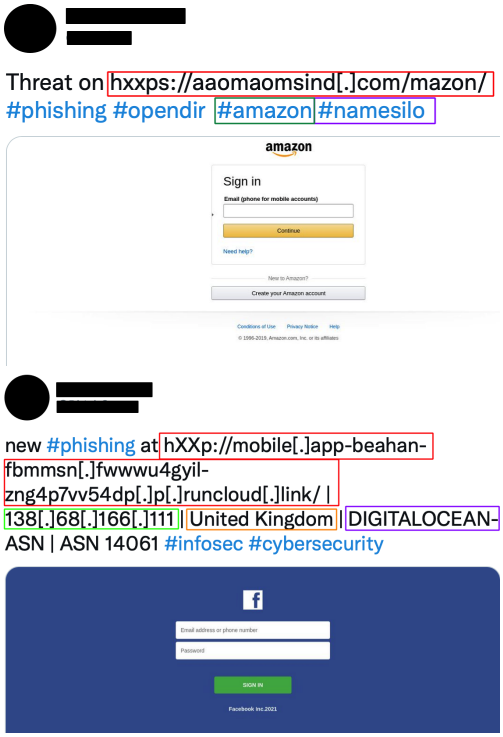

Fig. 4. Examples of information shared by phishing reports posted by Twitter accounts. **(On Top)** Report highlights the phishing URL (in red), the targeted organization (in green), and the hosting registrar (in purple). **On Bottom** In addition to the URL (in red) and registrar hosting (in purple), also shows the Location where the website originated from (in orange), as well as the IP address (in green)

which registrar hosted the website nor the IP address of the URLs, and relies only on user submission to populate it's feed. A valid submission only requires the user to provide the URL to be reported and then select the targeted organization from a list of options (with *Other* being a valid option for targets that are not present in the list). They can also provide an open ended response to indicate the contents of the phishing page/email. However this information does not appear anywhere on the feed. Downloading the comprehensive PhishTank feed (which at the time of our collection contained 10,622 URLs which were already verified as being phishing by the PhishTank community), we found that nearly 85% of URLs contained the *Other* label under targeted organization, thus providing no conclusive information regarding the organization that the phishing URL had targeted.

The feed also provided data about when the URL first appeared on PhishTank and when it was verified(the median verification time was around 12.96 minutes). However, as mentioned earlier, PhishTank's downloadable feed only provides URLs which have already been verified by their community of users. Thus, to determine the efficacy of the live feed (which also contains unverified URLs), we first monitored 1k new URLs taken from PhishTank to check what percentage of them are verified. Then through continued observation of these URLs, we found that the PhishTank community verified

nearly 724 of these URLs with a median verification time of 11.49 mins, marking 639 of them as phishing (VALID), and 85 of them as benign (INVALID). However, among the remaining 276 URLs, we found 119 of them to be phishing websites, and we could not observe 37 of them because they were already inactive. Interestingly, among the phishing websites which remained unverified, 53 of them seemed to originate from unconventional phishing domains [15], a family of phishing threats which use adversarial tactics to prevent detection by both registrars and anti-phishing engines alike. We verified the remaining 120 URLs as false positives, which were added to the 85 URLs that had been labelled by the verifiers as being *INVALID*. Thus, we find that these 1k URLs had a false positive rate of 20.5%. Considering that researchers often rely on the live PhishTank feed as a viable source for collecting and analyzing phishing URLs [50], this rate of false positives might add significant noise to their datasets. Also, PhishTank takes screenshots automatically when the URL is submitted (PhishTank does not ask the submitter to provide this information during submission), and if said website is already down then these screenshots do not contribute any useful information towards the appearance of these websites. We found that 29% of the URLs on PhishTank had screenshots which indicated that the website was already inactive before submission, a phenomenon we investigate more closely in section IV-D1. Moreover, PhishTank provides a label which indicates whether a URL is *Online*. However, by checking the labels of these 1k URLs, we found that for 33% of them, PhishTank provides incorrect information about the activity of the websites (It incorrectly identifies a website is **Online** when it is actually **Offline** or vice versa). Thus, using this indicator might provide incorrect information about the activeness of these phishing threats.

B. OpenPhish

Similar to PhishTank, we focus on 1k URLs which are collected from the OpenPhish feed. From these URLs, we found that about 39% of the URLs provided hosting registrar information, and 23% provided the IP address of the URLs. We also noticed that 74% of the URLs identified a relevant targeted organization. OpenPhish also reports when a URL first appeared in their feed. To the best of our knowledge, Openphish does not report the screenshot of the webpage, neither through their website, nor through their API access. Unlike Phishtank, Openphish does not report on the activity of the URL as well, i.e., whether a URL is online or not. Also it does not identify the category of the threat of the URL. How OpenPhish obtains URLs is ambiguous, as they on their FAQ page - *"OpenPhish receives millions of unfiltered URLs from a variety of sources on its global partner network."* [25]. However, we assume that these *partners* are curated by OpenPhish themselves, and thus the URL submissions might be more reliable than the open ended anonymous submission approach implemented by PhishTank. This is further corroborated by the low false positive rate of these submissions, as we identified only 41 (out of 1k URLs) which were incorrectly marked as being phishing. Later on, we sample from this set of URLs in Sections VI to track the activity of the reported URLs, as well as how quickly they are detected by anti-phishing engines, and how it compares to phishing reports provided by Twitter accounts.

