Clustering Potential Phishing Websites Using DeepMD5

Jason Britt, Brad Wardman, Dr. Alan Sprague, Gary Warner
Department of Computer & Inf. Sciences
University of Alabama at Birmingham
Birmingham, AL 35294

Abstract

Phishing websites attempt to deceive people to expose their passwords, user IDs and other sensitive information by mimicking legitimate websites such as banks, product vendors, and service providers. Phishing websites are a pervasive and ongoing problem. Examining and analyzing a phishing website is a good first step in an investigation.

Examining and analyzing phishing websites can be a manually intensive job and analyzing a large continuous feed of phishing websites manually would be an almost insurmountable problem because of the amount of time and labor required. Automated methods need to be created that group large volumes of phishing website data and allow investigators to focus their investigative efforts on the largest phishing website groupings that represent the most prevalent phishing groups or individuals.

An attempt to create such an automated method is described in this paper. The method is based upon the assumption that phishing websites attacking a particular brand are often used many times by a particular group or individual. And when the targeted brand changes a new phishing website is not created from scratch, but rather incremental upgrades are made to the original phishing website. The method employs a SLINK-style clustering algorithm using local domain file commonality between websites as a distance metric. This method produces clusters of phishing websites with the same brand and evidence suggests created by the same phishing group or individual.

1. Introduction

Phishing websites attempt to convince people to deliver their passwords, user IDs and other sensitive information by mimicking legitimate websites such as banks, product vendors, and service providers. Security workers at the victim brand, the one that is being imitated, must determine from a large collection of potential phishing sites which sites are phishing sites and targeted at their institution. These URLs generally are gathered from sources such as forwarded emails to the "abuse" email for the brand, or via emails from the Customer Service department from the brand. From this large collection of emails, URLs are extracted, but then must be reviewed to find the phishing sites which need to have an incident response action taken against them. Manual review of these assorted complaint-generated URL lists is a time-consuming process that in some cases is actually abandoned due to the significant manpower costs associated with the activity. Automated methods need to be created that group large volumes of phishing website data and allow investigators to focus their investigative efforts on the largest phishing website groupings that represent the most prevalent phishing groups or individuals.

An attempt to create such an automated method is described in this paper. The method is based upon the assumption that phishing websites attacking a particular brand are often used many times by a particular group or individual. And when the targeted brand changes a new phishing website is not created from scratch, but rather incremental upgrades are made to the original phishing website. The method employs a SLINK-style clustering algorithm using local domain file commonality between websites as a distance metric. This method produces clusters of phishing websites with the same brand and evidence suggests created by the same phishing group or individual.

The method has been evaluated against a collection of phishing sites in the UAB Phishing Data Mine [12] and has been shown to successfully create clusters of confirmed phishing URLs, grouped first by brand, and then by significant subgroups composed of many identical files. One common method of distributing phishing websites is by using "Phishing Kits", which are zip files used to store all the files and directory structures necessary to create a phishing website. These phishing kits are often used repeatedly by a single criminal or criminal group to create phishing sites. There is limited evidence showing the significant subgroups of identical files are in fact groups of highly related phishing kits or phishing kit families used or created by a small group or individual.

To implement such a clustering technique a SLINK-style algorithm [7] was implemented using a file set comparison method, Deep MD5 Matching [14], as a distance metric. The clustering technique was limited in its ability to process large volumes of phishing data as runtime increased substantially as the data set was increased. To reduce run time, the phish clustering technique was divided into two phases. Phase1 is used to reduce the number of comparisons while phase2 implements the SLINK algorithm. Additionally, to reduce the run time the two-phased approach was run over chronologically limited windows of one month each.

The resultant clusters are evaluated based upon brand and phishing kit relationships. Clusters are evaluated to see if cluster members agree on brand. A small number of the largest potential phishing clusters are evaluated to see if individual phishing kits relate to one or multiple phishing clusters. And a case in point example is presented showing the relationship between phishing kits.

