Behind Phishing: An Examination of Phisher Modi Operandi

D. Kevin McGrath, Minaxi Gupta
Computer Science Department, Indiana University, Bloomington, IN, U.S.A.
{*dmcgrath, minaxi*}*@cs.indiana.edu*

Abstract

Phishing costs Internet users billions of dollars a year. Using various data sets collected in real-time, this paper analyzes various aspects of phisher modi operandi. We examine the anatomy of phishing URLs and domains, registration of phishing domains and time to activation, and the machines used to host the phishing sites. Our findings can be used as heuristics in filtering phishing-related e-mails and in identifying suspicious domain registrations.

1 Introduction

According to a recent Gartner survey, 3.6 million U.S. adults lost a total of 3.2 billion dollars due to phishing in 2007 [16]. The survey projects that these numbers will continue to increase in the coming years because phishing is a lucrative business for the perpetrators. Phishers use a combination of tricks involving the Web, e-mail, and malicious software (a.k.a. malware) to steal personal identity data and financial account credentials. While detection of phishing e-mails and phishing sites have been researched, very little has been done to analyze the modi operandi of phishers. Understanding phisher modi operandi can help in better filtering of e-mails related to phishing and in taking down domains involved in phishing. They can also help in proactively tracking domain registrations to flag suspicious phishing-related activity. Specifically, this paper is motivated by the following questions:

- Do phishing URLs and domains exhibit characteristics that are different from other URLs and domains?
- To what extent are phishers registering new domains to put up phishing sites and how long does it take for such sites to become active phishing campaigns?
- What type of machines are phishers using to host phishing sites and how long does a typical phishing domain live?

We conduct our analysis using multiple data sets. The primary ones are real-time phishing URL feeds for over two months from two different sources. The secondary data sets include *whois* records and *zone files*. These contain domain registration and deregistration information, among other things. Additionally, we perform periodic DNS resolutions on phishing domains to gather the IP addresses of machines hosting them. This data is used to gain insights about the machines hosting the phishing sites. We use two additional data sets for comparison purposes. The first, `DMOZ`, is a collection of URLs in the Internet. This set is used to infer differences in the anatomy of phishing URLs and domains with respect to regular URLs and domains in the Internet. The second additional data set is a set of older phishing URLs. We use it to observe trends.

The main findings of our preliminary work include:

- Phishing URLs and domain names have very different lengths compared to other URLs and domain names in the Internet. Even the character frequency of phishing domain names is significantly different from English when the `DMOZ` URLs and domains follow the English letter character frequency very closely. Further, 50-75% of the phishing URLs contained the name of the brand they targeted. All these facts can be used to identify phishing URLs and domains.
- Phishers are misusing free Web hosting services as well as URL-aliasing services, such as, TinyURL [9]. This points to the need to better scrutinize the users of such services.
- Most domains registered for the purpose of phishing become active almost immediately upon registration. This implies that the window to track suspicious domain registrations from the perspective of phishing is very small.
- Many phishing domains were hosted on multiple machines spread across multiple countries. A significant percentage of these machines belonged to residential customers. These facts point to the use of botnets in hosting phishing sites.

The rest of this paper is organized as follows. Section 2 highlights related work. In Sections 3 and 4 describe the data collection methodology and an overview of the collected. Section 5 analyzes the data. Finally, Section 6 presents concluding remarks.

2 Related Work

Most of the past work in the area of phishing has dealt with detection, economic impact, phishing trends, and psychological aspects of phishing.

Phishing trends have been analyzed by Ramzan et al. [23] and RSA [24]. They also provide predictions on the type of phishing attacks that would be seen in the near future. The anti-phishing working group (APWG) also regularly publishes facts and figures on average lifetime of phishing sites, country of origin etc. [3]. Work in [5] by Dhamija et al. provides a psychology-based discussion of how people fall victim to phishing, while articles, such as, Goth's [11] discuss the economic impact of phishing related crimes.

Detection is a very important aspect in the fight against phishing. Fette et al. [6] looked into techniques for detecting phishing e-mails. They consider features, including, IP-based URLs, age of linked-to domains, number of links present in the e-mails, number of dots in the phishing URLs, and presence of JavaScript, to flag emails as phishing. Phishing e-mails are a special case of spam. Thus, techniques to defend against spam would help alleviating phishing as well, including those of Ramachandran et al. [22]. Other works have looked into detecting phishing sites based on site content using information retrieval algorithms, such as Wenyin et al. [28] and Zhang et al. [29]. Cranor et al.[4] analyzed the efficacy of toolbars in identifying phishing sites. These lines of works could benefit from the findings presented in this paper.

