Knowing your Enemies: Leveraging Data Analysis to Expose Phishing Patterns Against a Major US Financial Institution

Javier Vargas, Alejandro Correa Bahnsen, Sergio Villegas and Daniel Ingevaldson
Easy Solutions Research
Email: {jvargas, acorrea, svillegas, dsi}@easysol.net

***Abstract*—Phishing attacks against financial institutions constitutes a major concern and forces them to invest thousands of dollars annually in prevention, detection and takedown of these kinds of attacks. This operation is so massive and time critical that there is usually no time to perform analysis to look for patterns and correlations between attacks. In this work we summarize our findings after applying data analysis and clustering analysis to the record of attacks registered for a major financial institution in the US. We use HTML structure and content analysis, as well as domain registration records and DNS RRSets information of the sites, in order to look for patterns and correlations between phishing attacks. It is shown that by understanding and clustering the different types of phishing sites, we are able to identify different strategies used by criminal organizations. Furthermore, the findings of this study provide us valuable insight into who is targeting the institution and their *modus operandi*, which gives us a solid foundation for the construction of more and better tools for detection and takedown, and eventually for forensic analysts who will be able to correlate cases and perform focused searches that speed up their investigations.**

Keywords*—*Phishing detection; Data analysis; Clustering; Cybercrime; Feature extraction; Social engineering; Expectation maximization; Forensic analysis

I. Introduction

Phishing attacks have been a growing problem not only in the United States but worldwide [1]. According to the Anti-Phishing Working Group, during 2014 the number of unique phishing sites in the world reached an all time high of 247,713 [2], [3]. By definition, phishing is the act of defrauding an online user in order to obtain personal information by posing as a trustworthy institution or entity [4]. Phishing can also be understood as a social engineering methodology to attempt to manipulate an user, using fear and apprehension to perform certain action in order to obtain private information [5]. For end users, it is usually quite difficult to differentiate between legitimate and malicious sites because they are made to look exactly the same [6]. The objective of the phisher is to make a copy of the site and make it look as similar as possible in order to convince the user to enter their personal information, such as login credentials, banking passwords, and credit card information, among others.

Detection of phishing remains a top concern for financial institutions. The prevention of phishing attacks used to be done by managing blacklists, either by adding functionality in toolbars, appliances and search engines. Blacklists are constructed by a using range of techniques, including manual reporting, honeypots, or by crawling the web in search of known phishing characteristics [7], [8]. However, its expected that many malicious sites are not blacklisted, either because phishing pages are normally only active for approximately three days, with the majority lasting less than a day [8], [9]. To address this issue, machine learning methods have been employed to detect phishing attacks [10], some of them arriving to accuracy levels higher than 99% while having very low false positive rates [9]. Some of the methods that have been used to detect phishing are: support vector machines [11], streaming analytics [12], gradient boosting [13], random forests [14], latent Dirichlet allocation [15], online incremental learning [16], and neural networks [17], among others. The general objective of these machine learning models for phishing detection is to find patterns either in the emails, site URLs or the websites themselves, that helps differentiate between a legitimate and a malicious site.

However, the techniques that are normally used to detect phishing attacks are based on using machine learning models that need a huge amount of marked data to train classification models [13]. This implies that models are not able to detect new or very recent phishing strategies, as in order to train a model, information regarding new phishing attempts must be collected, which is very time consuming. Moreover, machine learning algorithms are built to maximize predictive power of a model and not to understand the underlying behavior behind a prediction [18]. Furthermore, normal studies on phishing attacks are made from an academic perspective in which the objectives are biased towards maximizing the prediction accuracy of phishing detection systems, but on the other hand, knowledge of how the phishers networks behave is uncommon [19], neither is knowledge on their strategies and methodologies [20]. This is a major weakness as these techniques are not giving vital information regarding patterns and correlations between attacks which may help us to build better tools, with the potential to dramatically increasing our ability to fight phishing attacks.

In this paper, using information from a major US financial institution, we apply data analysis in order to generate clusters of the different attacks targeting the institution. In particular, we extract features by analyzing the structure and content of the HTML code of the phishing sites [4], as well as the domain registration records and DNS RRSets extracted using Whois providers and a passive DNS database. With these features we group the attacks in clusters sharing similar characteristics by using the cluster methodology expectation-maximization [21].

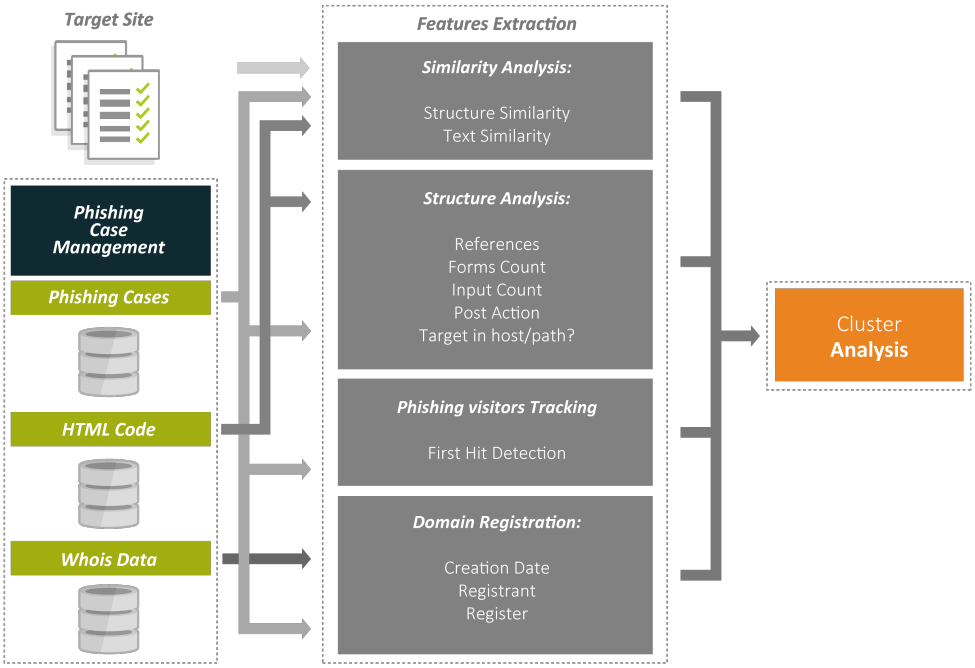


Fig. 1. Proposed framework for analyzing phishing sites.