2. Related Work

Phishing researchers have presented a number of classification methods for identifying phishing attacks. These methods can be categorized into three groups: email-, URL-, and content-based approaches. Email-based approaches are used to prevent the phishing attack from reaching the intended recipient. Some researchers classify the words in the email body to determine the legitimacy of the email [10]. Other email-based approaches use features derived from the email message such as the sender email, sender IP address, and non-matching URLs between the hyperlink and anchor tag [1]. These features are used to classify the email through machine learning algorithms [1][4]. One issue with email filters is the vast number of email hosting providers that do not provide their users with this type of protection.

In response, research has been conducted to determine phishing attacks through the browser. Content- and URL-based approaches have been suggested for detection by the browser. *Gyawali et al.* and *Ma et al.* proposed solutions to phishing identification by using features that can be derived from a URL [6][9]. These researchers demonstrated that URL-based methodologies can identify phishing URLs with high accuracy; however, such techniques can be attacked causing lower detection rates by shortening the phishing URLs or hosting the website in the root directory. Content-based approaches use the content of the website for detection. *Dunlop et al.* presented a method for determining the visual similarity between screenshots of phishing websites [3]. Other researchers have used components within the source code [2][11]. While such approaches have demonstrated good detection and false positive rates, there are attacks against these methodologies as well. Therefore, there have been a number of researchers that use combinations of all three categories [16][17][18].

Phishing website aggregation has been an area of interest for researchers that are proactively trying to determine the prevalence of the criminals behind the phishing attack [2][8][15]. Phishing actors used to create domains on the same IP blocks. In response, *Weaver and Collins* presented a clustering algorithm using the IP address or network hosting the phishing website as a measure of prevalence [15]. The researchers in *Wardman et al* suggested that domains compromised by the same attack may indicate the same phisher [13]. This research presents a method for aggregating phishing websites by phishing actor using a content-based approach that is based upon the phishing website's files. Utilizing a content-based approach based upon the phishing website's files is harder for prevalent phishing actors to avoid than approaches based upon domain and/or IP.

3. Data Set

The data set for this research was collected through the UAB Phishing Data Mine [12] from January 1st, 2011 to May 25th 2011. This data set consists of 265,611 potential phishing websites collected from a large spam-based URL provider, a large anti-phishing company (Internet Identity), and a number of other feeds including private companies, security companies, and financial institutions. The source of the URLs is either URLs contained in spam or URLs reported by the public to fraud alert email addresses.

Comparing phishing collections is an inexact science, due to a number of disagreements throughout the industry on how phishing pages should be counted. Many vendors and public sources count each occurrence as distinct if there is any variation in the URL, which leads to extreme "over-counting" in conditions where URLs are customized per user, or where randomization is combined with a wild-card DNS entry to allow every domain name to be unique. Virtual hosts, where the same directory path can be resolved for every domain hosted on a single IP address also lead to over-counting by some other sources. UAB uses a conservative counting mechanism that attempts to deduplicate URLs that are actually the same phishing content prior to counting. UAB's data is biased in favor of phishing against financial institutions and currently underrepresents gaming and social media phishing when compared to some other phishing collections.

The data consists of all files referenced in the potential phishing website that were hosted on the same domain as the potential phishing website. The website files were fetched using an automated web crawler that makes use of GNU's Wget[5]. After the files were downloaded, a hash value was generated for each file using the MD5 hashing algorithm. A combination of human and automatic labeling was employed by the UAB Phishing Data Mine to determine whether the website was a phish or legitimate website. The automatic labeling strategies include main page hash matching, and Deep MD5 [14] matching. These two automatic labeling strategies depend upon already detected and branded phish to brand incoming phish. If the potential phish is labeled as a phishing website by a human, then an associated brand is chosen by the human. If the potential phish is labeled as a phishing website using automatic confirmation then the confirmed phishing website is given the same brand as the phishing website used to confirm it. This data set contains 349 different spoofed organizations. Out of all of the potential phishing websites received, the data set consists of approximately 38% manually or automatically confirmed phish, 12% manually or automatically confirmed non-phish, 30% marked as unreachable or fetching errors, and 20% marked as unconfirmed. The data was split into five time windows based upon the month the potential phishing website was first observed.