The network characteristics of spam has been examined by both Anderson et al. [2] and Ramachandran et al. [21]. Work in [21] examines the network-level behavior of spammers, including, the IP address ranges that send most spam, common spamming modes, characteristics of spamming botnets, and the persistent of individual spamming hosts. Work in [2] focuses on the hosting infrastructure for scams. Only 2% of the sites they examine fall into the malicious category, which includes phishing and malware-hosting sites. Thus, none of these works deal with phishing explicitly. We make comparisons with the results found in [2] subsequently in the paper, where applicable.

Moore et al. [17] analyzes empirical data on phishing Web site removal times and the number of visitors these sites attract. The authors conclude that phishing Web site removal is part of the answer to phishing, it is not done fast enough to mitigate the problem. Garera et al. [8] investigate URL-anatomy and phishing pages with the goal of identifying phishing URLs and domains. In particular, they observe that the phishers sometimes use IP addresses instead of host names, their URLs often contain the name of organization being phished, and that the phishing domain names are sometimes misspelled versions of well-known domain names. While this work focuses on training a classifier that incorporates these features, we focus more on the quantification of how often these features occur in present-day phishing URLs. Further, our work also focuses on generic features of the phishing URLs and domains and their comparison with good URLs and domain names in the Internet. Ludl et al. [14] examine the effectiveness of multiple classes of detection methods.

3 Data Collection Methodology

Our primary data sets are the phishing URLs secured from two different sources: PhishTank [20] and our industry collaborator, MarkMonitor [15]. The PhishTank phishing URLs are either user submitted or obtained via external feeds. The user-submitted URLs are voted upon for verification purposes. The MarkMonitor phishing URLs are obtained from various large e-mail providers and ISPs. To verify them, they are passed through a filter which determines the likelihood that the URL is a phishing site. This filter performs URL and content analysis and determines the likelihood that the URL is pointing to a phishing site. For URLs with a high probability of being phishing URLs, MarkMonitor performs a manual check on the URL, any hour of the day or night. The unique, positively identified URLs are recorded along with the brand they target. All phishing URLs in our data are final, implying that no redirections need to be traversed to reach the phishing site.

We obtain the phishing URLs from each of these sources via a real-time *feed*. The feed from PhishTank adds phishing URLs every hour while the feed from MarkMonitor does so every five minutes. For every phishing URL, we extract the *effective second-level domain name* in real time. For example, in URL, `http://www.xyz.example.com/doc.html`, `example.com` is the second-level domain and `.com` is the generic top-level domain (gTLD). Similarly, in, `http://www.example.ac.au/`, `example.ac.au` is the effective second-level domain and `.au` is a country code TLD (ccTLD). There are a few exceptions to this general rule of extracting second-level domains. We account for those as well [7]. We will subsequently use the term *domain* to refer to effective second-level domains.

Upon extracting the domain for each phishing URL, we gather two additional pieces of information. First, we perform a *whois* lookup on it. The *whois* is a distributed database that contains contact information about the owner and registrar of the domain (including home

page URL), date of registration, last update, expiration, primary and secondary DNS servers, and any additional status information of the domain. Due to rate-limiting issues, we obtain the *thin* records, which contain only registrar-provided information, including, registrar's name and the date domain was registered. This constraint is serendipitous, for its mitigates any inaccuracy in the user-provided portion of *whois* records. Another reason to avoid the other possibility, the *thick* records, was to avoid parsing issues in the *whois* records, which tend to be plain text with hardly any agreed-upon formats. As the second piece of information, we look up the DNS records for each domain. This gives the IP addresses of the machines hosting the phishing sites. We perform the DNS lookups periodically, at a rate of once per 5 minutes, through the department's DNS server to get a sense of the entire infrastructure for each phishing domain and to infer their lifetimes. Further, we also geolocate the IP addresses obtained from DNS lookups using IP2Location [12] software to find out where the hosting machines were physically located.