C. Validity of URLs shared by phishing reporters

We have established that different phishing feeds share different volumes and variations of information and illustrated how they compare to the Twitter phishing reports. However, since both researchers and industrial entities rely on these feeds in some capacity, one of the most important aspect of these reports are the validity of the URLs that they share. In the previous section we have already determined that PhishTank and OpenPhish have a false positive rate of 20.5% and 4.1% respectively, based on our investigation of 1k URLs collected randomly from these feeds. In this section we evaluate the validity of URLs shared by the phishing reports posted on Twitter. We evaluated the false positive rate of the URLs found in the Twitter phishing reports by scanning these URLs on VirusTotal, as well as using manual observation and applying an ensemble machine learning approach. We report the methodology and findings of our evaluation below:

We used VirusTotal as an initial filter to reduce the number of phishing websites needed for manual evaluation. For URLs which had at least 2 detections a day after their appearance in our dataset, we marked them as *true positives*. We found nearly 31% of the tweets (n=5,109) containing 3,827 unique URLs which did not reach this threshold. Manually labelling such a large volume of URLs is not practical, and thus we used two machine learning based implementations, one being a tool developed by Papernot et al. [51] trained on UCI's Phishing Website Dataset (Mustafa et al. [52]), and the other being Sharkcop [53] [54] to automatically label these URLs. The two different tools were used together for consensus, i.e., a URL was only considered as phishing if both the tools detected as phishing. To gauge the effectiveness of these tools, we manually observed 200 URLs from our dataset and observed an accuracy of 94% from our setup. Any URL where the tools had disparate labels were put aside for manual labelling. In this way our setup was able to mark 2,464 URLs, among which it detected 1,619 URLs as phishing and 845 URLs as benign. The remaining 1,363 URLs (for which both the tools were not able to reach a con-census) were labelled by 4 independent coders. To make sure the coders did not directly interact with URLs which were potential phishing attacks, we provided them with screenshots of for each website, which also contained its URL. The coders verified 824 URLs as phishing and 539 URLs as benign. Thus, for URLs which had less than 2 detections on VirusTotal, we found 2,443 URLs to be phishing and 1,384 URLs to be benign. In total, we found 9,755 URLs to be phishing (87% of all unique URLs) which were shared by 15,241 tweets.

Therefore, it can be established that the URLs reported by these phishing reporters have a high true positive rate.

Functionality	Twitter Phishing Reports	PhishTank	OpenPhish
Submission method	Self-submission Self-verification	Community submission Community verification	Partner submission Self verification
Hosting Registrar	52% [a]/ 61%[wt]	No	39%
Targeted organization	47% [a]/ 53%[wt]	15%	74%
IP address	31% [a]/ 44%[wt]	No	23%
Screenshot shared	23% [a]/ 28% [wt]	71%	No
Threat type identified	36% [a]/ 42%[a]	No	No
Drive-by Downloads	8%	No	No
URL Activity status	No	Yes, but error rate of 33%	No
Dead on arrival rate	3.8%	24.2%	11.4%
Overlap with Twitter Reports	N/A	4%	13%
False positive rate	11% [a]/ 6% [a]	20.5%	4.1%

TABLE I: Summarizing the information shared by **Twitter Phishing Reports, PhishTank and OpenPhish a**=Respective stats of all Twitter reports including those from *Top poster*, and **wt=** Respective stats for all Twitter reports excluding those of top poster

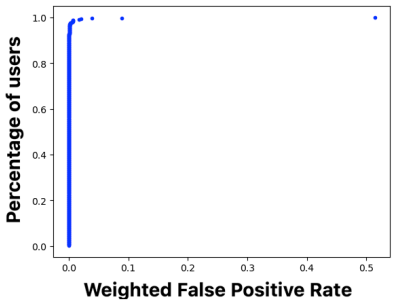

Fig. 5. CDF of the weighted false positive rate for each user

However, our dataset is skewed towards the *top reporter* who contributed to 48.2% of the tweets to our dataset. Interestingly, we found that out of 1,384 benign URLs, 712 URLs were posted by this user alone, which constitutes 11.3% of all unique URLs posted by this account (n=6,258, out of which 4,188 URLs were unique). As mentioned in Section IV-A, we found that the *top reporter* shares fewer details in their reports, compared to other users. Since the distribution of our dataset with respect to tweets shared by the reporters is non-uniform, with a large number of users only sharing one post, we construct a cumulative distribution of the *weighted* false positive rate based on how many phishing reports each user shared versus how many of these shared reports contained URLs which were false positives. We illustrate the distribution in Figure 5. We find that the *top reporter* is one of two outliers in the distribution, with the other user contributing to 10% of the false positives URLs found in our dataset. However, both these users also have a high true positive rate of 91% and 88%, respectively. Outside of these two outliers, most users have a *false positive* rate of less than 1%. Thus, the majority of these reporters are more reliable than PhishTank and OpenPhish with respect to the validity of the URLs that they report.

D. Comparison of other attributes

1) Dead on arrival rate (DoA): We identify a URL as being *dead on arrival* when said URL is already inactive when it first appears on a phishing report/feed. We randomly selected 1k URLs from our Twitter phishing report dataset, and checked how many of them were inactive when they first appeared in a report. We found only 3.8% of URLs posted by phishing reporters on Twitter exhibit this behaviour. In comparison, using the 1k URLs that we had already selected from Phishtank and OpenPhish in sections IV-A1 and IV-B respectively, we found 24.2% of those URLs on Phishtank *are dead on arrival*. This statistic was 11.4% for Openphish.

This indicates that URLs when posted on Twitter reports are more likely to be alive, and thus need immediate intervention from the targeted registrars and organizations.

