Towards Adversarial Phishing Detection

Thomas Kobber Panum
Aalborg University

Kaspar Hageman
Aalborg University

René Rydhof Hansen
Aalborg University

Jens Myrup Pedersen
Aalborg University

Abstract

Over the recent decades, numerous evaluations of automated methods for detecting phishing attacks have been reporting stellar detection performances based on empirical evidence. These performances often neglect the adaptive behavior of an adversary seeking to evade detection, yielding uncertainty about their adversarial robustness. This work explores the adversarial robustness of highly influential and recent detection solutions, by assessing their common detection strategies. Following discussions of potential evasion techniques of these strategies, we present examples of techniques that enable evasion through imperceptible perturbations. In order to enable and improve future evaluations for adversarial robustness, a set of design guidelines is proposed.

1 Introduction

Protecting digital infrastructure against malicious attacks has become essential as computational systems increasingly store and exchange private information of interest. This has motivated the design of initiatives to make the systems under attack fundamentally more secure, causing adversaries to adopt attacks that circumvent these initiatives by exploiting the social behavior of users of the systems rather than the system itself. A type of these attacks, phishing has had increased frequency in recent years [1, 22], and was described as the most widely adopted method for criminals to get unauthorized access to private networks in 2017.

Phishing attacks seek to exploit users by deceiving them, through some form of non-physical interaction, to release sensitive information for the benefit of the adversary [19]. Since their discovery, over two decades ago, numerous research initiatives have tried to design solutions for automating identification of these attacks, in order to actively prevent them from reaching their targets [37]. This has yielded solutions that demonstrate high accuracy for detecting these attacks, yet it has been highlighted that this performance is seemingly counter-intuitive and in contradicting to the observed volume of attacks [32].

Following this, it was deemed that the used evaluation methodologies influenced this phenomenon, emphasizing that a very limited set of methods accounted for the adaptive behavior of adversaries in their evaluation. Thereby, these solutions could potentially be relying on attributions of phishing attacks that are exploitable, and potentially enable adversaries to find attacks that evade detection. The absence of these considerations serves as the main motivation behind this work, as stated performance might be causing a false sense of security, as evaluations are unable to reflect the true adversarial setting that a detection solution faces in practice.

Firstly, we cover related work that has addressed problems of evaluation methods for the non-adversarial setting, and address the situation of varying definitions of phishing attacks (Section 2). Following this, we introduce a new terminology for phishing attacks, and their associated adversarial environment, that is independent of implementations and applications (Section 3). We then introduce a set of axioms for phishing attacks, that encapsulate the functional properties of the attacks, and serve as abstract guidelines for selecting information to use for inferring attacks in a given context (Section 4). Using the introduced terminology and axioms, we then assess two groups of existing work, highly influential and recent, by presenting four common strategies that the selected methods use for inferring attacks. The robustness of these strategies, to an adversary with an objective of creating phishing attacks that avoid detection, is then discussed and examples of perturbations that enable evasion are presented (Section 5). Based on the knowledge obtained throughout the assessment, we present a set of design guidelines for designers of future detection solutions to adopt, in order to enhance robustness to adaptive attacks and enable evaluations of their solution (Section 6). The contributions provided by this work can be summarized as:

- Propose a set of axioms for phishing attacks using a terminology that is independent of the application environment.
- Demonstrate and discuss the adversarial robustness of

common detection strategies among highly influential and recent detection solutions.

- Following the assessment, put forward a set of design guidelines to enable and improve evaluations of adversarial robustness.

The implementation of the experiments for the conducted assessment, including reproductions of detection solutions and perturbation methods, is provided with open access at https://github.com/tpanum/towards-adversarial-phishing-detection.

2 Background

In 1995 criminals performed a large scale attack on users of the chat service America Online (AOL), that involved tricking users into sharing their passwords, as the criminals exploited software to impersonate staff members of AOL [38,45]. This incident is often associated with the origin of phishing attacks, and despite numerous initiatives to combat the attack, certain sources state that phishing attacks have never been more frequent [1,5,36,41]

Marchal et al. highlighted that this fact is counter-intuitive to the the fact that existing phishing detection solutions are reporting detection accuracy of over 99.9% [32,44]. They suggested that the cause might be design limitations of the methods, making them infeasible to deploy in real-world settings, or that evaluations of these methods are biased [32].

Following this, they propose a systematic methodology with recommendations for future designers of detection solutions, that covers the topics of: data usage, evaluation metrics and temporal resilience. Within the scope of temporal resilience, they emphasize that these solutions are likely to see active attempts of evasion over time and deem that detection solutions should seek to become robust against these attempts. We refer to the ability for the solutions to resist these evasion attempts as adversarial robustness and cover it in greater detail in Section 3. Marchal et al. finds that a limited set of methods have directly addressed their adversarial robustness, which could potentially cause them to be open to evasion in the setting with an adversary.

While phishing attacks have been studied thoroughly, a recent meta-analysis of scientific publications showed that a variety of definitions of phishing attacks exist across the literature [24]. The cited analysis examined 536 publications containing 113 definitions, highlighting that definitions have varied globally across time and internally among research groups. Taking all of these definitions and their variations into account, Lastdrager et al. arrived at the following definition of phishing:

Definition 1 (from [24]). *Phishing is a scalable act of deception whereby impersonation is used to obtain information from a target.*

Here, it is essential to clarify that scalable refers to the relative effort for an adversary to perform the attack. Lastdrager et al. states that this formulation encapsulates highly targeted attacks, often referred to as spear phishing, while disallowing face-to-face interaction and phone calls as valid measures for conducting phishing attacks. Impersonation as a measure for obtaining information can be exemplified by falsely claiming to be a policeman in order to see an identity card with sensitive information.