To correlate the phishing activity with the registration of the domains, we utilize the *zone files* for several popular gTLDs: `.com`, `.net`, `.info`, and `.biz`. (Getting zone files from ccTLDs was not possible.) We have been collecting these zone files nightly since 10/10/2007. Having zone files from a time prior to our phishing feeds allows us to track domain registrations more effectively. In addition to storing the raw zone files, we store differences from the previous day. This is necessary as the biggest TLD, `.com`, itself contains information about over 65 million domains in each zone files [26]. Further, it provides an easy way to find new domain registrations, as well as domains which have been dropped.

Finally, we have two additional data sets that we use for comparison purposes. The first data set, `MarkMonitor-2006`, contains phishing URLs from MarkMonitor from 2006. We have zone files for the corresponding duration as well but no DNS resolutions (hence, no IP addresses) or *whois* records. We use this data set to observe trends. The second, `DMOZ`, is from the Open Directory Project [18]. This project contains user submitted URLs and is the largest and most comprehensive directory of the Web. Our input data, collected on October 28, 2006, has over 9 million unique URLs and 2.7 million unique effective second-level domain names. We use this data set to observe differences in the anatomy of phishing URLs with those of good URLs in the Web.

4 Data Overview

Both of our primary data sets, `PhishTank` and `MarkMonitor`, are each for 71 days each with a gap of eight days in between. This gap in the data collection was due to a software upgrade causing errors in the collection script which went unnoticed during the holiday break. Table 1 shows an overview of both data sets. There are several things to note from this table. First, there are 21 gTLDs and close to 250 ccTLDs in the Internet today and most of the popular gTLDs and ccTLDs are represented in our data. As expected, the `.com` TLD accounts for at least 1/3rd of all the phishing domains as well as 1/3rd of the phishing sites (denoted by unique URLs) in both data sets. A significant number of phishing domain names are also simply IP addresses of machines hosting them. (This fact has been used as a heuristic by work aiming to filter phishing e-mails [6].) Further, though most phishing domains host a small number of phishing sites, within `MarkMonitor` six host a 100 or more sites. There are 17 such domains in PhishTank.

	`PhishTank`	`MarkMonitor`
Start date	11/30/2007	11/30/2007
Collection days	71	71
TLDs	144	116
gTLDs	10	6
ccTLDs	134	108
Unique domains	17,105	7,394
`.com`	5,749	2,889
other gTLDs	2,031	1,136
ccTLDs	5,355	2,284
IP	3,970	1,035
Phishing URLs	44,320	25,304
`.com`	15,526	11,314
other gTLDs	5,017	4,023
ccTLDs	17,131	8,319
IP	6,646	1,648
Brands	n/a	207

Table 1: Overview of primary data sets. The MarkMonitor data is higher quality because it does not contain duplicate phishing URLs or false positives, which PhishTank does at times due to its community-driven nature.

Our primary data sets have some intersection. 14% of phishing URLs contained in `PhishTank` are also contained in `MarkMonitor`. Similarly, 12% of `PhishTank` domains are also contained in `MarkMonitor`. Finally, 599 of the cases where domain names were IP addresses occur across the two data sets.

Table 2 provides an overview of the `MarkMonitor-2006` data set, which we use for comparison purposes.

Start date	1/1/2006
Collection days	211
TLDs	168
Unique domains	27,707
Phishing URLs	189,239
Brands	564

Table 2: Overview of MarkMonitor data set from 2006.

5 Analysis

Our analysis focuses on the following aspects of phishing: anatomy of phishing URLs and domains, use of newly registered domains in phishing, time between domain registration and its use in phishing, infrastructure used to host phishing sites, and the lifetime of phishing domains. Next, we describe each.

5.1 Anatomy of Phishing URLs

Each URL has two main components: a *fully qualified domain name (FQDN)* and *path*. For example, in URL, `http://www.xyz.example.com/dir/doc.html`, `www.xyz.example.com` is the FQDN and `dir/doc.html` is the path. The FQDN also has sub-components. In this example, `xyz.example.com` is the subdomain and `example.com` is the domain. More than 3/4th of the phishing URLs in each of our data sets contained subdomains. We now analyze the anatomy of phishing URLs. Our goal is find out if characteristics of URLs themselves can be used as a factor in identifying phishing campaigns.