4. Algorithms

A two phased approach is used to cluster the potential phishing websites. Phase1 creates website clusters based upon an exact match of the MD5 value of the main index page of the phishing website. All pages are placed into phase1 groups based upon their main page's MD5 value. If a page has no matches, it is placed into a phase1 group consisting of only itself. Phase2 employs a SLINK-style algorithm [7] using Deep MD5 Matching as the distance metric between phase1 clusters.

4.1. Deep MD5 Score

Deep MD5 generates a score using the count of candidate one's files (count1), the count of candidate two's files (count2), and the number of matching MD5 values between candidate one and candidate two (overlap). A Kulczynski 2 coefficient is then applied to count1, count2, and overlap to generate the Deep MD5 score.

$$Deep\ MD5\ Score = 0.5\left(\frac{overlap}{\text{count1}}\right) + 0.5\left(\frac{\text{overlap}}{\text{count2}}\right)$$

For example two websites, website X and website Y, could be compared using Deep MD5 Matching. If website X's html code makes references to local domain files {a,b,c,d,e} and website Y's html code makes references to local domain files {a,b,f,g} then the overlap count between the two websites' file sets is two (overlap). Website X's file count is five (count1) and website Y's file count is four (count2). Then the Deep MD5 score is 0.5(2/5) + 0.5(2/4) or 0.45.

4.2. SLINK Clustering Algorithm

The SLINK clustering algorithm is a graph theoretic clustering algorithm. The graph has vertices (potential phishing websites). For each pair of vertices a score (DeepMD5 score) is generated and an edge is drawn when the score exceeds some threshold. After all edges have been created each connected component is a cluster.

The Phase2 clustering algorithm depends on setting an appropriate threshold between 0 and 1. After performing a statistical analysis of the DeepMD5 threshold, a 0.8 threshold was chosen to make sure clusters consisted of highly similar sites. Phase2 clustering implements the SLINK-style algorithm by selecting a phase1 candidate as a seed for the phase2 cluster from a phase1 representative list. Next, an MD5 similarity coefficient is generated between the phase1 candidate and all other phase1 representatives. If the similarity coefficient meets or exceeds the phase2 clustering threshold, then it is added to the phase2 cluster and removed from the phase1 representative list. The above operation is then recursively applied to every new member of the phase2 cluster. When all members of the phase1 representative list that match the current phase2 cluster have been assigned, a new representative is chosen to create the next phase2 cluster. The process repeats until all phase1 representatives have been assigned to a phase2 cluster.

4.3. Associating Phishing Kits

To determine how similar phishing website cluster members are to a particular phishing kit a file set comparison was made between the phishing kit files and phishing website files. A phishing kit can contain files such as php or other scripting files that dynamically construct html to display or implement functionality that is not observable in the web browser, such as emailing the stolen personal information to the criminal. Since phishing kits contains many files that cannot be seen when the phishing website is downloaded the Kulczynski 2 coefficient would not be suitable because half of the score is derived by using the larger phishing kit file set count as a denominator. Instead the Simpson coefficient was used, which generates a score by taking the number of files shared by the phishing kit and the potential phishing website (overlap) and dividing by the potential phishing website file count (count1).

Phishing kits were chosen from the UAB Phishing Data Mine collected between July 1st 2010 and November 30th 2011 consisting of 27,801 phishing kits. Many of the phishing kits are duplicates or very similar to one another. A similarity score is generated by comparing a phishing website and phishing kit. An association between the two is made only if the Simpson coefficient is 0.8 or greater. The 27,801 phishing kits had 1,485,774 files causing a high run time when creating connections between all phishing kits and suspected phishing websites. Hence, connection scores are only calculated between suspected phishing websites in the top 24 clusters and the 27,801 phishing kits to maintain a reasonable runtime.