Finally, by using the different clusters, we are able to gain valuable insights into who is targeting the institution, and more importantly, what is their *modus operandi*, so that the detection and takedown of phishing sites can potentially be improved by incorporating knowledge of the enemy.

The remainder of the paper is organized as follows. In Section II, we discuss current approaches to understand phishing attacks and attackers. Then, in Section III, we present our proposed methodology to cluster phishing cases. Afterwards, in Section IV, we present our results by clustering a collection of phishing attacks targeting a major financial institution in the US. Moreover, we analyze the criminal organizations using the created clusters and the DNS RRSets in Section V. Finally, conclusions and discussions of the paper are presented in Section VI.

II. Related Work

Until very recently work on understanding phishers was non existent. Primarily, studies have focused on extracting strategies by leaving doors open and collecting information on the phishing attacks, mainly done by using honeypots [22], [23]. The objective behind this strategy is to learn the different tactics that phishers use. The forensic data collected is then analyzed to shed light on the technical methods used by attackers.

Other work focuses on understanding phishing and its role in organized crime and money laundering [20]. Among the top strategies used to expose the fraudsters is sending fake credentials to phishing sites, so that phishers can be identified when they try to authenticate in the target portal. This approach can also be understood as fake auditing, in which phishers are allowed to use the indistinguishable fake credentials, so they can be traced and their behavior profiled [19]. In other studies, the focus is to understand the markets where the stolen information is sold [24]. These kinds of studies normally do not distinguish where or how the credentials are obtained by thieves. They could have been stolen using keyloggers, malware, or phishing [25]. Nevertheless, analyzing the underground markets of stolen credentials is a powerful way to understand the motivation behind a phisher, after all, an attacker's objective is not solely to steal someone's personal information, but to sell or use the information for financial gain [26].

Recently, other authors have focused on correlating the behavior of a phisher's activity and websites to be able to identify different types of phishing campaigns [27]. By correlating different phishing websites, researchers are able to identify websites that were created using the same phishing kit [28]. This correlation is very important as it gives law enforcement agencies and the affected institution ways to prioritize the allocation of resources to further investigate and take down phishing sites.

III. Methodology

In this section we present our methodology for clustering the phishing attacks based on analyzing the content and structure of the phishing sites' HTML code, and also by using the domain and DNS RRSets information of the sites. In Fig. 1, we present our proposed framework for analyzing the phishing attacks. First we explain the particularities of the data that we used. Then, we explain the different methods for extracting the features out of the phishing cases. Lastly, we briefly discuss the clustering algorithm and the method for extracting useful information out of the clusters.

A. Data

For this study we used the data collected by the company along more than a year of tracking and takedown of phishing cases for a major financial institution in the US. This data contains all the details regarding the management of each case, including the Whois details of the domain where the attack was hosted, as well as related RRSet records and the HTML code of the phishkit used by the criminal. The service provided by the company also deploys some tools that track the browsing activity of phishing sites; when available we used such data to identify the visitors of each case.

The handling details for each phishing incident are very diverse and it is common to find cases where the system had problems gathering all the data we needed for the purposes of this study, such as in geolocation-protected attacks where the HTML code could not be extracted because the criminal setup htaccess files to prevent visitors from accessing the attack when comming from certain areas or from specific IPs addresses. Hence, the data was filtered to include only the most complete cases, and the final analysis was performed on 3,030 phishing cases distributed over the last four months of 2015.

B. Feature Extraction

In order to cluster the aforementioned phishing cases, we extracted several features using four types of techniques. First, we checked the similarity of the phishing site with the target site. Then we investigated the structure and deployment details of the phishing case. Afterwards, we searched for the first hit to a phishing site. Finally, we analyzed the domain registration records of the phishing site.

1) Similarity Analysis: In order to create a set of features that describe how similar is the phishing site when compared to the targeted web page, we extracted two features evaluating the similarity of the web page structure and and its text content.

- *Structure Similarity:* We evaluate the similarity of the structure by comparing the HTML source code. This method is based on comparing the original web site with each phishing site by attribute matching of the HTML tags [4]. In particular, we are interested in counting the number of matching tags and mismatching tags of the phishing sites. This method allow us to have an estimation of a web page's structure-similarity with the targeted web page. It is calculated using the algorithm described in Algorithm 1.
- *Text Similarity:* We calculate the cosine similarity [29] of the text part of the phishing and the target site. The text similarity is evaluated using the following equation:

$$TS = \cos(\theta) = \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\| \, \|\mathbf{B}\|} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2} \sqrt{\sum_{i=1}^{n} B_i^2}},$$

where A_i and B_i as the elements of the documents A and B respectively.

Algorithm 1 HTML code comparison Algorithm [4]

Input: Vector A, Vector B
Output: Structure Similarity Index

Initialization :
$i = 0$;
$j = 0$;
$match = 0$;
while $A(i)$ is empty **do**
  while $B(i)$ is empty **do**
    if $(A(i) == B(i))$ **then**
      if $(A(i).attribute == B(i).attribute)$ **then**
        $match = match + 1$;
        $i = i + 1$;
      else
        $j = j + 1$;
      end if
    else
      $j = j + 1$;
    end if
  end while
  $i = i + 1$;
end while
$SSI = match/.totaltags$
return SSI

2) Structure Analysis: In order to have a set of indicators about how each phishing page was constructed and deployed, we extracted a set of features as follows:

- *Forms Count:* A simple count of how many HTML Forms are included in the page.
- *Forms Data Count:* A simple count of how many data fields are included in the HTML Forms, including both visible and invisible fields.
- *Post Action:* The string describing the location where the main form-data will be submitted.
- *Is it Logon Form:* Whether the name of the main form matches the same name used in the original page or not.
- *Target in Path:* True if the phishing URL contains explicit references to the attacked brand in its path.
- *Target in Host:* True if the phishing URL contains explicit references to the attacked brand in its host as part of the domain name.
- *Is it WordPress:* True if the URL contains keywords typical of compromised WordPress sites.
- Is it User Folder: True if the path belongs to a per-user web directory [30].