URL and domain name lengths: We begin by comparing the length of URLs and domain names in the `DMOZ` data set to those in `PhishTank` and `MarkMonitor`. While considering domain name lengths, we ignored the TLDs since they are common across all domains, phishing or otherwise. Figures 1 and 2 show the percent of URLs and domains at various lengths respectively. We note that the distribution of the length of phishing URLs as well as domain names is quite different for `DMOZ` than `PhishTank` or `MarkMonitor`. While the URL lengths peak at 22 characters for `DMOZ`, they peak at 67 for `PhishTank` and at 107 for `MarkMonitor`. Also, very few URLs in `DMOZ` have a length longer than 75 characters while an insignificant percentage of URLs in `PhishTank` and `MarkMonitor` have lengths longer than 150 characters.

Phishing domains (without the TLD portion) tend to be shorter than regular domains contained in `DMOZ` data (see Figure 2): while the `DMOZ` domains peak at 10 characters, phishing domains peak at 7 characters. Further, the peak is more pronounced than `DMOZ` domains for both `PhishTank` and `MarkMonitor`. This indicates that longer URLs and shorter domain names could be used as two of the heuristics to identify phishing.

Domain name character composition: Relative letter frequencies of characters in English language are well known. We were curious how letter frequencies of domain names hosting phishing sites compared. In Figure 3, we show the relative letter frequencies in English, the `DMOZ` data set, and our primary data sets, `PhishTank` and `MarkMonitor`. We stripped off the TLDs from the domain names from this analysis as well because the

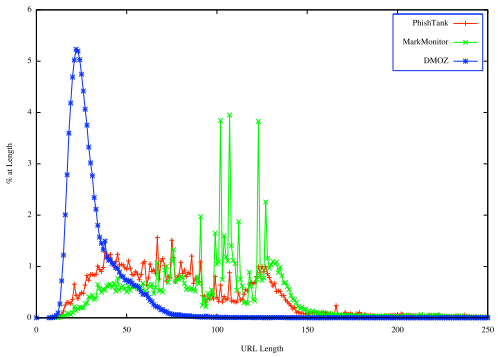

Figure 1: Distribution of URL lengths.

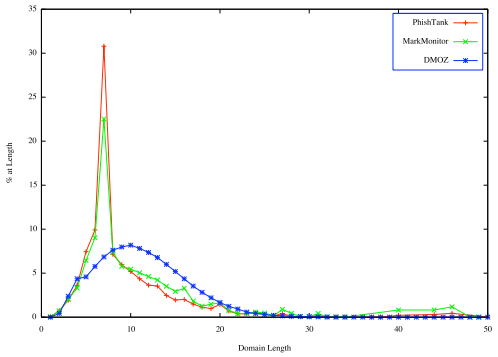

Figure 2: Distribution of domain name lengths.

TLDs are common across all domains. We find that the `DMOZ` domains resemble English very closely while both `PhishTank` and `MarkMonitor` have less pronounced peaks at each of the vowels. This implies that phishing domains tend to use fewer vowels. Another noteworthy thing is the relative popularity of letters of the English language: while letters 'a', 'c', and 'e' have significantly different probability of appearing in an English document or a `DMOZ` domain name, they have very similar probabilities of occurrence in phishing domains. Both these characteristics can be used as additional heuristics to flag domains involved in phishing.

Next, we looked at the number of unique characters within the domain name. Our goal was to check if the phishing domains differ significantly from the `DMOZ` domains. Figure 4 shows a comparison. The phishing domains have fewer number of unique characters in their domain name than the `DMOZ` domains: while the number of unique characters in phishing domains peaks at 6 characters, the `DMOZ` domains do not peak until 9 characters. This indicates that domain names with fewer unique characters may be an indicator of phishing. While URL-based tests can only be done once a phishing campaign has started, domain name-based tests can be applied even to new domain registrations. This can help thwart phishing activity that relies on new domain registrations.

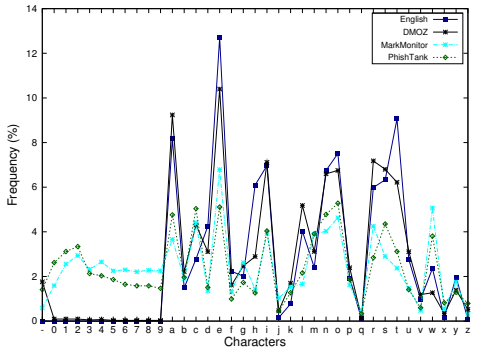

Figure 3: Comparison of letter frequencies in English, DMOZ, PhishTank, and MarkMonitor.