2) Overlap between reported URLs and other phishing feeds: We checked if each of Twitter phishing report URL which was determined as true positive in Section IV-C were also available on OpenPhish and PhishTank using their respective APIs. Prior literature [27] has noted that URLs might keep appearing and disappearing from these phishing feeds based on if they are still active or not. Thus, we keep checking for the URLs in both OpenPhish and PhishTank every 30 minutes till after a week of their first appearance in their respective feeds. We find that a low number of URLs overlap with entries on Openphish and PhishTank, with the former having only 13% of URLs overlapping with the Twitter phishing reports, and the later a mere 4%. This indicates that a lot of true positive URLs posted by the phishing reporters on Twitter do not appear in either PhishTank or OpenPhish. Interestingly, 5.8% of the overlapping URLs that showed up in Openphish did so at a median time of 6 hours after being posted on Twitter. The same statistic stands at 1.3% for Phishtank. Considering that domain registrars and even anti-phishing engines often rely (at least partially) on these phishing feeds to identify URLs, failing to cover a large percentage of newer phishing attacks that we found in Twitter reports can be detrimental for user

Feature	Type	Min	Max	Mean	Median
Followers	Count	0	127,692	1703.78	472
Posts	Count	1	7,958	23.61	1
Likes	Count	0	205	0.45	0
Retweets	Count	0	161	2.07	0
Listed count	Count	0	6,770	83.77	7
Age (in days)	Count	42	5,298	2,325.66	2,129
Detections	Count	0	23	4.09	2
Verified	Boolean	Total accounts = 15			

TABLE II: Descriptive statistics of Twitter accounts who shared phishing reports.

protection. Thus our findings indicate that the phishing reports are an untapped resource for quickly acquiring a vast breadth of information about new phishing websites when compared to PhishTank and Openphish. We summarize the functionalities exhibited by the reports from each of these phishing feeds in Table I. In the next section, we determine how other users on Twitter interact with these phishing reports.

V. Phishing Report Interactions

We collected comments posted on each of the Twitter phishing report tweets, and found that only 2,285 of them got at least one reply , which is only around 14% of the total number of tweets in our dataset. Moreover, very few of these interactions came from the hosting registrars or the targeted organizations (752 out of 2285 conversations with at least 1 reply, 4% overall). This is despite the fact that 55.2% of these reports contain a hashtag citing these concerned services. We see that for services which were tagged in more than 100 phishing reports, only 2 of them were able to reply to about 30% of the tweets that they were tagged in, with 5 targets not replying to any of these tweets. Figure 6 illustrates the cumulative distribution of *explicit* interactions (comments) by the registrars/ targeted organizations with the Twitter phishing reports, which indicates that these services (targets) have indeed have very low interaction with these reports, despite these reports containing URLs which have a high chance of being true positives. We noticed that the median time for getting a reply from the domain registrars is 103 minutes, whereas the same from targeted organizations is is 171 minutes. We term this form of interaction from the registrars/targets as *explicit interaction*, because in these cases, we can say for sure that the target has noticed the report. Later on in Section VI we explore how this interaction influences the pace at which these reported websites go offline, and how it compares to reports which do not receive any explicit interaction.

A. Likes and Retweets

We have already seen in Section IV-C that there is a high chance that the posts shared by the phishing reporting users contain legitimate phishing URLs. Additionally, the URLs posted by these accounts have a low overlap with the URLs posted in PhishTank and OpenPhish. Thus, the visibility of these Twitter reports is vital to recognize new phishing websites. However in the previous section we have

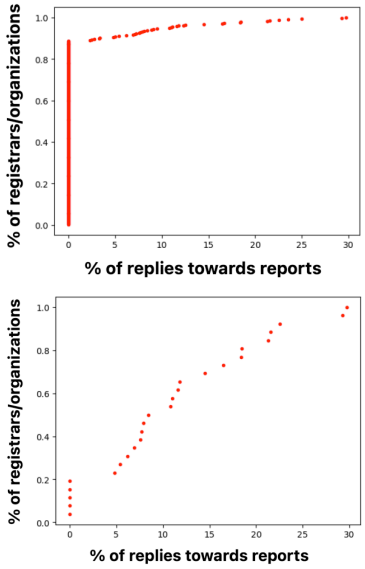

Fig. 6. CDF of replies (explicit interactions) provided by targeted domain registrars/organizations when **a)** considering all tagged registrars/targets and **b)** considering only registrars/targets who were tagged in atleast 100 tweets.

already determined that these reports receive very explicit interactions from the targeted domains and organizations. Another approach to measure the popularity of these reports is by tracking the number of likes(favourites) and retweets [55] that they get.

We found that these posts have very few interactions in the form of likes (median=0) and retweets (median=2) as well. In fact nearly 82% of the tweets in our dataset (n=13,511) did not receive any likes, and 58% of tweets in our dataset (n=9,596) did not get retweeted. The cumulative distribution of the number of likes(favourites) and retweets received by the report tweets is illustrated in Figure 9. Thus, the lack of this form of interaction further limits the propagation of these phishing reports through the Twitter community. But, we find that the total number of retweets (n=34,284) is 4.5 times more than the total number of likes received (n=7,527) when collectively considering all the tweets in our dataset. This indicates that the users who do interact with these posts have the intention of sharing the information along to their peers, which might improve their visibility.

Now, we are interested in determining what percentage of likes/ retweets received by these tweets are from users in the technological communities, especially those who are in the field of computer security. We do so by examining the profile descriptions of the users who have liked and retweeted the tweets in our dataset. Since it is impossible to qualitatively analyze this text of all such users, we assigned four coders to go through 500 randomly selected users (who liked/retweeted the reports) to identify which of their profile descriptions contain text which indicated they belong to/interested in the computer security community. We further used the profile descriptions of these users to create a Word-cloud as illustrated

in Figure 7. We obtain the top 20 most frequently occurring words and their combinations and match it with the profile descriptions of users who liked (n=7,527) and/or retweeted (n=34,284) the phishing reports. We find about 37% of likes/ retweets came from users who are interested/work in computer security. Do note that our findings are based on the keywords that we had selected from the world cloud, and also about 14% of the users had an empty or irrelevant profile description. Thus, realistically, the number of security focused users who interact with these tweets might be even higher. Even then, a large number of these interactions came from individuals belonging to the security community, which might increase the chances of the reports to be noticed by a registrar/targeted organization.