3) Phishing visitors Tracking: As described in Section III-A, phishing tracking systems deployed by the company include tools that track activity related to phishing sites. Using that information we can build a fair estimator of who were the first visitors/hits of a phishing site. Therefore, we created the following features: Country of the first hit, region of the first hit, country of the second hit, region of the second hit. Hopefully the first visitor of a phishing page will be highly related to the criminal gang.

4) Domain registration: From the Whois records stored during the management of phishing cases we were able to determine the time elapsed between the domain registration and the phishing event.

C. Clustering

After the different sets of features were extracted from the phishing sites, we proceeded to cluster them using the expectation maximization [21] algorithm. The objective of this algorithm is to find groups of phishing sites that are similar to one another, but not to other sites. We used the WEKA [31] implementation of the algorithm. This implementation estimates the number of clusters using cross validation. In particular, the algorithm sets the number of clusters to one, then it performs k-fold cross validation and estimate the loglikelihood of each fold. Then if the average loglikelihood has increased, the number of clusters is increased by 1 and the process is repeated until no further increases is found.

With this methodology, we could find groups of phishing sites depending on the way the phishing site is created, and where and how the site domain were registered, giving more and better tools to analysts to understand the behavior of attackers.

D. Evaluating Clustering Features Importance

After the clusters are created, its very important to understand which features are the most relevant to cluster the phishing attacks. There are several methods to select from beforehand before clustering takes place (See [32] or [33]). However, as the purpose of this study is to explore clustering models as a tool to describe the nature of the attacks, we are more interested in understanding the importance of the features after the cluster is built.

For this, we used a variation of the method presented in [34], to build a machine learning classifier, in our case we selected extremely randomized trees [35], using the same features and defining the target of the classifier as the cluster that each of the phishing sites belong. Then using the mean decrease impurity [36], we estimated the feature importance in order to understand the underlying behavior of the features in the clustering procedure.

IV. Finding Similarities in Phishing Attacks

In this section we describe the results of applying the proposed methodology to a collection of phishing attacks targeting a major financial institution in the US. We started with the hypothesis that every criminal has different motivations and therefore is interested in capturing different types of information, and also over the time they manage to develop their own unique attacking style and *modus operandi*. For that reason we performed an initial cluster analysis that will segment all the attacks into groups of cases exhibiting similar appearance and structure. Then we decided to perform a fine-grain analysis and investigate each cluster in depth. For that we executed one more cluster analysis over each of the initial clusters and created different sub-clusters using a different set of features that may provide insight about the structure and deployment details of each case.

Fig. 2. Distribution of the phishing sites into the three clusters.

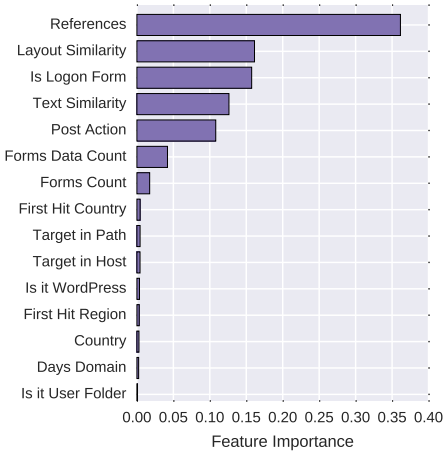


Fig. 3. Clustering feature importance.

A. Clustering the Phishing Attacks

We apply the methodology described in Section III. We first estimate an initial sets of clusters using the expectation-maximization algorithm. We found that the phishing sites were easy to cluster, taking into account the page structure. An example of a typical target website is shown in Fig. 5. Usually a phishing attack is expected to mimic the appearance of the targeted page, however we found several styles and motivations among the groups extracted by cluster analysis. In particular, using the features structure similarity, text similarity and number of references, we found three clusters of phishing sites. Fig. 2 shows the distribution of the different clusters.

Then we gauged the importance of the different features by evaluating the predictive power of each feature in the cluster class, using the method described in Section III-D. In Fig. 3 the importance of the different features is shown. It is observed that the features that help most to differentiate across the clusters are the number of references to the attacked domain and the structure and text similarity features. This gives us an indicator that it is more important to segment the phishing attacks by the web page structure, than the domain registration or the country of the attack.

Accordingly, we analyze the composition of the clusters using the most important features, namely, number of references, structure similarity and text similarity. In Fig. 4, we