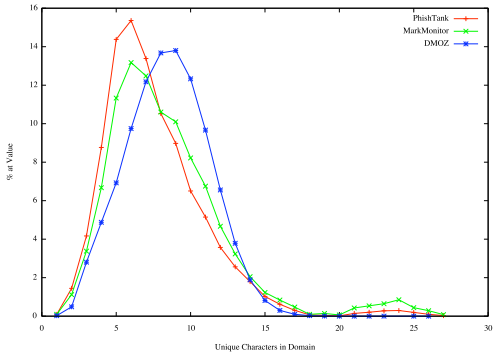

Figure 4: Number of unique characters within the domain name.

Presence of brands in URLs: Phishing URLs often contain the names of the brands they target. It could be to lure users or for accounting purposes. To quantify this phenomenon, we used `MarkMonitor-2006` data set to collect a list of brand names. These brand names are in the form of official company name. For the 51 brand names that were targeted by 97.88% of the phishing URLs in that data set, we manually derived their domain names, including any obvious canonicals. For example, for the hypothetical "Fourth Ninth Bank Inc.", if we found that domain names `fourthninth.com` and `49.com` are canonicals for this brand because they take a customer to the same site, we added it our list of domain names for that brand. This gave us as total of 79 domain names to look for in our primary data sets, `MarkMonitor` and `PhishTank`.

The domain names for the brands showed up both in host name and path portions of the URLs, sometimes with the TLD included and sometimes without. We categorize them in Table 3. The five categories in this table are given below. Clearly, a large percentage of phishing URLs and FQDNs contain brand names in them, more often with TLD than not. This fact can be used as a criterion for filtering phishing e-mails.

- *FQDN without TLD:* The brand domain name was a part of FQDN but without TLD.
- *FQDN with TLD:* The brand domain name was a part of FQDN with TLD.
- *Path without TLD:* The brand domain name was a part of path without TLD.
- *Path with TLD:* The brand domain name was a part of path with TLD.
- *Path and FQDN:* The brand domain appeared in both path and FQDN, regardless of whether the TLD was present or not.

Category	`PhishTank`	`MarkMonitor`
FQDN without TLD	5.71%	1.49%
FQDN with TLD	13.86%	20.42%
Path without TLD	10.99%	8.06%
Path with TLD	11.46%	8.44%
Path and FQDN	10.57%	39.39%
No brand found	47.42%	22.21%

Table 3: Percent of URLs with brand domain name in each data set.

Misuse of URL-aliasing services and free Web hosting services: Long URLs pose difficulties when they are cut and paste. Starting with TinyURL [9], many URL-shortening services have become popular. They create short URLs, enhancing the ease with which URLs can be shared. Most URL-shortening services are free. Some even allow tracking the clicks to the shortened URLs. We tested the phishing URLs in both our data sets against a recent list of URL-shortening services [19]. Both our data sets showed evidence that phishers are exploiting these services. In `MarkMonitor`, a total of 41 such URLs were found. 34 of these exploited TinyURL. The rest exploited 5 other URL-shortening services. The numbers were much higher for `PhishTank`, where a total of 176 such URLs were found. 116 of these exploited TinyURL. The rest exploited 10 other URL-shortening services. Though these numbers are not large, clearly, phishing is abusing URL-shortening services.

Comparing this trend with respect to our older data set, `MarkMonitor-2006`, we find that URL-shortening services were exploited in 2006 as well. Over a period of 7 months, we found 123 cases exploiting 10 different URL-shortening services. Most of these misused TinyURL and NotLong, another service that was not exploited as much in our newer data sets.

We also found evidence that phishers were misusing free Web hosting services as well. Specifically, we found that 17 different Web hosting services contained a total of 671 phishing sites from out `MarkMonitor` data set. The most exploited Web hosting service, `land.ru`, contained 101 phishing sites. Both these findings point to the need for better scrutinizing the users of such services.

5.2 Registration of Phishing Domains

Many of the domains used for phishing are registered exclusively for phishing. Our goal in this section is to estimate to what extent are new domains being registered for phishing. We also want to find out how long phishers wait to serve the phishing site on the newly registered domain. This knowledge can be used to watch new domain registrations for defending against phishing activity.