5. Results and Discussion

The SLINK-style clustering algorithm generated 185,892 clusters for the five monthly data windows combined. Of which there are 162,206 singleton clusters and only 22,904 multi-member clusters. Out of the 22,904 multi-member clusters there are 14,129 clusters where all cluster members have been assigned a brand. The 14,129 multi-member branded clusters are going to be the focus as singleton clusters are would not be helpful to an investigation and branded clusters establish a ground truth allowing for evaluation.

5.1. Branding

In the 14,129 multi-member branded clusters there are 199 brands. Out of these 14,129 multi-member branded clusters there were only seven where all members did not have the same brand. Approximately 99.5% of multi-member branded phishing website clusters are pure brand and approximately 0.5% contain multiple brands or are cross-branded.

Measure	January	February	March	April	May
Homogeneity	0.9998	0.9994	0.9996	0.9992	0.9989
Completeness	0.5551	0.4123	0.4656	0.4665	0.4812

Table 1: Phish Branding Measures

In Table 1, the completeness scores reflect that several different website templates may be used by criminals to imitate a single brand. The high homogeneity score reflects the high brand purity of the clusters.

5.2. Cross Branded-Clusters

The seven cross-branded clusters are the result of a shared structure that produces a similar look and feel between the cluster members, but targets different brands. These websites use the same template for creating phishing websites for different brands. For example, the three websites in Figure 1 are from one of the seven cross branded clusters and all have the same look and feel. The cluster contained four websites branded MasterCard, two websites branded Key Bank, and the rest of the websites were branded Bank of America.

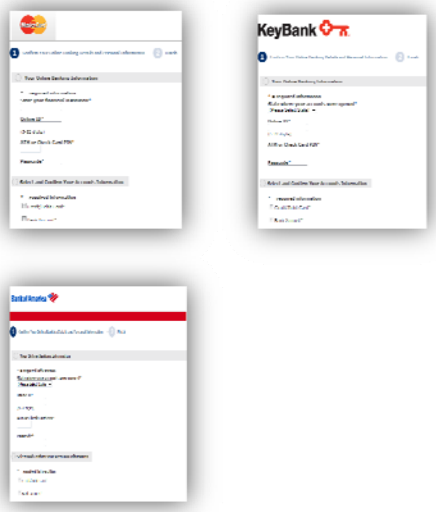

Figure 1: Example Master Card, Key Bank, and Bank of America Branded Phish

All the examples are very similar except for the logo at the top of the page. Observing the similarity scores and file counts between the three samples demonstrates that most of the files in the website file sets are the same. The main index page for each of the three sample files is different, as well as various other files consisting of several .gif, a .js, and an .ico. The phishers have changed only the images or fields listed on each website creating a generic phishing website. Therefore, generic phishing websites may share a similar structure that creates a similar look and feel, but target different brands.

5.3. Phishing Kit

Out of the largest 24 potential phishing website clusters there were eleven potential phishing website clusters where the files from a cluster member were highly correlated with an archived phishing kit. All eleven of the potential phishing website clusters that had a phishing kit associate to a member has the property that every member is a branded phishing website. There are four clusters for Financial Institution A, three clusters for Payment Processor A, three cross-branded clusters, and one cluster for Financial Institution B. All of the eleven phishing clusters have many different kits relating to most members and all of these clusters have some kits that related to only a few members.