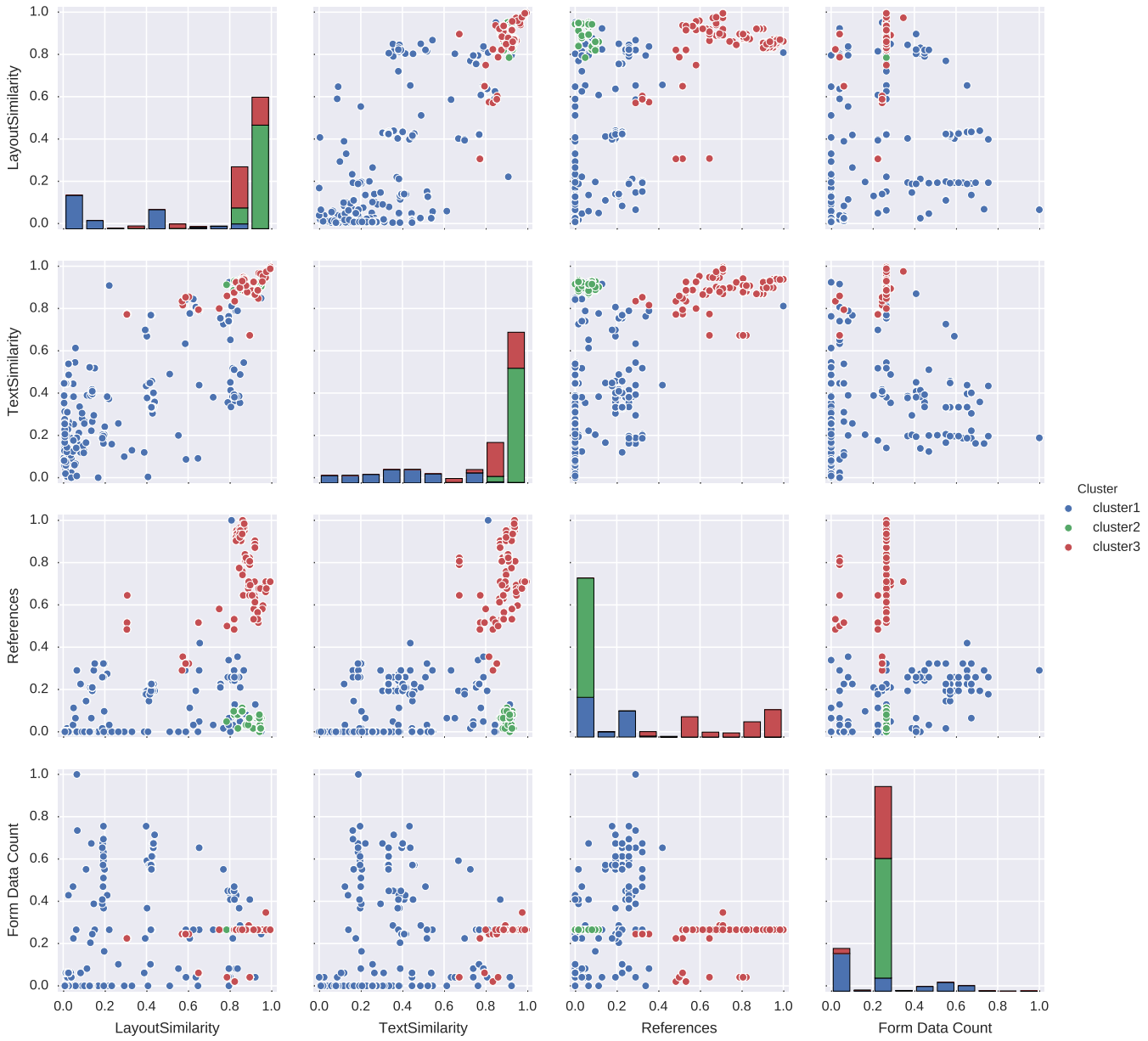


Fig. 4. Analysis of the clusters of phishing attacks. Cluster2 and cluster3 are very similar in the sense that both are characterized by having a very high layout and text similarities with the targeted site. However, when observing the number of references, cluster3 is characterized by having a very high number of references to the original site. Lastly, cluster1 represents attacks that are not intended to visually replicate the target site.

show the scatter plots of the clusters among selected features. We found that cluster2 and cluster3 are very similar in the sense that both are characterized by having a very high layout and text similarities with the targeted site. However, when observing the number of references, cluster3 is characterized by having a very high number of references to the original site, meaning that in these attacks, the phisher is replicating the site but is maintaining an explicit reference to the site's images, JavaScript and CSS codes.

On the other hand, in the cluster2, the attackers are replicating the web page, but they are also downloading all the images and code, and hosting them in their own server. This is very interesting, as it allow us to quickly separate between two attack types that may look the same, but actually are using a very different strategy. In practice, such attacks maintain a close resemblance to the original site and their appearance is exactly the same than the example shown in Fig. 5. However, although it is not a very relevant feature, you can see in Fig. 4 that attacks in cluster3 always keep the same number of form-fields than the original page, while cluster2 contains attacks with different counts.

Cluster1 on the other hand, represents attacks that are not similar to the target page or are do not try to visually replicate it, for example in Fig. 6, the attacker is pasting images of the target site and is just using login and password input fields. Also, in Fig. 7, we show an example of a phishing

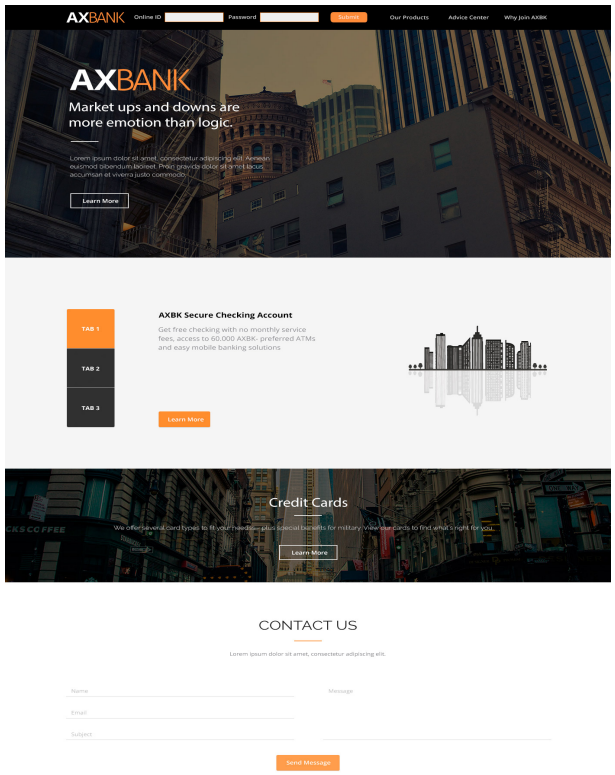


Fig. 5. Example of a typical target website.

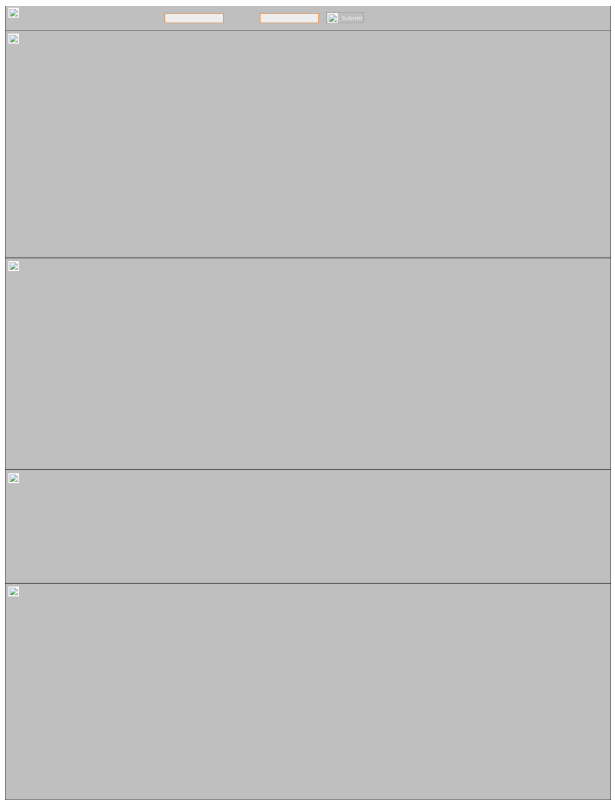


Fig. 6. Example of a phishing site were the phisher is pasting images of the target site and just using the login and password input boxes.