In addition to our primary data sets, `MarkMonitor` and `PhishTank`, we used the *whois* data and the *zone files* for this analysis. The *whois* records we collected contain the creation and expiration dates of domains. These fields are trustworthy since they are registrar provided. In practice, we have found that the registrars often times do not provide information on domain deregistrations. Further, more than half of *whois* queries fail either because the servers are not available on standard ports[1] for certain TLDs or they do not respond. To overcome this shortcoming, we supplement *whois* data with the zone files that we have been collecting since 10/10/2007. The limitations of zone files are two-fold. First, we do not have zone files available for all the TLDs. We have them for the biggest TLD, `.com`, and `.net`, `.biz`, and `.info`. Second, they provide no information for domains registered prior to 10/10/2007 since that is when we started collecting them. Also, the zone files are very large, each `.com` requiring approximately 5.3GBytes of storage in uncompressed format. Recall that we process them as soon as they are received each night and store the differences, both in terms of added and deleted domains. This allows us to go through them relatively quickly.

We merged the phishing domains contained in `MarkMonitor` and `PhishTank` for this analysis so we can get a collective view of phishing activity. Of the total $20,313$ unique domains across both the data sets, *whois* data and zone files gave us at least partial information on approximately 50% domains. Using this data, we first examined the length of time between domain registration and the first reported phishing site hosted on the domain. We refer to this as *time to activation* (TTA). We have the registration dates for phishing domains from both *whois* and zone files for $6,969$ domains (34% of unique domains). Figure 5 shows the TTA for these domains. While a few domains stay dormant for a long time before they are used for phishing, the distribution of TTA is virtually a delta function centered at 0. This implies that most domains registered for the purpose of phishing are put to use almost immediately. It is possible that of the rest, many are not registered for phishing. In fact, they might be hacked and then exploited for phishing. These results show that the window to identify suspicious domain name registrations from the perspective of phishing is very small. (Oddly, we also find a few TTA values which are negative. To understand why this was occurring, we consulted the full zone files. These negative values were due to domains which were registered prior to our zone file collection period, but were removed during the collection period, and then re-added. There were 157 such domains.)

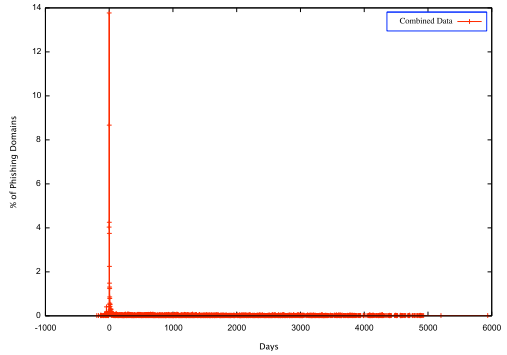

Figure 5: Time to activation of phishing domains.

Domain tasting allows a domain registrant to register a domain and return it within 5 days without incurring any financial liability. This practice has been heavily debated in the Internet community and their are talks about ending it due to its misuse [27]. One of the ways in which domain tasting is feared to be misused is in phishing. A recent study by the APWG [1] discusses two studies which attempt to determine the prevalence of domain tasting. In the first, 793 domains used for phishing were tested for domain tasting. Approximately 20% of them were found to be tasted. In the second study, the converse view was used: all tasted domains for a large gTLD were checked for use by phishers. Of the approximately 3 million domains checked, 10 were found to have been used for phishing. We examine if there is evidence of domain tasting in our data set. Of the $6,969$ domains we have registration dates for, we have drop dates for $2,452$. For these domains, we checked if domain tasting was used for phishing. We found that $1,238$ domains were dropped during the registration grace period of 5 days. This is approximately 6.1% of all domains used for phishing between the two data sets. While it is possible that the registrars actively dropped these domains due to their involvement with phishing, it is likely that many of these domains may actually have exploited domain tasting.

5.3 Phishing Infrastructure

We now examine the infrastructure used for hosting phishing sites. We use the union of the phishing domains from our primary data sets, `MarkMonitor` and `PhishTank`, and the IP addresses got from the DNS

[1] Some registrars, including a big one, *GoDaddy*, prefer to provide the information only over *http*, even when they are required to over port 43.

resolution of phishing domains in this analysis. We find the geographical location of each IP address using IP2Location [12]. Further, for each IP address, we try to infer if it belongs to a residential machine. We do so by using the list of top residential ISPs in the United States [10].