Measure	January	February	March	April	May
Cluster Count	2	3	4	1	1
Homogeneity	0.061	0.111	0.138	0.000	0.000
Completeness	1.000	1.000	0.999	1.000	1.000

Table 2: Phish Kit Measures

Table 2 presents homogeneity and completeness scores for phish kits relating to phishing website clusters. The high completeness scores show phishing kits usually associate with members from a single phishing cluster. The low homogeneity scores show that the relationship of kits to clusters is many kits relate to one phishing cluster; a typical cluster is related to many kits. However, the small number of phishing clusters per window may be skewing the results. Especially given there was only a single phishing cluster in this evaluation for April and May and less than four clusters for each of the other monthly windows. Further work needs to be done to provide stronger support for phishing kit completeness within a phishing cluster.

Out of the 27,801 phishing kits compared to the top 24 clusters there are 8,489 kits that associated to at least a single member of a phishing cluster. The 8,489 phishing kits related to 6,458 phishing cluster members in one of the eleven phishing clusters. Six phishing kits related to more than one of the eleven branded phishing clusters within the same monthly window. One of the kits related to a February window Financial Institution A phishing cluster and a cross-branded phishing cluster. Five kits related to a March window Financial Institution A phishing cluster and a March window cross-branded phishing cluster.

There is a “one to many” relationship between phishing website clusters and phishing kits. Phishing kits were not de-duplicated and there can be multiple incremental versions of the same phish kit in the UAB Phishing Data Mine. The large number of kits relating to a large number of a phishing cluster’s members and not another phishing cluster could be the result of duplicate or very similar phishing kits. The fact that out of 8,489 phishing kits only six related to multiple phishing clusters in the same time window and the high completeness measures suggest that the kits relating to a single phishing cluster are strongly associated to each other. This supports the idea that the phishing kits are duplicates or updated versions of the same phishing kit.

5.4. Sample Phish Cluster Analysis

To show how cluster members relate to each other, other phishing cluster members, and to phishing kits an example is given. Three URLs have been chosen from one of the largest financial institution phish clusters (Phish Cluster 1) for January. The cluster has 364 members, all of which are phishing the same financial institution brand. Three phishing websites have also been chosen from a smaller financial institution phishing cluster (Cluster 2) of the same brand as Phish Cluster 1. Phish Cluster 2 has 115 members. Three phish kits have been chosen that associate to the members of Phish Cluster 1 and Phish Cluster 2. Phish Kit 1 associates with Phish 1, Phish Kit 2 and associates with 170 out of Phish Cluster 1’s 364 members. Phish Kit 2 associates with Phish 2 and Phish 3. Also Phish Kit 2 associates with 328 out of Phish Cluster 1’s 364 members. Phish Kit 3 associates with Phish 4, 5, and 6. Phish Kit 3 associates to 14 out of Phish Cluster 2’s 115 members. The following chart was generated to compare and contrast the local domain files for each phishing website.