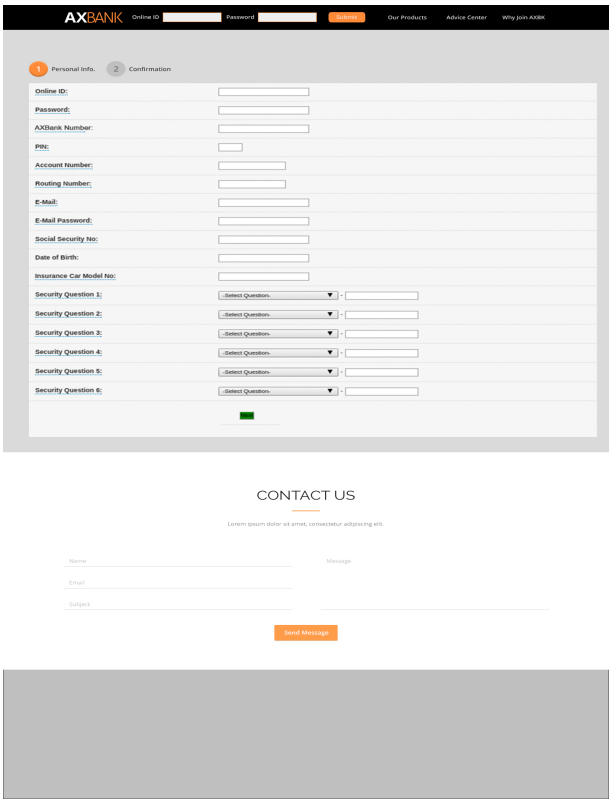


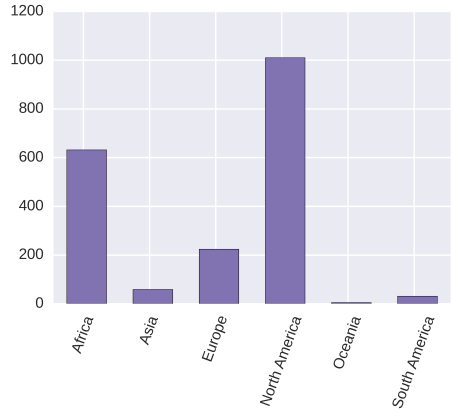
Fig. 7. Example of a phishing site that leaves the header of the target site equal, but then ask a lot of personal information other than the login and password.

site that leaves the header of the target site equal, but then ask a lot of personal information other than the login name and password. These kinds of phishing sites represents 29.7% of all attacks. Form data count exhibits more variability than previous clusters.

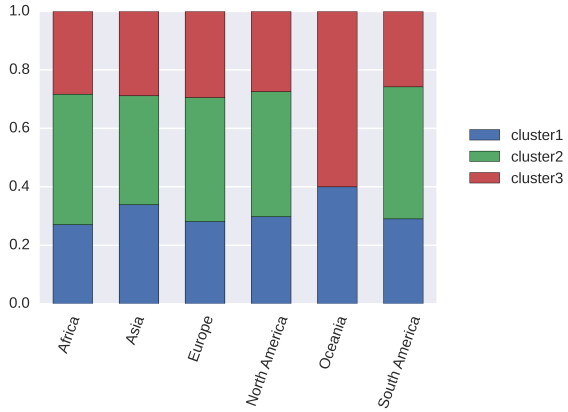
Despite not being one of the most important features for this stage of our study, we also analyzed the region from were the first connection to the phishing site is coming from. In Fig. 8a, we observed that most of the connections are from North America, but most interestingly, almost one-third of the connections are from Africa, most of them from Nigeria. Then, in Fig. 8b, we compare how the sources of the connections are distributed among the clusters, we observed that there is no discrimination of the regions within the clusters and the general population.

B. Clustering the Attackers

Now that we were able to separate the phishing attacks in three groups of different structural style, we wanted to find out if clustering analysis would be suitable to discriminate between gangs of attackers. Consequently, we decided to analyze each cluster independently, keeping those features that provide fine-grain detail of the layout and adding others that add information about the deployment details, such as evidence of being deployed in a WordPress site or explicit reference to the targeted brand in the URL. In the following sections we detail the findings of this analysis.



(a) Distribution of the region of the first connection.



(b) Distribution of the source of the connections between clusters.

Fig. 8. Analysis of the phishing attacks by country.

1) Cluster 1: As described in the previous section, this cluster includes phishing cases in which the page structure differs significantly from the target site, and the count of references to the original page varies a lot among the attacks.

The clustering procedure found four sub-clusters. The phishing cases are distributed as 50.37%, 16.99%, 29.43% and 3.20%, respectively. As can be observed in Fig 9, the most important aspects of the sub-clusters separation is the quantity of the requested fields, the post action, the way they host the phishing site and the origin location of the first visit to the phishing site.

Sub-Cluster 1-1. In this group we found phishing sites that are less similar to the targeted site in terms of structure. In several cases we found only blocks of images that represent the original site. However, in many cases the phishing sites in this sub-cluster, the sites do not offer much detail about the attacker because the pages are not referencing the target site or, the case management systems were not able to collect further information.