There were a total of $20,313$ unique phishing domains in our combined data set. Of these, 25% did not resolve. Most of these domains came from our `PhishTank` data set, which due to its community-driven nature, is slow to remove inactive phishing domains from its feeds. We ignore these domains from the analysis presented in this section. The remaining $15,380$ domains resolved to $20,466$ unique IP addresses. Less than half of these domains were hosted on a single machine. The rest, 54.4%, were hosted on multiple IP addresses. In fact, 95 phishing domains were hosted on 100 machines or more, with the domain with the highest number of machines hosting it hitting $1,320$. In contrast, work in [2] found that 94% of the Web sites contained in general spam are hosted on a single IP address. Thus it appears that the phishing domains are better provisioned than general spam domains.

The countries machines hosting phishing sites belonged to were diverse. At least 20% of the phishing domains were hosted in multiple countries. Overall, 45% of the machines hosting phishing sites were in the U.S., 35% in Europe, and 10% in Asia. Romania alone accounted for 13.5% of hosting IPs, and is responsible for more IPs than any country except the U.S. In contrast, the work in [2] found the numbers for the U.S., Western Europe (in contrast to all of Europe), and Asia to be, respectively, 57%, 14%, and 13%. Another interesting fact is that most of the top 95 domains which were hosted on a 100 or more machines were registered in `.cn`, `.com`, or `.us` TLDs when the machines hosting these sites did not always belong to the corresponding countries. Thus, the phishing domains were not always hosted in the country they were registered in. Further, many of the IP addresses of the machines hosting phishing sites belonged to residential customers. Specifically, 14% of the IP addresses of the machines hosting phishing sites belonged to customers of residential ISPs in the United States. We plan to examine these issues in more detail in the future.

While many phishing domains were hosted on multiple machines, many machines hosted multiple domains as well. This is not unexpected, given that many domains in the Internet are hosted on various hosting services [25]. In fact, the top 20 IP addresses that hosted multiple phishing domains hosted a total of 96.5% of the domains.

5.4 Lifetimes of Phishing Domains

For a given phishing domain, a logical question to be asked is how long does it last? They are two ways to determine this. The first is by tracking how long a give URL maintains reachability. This approach was taken by the work in [2], which investigated the lifetime of Web sites contained in spam in general. Given that the most common phishing site take down practice involves the registrar suspending the phishing domain, an alternative is to track how long the DNS name corresponding to a phishing site resolves to IP addresses. Our periodic DNS resolutions allowed us to use the latter approach to track the lifetime of phishing domains. Next, we present those results.

Due to the possibility of outdated information in the `PhishTank` data set, which could lead to reregistration of a phishing site by a new innocent party, we perform this analysis on only the `MarkMonitor` data set. We also note that our findings are a lower bound on the lifetime of phishing domains because there is some time lag before they are reported in our data sets. We find that on an average, a phishing domain lasts 3 days, 31 minutes and 8 seconds. Thus, phishing domains last for a much shorter duration than the scam domains, as reported in [2], which found that more than half of the scam sites lasted more than a week. Surprisingly, hardly any phishing domains fall at the average. While most are very short-lived, some last many days. Specifically, about 1/3rd of the phishing domains last 55 minutes while a quarter last almost 12 days. This indicates that a significant fraction of phishing domains either remain undetected or their take-down cannot be accomplished quickly[2].

6 Concluding Remarks

Our preliminary analysis points to some disturbing trends in phishing. It shows that residential machines are potentially being used to host phishing sites. Even free Web hosting and free URL-shortening services are being exploited by phishers. More needs to be done to understand the infrastructure phishers are putting up. An aspect that we plan to examine in the future is that is that of DNS misuse. In particular, we are currently collecting data to understand the extent to which *double-flux* [13], where phishers put up fast-changing DNS servers to avoid take-down, is being used in phishing.

Acknowledgments

This work would not have been possible without the real-time feed of phishing URLs. We thank PhishTank and MarkMonitor for these data sets. John LaCour from MarkMonitor and Laura Mather from the APWG have helped shape the issues addressed in this paper. Moheeb Rajab helped much with the presentation of this paper. All their help is greatly appreciated.

[2]MarkMonitor is often times involved in taking down phishing sites but they are not involved in all such attempts.

References

[1] Greg Aaron, Dmitri Alperovitch, and Laura Mather. The relationship of phishing and domain tasting. White Paper, September 2007.