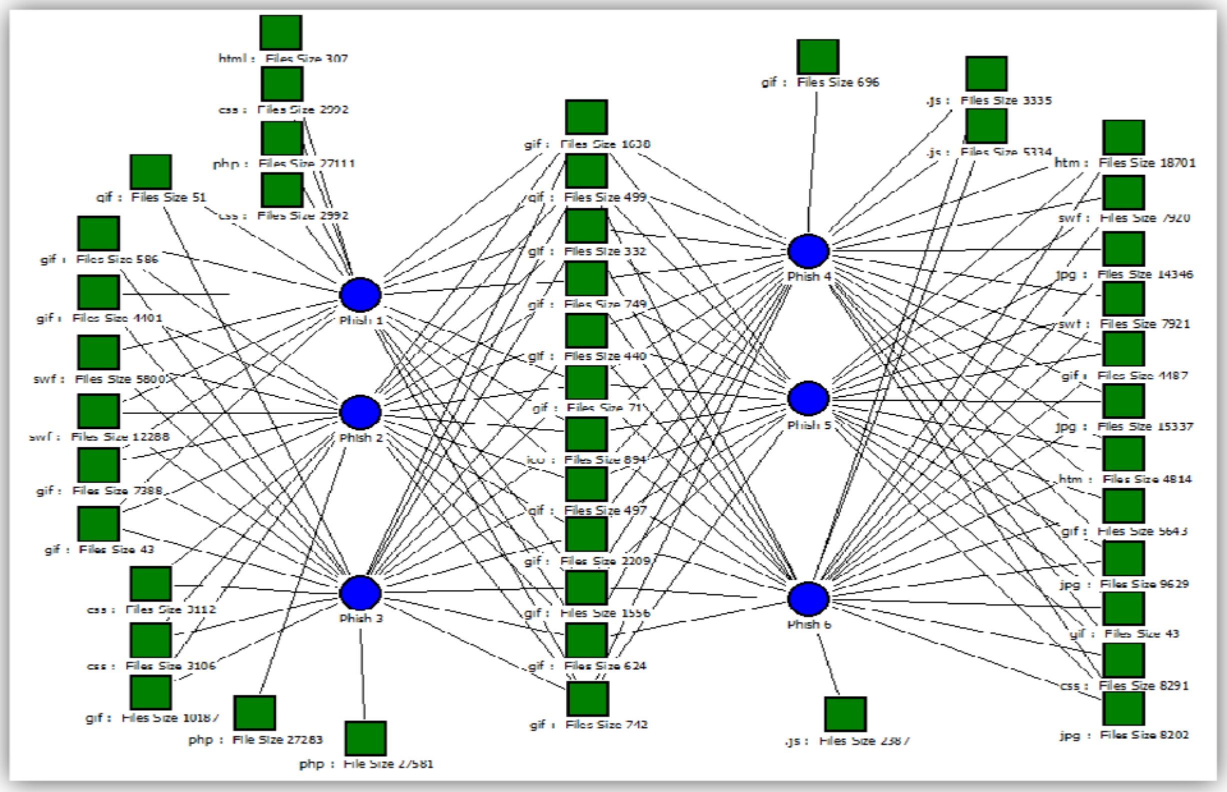

Chart 1: Comparing Phishing Websites' Local Domain Files

The three circles on the left represent the three phishing websites from Phish Cluster 1 and the three circles on the right represent the three phishing websites from Phish Cluster 2. The squares represent the local domain files referenced by the phish and the links between the circles and squares represent the use of the local domain file by the phishing website. The squares are denoted with the file extension followed by the size of the file in bytes.

The only file that is different between Phish 2 and Phish 3 is the php file, which is the file pointed to by the URL received for each phish. Both Phish 2 and Phish 3 have the exact same subset of Phish Kit 2's files and matching directory structures and have the same directory structure as Phish Kit 2. It appears that Phish Kit 2 or a very similar phish kit was used to create both Phish 2 and Phish 3. Phish 1 includes an html and php file. The URL received in the phishing feed references the html file. The html file in Phish 1 is a redirect to the php file. Phish 1 most likely represents a small upgrade to the template to include a re-direct page. Phish 1 is associated to Phish Kit 1 and Phish 1's directory structure matches Phish Kit 1's directory structure. It appears that Phish Kit 1 or a very similar phish kit was used to create Phish 1. Phish Cluster 2's phish (Phish 4,5,and 6) does not use a php page visible to the web. Phish Cluster 2's phish also use .jpg image files instead of only using .gif files like Phish Cluster 1's members. All three of Phish Cluster 2's members are associated to Phish Kit 3, and match Phish Kit 3's directory structure2. It appears that Phish Kit 3 or a very similar phish kit was used to create Phish 4, 5, and 6.

Both Phish Kit 1 and Phish Kit 2 have a directory named "KeNiHack" that contain the local domain files except for html and php files. Phish Kit 1 has a very similar directory structure to Phish Kit 2. It appears that Phish Kit 1 and Phish Kit 2 are related somehow. A clue to their relationship is given by their zip file names. Phish Kit 1 has the same zip file name as Phish Kit 2 except it includes the suffix "update". Phish Kit 1 appears to be an upgraded version of Phish Kit 2.