Fig. 7. World cloud of the most frequent words found in the profile description of Security focused users who liked/retweeted the phishing report posts

B. Followers

We find that the accounts in our dataset have a median follower count of 472 and median listed count of 7. Despite our previous findings that the phishing reports receive low explicit interaction, as well as very few likes and retweets, the large majority of phishing reporting accounts have a decent number of followers, with 523 accounts having more than 100 followers. On the other hand, we checked the *listed count rate* or LCR, which is the percentage of users who listed an account divided by the total number of followers the user has. Listed count of an account is considered as a metric for their credibility [56], i.e. users tend to list accounts who they rely on for information regarding specific topic(s). Interestingly, we find that three users have a higher LCR than their total no. of followers. On the other hand, 93% (n=652) of the reporters have a LCR of less than 10%, with 40% of accounts (n=283) having an LCR of less than 1%. This indicates that despite the users having a decent number of followers, most of them are either not recognized or considered to be a creditable source for providing phishing information, as indicated by their low LCR . Incidentally, the *top reporter* account has an LCR rate of only 2.9% despite contributing the majority of the URLs to our dataset. Using the keywords that we had found from the profile descriptions of security related users in Section V-A, we find that at least 33% of the users belong to the security community. While it is interesting to see that a majority of the users that follow these accounts have a background in computer security, the number of unique users in this field who actually interact with these tweets through likes and retweets is only 5%. This indicates that Twitter phishing reports, despite their effectiveness, are presently not considered as a popular resource among the individuals of their relevant community.

C. Targeted domains/organizations as followers

We have already observed that the domain registrars and organizations which are targeted by the reported URLs have very low *explicit interactions* (posting comments) with said reports. But since we have already established that these reports are reliable and provide a lot of information about the phishing website, it is very important that these reports are *discoverable*, i.e., the targeted entities can notice these reports such that they can expedite the process of removing the URLs. The most convenient way to discover such new reports is to follow the phishing report accounts, as posts from these accounts will then show up in the personalized feeds of their followers. Out of the 349 registrars and organizations that were tagged by these reports, we found that 303 of them (87%) had an active Twitter account. Using the Twitter API, we determined the number of official Twitter accounts belonging to the domain hosting services and targeted organizations which followed one or more of the 701 Twitter phishing reporting accounts which we have identified from our dataset. We found only 31 hosting domains/ targeted organizations (10.2%) which follow at least one of the phishing reporting accounts, with only one such account following a maximum of 12 phishing reporters. Figure 8 illustrates the distribution of the domains/ organizations across the number of phishing report accounts that they follow. While it is difficult to ascertain how and whether registrars/ organizations keep a track of URLs shared by these phishing reports, our findings imply that a large majority of targeted domains/ organizations do not follow these reporting accounts, either because they are not aware of them, or similar to security focused users on Twitter, just do not consider them a popular source for obtaining phishing reports.

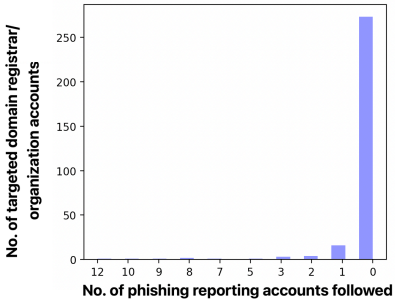

Fig. 8. Distribution of the domains/organization accounts across the number of phishing report accounts that they follow.

D. Age of accounts

The age of a Twitter account determines how long it has been active on Twitter. We found that the phishing report-

ing accounts in our dataset have been active for a median period of 2,129 days (5.83 years), with only 83 accounts (12%) having an age of less than a year. Prior literature has recognized accounts which tend to distribute spam and misinformation to have low account age [57], [58], and thus the longevity of these accounts can be considered as yet another feature/indicator which can be used by hosting registrars and anti-phishing tools to determine whether they should rely on a Twitter phishing report account.

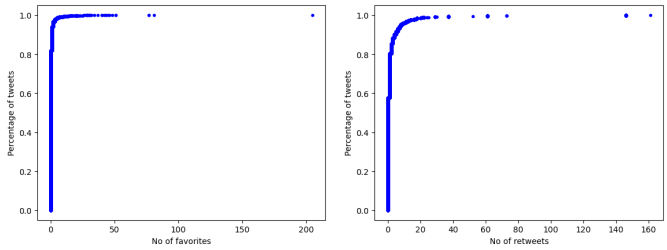

Fig. 9. CDF of retweets and favourites for Twitter reports

VI. Activity of URLs in Phishing Report

We continuously checked whether each unique URL found in the Twitter phishing reports was active and found that throughout the duration of the study, 39% of URLs reported by the accounts were still active after a day, and 31% after a week. Since only 752 such reports (containing 671 unique URLs) received a reply from the registrar/ targeted organization (4% of all reports in our dataset), we compare the URLs reported in these reports with the same number of randomly selected unique URLs included in reports which did not receive a reply from the registrar/ targeted organization. Note that for the latter group, we only selected URLs which had become inactive. We performed a Mann-Whitney U Test [59] on both groups of URLs, and found that URLs found in reports which get a reply from the targeted organization were significantly more likely to become inactive at a quicker pace than URLs in reports which do not get a reply ($p<0.01$). Statistically, URLs in posts which got a reply from an organization all became inactive within a median time period of 403 minutes. In fact, targets which explicitly interact with these reports often acknowledge the removal of the phishing website, as illustrated in Figure 10