[2] David S. Anderson, Chris Fleizach, Stefan Savage, and Geoffrey M. Voelker. Spamscatter: Characterizing internet scam hosting infrastructure. In *USENIX Security*, 2007.

[3] APWG. Anti-phishing working group. Electronic, 2008.

[4] Lorrie Cranor, Serge Egelman, Jason Hong, and Yue Zhang. Phinding phish: An evaluation of anti-phishing toolbars. In *Network & Distributed System Security (NDSS) Symposium*, 2007.

[5] Rachna Dhamija, J. D. Tygar, and Marti Hearst. Why phishing works. In *ACM Computer/Human Interaction Conference (CHI)*, 2006.

[6] Ian Fette, Norman Sadeh, and Anthony Tomasic. Learning to detect phishing emails. In *ACM International conference on World Wide Web (WWW)*, 2007.

[7] Mozilla Foundation. Public suffix list. http://publicsuffix.org/list/, 2008.

[8] Sujata Garera, Niels Provos, Monica Chew, and Aviel D. Rubin. A framework for detection and measurement of phishing attacks. In *ACM Workshop on Recurring Malcode (WORM)*, 2007.

[9] Gilby Productions. TinyURL. http://tinyurl.com/.

[10] Alex Goldman. Top 23 U.S. ISPs by subscriber: Q3 2007. http://www.isp-planet.com/research/rankings/usa.html, 2007.

[11] G. Goth. Phishing attacks rising, but dollar losses down. *IEEE Security & Privacy*, 3(1):8–, Jan.-Feb. 2005.

[12] Hexasoft Development Sdn. Bhd. IP2Location geolocation service. http://www.ip2location.com/, February 2008.

[13] ICANN Security and Stability Advisory Committee. SAC advisory on fast flux hosting and DNS. http://www.icann.org/committees/security/sac025.pdf, January 2008.

[14] Christian Ludl, Sean McAllister, Engin Kirda, and Christopher Kruegel. On the effectiveness of techniques to detect phishing sites. In *DIMVA*, 2007.

[15] MarkMonitor, Inc. http://www.markmonitor.com, 2008.

[16] Tom McCall. Gartner survey shows phishing attacks escalated in 2007. http://www.gartner.com/it/page.jsp?id=565125, December 2007.

[17] Tyler Moore and Richard Clayton. An Empirical Analysis of the Current State of Phishing Attack and Defence. In *Workshop on the Economics of Information Security*, 2007.

[18] Netscape. Open directory project. http://www.dmoz.org.

[19] Palin Ningthoujam. Url toolbox: 90+ url shortening services. http://mashable.com/2008/01/08/url-shortening-services/.

[20] OpenDNS. PhishTank. http://www.phishtank.com/, 2008.

[21] Anirudh Ramachandran and Nick Feamster. Understanding the network-level behavior of spammers. In *ACM SIGCOMM*, 2006.

[22] Anirudh Ramachandran, Nick Feamster, and Santosh Vempala. Filtering spam with behavioral blacklisting. In *ACM Conference on Computer and Communications Security (CCS)*, 2007.

[23] Zulfikar Ramzan and Candid Wüest. Phishing attacks: Analyzing trends in 2006. In *Conference on Email and Anti-Spam (CEAS)*, 2007.

[24] RSA Security. Phishing special report: What we can expect for 2007. White Paper, 2006.

[25] Craig Shue, Andrew Kalafut, and Minaxi Gupta. The Web is Smaller than it Seems. ACM SIGCOMM Internet Measurement Conference (IMC), 2007.

[26] VeriSign. The domain name industry brief. http://www.verisign.com/static/043194.pdf.

[27] Mark W. Is Google bringing an end to domain tasting? http://www.workboxers.com/2008/01/25/is-google-bringing-an-end-to-domain-tasting/.

[28] Liu Wenyin, Guanglin Huang, Liu Xiaoyue, Zhang Min, and Xiaotie Deng. Detection of phishing webpages based on visual similarity. In *Special interest tracks and posters of the ACM International Conference on World Wide Web (WWW)*, New York, NY, USA, 2005.

[29] Y. Zhang, J. Hong, and L. Cranor. Cantina: A content-based approach to detecting phishing web sites. In *ACM International Conference on World Wide Web (WWW)*, 2007.