Sub-Cluster 1-2. The phishing attacks in this group attempt to collect more information other than the login credentials, such as the user address, ATM PIN, and credit card numbers. Furthermore, these attacks are hosted in legitimate domains that have been compromised. This is evidenced by the fact that domain names are still active and belong to small non-profit or commercial organizations, and also personal websites, with the common denominator being that they use content management systems. Another important aspect of these attacks is that their first visitors where mostly localized in Nigeria and the US, usually in Lagos and Florida.

Sub-Cluster 1-3. The third group includes pages designed to capture much more information than in other kinds of attacks. They ask for challenge questions, and for customer details in a clear attempt get as much information as necessary to get in control of the user account. We could expect these attacks to be an evolution of some of the set contained in the previous cluster, since they show a continuation in the dates of the event, and also because most of their first visitors come from Lagos and Florida.

Sub-Cluster 1-4. This group initially behaves similarly to the previous one regarding the information that is captured. However, these attacks display a higher level of sophistication. For these cases, the attackers procured images to use as a replacement for all the text related to the collection of information. In other words, they embedded specially-designed images to display text words like 'password', 'user name' or 'credit card', instead of using plain text. This is a clear attempt to bypass phishing detection systems based on semantic analysis of the content. Interestingly, first visitors of these attacks were normally based in The Netherlands.

2) Cluster 2: As detailed before, the second cluster aggregates the phishing attacks with high resemblance to the original site. But also, those attacks were phishers that are hosting all the content of their site, therefore, avoiding references to the original web page.

The clustering procedure found four sub-clusters. The sub-clusters are distributed as 15.81%, 20.64%, 24.72% and 38.82%, respectively. The most important features of these sub-clusters is the country of the first hit, if the site is in WordPress, and the days since the domain creation, as is presented in Fig. 10.

Sub-Cluster 2-1. In this first sub-group we found attacks are primary related by the structure of the phishkit and the post action used by the rogue site to collect data. Notably, the first connection to the phishing sites is coming mostly from Lagos, which may relate these attacks to the ones of the sub-clusters 1-2 and 1-3. Moreover, the attackers are also using domains that have been online for more than two years and are registered to small companies or personal pages.

Sub-Cluster 2-2. In this cluster we found cases with extremely low number of references to the targeted site, and some are even hosting JavaScript codes. Aside from being very similar to the HTML structure, these attacks are using mostly WordPress. We could tell that the attacks are part of organized phishing campaigns from criminal gangs specialized in exploiting WordPress vulnerabilities. This allows the attackers to have access to legitimate hosts to stage their phishing pages.

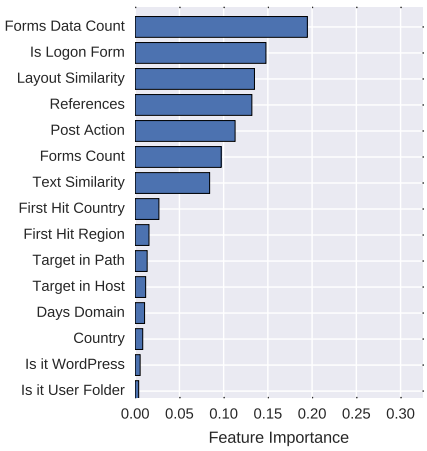


Fig. 9. Feature importance of Sub-Cluster 1. The most important aspects of the sub-clusters separation is the quantity of the requested fields, the post action, the way to host the phishing site and the origin location of the first visit to the phishing site.

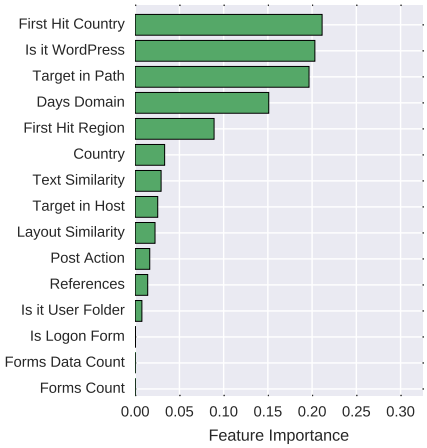


Fig. 10. Feature importance of Sub-Cluster 2. The most important features of this sub-clusters are the country of the first hit, if the site is in WordPress, and the days since the domain creation.

The last two clusters include a series of attacks correlated by site structure and the post action. We found that phishing cases in sub-cluster 4 tend to have domains registered up to four times longer than the domains used by phishers in sub-cluster 2-3. Unfortunately, both clusters lack of a clear pattern that allows us to deeply analyze the attack or the strategies of the attacker.

3) Cluster 3: Cluster 3 is the one grouping that sets up simplest and most easy-to-deploy attacks. This group is expected to be a more diverse group of attacks operated by different criminal organizations focused on zero-day strategy.

The clustering procedure found four sub-clusters. The sub-clusters are distributed as 9.32%, 53.59%, 27.27% and 10.82%, respectively. The most important features, as shown in Fig. 11, of this sub-clusters are whether the target site name is in the phishing URL path, the post action, and form data count.

Sub-Cluster 3-1. In this cluster we find attacks that are using old compromised sites and that slightly alter the quantity

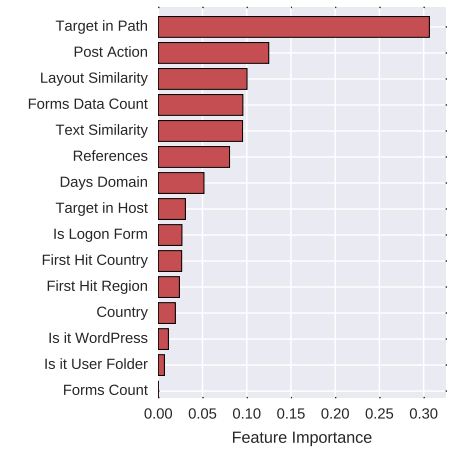


Fig. 11. Feature importance of Sub-Cluster 3. The most important features of this sub-clusters are if the target site name is in the phishing URL path, the post action, and the number of requested fields.

of data fields included (both hidden and visible) in the login form when compared with the original site. One important fact about this sub-cluster is that the phishing URLs avoid making any references to the name of the targeted financial institution.