Despite this definition being able to express commonalities among historical definitions for various environments, it does not cover the properties of the environments hosting these attacks. We seek to establish more clarity of the problem of phishing detecting and address this gap, defining terminology capable of expressing the shared properties of these environments in conjunctions with elements of phishing detection. Following this, we decompose Definition 1 into a set of axioms of phishing attacks expressed using the established terminology. This will serve as language for clarifying the adversarial setting in which phishing detection solutions exist, and the challenges that arise when seeking adversarial robustness.

3 Terminology

Phishing attacks are known to exist across multiple environments, such as: instant messaging, websites, and emails. These environments share common properties that enable them to host phishing attacks. Throughout this section we seek to derive these properties, by establishing a terminology for phishing attacks, their related entities, and the interference conducted by detection solutions.

Definition 1 clearly states that the objective for the adversary conducting the phishing attack, is to obtain information. In order for this to be feasible, the environment must have some ability to exchange information across certain subjects. We denote the exchanged information as *messages* and the method of exchange as a *channel*. Each message has some *content* and a pair of subjects that reflect the *sender* and the *recipient*. We refer to an environment with these abilities as a *messaging environment*, which effectively serves as the foundation for phishing attacks to exist.

Attacks are carried by messages and are only functional when recipients receive and read them. This fact serves as the motivation for the design of phishing detection solutions, that seek to effectively filter out messages being sent that contain attacks, such that they never reach their recipient. Throughout this work we refer to these solutions as *detectors* and *detection solutions* interchangeably.

Naturally, the objective for these detectors is to categorize messages as benign or phishing with limited misclassifications. In order for a detector to categorize a message as phishing, it relies on a set of *attributes* that messages with phishing attacks are expected to have. We refer to these attributes as phishing attributes, for which each detector has a set of phishing attributes that effectively serves as the requirements for a message to be considered a phishing attack.

Let a detector be a function, that maps a message m to a set of phishing attributes from a set of candidate attributes $\mathcal{A}_\mathcal{D}$, such that $\mathcal{D}(m) \subseteq \mathcal{A}_\mathcal{D}$. The candidate set $\mathcal{A}_\mathcal{D}$ serves as a specification of features that a message must have to be considered phishing for the respective detector. A message m is considered to be a phishing attack if, and only if, the detector yields the entire set of phishing attributes from the candidate set, such that $\mathcal{D}(m) \supseteq \mathcal{A}_\mathcal{D}$. Note that this discrete formalization does not directly encapsulate applications relying on probabilistic approaches. However, ultimately these solutions are used for classification, effectively forcing them to be discrete functions as they predict discrete classes. See Section 5.4. Importantly, this formulation intents to emphasize the scope of information used for inferring attacks, thus excluding more complex compositions of logic for simplicity purposes.

Given that detectors need to specify a candidate set of phishing attributes, the extent to which these definitions reflect the ground-truth set of attributes is uncertain, as obtaining completeness of this set is deeply philosophical. However, observations of successful attacks can serve as samples of empirical evidence of messages that contain the ground-truth set of attributes. Here we let the ability to obtain ground-truth categorizations of messages be expressed by an oracle function that maps a message m, to a set of phishing attributes from the ground-truth candidate set of phishing attributes $\mathcal{A}_{GT}$, such that $\mathcal{O}(m) \subseteq \mathcal{A}_{GT}$. Similar to detectors, a message m is a phishing attack if the oracle yields the ground-truth set of phishing attributes in its entirety.

Using this formulation, the natural objective for adversaries is to find a message m that satisfies the constraint $\mathcal{O}(m) \supseteq \mathcal{A}_{GT}$. When a detection solution is introduced into a previously undefended environment, it challenges the adversaries by requiring that the message must also be incorrectly classified by the detector. Thereby, the given message m must also satisfy $\mathcal{D}(m) \subset \mathcal{A}_\mathcal{D}$ to be a valid attack. We refer to this constraint as *circumvention*.

Occurrences of messages that satisfy circumvention is caused by *over-attribution*. Over-attribution is an inherit problem of the assumptions of phishing attacks adopted by a given detector. Concretely, it occurs when the set of phishing attributes used by the detector $\mathcal{A}_\mathcal{D}$ includes attributes that are unrelated to the ground-truth, such that $\mathcal{A}_\mathcal{D} \not\subseteq \mathcal{A}_{GT}$. Thereby, the detector relies on attributes of the message that are unrelated to its ability to carry an attack.

Conversely, when the detector's candidate set of attributes is a subset of the ground-truth candidate set, $\mathcal{A}_\mathcal{D} \subset \mathcal{A}_{GT}$, it can cause the detector to have overly defensive behaviour. We refer to this behaviour as *under-attribution*, and it can cause benign messages to falsely be considered attacks. Such situations are highly undesirable, as it can cause the detectors to become inapplicable for practical settings.

4 Axioms

The fact that the true attributes of phishing attacks $\mathcal{A}_{GT}$ are not directly observable, remains the core challenge for the process of designing detection solutions. Knowing this has driven the community to create numerous definitions [24], which is undesirable for establishing common progress. As a measure to improve upon this situation, and as an attempt to establish a common perception of the problem of phishing, we propose a set of axioms. These axioms serve as abstractions of phishing attributes, which detection solutions should explicitly account for in their method of inference. We derive these axioms by examining and decomposing Definition 1.

Definition 1 initially states *Phishing is a scalable act [...]*, emphasizing that the *act* (of phishing) has to be *scalable*, giving name to the first decomposed axiom.