On the other hand, for URLs which did not get a reply, we found the median time of removal to be at 1,172 minutes. However, the latter group of URLs (which did not get a reply) can also be bifurcated into two more groups. Earlier we have seen that 52% of these reports use a hashtag which cites the registrar or targeted organization. Thus to determine if there is a difference between the activity time of URLs which contained relevant hashtags versus those which did not, we randomly selected 500 posts (each containing unique URLs) from the two groups, and performed a Mann-Whitney U test again. Our results indicate that posts which tag the hosting or targeted organization were significantly more likely to be removed at a quicker pace than posts which do not contain such hashtags ($p<0.01$). The median time of removal for URLs

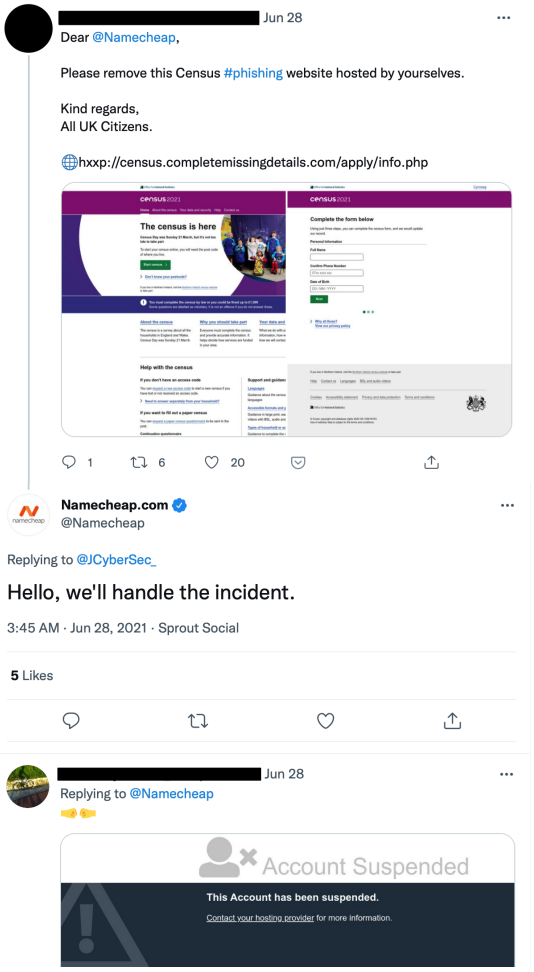

Fig. 10. The effect of explicit interaction from the registrar (On Top) A registrar acknowledges to look into a phishing report. (On Bottom) The reporter confirms that the URL has been removed.

which contained a relevant hashtag was at 847 minutes, while those which did not had a median time of removal of 1,591 minutes. It is to be noted that all URLs which garnered a reply from the target had relevant hashtags. However these reports were only 8% of the overall tweets which had used hashtags (752 out of 8,572 posts), indicating that the majority of reports with hashtags do not receive an explicit interaction from the target or hosting organizations. Thus, we were not able to determine the factors or characteristics of the reports which encourage a targeted registrar/organization to reply to them. But, we hope to persue these granular details in a future study.

A. Removal rate comparisons with other phishing feeds

URLs in Phishing report tweets become inactive very quickly when they receive a reply from the targeted registrar/ organization, with the 671 unique URLs which received an explicit interaction being removed within a median time of 403 minutes after the reports were shared. Comparing this time with the same number of (true positive) URLs chosen from PhishTank and OpenPhish, and also performing respective

Mann-Whitney U Tests between removal times between them and the Twitter phishing reports, we notice that URLs found on PhishTank are removed at a median time of 132 minutes, having a significant edge over Twitter URLs ($p<0.05$). The same is seen for OpenPhish URLs, which are removed at a median time of 71 minutes ($p<0.01$). Our findings thus suggest that URLs appearing in these feeds get removed much faster than those found in the Twitter phishing reports. However, in Section IV-D1, we have already found that several URLs submitted on PhishTank and OpenPhish are *dead on arrival*, when compared to URLs found in Twitter phishing reports. This when added to the fact that there is minimal overlap of URLs included in the Twitter Reports with the two other phishing feeds, further suggests that: (a) Phishing reports on Twitter are a viable solution for finding new phishing URLs which are not found on atleast two other popular phishing feeds, and (b) Registrars and targeted organizations are slower at removing websites which appear on Twitter phishing reports compared to those found on PhishTank and OpenPhish, something that can be easily improved upon by including these Twitter reports in their phishing identification and detection workflow. We also explore the reception of anti-phishing engines towards URLs shared in Twitter reports in Section VI-C, where we compare the coverage of URLs in phishing reports by these tools, compared to URLs found on PhishTank and OpenPhish.

B. URLs which remained active after a week

Around 31% of unique URLs (n=3,453 URLs) remained active even after a week. It is interesting to note that none of the 671 URLs which were part of the posts that targets replied to were found in this category, suggesting that an *explicit interaction* from the targeted registrar/ organization leads to the removal of the website. Almost 67% of the URLs which did not get removed after a week (n=2,311 URLs) were those which did not have any relevant hashtags. Thus, we hypothesize that the lack of such hashtags might make it difficult for the registrar or targeted organization accounts to search for them/index them, compared to the reports which already have a hashtag. The higher rate of removal for URLs which were shared by reports which contained relevant hashtags further hints at the phenomenon of *Implicit* interactions between the target registrars/ organizations with these phishing reports, indicating that the latter can investigate (and remove these URLs) even without explicitly interacting with the reports. However, this assumption is not comprehensive, as URLs contained in the reports which contained hashtags might have also shown up in phishing feeds other than PhishTank and OpenPhish, which might have been the primary reason for their removal. We can rectify this in a future study by focusing only on URLs which appear exclusively in these phishing reports.