Sub-Cluster 3-2. The second cluster is mostly characterized by the post action used to collect data. Moreover, the attackers are always using the same number of data fields in the login form than the impersonated page. It is also quite relevant that the phishing site domains tend to be new, or domains up to three years old, which include the name of the targeted brand in their sub-domains. This may imply that the criminal gangs are used to buying domain names and reusing them in their attacks, and that the criminal gang has been operating similarly for a long time, or that maybe there is an underground black market for sub-domains.

Sub-Cluster 3-3. In this group we found attacks that maintain the same structure, number of input forms, and post action of the targeted site. However, these attacks are usually hosted on WordPress sites, or potentially brute forced Cpanels. Lastly, the first visits to these sites are often from the US.

Sub-Cluster 3-4. The last cluster is one of the smaller of the sub-clusters. It groups a set of special cases where even if the HTML structure and the content of the attack keeps a very close relation with the impersonated site, the attackers pruned the original form to include only two data fields. Several of this cases used IP addresses instead of domain names. It is also important to note that in several cases, the sites are hosted in per-user web directories at a major educational institution in the US, as well as several now-suspended or dormant domains.

V. Profiling Criminal Organizations

In the previous section we showed how we were able to fine-grain group phishing attacks based only in a set of features describing the building blocks of the attack and some simple info gathered from Whois data. But, how effective is our approach to identify criminal organizations? There is no way to effectively answer such a question, unless we work together with law enforcement agencies and help them to gather the

evidence to prosecute criminals. However, DNS RRSets can help us to build a wild estimate on how good our approach may be in practice. In the following subsections we summarize our observations on how the clustering analysis is actually grouping correlated attacks, how we used RRSet analysis and additional Whois data that allowed us to make very interesting observations.

A. Domain Owners

In Section IV-B3 we described a sub-cluster that grouped attacks mostly hosted in just-created domains or no-older-than-three-years-old domains. A quick review of the RRSet for each of these domains showed that most of them were associated with a very small number of IP addresses, with occurrence frequencies separated by months and lasting only a few days. But more interestingly, when the records associated with those IP address were queried, we found that several domains associated with the IP addresses were also domains associated with phishing cases, most of them contained by the same cluster. Some domains were also found to be associated with phishing cases that were excluded during the data clean-up phase, and others were found in other clusters, but they were clearly outliers according to the features-importance analysis. We were also able to identify among the records, domains including direct reference to the brand of our customer which had not been previously registered in the company's case management system, as well as domains clearly intended to be used in attacks on other major financial institutions.

B. Rogue Web Masters

Other cases that caught our attention were those where the attacker were using sub-domains of legitimate active domains. Our guess was that there may be criminals specialized on setting up large numbers of allegedly legitimate web pages and use them to support their phishing campaigns.

After browsing for a while across DNS records, we found a "legit" domain that was related to other "legit" domains that were implicated in phishing or redirect cases. Also we found in the Whois records that the registrant for some of those domains was the same person. Even more interestingly, we found that the pages hosted in those domains contained explicit references to one of the analyzed domains. In this case it seems that the referenced page belongs to an individual who offers web master services, while the other domains belonged to his clients. Whether this individual is the criminal or his servers got compromised together with the credentials for the hosting of his clients, we don't know. What we do know now is that being in control of a set allegedly legit pages may be a good alibi, this way, the day your page gets compromised and you are contacted by an anti-phishing team notifying you about the incident, you can act like 'the good guy' and be very responsive to take down the attack.

VI. Conclusions and Future Work

In this paper we have shown the importance of performing analysis to look for patterns and correlations between phishing attacks. As we were able to classify phishing cases based on the strategy used by the attacker, it became clear that there are three main clusters of phishing sited based on the HTML structure and similarity with the target site. The first cluster grouped was when the phisher wants a fake site that does not resemble or reference the original site. Clusters 2 and 3 represent cases were the attacker is copying the financial institution's website to the best of his ability. They differ in how the phishing site is hosted; in cluster 2 the attackers are hosting all the page contents themselves, on the other hand, in cluster 3, the attacker is referencing most of the content to the original site.

Afterwards, we performed an additional analysis of each cluster. Our objective was to create clusters of the phishers. By using information regarding the location of the attacker the type of phishkit used and Whois information of the domain, we found a total 12 sub-clusters. This procedure helps us to understand the attackers' strategies, locations and motivations, which proved extremely useful for improving the ways we combat attacks.

Finally, we analyzed the DNS RRSets in combination with the clusters characteristics. This enabled us to clearly identify two criminal organizations. The first one was focused on owning several domain names and redirecting them to the same machines. It gave us interesting insights regarding how these attackers operate. The second group corresponds criminals that are infiltrating websites and creating fake sub-domains of legitimate active domains. Furthermore, in some cases we found that the pages hosted in those domains contained explicit references to other of the analyzed domains. This implies that the attackers are the owners of the legitimate domain or have control over it.

The work described here proved to be valuable to gain insight on how criminals operate, and gives a solid foundation toward the construction of more and better tools for forensic analysts, that help them to correlate cases and perform focused searches that speed up their investigations.

In the future we plan to expand this study to automate the way the clusters and sub-clusters are created, so we can understand more rapidly what is happening during an attack on an institution. By doing this, we can also understand how consistent are the clusters of attackers, and how anti-phishing strategies should be designed to protect each targeted site. Finally, we are also on the path to include the insights collected in this study into our detection and takedown systems. It is expected that by having a deep understanding of the phishers' strategies, our fight against phishing will be as potent as ever.

References

[1] M. Zareapoor and K. Seeja, "Text Mining for Phishing E-mail Detection," in *Intelligent Computing, Communication and Devices*. Springer, 2014, pp. 64–71.

[2] APWG, "Global Phishing Survey : Trends and Domain Name Use in 1H2014," Tech. Rep. September, 2014.

[3] ——, "Global Phishing Survey : Trends and Domain Name Use in 2H2014," Tech. Rep. May, 2015.

[4] S. Roopak and T. Thomas, "A Novel Phishing Page Detection Mechanism Using HTML Source Code Comparison and Cosine Similarity," in *2014 Fourth International Conference on Advances in Computing and Communications*, 2014, pp. 167–170.

[5] S. Marchal, R. State, and T. Engel, "PhishScore: Hacking Phishers Minds," in *CNSM*, 2014, pp. 46–54.