Axiom Scalable. Being scalable in the context of phishing attacks means that the method of carrying out the attack should be inexpensive. Importantly, this axiom does not address the volumes of attacks, and thereby the more targeted variations of email phishing, such as spear phishing, that also satisfy this axiom [24].

Remark. Cost is context dependent, and the boundary for inexpensive is largely determined by a threat model for the respective environment. Examples of scalable attacks are phishing attacks conducted using inexpensive channels, such as email, as opposed to face-to-face communication which is considered expensive. For most practical environments, the use of a certain channel is often associated with a foreseeable and invariant cost. If such a cost is considered inexpensive, solutions acting in an environment, that uses solely such a channel, can implicitly satisfy this axiom.

Following this, additional specifications of the mentioned *act* are covered by: *[...] of deception whereby impersonation is used [...]*. Here, impersonation is described as a method of deception, serving as a functional dependency of the mentioned *act*. We decompose this functional dependency of impersonation into an axiom of the same name.

Axiom Impersonating. An essential ability of phishing attacks, is the ability to deceive victims into believing that the sender's identity, of a message carrying an attack, is genuine and benign. Adversaries exploit identities across various abstractions of subject identities, varying from identities of specific subjects to mimicking a class of subjects. Exemplifying this in a context of websites, an adversary might seek directly replicate the appearance of a specific bank, e.g. Bank of America, or alternatively construct an appearance that resemble a generic identity of banks as a class.

Remark. This axiom implies that recipients of messages are to a certain degree responsible for validating the identities of senders. Additionally, their ability to do this must be imperfect in order for adversaries to exploit this axiom.

Lastly the definition states that the adversarial objective of the attack is: *[...] to obtain information from a target*. This objective suggests that the attack should induce some action that leads to the exchange of information, we capture this by the following axiom.

Axiom Inducive. As phishing attacks seek to exploit the users of a system rather than the system itself, it is necessary for the recipient of the attack message to conduct some action that allows for the attacker to fulfill his objective of obtaining information. This axiom encapsulates the fact that users must act upon interpreting the received message, that cause the adversary to obtain desired information.

5 Assessment of Existing Methods

Numerous initiatives from academia and industry have proposed methods for detecting phishing attacks without human intervention [23]. These solutions have reported impressive performance measures based on historical observations of phishing attacks. It is often implicit or unknown to which extent these observations reflect the posterior measures that adversaries are likely to adopt for evading a detection solution, while maintaining functional attacks. This naturally yields uncertainty about to which extent the work accounts for evasion techniques. A study, that has analyzed lateral phishing attacks at large-scale, suggests that adversaries are willing to invest additional time into avoid being detected by their victims (opposed to a detection solution) [1] [18].

Therefore, we seek to assess the ability of existing work to perform under these conditions, through an assessment of their adversarial robustness. For the assessment, we include methods of two categories, namely *highly influential* and *recent*. For highly influential methods, we aggregated the union of the ten most cited (or highest ranked) publications among the search results from a series of well-established search engines commonly used by the scientific community [2]. Throughout these searches we used search queries related to phishing detection [3]. Most of the publications within this group have impacted a network of succeeding solutions that either adopt a similar methodology or directly extend the given method. As a measure to explore if adversarial robustness has changed over time, and acknowledging that high citation counts favors older publications, we furthermore manually select a group of recently published methods that use novel methodologies for inference. The full list of methods selected for the assessment can be seen in Table 1.

The selected methods are designed for a limited set of messaging environments, suggesting that these environments, namely the web and email, remain of highest interest. The web is an environment in which websites are exchanged across publishers and consumers, through servers and clients typically using the HTTP protocol as a channel. Email is an environment for exchange of text-based messages sent across the channels of SMTP, POP3 and IMAP. We speculate that this dominance of environments is largely caused by the large volume of phishing attacks in these environments, serving as a natural starting point for solutions seeking to detect attacks. Importantly, as suggested by the proposed axioms in Section 4, attacks are not strictly limited to only exist in these environments.

Examining the selected methods is challenged by the fact that none of them provide a publicly accessible implementation. Additionally, reproducing an implementation of these methods is challenged by the the fact they often rely on private datasets or third-party components that are unrecoverable, such as older search engines. This inherently makes evaluations of adversarial robustness difficult, as perturbations for evasion requires access to the output of an implementation to validate their performance. Additionally, most of the methods do not explicitly state the entire set of attributes that their method relies on for classifying a message as a phishing attack. In respect to our notation, this effectively leaves the candidate set of phishing attributes, for a given detector, $\mathcal{A}_{\mathcal{D}}$ to be unknown. We address this by only assessing common attributes and methodologies that are thoroughly covered across multiple methods which expressed core ideas of the inference design.

These common attributes and ideas are expressed as *strategies*, for which we identified four among the selected methods: Visual Similarity (VS), Statistical Modelling (SM), Reverse Search Credibility (RSC), Channel Meta-information (CI). We seek to either show perturbations that enable evasions of the given strategy, or discuss fundamental problems that are likely to enable evasions.

Only two of the recently proposed methods [3, 11] have been evaluated with respect to adversarial robustness. We attempt to reproduce one of the solutions and find evidence suggesting that the reported adversarial robustness is flawed. More details are contained within Section 5.4.

5.1 Visual Similarity

Human perception is often the centrepiece of impersonation, known to be an axiom of phishing attacks, as it is exploited to deceive recipients into misinterpreting the identity of the sender. Certain solutions use a strategy that seeks to detect attacks by mimicking human perception of messages' visual identity and ideally is able to differentiate between messages that appear to have similar and unique visual identities. These similarity measures are then used in conjunction with appearances of known benign messages, to detect the visual similarity of future messages. If a message is similar to the set of known benign messages, while originating from a different source, then it is considered to be impersonation and for some

[1] Importantly, as more time is invested into individual attacks, fulfillment of the axiom of scalability decreases.