Comparing Phish Kit 1 and Phish Kit 2 to Phish Kit 3 there are the same noticeable file differences seen in Phish Cluster 1's members when compared to Phish Cluster 2's members. Also, Phish Kit 3 does not even contain a directory named "KeNiHack" and does not share a similar directory structure with Phish Kit 2 and Phish Kit 3. Examining these six phishing websites across two different phishing website clusters shows that Phish 2 and 3 are created by Phish Kit 2 and are present in the same cluster as Phish 1 that was created by an update version of Phish Kit 2, Phish Kit 1. Phish Kit 1 and Phish Kit 2 are a part of the same phish kit family. Also, while members of Phish Cluster 1 and Phish Cluster 2 share similar files they were not created by the same phishing kit. In this example Slink-style Deep MD5 clustering creates cluster based upon brand and phish kit family.

6. Limitations

SLINK Clustering using Deep MD5 as a distance metric does not deal well with all phishing websites. There are several cases where it does have issues. The first case is when there are large numbers of innocuous files such as single pixel image files or common scripts such as web statistics. Large numbers of innocuous files lead to higher DeepMD5 scores where the websites being compared may not have very strong structural similarities, but rather only share innocuous files that are not vitally important to the website.

Another problem case is small file count websites and the UAB Phishing Data Mine contains many small file count potential phishing websites. The Deep MD5 comparison technique is not effective at comparing and linking phishing websites with small file counts [14]. Since the Deep MD5 comparison is used as a distance metric for SLINK clustering the result is a tendency to place phishing websites with small file counts in singleton clusters.

While not necessarily a limitation it should be noted that SLINK Clustering using the Deep MD5 distance metric does not produce clusters based upon similar look and feel, but rather based upon website structure as given by websites utilizing the exact same local domain files. The algorithm does produce clusters where the members share a similar look and feel because they share so many local domain files. However, there can be potential phishing websites that share a similar look and feel that are not clustered together by this algorithm.

7. Future Work

The Deep MD5 distance metric does not perform well when evaluating small file count phishing websites. Using another structural distance metric with a SLINK-style algorithm to cluster only small file count phishing sites instead of the Deep MD5 algorithm may lead to a reduction in the number of small clusters. Syntactical Fingerprinting [14] could be used as a distance metric. Syntactical Fingerprinting utilizes the main page's html code to generate multiple keys for the page. Since potential phishing websites with small file counts will still have a main page the Syntactical Fingerprinting could still be effectively used.

The "one to many" relationship between phishing website clusters and phishing kits needs to be explored further. The limited relations generated between the largest eleven phishing clusters and phish kits shows phishing kits almost always relate to only a single phish cluster. The sample phish cluster analysis section shows on a very limited example that the phish kits that relate to particular phish cluster are from the same phish kit family. To gain more evidence of clustering based on phish kit family further work needs to be done. Creating a clustering algorithm that clusters different versions of the same phishing kit together could result in phishing kit family clusters. Relating phish clusters to kit clusters could provide a run time reduction to allow for relating larger numbers of phishing clusters to phishing kits. Analyzing how the resulting phishing kit family clusters tie to phishing website clusters may show more substantial evidence that SLINK-style DeepMD5 clusters phishing websites based on phishing kit family.

8. Conclusion

SLINK-style DeepMD5 clustering generates clusters where members are highly consistent in brand. Comparing phishing kits to potential phishing website clusters shows phishing kits are only tied to actual phishing website clusters. It also shows phishing kits rarely relate to more than one phishing cluster, which suggest there is some relationship between the phishing kits. There is limited evidence showing the phishing kits relating to the same phishing cluster are from the same phishing kit family. While not substantial there is evidence showing SLINK-style DeepMD5 clustering groups phishing websites based on phish kit families which would allow investigators to target prevalent phishing groups and individuals.