1) The case of unconventional phishing URLs: Work by Saha Roy et al. [15] explored a new category of phishing URLs which use free hosting domains to remain undetected from anti-phishing tools, and are similarly not removed by registrars for a long period of time, if at all. In our phishing report dataset, we found 5% (n=631) URLs belonging to this category, out of which 53 received a reply from the registrar/ targeted organization. We found that all of these 53 URLs which got a reply were removed at a median time of 319 minutes, which is much quicker (by several days) than previously established removal rates for this category of phishing attacks. We thus extended our analysis by selecting 100 random URLs belonging to this category from both PhishTank and OpenPhish,and noticed that such URLs on PhishTank were removed after a median time of 1047 minutes after appearing, whereas the statistics for OpenPhish was 892 minutes. Thus, our preliminary analysis suggests that Twitter phishing reports are a much faster way of making sure these evasive family of phishing attacks are removed by the hosting registrars, compared to sharing them on OpenPhish and/or PhishTank.

C. VirusTotal coverage

In this section, we investigate how quickly anti-phishing URLs pick up on URLs shared by Twitter phishing reports and how this coverage compares to URLs found on OpenPhish and PhishTank.

Since we only have 752 posts on Twitter containing 671 unique URLs which received *explicit interaction* from the *targets*, it would not be fair to compare them with a large volume of URLs from the other scenario, i.e., Twitter posts without explicit interaction posts. Thus, we sample 500 random tweets from each of these sets. Since most phishing URLs and are online for less than a day, we tracked how many anti-phishing tools detected these URLs at an interval of every 30 mins through a period of 24 hours. We illustrate the detection of the URLs through time in Figure 11. To avoid congestion in the figure we extended the time bins to 1 hr instead of 30 mins.

Our results indicate that URLs on Openphish and PhishTank are detected by more engines within a short time after their appearance compared to those included in Twitter phishing reports. However, we see that reports once explicitly interacted upon by targeted registrars and organizations, see a rapid rise in detection rate by anti-phishing engines, going almost head to head with the other phishing feeds, if not exceeding them. However, URLs which did not get explicit interactions tend to consistently have lower anti-phishing tool detection throughout the day. We have already noticed in Section VI that the majority of the URLs included in Twitter phishing reports were alive or had a very slow rate of removal. We see here that they are very sparsely detected by anti-phishing tools as well. Thus, while it is hard to determine how information from different phishing feeds/blocklists is utilized by anti-phishing engines to detect newer phishing threats, it is evident from our analysis that there is a large possibility of URLs which are shared in Twitter reports to be covered by fewer anti-phishing engines compared to other more prominent blocklists. Considering the large volume of true positive phishing URLs which are shared by these Twitter reports, this detection gap can result in these

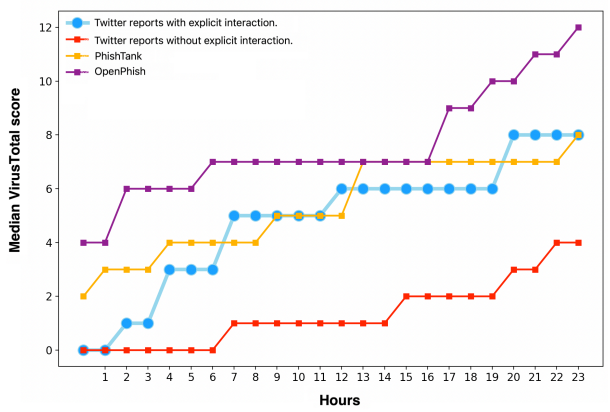

Fig. 11. Tracking median VirusTotal scores for the reported URLs through their first day of appearance for phishing reports which received a comment from the registrar/ target organization, reports which do not, as well as PhishTank and Openphish.

attacks successfully evading traditional anti-phishing solutions for a longer period of time.

VII. Limitations and Future Work

Our work presents a preliminary picture about the effectiveness of Twitter reports as an anti-phishing knowledge-base. In this section we address some of the limitations of our work which can be addressed in a future publication. Firstly, to reliably analyze the qualitative content shared by the tweets used in this study, we only examined reports which were in English. We also limit our data collection to tweets which contained four terms/ hashtags - 'hxxp', 'hxxps', #phishing and #scam. Thus, our dataset is not exhaustive, with the possibility of more phishing reports existing in other languages or having different keywords, which we plan to cover in a future publication. Also, based on the tweets in our dataset, we determined that the information shared through Twitter reports are fairly reliable. However, our methodology for indexing these reports using only a few keywords (such as #phishing, hxxp/hxxps) is not robust, as adversaries can easily poison the feed by sharing false reports using the same keywords. Thus, there is a need to identify auxiliary features for the reporting account (such as account age) to ascertain their trustworthiness. In Section V, we found that Twitter reports which get *explicit* interaction from hosting registrars or targeted organizations tend to get removed at a much faster pace than those that did not. However, we could not conclusively identify the factors which influence this interaction in the first place. We intend to perform a more granular study of the reports (that were interacted with) in the future, which would provide us with a better comprehension in this regard.