[2] Search engines: Web of Science, Scopus, IEEE Xplore, Google Scholar.

[3] The searches were conducted during September and October 2019 and used the queries "phishing detection" and "phishing classification".

	Influential															Recent		
	Fu et al. (2006) [15]	Pan et al. (2006) [34]	Zhang et al. (2007) [48]	Fette et al. (2007) [14]	Garera et al. (2007) [16]	Dunlop et al (2010) [12]	Aburrous et al. (2010) [4]	Whittaker et al. (2010) [44]	Le et al. (2010) [25]	Xiang et al. (2011) [47]	Islam et al. (2013) [21]	Mohammed et al. (2013) [33]	Chen et al. (2014) [10]	Abdelhamid et al. (2014) [2]	Chiew et al. (2015) [27]	Mao et al. (2017) [31]	Corona et al. (2017) [11]	Abdelnabi et al. (2019) [3]
VS	•												•		•	•	•	•
SM		•	•	•			•	•	•	•	•	•		•		•	•	•
RSC			•			•				•					•			
CI					•				•									
ENV	*W*	*W*	*W*	*E*	*W*	*W*	*W**	*W*	*W*	*W*	*E*	*W*	*W*	*W*	*W*	*W*	*W***	*W*

Table 1: Methods selected for assessment and their identified strategies: Visual Similarity (VS), Statistical Modeling (SM), Reverse Search Credibility (RSC), and Channel Meta-information (CI). Messaging environments (*ENV*) cover email (*E*) and web (*W*), for which *W** is restricted to only e-banking websites and *W*** being phishing websites created using phishing toolkits.

methods this is the only attribute required to be considered a phishing attack. This strategy is expressed in the following derived attribute of phishing attacks:

Phishing Attribute 1. Sharing visual identity with an already observed benign message while originating from a different source.

Fu et al. implements this strategy by measuring the visual similarity using the Earth Mover's Distance for pixel intensity values of rendered websites [15]. Chen et al. use Normalized Compression Distance for byte-representations of rendered websites' pixel intensities [9,10]. Mao et al. introduce a method that implements this attribute by comparing aggregations of a page's HTML elements, including respective CSS styles for each element, thus assuming similarities across these aggregations are identical to their rendered representations [30, 31]. Corona et al. introduce a two-fold method for detecting attacks, for which one component uses image descriptors, in the form of Histogram of Oriented Gradients, and color histograms to measure for visual similarity among websites that host phishing attacks from phishing kits [11].

We argue that Phishing Attribute 1 is a direct adoption of Axiom Impersonating, thereby serving as a useful attribution for inferring attacks. However, measuring the correctness of models, that seek to mimic human perception, is difficult and the inability to do so can lead to potential imperfections. These imperfections can potentially serve as an opportunity for exploitation that would enable adversaries to create attacks that circumvent detection. For demonstrative purposes, we employ a perturbation technique that yield seemingly imperceptible changes, and thereby are expected not to affect Phishing Attribute 1. However, this perturbation technique significantly changes the similarity values of NCD [9, 10]. The technique is based on the fact that colors, in their binary representation, are completely distinct while color perceptions of humans are more fluid [13].

Thereby, conducting color perturbations that are small in the perception space of humans still yield large distinctive changes in the binary changes. We exploit the HSL (hue, saturation, light) color space, for which changes in its continuous values reflect human perception better than similar changes conducted in the frequently used RGB (red, green, blue) color space. For the implementation of this color space, we use HSLuv [8], and perturb images of websites by increasing saturation values by 1% and projecting the values to respective numerical limits. Let the similarity measure for two websites x_i and x_j be specified by $\text{NCD}(x_i, x_j) \rightarrow \mathbb{R} \in [0;1]$, for which higher values reflect more similarity. Additionally, let a given appearance of a website be x, for which x' is the perturbed variant of x. The experiments showed that the perturbation technique dropped the similarity scores significantly, effectively causing $\text{NCD}(x, x') \approx 0$, for the most popular websites of the Tranco list [26]. Given that phishing attacks often impersonate popular websites, this suggests that this perturbation technique could potentially lead to a consistent method for circumventing detection. Examples of perturbations, and the ability of the technique to influence the similarity score, can be seen in Figure 1.

While our perturbation technique for NCD illustrates that specifically NCD is not adversarial robust, we argue that similar imperfections will exists for *any* method using the VS strategy until human perception have been effectively verified to be reproduced in a computational setting.

5.2 Reverse Search Credibility

Search engines are a fundamental tool for finding and ranking information from the Internet using search queries of provided

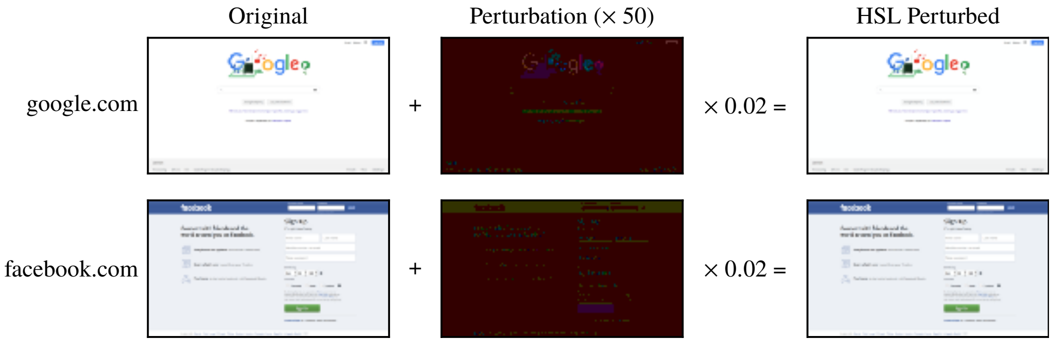

Figure 1: Visual appearances of the two most popular websites from the Tranco list [26], being perturbed by tiny shifts in the HSL color space. For visualization purposes, the perturbations are enhanced by a multitude. These perturbations cause significant drops, $\mathrm{NCD}(x,x') - \mathrm{NCD}(x,x) = -0.96 \pm 0.01$, in the visual similarity scores of NCD while being seemingly imperceptible.