9. References

[1] Abu-Nimeh, S., Nappa, D., Wang, X., & Nair, S. (2007). A Comparison of Machine Learning

Techniques for Phishing Detection. eCrime Researchers Summit (pp. 60-69). Pittsburgh, PA: APWG.

[2] Basnet, R., Mukkamala, S., & Sung, A. H. (2008). Detection of Phishing Attacks: A Machine Learning Approach. Studies in Fuzziness and Soft Computing 226 (pp. 373-383). Springer-Verlag.

[3] Dunlop, M., Groat, S., & Shelly, D. (2010). GoldPhish: Using Images for Content-Based Phishing Analysis. The Fifth International Conference on Internet Monitoring and Protection (pp. 123-128). IEEE.

[4] Fette, I., Sadeh, N., & Tomasic, A. (2007). Learning to Detect Phishing Emails. WWW 2007 (pp. 649-656). Banff, Alberta, Canada: ACM.

[5] GNU wget. GNU Project Free Software Foundation(FSF). http://www.gnu.org/software/wget/wget.html.

[6] Gyawali, B., Solorio, T., Montes-y-Gomez, M., Wardman, B., & Warner, G. (2011). Evaluating a Semisupervised Approach to Phishing URL Identification in a Realistic Scenario. Conference on Email and Anti-Spam. Perth, Western Australia, Australia: ACM.

[7] Han, J., & Kamber, M. (2001). Data Mining: Concepts and Techniques. San Diego, CA, USA: Academic Press.

[8] Irani, D., Webb, S., Griffin, J., & Pu, C. (2008). Evolutionary Study of Phishing. eCrime Researchers Summit. Atlanta, GA: IEEE.

[9] Ma, J., Saul, L., Savage, S., & Voelker, G. (2009). Beyond Blacklists: Learning to Detect Malicious Web Sites from Suspicious URLs. KDD '09. Paris, France: ACM.

[10] Saberi, A., Vahidi, M., & Bidgoli, B. M. (2007). Learn to Detect Phishing Scams Using Learning and Ensemble Methods. Web Intelligence and Intelligent Agent Technology Workshops (pp. 311-314). Silicon Valley, CA: IEEE.

[11] Suriya, R., Saravanan, K., & Thangavelu, A. (2009). An Integrated Approach to Detect Phishing Mail Attacks : A Case Study. SIN '09 (pp. 193-199). North Cyprus, Turkey: ACM.

[12] UAB Phishing Data Mine http://www.cis.uab.edu/UABSpamDataMine

[13] Wardman, B., Shukla, G., & Warner, G. (2009). Identifying Vulnerable Websites by Analysis of Common String in Phishing URLs. eCrime Researchers Summit. Tacoma, WA: IEEE.

[14] Wardman, B., Stallings, T., Warner, G., & Skjellum, A. (2012, February 8). thecenter.uab.edu. Retrieved February 8, 2012, from http://thecenter.uab.edu/media/2011/12/High-Performance-Content-Based-Phishing-Attack-Detection.pdf

[15] Weaver, R., & Collins, M. (2007). Fishing for Phishes: Applying Capture-Recapture Methods to Estimate Phishing Populations. eCrime Researchers Summit. Pittsburgh, PA: APWG.

[16] Whittaker, C., Ryner, B., & Nazif, M. (2010). Large-Scale Automatic Classification of Phishing Pages. Network and Distributed Systems Security Symposium. San Diego, CA.

[17] Xiang, G., & Hong, J. (2009). A Hybrid Phish Detection Approach by Identity Discovery and Keywords Retrieval. WWW 09' (pp. 571-580). Madrid, Spain: ACM.

[18] Zhang, Y., Hong, J., & Cranor, L. (2007). CANTINA: A Content-based Approach to Detecting Phishing Web Sites. International Conference on World Wide Web. Banff, Alberta, Canada.