VIII. Conclusion and Discussion

In this study, we establish phishing reports posted on Twitter as a new and reliable resource for sharing information about phishing attacks. When compared to two other open-source phishing feeds - PhishTank and OpenPhish, our findings indicate that reports on Twitter tend to share more information about phishing URLs, cover an extra threat category (drive-by downloads) and tend to have lower number of false positives. However, we found very few instances where the concerned domain registrars or organizations acknowledged these reports, and our results further indicate that these reports have very low visibility and interest among the online security community overall, with the URLs also being detected by very few anti-phishing engines. This results in a majority of the reported URLs remaining only for an extended period of time compared to entries found on PhishTank and OpenPhish. URLs shared on Twitter also have low overlap with those found on the PhishTank and OpenPhish, indicating the scope for identifying newer phishing threats through this Twitter phishing reports which may not be found in other blocklists.

In summary, our work aims to raise awareness about the reliability, volume of information and exclusivity of the reports shared on Twitter, which can expedite the process of moderation and removal of newer phishing threats. Also, considering the low rate of false positives, security researchers can especially benefit from extracting information from these reports and utilize it for ground-truth labelling. Thus, there is a need to integrate this resource in prevalent phishing moderation and research workflows, as well as motivate further research towards analyzing these phishing report accounts on Twitter and determining if similar useful knowledge-bases can be encountered within other Online social media networks.

References

[1] F. Simon Chandler, "Google registers record two million phishing websites in 2020," https://www.forbes.com/sites/simonchandler/2020/11/25/google-registers-record-two-million-phishing-websites-in-2020/?sh=23b004881662.

[2] M. Rosenthal, "Must-know phishing statistics: Updated 2020," https://www.tessian.com/blog/phishing-statistics-2020/.

[3] O. K. Sahingoz, E. Buber, O. Demir, and B. Diri, "Machine learning based phishing detection from urls," *Expert Systems with Applications*, vol. 117, pp. 345–357, 2019.

[4] R. S. Rao and A. R. Pais, "Detection of phishing websites using an efficient feature-based machine learning framework," *Neural Computing and Applications*, vol. 31, no. 8, pp. 3851–3873, 2019.

[5] R. Hassanpour, E. Dogdu, R. Choupani, O. Goker, and N. Nazli, "Phishing e-mail detection by using deep learning algorithms," in *Proceedings of the ACMSE 2018 Conference*, 2018, pp. 1–1.

[6] N. Abdelhamid, F. Thabtah, and H. Abdel-jaber, "Phishing detection: A recent intelligent machine learning comparison based on models content and features," in *2017 IEEE international conference on intelligence and security informatics (ISI)*. IEEE, 2017, pp. 72–77.

[7] P. Yang, G. Zhao, and P. Zeng, "Phishing website detection based on multidimensional features driven by deep learning," *IEEE Access*, vol. 7, pp. 15 196–15 209, 2019.

[8] P. Yi, Y. Guan, F. Zou, Y. Yao, W. Wang, and T. Zhu, "Web phishing detection using a deep learning framework," *Wireless Communications and Mobile Computing*, vol. 2018, 2018.

[9] L. A. T. Nguyen, B. L. To, H. K. Nguyen, and M. H. Nguyen, "Detecting phishing web sites: A heuristic url-based approach," in *2013 International Conference on Advanced Technologies for Communications (ATC 2013)*. IEEE, 2013, pp. 597–602.

[10] G. Sonowal and K. Kuppusamy, "Phidma–a phishing detection model with multi-filter approach," *Journal of King Saud University-Computer and Information Sciences*, vol. 32, no. 1, pp. 99–112, 2020.

[11] J. Sreedharan and R. Mohandas, "Systems and methods for risk rating and pro-actively detecting malicious online ads," apr 5 2016, uS Patent 9,306,968.

[12] A. Oest, Y. Safaei, P. Zhang, B. Wardman, K. Tyers, Y. Shoshitaishvili, and A. Doupé, "Phishtime: Continuous longitudinal measurement of the effectiveness of anti-phishing blacklists," in *29th {USENIX} Security Symposium ({USENIX} Security 20)*, 2020, pp. 379–396.

[13] B. Liang, M. Su, W. You, W. Shi, and G. Yang, "Cracking classifiers for evasion: a case study on the google's phishing pages filter," in *Proceedings of the 25th International Conference on World Wide Web*, 2016, pp. 345–356.

[14] P. Rajivan and C. Gonzalez, "Creative persuasion: a study on adversarial behaviors and strategies in phishing attacks," *Frontiers in psychology*, vol. 9, p. 135, 2018.

[15] S. S. Roy, U. Karanjit, and S. Nilizadeh, "What remains uncaught?: Characterizing sparsely detected malicious urls on twitter."

[16] A. AlEroud and G. Karabatis, "Bypassing detection of url-based phishing attacks using generative adversarial deep neural networks," in *Proceedings of the Sixth International Workshop on Security and Privacy Analytics*, 2020, pp. 53–60.

[17] Twitter, https://twitter.com/home.

[18] "1.4 million phishing websites are created every month," https://zd.net/30T55DW, 2020.

[19] I. Vayansky and S. Kumar, "Phishing–challenges and solutions," *Computer Fraud & Security*, vol. 2018, no. 1, pp. 15–20, 2018.

[20] M. Junger, L. Montoya, and F.-J. Overink, "Priming and warnings are not effective to prevent social engineering attacks," *Computers in human behavior*, vol. 66, pp. 75–87, 2017.

[21] J.-W. H. Bullée, L. Montoya, W. Pieters, M. Junger, and P. Hartel, "On the anatomy of social engineering attacks—a literature-based dissection of successful attacks," *Journal of investigative psychology and offender profiling*, vol. 15, no. 1, pp. 20–45, 2018.

[22] T. Ahmad, "Corona virus (covid-19) pandemic and work from home: Challenges of cybercrimes and cybersecurity," *Available at SSRN 3568830*, 2020.

[23] P. Jason Cohen, "Phishing attacks increase 350 percent amid covid-19 quarantine," https://bit.ly/3HiprGy.