keywords. This strategy is based on the assumption that search engines only display trustworthy and credible websites in their search results. Consequently, the absence of given websites in search results can be attributed to a website being a phishing attack, expressed in the following derived attribute:

Phishing Attribute 2. Absence of a given website in the most relevant search results returned by querying search engines with a signature derived from the given website.

Zhang et al. and Xiang et al. implement this attribute by using the term frequency of infrequent words, from a corpus of websites, for creating signatures of individual websites used for querying the Google search engine [47, 48]. Dunlop et al. expand upon this strategy by deriving the text of websites using Object Character Recognition (OCR), in order to be more robust against imperceptible text to image transformations of website content [12]. Chiew et al. introduce a method extracting logos of websites, using them as signatures for reverse image searches [27].

These studies demonstrate promising empirical evaluations using real-world data. However, we question their adversarial robustness as Phishing Attribute 2, suffer from some fundamental assumptions that can be exploited by an adversary. Firstly, we argue that the related attribute of the strategy does not align with any of the proposed axioms in Section 3, thereby causing over-attribution. Secondly, the strategy puts forward strong assumptions on the algorithmic functionality of the search engines, which remain undocumented due to commercial interest, thus creating uncertainty about the functionality of crucial components for the design.

This uncertainty could potentially lead to exploits that allow for circumvention. Concretely, the solutions relying on the text documents of websites could suffer from injected rare words that would affect the extracted signature without altering the rendered interpretation for the user in the browser [12]. Effectively this would cause the appearance of the website to remain unchanged while the signature could be altered to the desire of the adversary. Additionally, for solutions relying on OCR to extract signatures, it has been shown that OCR systems are vulnerable to imperceptible noise that gains the adversary some control of the set of recognized characters [39].

For these reasons, we find this strategy insufficient for achieving adversarial robustness.

5.3 Channel Meta-information

Phishing attacks are carried by messages and require that these messages are exchanged in order to reach their target. The channel responsible for this exchange typically relies on user-controlled information that could potentially carry attributions of phishing attacks. This strategy is based on the assumption that attacks can be inferred purely based on meta-information of messages, and thereby independently of message content. Effectively, this restricts the set of allowed attributes to be within a certain domain of information that the channel exposes.

A common implementation of this strategy for websites, is to infer attacks solely based on similarities across the Uniform Resource Locators (URLs) [16, 25]. Unknown URLs are then compared to a set of URLs from known benign websites, in order to infer attacks, as resemblance is a sign of impersonation, as derived in the following attribute.

Phishing Attribute 3. URLs resembling a URL from a known benign source.

Garera et al. implement this attribute by creating a statistical model that uses lexical information contained in URLs, in conjunction with Google PageRank information, in order to proactively prevent attacks prior to visiting websites [16]. Le et al. propose a method that uses lexical features of URLs while being resistant to common obfuscation techniques used by adversaries for client-side inference of attacks [25].

We argue that solutions adopting this strategy, using only content-independent-attributes, are prone to under-attribution. This stems from the inability to include attributions that align with Axiom Inducive. Additionally, we question to which extent Axiom Impersonating can be fully determined from content-independent attributes. However, we acknowledge that content-independent attributes can influence Axiom Impersonating.

5.4 Statistical Modeling

Deriving a set of concrete phishing attributes for a given messaging environment that prove to be useful and robust is the fundamental challenge of phishing detection. This difficulty has lead many researchers to learn these attributes using data of messages and statistical modeling. In particular, machine learning, a class of statistical modeling, has a significant presence in the selected methods. Machine learning is based on the popular approximation technique Empirical Risk Minimization [43], where the objective is to learn some probability distribution by minimizing the risk, typically represented as the weighted sum of some goodness of fit measure over the used data points. The learned probability distribution is then expressed as a model, that can be used for inference.

For phishing detection it is desired to learn some probability distribution of a given message containing a phishing attack, with respect to some set of information desired for inferring attacks. This information serves as a bound of information in which attributes of phishing can be learned by solving the stochastic optimization problem of minimizing the objective function. Naturally, the ability to approximate this probability distribution influences the chances of finding a distribution with low uncertainty to be useful for inference. Additionally, even if the ideal probability distribution is found, it might not even be useful for inference, if the information used for inference is seemingly uncorrelated with phishing attacks.

For certain models, including information that is strongly uncorrelated to the learning objective can hinder the ability of uncovering a useful probability distribution, thereby requiring manual labor for removing them prior to the learning process. However, progress over the last decade has allowed for more flexible models that are less sensitive to inclusion of uncorrelated information, such as Deep Neural Networks (DNNs). After the learning process is over for the statistical model, information that is associated with increased probability of the presence of attacks can effectively be addressed as phishing attributes of phishing attacks.

This strategy does thereby not impose the of use a specific attribute of phishing, it only expresses that attributes should be uncovered using patterns contained within data points.