[6] R. Dhamija, J. D. Tygar, and M. Hearst, "Why Phishing Works," in *SIGCHI Conference on Human Factors in Computing Systems*, 2006, pp. 581–590.

[7] J. Zhang, P. Porras, and J. Ullrich, "Highly Predictive Blacklisting," in *17th USENIX Security Symposium*, 2008, pp. 107–122.

[8] J. Ma, L. K. Saul, S. Savage, and G. M. Voelker, "Beyond Blacklists : Learning to Detect Malicious Web Sites from Suspicious URLs," *World Wide Web Internet And Web Information Systems*, pp. 1245–1253, 2009.

[9] C. Whittaker, B. Ryner, and M. Nazif, "Large-Scale Automatic Classification of Phishing Pages," *NDSS '10*, 2010.

[10] S. Abu-Nimeh, D. Nappa, X. Wang, and S. Nair, "A comparison of machine learning techniques for phishing detection," in *Proceedings of the anti-phishing working groups 2nd annual eCrime researchers summit on - eCrime '07*, 2007, pp. 60–69.

[11] G. L'Huillier, A. Hevia, R. Weber, and S. Rios, "Latent semantic analysis and keyword extraction for phishing classification," in *International Conference on Intelligence and Security Informatics*, 2010, pp. 129–131.

[12] S. Marchal, J. Francois, R. State, and T. Engel, "PhishStorm: Detecting Phishing With Streaming Analytics," *IEEE Transactions on Network and Service Management*, vol. 11, no. 4, pp. 458–471, 2014.

[13] S. Marchal, K. Saari, N. Singh, and N. Asokan, "Know Your Phish: Novel Techniques for Detecting Phishing Sites and their Targets," oct 2015.

[14] R. Verma and K. Dyer, "On the Character of Phishing URLs: Accurate and Robust Statistical Learning Classifiers," in *ACM Conference on Data and Application Security and Privacy*, 2015, pp. 111–121.

[15] V. Ramanathan and H. Wechsler, "Phishing detection and impersonated entity discovery using Conditional Random Field and Latent Dirichlet Allocation," *Computers & Security*, vol. 34, pp. 123–139, 2013.

[16] J. Ma, L. K. Saul, S. Savage, and G. M. Voelker, "Identifying Suspicious URLs : An Application of Large-Scale Online Learning," in *International Conference on Machine Learning*, Montreal, Canada, 2009, pp. 681–688.

[17] R. M. Mohammad, F. Thabatah, and L. McCluskey, "Predicting phishing websites based on self-structuring neural network," *Neural Computing and Applications*, vol. 25, no. 2, pp. 443–458, 2014.

[18] C. M. Bishop, *Pattern Recognition and Machine Learning*, ser. Information science and statistics. Springer, 2006, vol. 4, no. 4.

[19] S. Gajek and A.-R. Sadeghi, "A forensic framework for tracing Phishers," in *3rd IFIP International Federation of Information Processing*, vol. 262, 2008, pp. 23–36.

[20] D. Birk, S. Gajek, F. Grobert, and A.-R. Sadeghi, "Phishing Phishers - Observing and Tracing Organized Cybercrime," in *Second International Conference on Internet Monitoring and Protection (ICIMP 2007)*, 2007.

[21] A. Jain, R. Duin, and J. Mao, "Statistical pattern recognition: A review," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, no. 1, pp. 4–37, 2000.

[22] D. Watson and J. Riden, "The honeynet project: Data collection tools, infrastructure, archives and analysis," in *OMBAT Workshop on Information Security Threats Data Collection and Sharing, WISTDCS 2008*, 2008, pp. 24–30.

[23] R. Schmitz, "A novel anti-phishing framework based on honeypots," in *2009 eCrime Researchers Summit*, 2009, pp. 1–13.

[24] O. Angelopoulou, S. Vidalis, and I. Robinson, "Who are you Today? Profiling the ID Theft Fraudster," in *European Conference on Information Warfare and Security*, 2012.

[25] T. Holz, M. Engelberth, and F. Freiling, "Learning more about the underground economy: A case-study of keyloggers and dropzones," in *Lecture Notes in Computer Science*, 2009, vol. 578, pp. 1–18.

[26] S. McCombie and J. Pieprzyk, "Winning the phishing war: A strategy for Australia," in *2nd Cybercrime and Trustworthy Computing Workshop*, 2010, pp. 79–86.

[27] B. Wardman, "A Series of Methods for the Systematic Reduction of Phishing," Ph.D. dissertation, University of Alabama, 2011.

[28] B. Wardman, G. Warner, H. Mccalley, S. Turner, A. Skjellum, and A. South, "Reeling in Big Phish with a Deep MD5 Net," *Journal of Digital Forensics, Security and Law*, vol. 5, no. 3, pp. 33–56, 2010.

[29] A. Singhal, "Modern Information Retrieval: A Brief Overview," *IEEE Computer Society Technical Committee on Data Engineering*, vol. 24, no. 4, pp. 35–42, 2001.

[30] Apache Software Foundation, "Per-user web directories," 2016.

[31] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The WEKA data mining software," *ACM SIGKDD Explorations*, vol. 11, no. 1, pp. 10–18, 2009.

[32] M. Dash and H. Liu, "Feature Selection for Clustering," *Knowledge Discovery and Data Mining. Current Issues and New Applications*, vol. 1805, no. 8, pp. 110–121, 2000.

[33] V. Roth and T. Lange, "Feature selection in clustering problems," *Advances in neural information processing systems*, vol. 16, pp. 473–480, 2004.

[34] M. Ceccarelli and A. Maratea, "Assessing Clustering Reliability and Features Informativeness by Random Permutations," in *Knowledge-Based Intelligent Information and Engineering Systems*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2007, pp. 878–885.

[35] P. Geurts, D. Ernst, and L. Wehenkel, "Extremely randomized trees," *Machine Learning*, vol. 63, no. 1, pp. 3–42, mar 2006.

[36] G. Louppe, L. Wehenkel, A. Sutera, and P. Geurts, "Understanding variable importances in forests of randomized trees," in *Neural Information Processing Systems*, 2013, pp. 1–9.