[24] "PhishTank," https://www.phishtank.com/faq.php, 2020.

[25] Openphish, "Phishing feed," "https://openphish.com/faq.html".

[26] S. Sheng, B. Wardman, G. Warner, L. Cranor, J. Hong, and C. Zhang, "An empirical analysis of phishing blacklists," 2009.

[27] S. Bell and P. Komisarczuk, "An analysis of phishing blacklists: Google safe browsing, openphish, and phishtank," in *Proceedings of the Australasian Computer Science Week Multiconference*, 2020, pp. 1–11.

[28] T. Moore and R. Clayton, "Evaluating the wisdom of crowds in assessing phishing websites," in *International Conference on Financial Cryptography and Data Security*. Springer, 2008, pp. 16–30.

[29] M. Firefox, "How does built-in phishing and malware protection work?" https://support.mozilla.org/en-US/kb/how-does-phishing-and-malware-protection-work.

[30] Apple, "Safari and privacy," https://www.apple.com/legal/privacy/data/en/safari/.

[31] "Friends of PhishTank [Infographic]," https://www.phishtank.com/friends.php, 2020.

[32] "Email Security – Defanging URLs," https://ibm.co/3of95qz, 2021.

[33] Twitter, https://developer.twitter.com/en.

[34] VirusTotal, https://www.virustotal.com/gui/home/upload.

[35] "VirusTotal," https://www.virustotal.com/gui/home/, 2020.

[36] R. Masri and M. Aldwairi, "Automated malicious advertisement detection using virustotal, urlvoid, and trendmicro," in *2017 8th International Conference on Information and Communication Systems (ICICS)*. IEEE, 2017, pp. 336–341.

[37] Y. Tanaka, M. Akiyama, and A. Goto, "Analysis of malware download sites by focusing on time series variation of malware," *Journal of computational science*, vol. 22, pp. 301–313, 2017.

[38] B. Sun, M. Akiyama, T. Yagi, M. Hatada, and T. Mori, "Automating url blacklist generation with similarity search approach," *IEICE TRANSACTIONS on Information and Systems*, vol. 99, no. 4, pp. 873–882, 2016.

[39] H. Wang, J. Si, H. Li, and Y. Guo, "Rmvdroid: towards a reliable android malware dataset with app metadata," in *2019 IEEE/ACM 16th International Conference on Mining Software Repositories (MSR)*. IEEE, 2019, pp. 404–408.

[40] R. Penman, "Python-whois," https://pypi.org/project/python-whois/.

[41] J. Lehmann, B. Gonçalves, J. J. Ramasco, and C. Cattuto, "Dynamical classes of collective attention in twitter," in *Proceedings of the 21st international conference on World Wide Web*, 2012, pp. 251–260.

[42] A. Ferreira and G. Lenzini, "An analysis of social engineering principles in effective phishing," in *2015 Workshop on Socio-Technical Aspects in Security and Trust*. IEEE, 2015, pp. 9–16.

[43] N. Provos, P. Mavrommatis, M. Rajab, and F. Monrose, "All your iframes point to us," 2008.

[44] P. Peng, L. Yang, L. Song, and G. Wang, "Opening the blackbox of virustotal: Analyzing online phishing scan engines," in *Proceedings of the Internet Measurement Conference*, 2019, pp. 478–485.

[45] "2021 Cybersecurity threat trends: phishing, crypto top the list," https://umbrella.cisco.com/info/2021-cyber-security-threat-trends-phishing-crypto-top-the-list, 2021.

[46] "What is a Trojan? Is it a virus or is it malware?" https://us.norton.com/internetsecurity-malware-what-is-a-trojan.html, 2020.

[47] "Mobile Malware," https://usa.kaspersky.com/resource-center/threats/mobile, 2020.

[48] "Cryptojacking – What is it?" https://www.malwarebytes.com/cryptojacking, 2020.

[49] "What are Browser Hijackers?" https://us.norton.com/internetsecurity-malware-what-are-browser-hijackers.html, 2020.

[50] P. Peng, L. Yang, L. Song, and G. Wang, "Opening the blackbox of virustotal: Analyzing online phishing scan engines," in *Proceedings of the Internet Measurement Conference*, 2019, pp. 478–485.

[51] N. Papernot, "Detecting phishing websites using a decision tree," "https://github.com/npapernot/phishing-detection".

[52] R. M. A. Mohammad, "Phishing websites data set," https://openphish.com/faq.html.

[53] C. H. Tung, "Sharkcop," https://github.com/CaoHoangTung/sharkcop.

[54] T. D. Swig, "Sharkcop," https://bit.ly/3spN1wz".

[55] L. McShane, E. Pancer, M. Poole, and Q. Deng, "Emoji, playfulness, and brand engagement on twitter," *Journal of Interactive Marketing*, vol. 53, pp. 96–110, 2021.

[56] B. Kang, J. O'Donovan, and T. Höllerer, "Modeling topic specific credibility on twitter," in *Proceedings of the 2012 ACM international conference on Intelligent User Interfaces*, 2012, pp. 179–188.

[57] H. Gupta, M. S. Jamal, S. Madisetty, and M. S. Desarkar, "A framework for real-time spam detection in twitter," in *2018 10th International Conference on Communication Systems & Networks (COMSNETS)*. IEEE, 2018, pp. 380–383.

[58] W. Herzallah, H. Faris, and O. Adwan, "Feature engineering for detecting spammers on twitter: Modelling and analysis," *Journal of Information Science*, vol. 44, no. 2, pp. 230–247, 2018.

[59] P. E. McKnight and J. Najab, "Mann-whitney u test," *The Corsini encyclopedia of psychology*, pp. 1–1, 2010.