Whittaker et al. use a random forest classifier that was trained using more than 3000 features in order to infer websites containing phishing attacks. Abdelnabi et al. exploit the potential of DNNs, by using a model named WhiteNet that has more than 100M trainable parameters for translating pixel intensity values of websites' appearances into a set of visual metrics.

Despite these methods showing promising results during evaluation, the ability to interpret the inference conducted by these models is difficult, thereby challenging the ability to validate and uncover the underlying phishing attributes. Validating these attributes is valuable, as the methods are prone to learn bias in high-dimensional spaces [20]. In addition, approaches such as DNNs also suffer from a fundamental problem named adversarial examples, for which tiny perturbations of legitimate input cause unexpected large changes in the predictions of the model [7,42]. A perturbation is expressed as some noise $\delta \in \mathbb{R}^D$ for a given model input $x \in \mathbb{R}^D$, such that its perturbed variant is given by $x' = x + \delta$. Typically a threat model for these attacks is defined by a perturbation bound ϵ, such that for any given noise δ it must satisfy $\epsilon \leq \|\delta\|_p$ for some p-norm. To reduce this problem of these perturbations, and thereby making models more robust, it was proposed to use a training technique named adversarial training that includes perturbed input into the training process [29]. This technique has shown to reliably increase robustness.

WhiteNet uses a variation of this technique, to suit the training objective of metric learning, and improve the robustness of the model. The evaluation of the original model reports an accuracy of 65% (closest match) against adversarial examples generated using the Fast Gradient Sign Method (FGSM) [17] ($\epsilon \leq 0.01$). This accuracy is considered, in contrast to other applications being attacked by adversarial examples [6], relatively high in relation to the 81% accuracy for the original data. Using the adversarial training lifts this accuracy to 71% against adversarial examples.

This decreased performance led us to hypothesize that the reported high robustness could stem from two causes: the Siamese Neural Network (SSN) architecture used by WhiteNet has some inherent robust properties or the evaluation was performed incorrectly. Importantly, SSNs use a fundamentally different training procedure than typical machine learning classifiers, as the loss function takes triplets of data points as parameters.

As a measure to explore our hypotheses, we replicate the WhiteNet model using a similar data set of 37043 websites across 2449 domains, gathered using the Kraaler tool [35]. Our implementation achieves a significantly lower accuracy of 24.6% against adversarial examples prior to adversarial training, with an increase to 30.8% after using adversarial training, as seen in Table 2. During the generation of the attacks we adopted a larger batch size for attacks, as it have previously been discovered that the sampling of triplets can greatly influence the calculated loss during training [46]. Given that the loss function is also used for attacking, we hypothesize that a similar importance should be accounted for during attacks. We speculate that the increase in batch size, and thereby better sampling, enabled us to create substantially stronger attacks. Following these results, we conclude that the reported robustness measures of WhiteNet are not representative for actual robustness towards adversarial examples.

Model	Unperturbed	$\epsilon = 0.005$	$\epsilon = 0.01$
Traditional Training			
WhiteNet	81.0%	72.8%	62.5%
WhiteNet *(replica)*	87.8%	**30.0%**	**24.6%**
Adversarial Training			
WhiteNet	81.0%	79.0%	73.1%
WhiteNet *(replica)*	90.3%	**33.3%**	**30.8%**

Table 2: Precision (closest match) for WhiteNet and our replica model across perturbations created using the FGSM attack for various threat models ϵ.

6 Design Guidelines

Assessing existing detection solutions in Section 5 highlighted problems related to the ability of evaluating adversarial robustness and attaining it, concretely: inaccessible implementations, implicit attributes, and over-attribution. As a measure to prevent future detection solutions from inheriting these problems, we propose three design guidelines: Accessible, Explicit Attributes, and Axiom Alignment.

Accessible. Adversaries are adaptive by nature, this fact should be reflected in the ability for the scientific community to be able to continuously evaluate solutions as new attacks emerge. Currently, most of the methods we have assessed do not provide widely available implementation, thus making it challenging to independently evaluate their performance. For certain methods it is infeasible to even reproduce the results, namely the data driven approaches that use private datasets in conjunction with not sharing the trained model weights.

To combat this phenomenon, we encourage that more authors make their methods easily accessible to the community. Ideally, this would be open access to the implementation, used throughout the original evaluations, or as a bare minimum ensure reproducibility. For methods relying on statistical modeling, this would include either making training data, or the found model weights, widely available. We deem that higher accessibility of solutions could contribute to the establishment of a community for perturbation techniques, that will prove useful for systematic evaluations of adversarial robustness.

Explicit Attributes. Defining phishing attacks has been shown to be difficult, causing a variety of definitions to exist [24]. When designing a phishing detection solution, the fundamental task is to design some mechanism capable of quantifying attacks, typically based on some intuition of attacks. If the adaption of the intuition of attacks into concrete design decisions remain unclear, it can potentially disguise strong assumptions of attacks that can be violated and exploited in the adversarial setting.

We suggest that designers of detection solutions explicitly state, to the best of their ability, which information is considered as attribute(s) of a given message being a phishing attack for their respective domain. For statistical modeling, we suggest either using models for which causality can be directly studied at test-time or adopt methods for exploring the underlying attributions during inference [28, 40]. This should be conducted to reduce the risk of the model inheriting bias from the underlying training data, causing undesired effects for the generalization performance. This guideline seeks to ensure that assumptions adopted throughout the design, namely the phishing attributes, become more explicit and thereby make the identification of potential cases of bias, over-attribution, or under-attribution, more effective.

Align with Axioms. Throughout the introduced terminology we have covered the consequences of having phishing attributes that are unaligned with true set of attributes, namely over-attribution and under-attribution. Unfortunately, the assessment highlighted that some of the common strategies adopted by the selected methods could be affected by these consequences. As a first measure for combating this phenomenon, we introduced a set of axioms of phishing attacks in Section 4. Effectively, these axioms serve as abstract phishing attributes, independent of messaging environments, that one has to account for and transform into concrete information for inference in a given environment of application. Thereby, we suggest that designers explicitly document the relationship between their phishing attributes used for inference and the proposed axioms. Additionally, it must be ensured that the used set of phishing attributes cover the full set of axioms.

7 Conclusion

Detection solutions for identifying phishing attacks are continuously challenged by adversaries trying to adapt their attacks to evade detection. Across the influential and recent methods, most of these solutions do not account for this challenge in their evaluation, yielding uncertainty about their adversarial robustness. In order to clarify the conditions of this adversarial setting, we introduced a terminology that is independent of environment and application for respective methods. Based on a consensual definition of phishing, we presented three axioms of phishing attacks, that any detection solution should account for to avoid using incorrect attributes for inference. Following this, the adversarial robustness of highly influential and recent work were assessed by decomposing their methods of inference into a set of strategies. The ability to evade detection for the respective strategies was then discussed, and examples of perturbations that enabled evasions for certain methods were discovered. These findings let us to define a set of design guidelines for the community of phishing detection to adopt to both enable and improve evaluations of adversarial robustness.

References

[1] Greg Aaron. Phishing Acitivity Trends 1st Quarter 2018. Technical report, Anti-Phishing Working Group APWG, 2018.

[2] Neda Abdelhamid, Aladdin Ayesh, and Fadi Thabtah. Phishing detection based associative classification data mining. *Expert Systems with Applications*, 41:5948–5959, 10 2014.

[3] Sahar Abdelnabi, Katharina Krombholz, and Mario Fritz. WhiteNet: Phishing Website Detection by Visual Whitelists, 2019.

[4] Maher Aburrous, M. A. Hossain, Keshav Dahal, and Fadi Thabtah. Intelligent Phishing Detection System for E-Banking Using Fuzzy Data Mining. *Expert Syst. Appl.*, 37(12):7913–7921, December 2010.

[5] Anti-Phishing Working Group. Phishing Activity Trends Report 2016, February 2017.

[6] Anish Athalye, Dimitris Tsipras, Logan Engstrom, and Aleksander Mądry. RobustML. https://www.robust-ml.org/, 2018. Accessed: 2020-07-03.

[7] Battista Biggio, Igino Corona, Davide Maiorca, Blaine Nelson, Nedim Šrndić, Pavel Laskov, Giorgio Giacinto, and Fabio Roli. Evasion Attacks against Machine Learning at Test Time. *Lecture Notes in Computer Science*, page 387–402, 2013.

[8] Alexei Boronine. HSLuv. https://www.hsluv.org/, 2012. Accessed: 2020-05-19.

[9] Teh-Chung Chen, Scott Dick, and James Miller. Detecting Visually Similar Web Pages: Application to Phishing Detection. *ACM Trans. Internet Technol.*, 10(2):5:1–5:38, June 2010.

[10] Teh-Chung Chen, Torin Stepan, Scott Dick, and James Miller. An Anti-Phishing System Employing Diffused Information. *ACM Trans. Inf. Syst. Secur.*, 16(4):16:1–16:31, April 2014.

[11] Igino Corona, Battista Biggio, Matteo Contini, Luca Piras, Roberto Corda, Mauro Mereu, Guido Mureddu, Davide Ariu, and Fabio Roli. DeltaPhish: Detecting Phishing Webpages in Compromised Websites. *CoRR*, 2017.

[12] M. Dunlop, S. Groat, and D. Shelly. GoldPhish: Using Images for Content-Based Phishing Analysis. In *2010 Fifth International Conference on Internet Monitoring and Protection*, pages 123–128, May 2010.

[13] Mark D. Fairchild. *Color Appearance Models*, chapter 10, pages 199–212. John Wiley & Sons, Ltd, 2013.

[14] Ian Fette, Norman Sadeh, and Anthony Tomasic. Learning to Detect Phishing Emails. In *Proceedings of the 16th International Conference on World Wide Web*, WWW '07, page 649–656, New York, NY, USA, 2007. Association for Computing Machinery.

[15] Anthony Fu, Liu Wenyin, and Xiaotie Deng. Detecting Phishing Web Pages with Visual Similarity Assessment Based on Earth Mover's Distance (EMD). *Dependable and Secure Computing, IEEE Transactions on*, 3:301–311, 11 2006.

[16] Sujata Garera, Niels Provos, Monica Chew, and Aviel D. Rubin. A Framework for Detection and Measurement of Phishing Attacks. In *Proceedings of the 2007 ACM Workshop on Recurring Malcode*, WORM '07, page 1–8, New York, NY, USA, 2007. Association for Computing Machinery.

[17] Ian Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and Harnessing Adversarial Examples. In *International Conference on Learning Representations*, 2015.

[18] Grant Ho, Asaf Cidon, Lior Gavish, Marco Schweighauser, Vern Paxson, Stefan Savage, Geoffrey M. Voelker, and David Wagner. Detecting and characterizing lateral phishing at scale. In *28th USENIX Security Symposium (USENIX Security 19)*, pages 1273–1290, Santa Clara, CA, August 2019. USENIX Association.

[19] Markus Huber, Stewart Kowalski, Marcus Nohlberg, and Simon Tjoa. Towards Automating Social Engineering Using Social Networking Sites. In *2009 International Conference on Computational Science and Engineering*, page nil, - 2009.

[20] Andrew Ilyas, Shibani Santurkar, Dimitris Tsipras, Logan Engstrom, Brandon Tran, and Aleksander Madry. Adversarial Examples Are Not Bugs, They Are Features. In *Advances in Neural Information Processing Systems 32*, pages 125–136. 2019.

[21] Rafiqul Islam and Jemal Abawajy. A Multi-Tier Phishing Detection and Filtering Approach. *J. Netw. Comput. Appl.*, 36(1):324–335, January 2013.

[22] John-Paul Power. Latest Intelligence for March 2018. Technical report, Symantec, 2018. https://www.symantec.com/blogs/threat-intelligence/latest-intelligence-march-2018, Last accessed on 05-16-2020.

[23] M. Khonji, Y. Iraqi, and A. Jones. Phishing detection: A literature survey. *IEEE Communications Surveys Tutorials*, 15(4):2091–2121, Fourth 2013.

[24] Elmer Lastdrager. Achieving a Consensual Definition of Phishing Based on a Systematic Review of the Literature. *Crime Science*, 3(1):9, 2014.

[25] Anh Le, Athina Markopoulou, and Michalis Faloutsos. Phishdef: URL names say it all. *IEEE INFOCOM*, 09 2010.

[26] Victor Le Pochat, Tom Van Goethem, Samaneh Tajalizadehkhoob, Maciej Korczyński, and Wouter Joosen. Tranco: A research-oriented top sites ranking hardened against manipulation. In *Proceedings of the 26th Annual Network and Distributed System Security Symposium*, NDSS 2019, February 2019.

[27] Kang Leng Chiew, Ee Hung Chang, San Nah Sze, and Wei King Tiong. Utilisation of website logo for phishing detection. *Computers & Security*, 54, 08 2015.

[28] Scott M Lundberg and Su-In Lee. A Unified Approach to Interpreting Model Predictions. In *Advances in Neural Information Processing Systems 30*, pages 4765–4774. 2017.

[29] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards Deep Learning Models Resistant to Adversarial Attacks. In *International Conference on Learning Representations*, 2018.

[30] Jian Mao, Jingdong Bian, Wenqian Tian, Shishi Zhu, Tao Wei, Aili Li, and Zhenkai Liang. Detecting Phishing Websites Via Aggregation Analysis of Page Layouts. *Procedia Computer Science*, 129:224–230, 2018.

[31] Jian Mao, Wenqian Tian, Pei Li, Tao Wei, Liang, and Zhenkai Liang. Phishing website detection based on effective css features of web pages. In *Wireless Algorithms, Systems, and Applications*, pages 804–815. Springer International Publishing, 2017.

[32] Samuel Marchal and N. Asokan. On Designing and Evaluating Phishing Webpage Detection Techniques for the Real World. In *11th USENIX Workshop on Cyber Security Experimentation and Test (CSET 18)*, Baltimore, MD, 2018. USENIX Association.

[33] Rami Mohammad, T. Mccluskey, and Fadi Thabtah. Intelligent rule based phishing websites classification. *IET Information Security*, 8, 01 2013.

[34] Y. Pan and X. Ding. Anomaly Based Web Phishing Page Detection. In *2006 22nd Annual Computer Security Applications Conference (ACSAC'06)*, pages 381–392, 2006.

[35] T. K. Panum, R. R. Hansen, and J. M. Pedersen. Kraaler: A User-Perspective Web Crawler. In *2019 Network Traffic Measurement and Analysis Conference (TMA)*, pages 153–160, 2019.

[36] PhishLabs. Phishing Trends and Intelligence Report 2019, 2019.

[37] Swapan Purkait. Phishing Counter Measures and Their Effectiveness - Literature Review. *Information Management & Computer Security*, 20(5):382–420, 2012.

[38] Koceilah Rekouche. Early phishing, 2011.

[39] Congzheng Song and Vitaly Shmatikov. Fooling OCR Systems with Adversarial Text Images, 2018.

[40] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic Attribution for Deep Networks. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, ICML'17, page 3319–3328, 2017.

[41] Symantec. Internet Security Threat Report. Technical report, Symantec, 2018.

[42] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. In *International Conference on Learning Representations*, 2014.

[43] V. Vapnik. Principles of risk minimization for learning theory. In *Proceedings of the 4th International Conference on Neural Information Processing Systems*, NIPS'91, page 831–838, San Francisco, CA, USA, 1991. Morgan Kaufmann Publishers Inc.

[44] Colin Whittaker, Brian Ryner, and Marria Nazif. Large-scale automatic classification of phishing pages. In *NDSS '10*, 2010.

[45] Adam Wright, Skye Aaron, and David W. Bates. The Big Phish: Cyberattacks Against U.S. Healthcare Systems". *Journal of General Internal Medicine*, 31(10):1115–1118, Oct 2016.

[46] Chao-Yuan Wu, R. Manmatha, Alexander J. Smola, and Philipp Krahenbuhl. Sampling Matters in Deep Embedding Learning. *2017 IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.

[47] Guang Xiang, Jason Hong, Carolyn P. Rose, and Lorrie Cranor. CANTINA+: A Feature-Rich Machine Learning Framework for Detecting Phishing Web Sites. *ACM Trans. Inf. Syst. Secur.*, 14(2):21:1–21:28, September 2011.

[48] Yue Zhang, Jason I. Hong, and Lorrie Faith Cranor. CANTINA: a content-based approach to detecting phishing web sites. In *Proceedings of the 16th International Conference on World Wide Web (WWW 2007)*, pages 639–648, 01 2007.