Towards Adversarial Phishing Detection

T. K. Panum, K. Hageman, R. R. Hansen, J. M. Pedersen

13th USENIX Workshop on Cyber Security Experimentation and Test

Long Paper (Position) – August 10, 2020

Motivation

Phishing Attacks

- Advances in technical security measures cause users to be victims of exploits
- Phishing attacks have exploited users for over two decades
- Numerous counter-measures have been developed to fight the problem

Contradictory Effectiveness (Marchal et al., 2018)

- Multiple reports claim frequency of attacks remain high (or increasing)
- State-of-the-art detection solutions report impressive evaluation measures[1]
  - Causes: Biased evaluations and infeasible deployment

Adversarial Robustness

- Few methods evaluate their performance on attacks that seek to actively evade the proposed detection solution

[1] Accuracy of $\geq 99.9\%$. False Positive Rates of $\leq 1\%$.

Adversarial Robustness

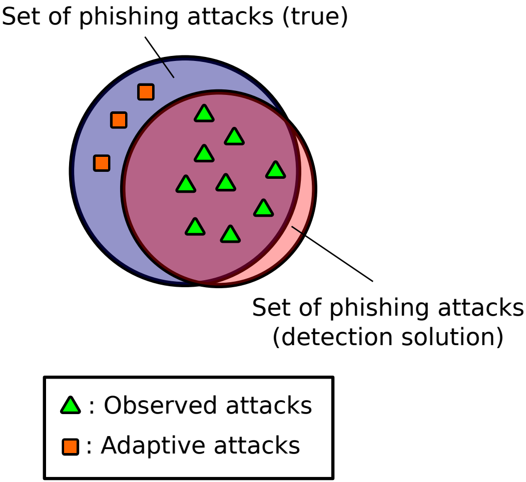
Adaptive attacks

Adaptive phishing attacks are attacks that remain undetected for a certain detection solution, yet maintain the functional properties of phishing attacks

Exists due to discrepancy between model and reality

Adversarial Robustness

Given solutions are likely to face adaptive attacks in a practical setting, evaluations should seek quantify their performance towards these (Ho et al., 2019)



Phishing Environments

- Attacks have existed across multiple environments
- We formalize the shared properties of such environments as:

Environment for Phishing Attacks

A *messaging environment* for which *messages* within this environment can fulfill the three axioms:

Impersonating, Inducive, and Scalable.

Messaging Environment

An environment for which *messages* can be exchanged using a *channel* across multiple senders and recipients

Message

Contains some *content* and relate to a *sender* and *recipient*

Axioms[1]

Lastdrager et al.'s Definition of Phishing Attacks

Phishing is a scalable act of deception whereby impersonation is used to obtain information from a target.

Impersonating

Should deceive the recipient into trusting the fake identity of the sender

Inducive

Should induce some form of action that yields the attacker to obtain information

Scalable

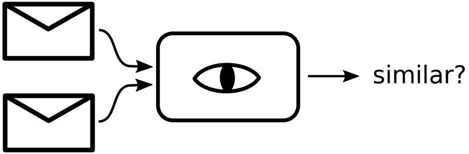
Crafting the attack should be inexpensive (time, $)

[1]These are merely abstract classes of information required to infer phishing, and does thereby not put logical constaints on the ability to obtain this information for concerete applications.

Assessment of Adversarial Robustness

- Examine the extend of which existing detection solutions have accounted for adversarial robustness
  - Selected work cover influential- and recent publications
- Derived a four of commonly used strategies for detecting attacks:
  - Visual Similarity, Reverse Search Credibility, Channel Meta-information, Statistical Modeling
- Discuss these strategies and their ability to account for the identified axioms
- Demonstrate techniques for creating perturbations that enable attacks to avoid detection

Visual Similarity



Phishing Attribute

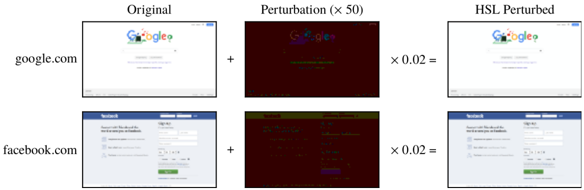
Sharing visual identity with an already observed benign message while originating from a different source.

Axioms

- ✓ Impersonating
- ÷ Inducive
- ✓ *Scalable*

- Based on reflecting human perception in a computational setting
- Known to be a challenging and unsolved problem
- Incomplete coverage of axioms

Example: Normalized Compression Distance (Chen et al.)



- Compare visual similarity as intersection over union of byte compressions

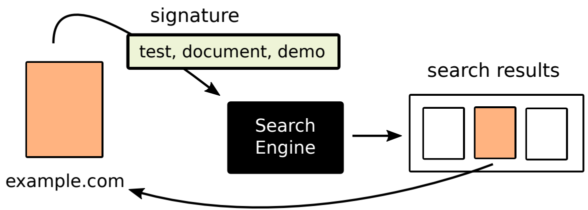
Simple attack

1. Use a color space that align closely with human color perception
2. Perturb all colors by small steps ($\pm 1\%$)

- Our attack is remain imperceptible yet effectively breaks NCD:

$$\mathrm{NCD}(x, x') - \mathrm{NCD}(x, x) = -0.96 \pm 0.01$$

Reverse Search Credibility



if found → not phishing **else** → phishing

Phishing Attribute

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

- Relies *credit scoring* using search engines
- Search engines are black boxes → Uncertainty

Axioms

? Impersonating
? Inducive
✓ *Scalable*

Channel Meta-information

Strategy

Constrain information used for inference to only be within the scope of the *channel*, ignoring the content of the respective *messages*.

Phishing Attribute *(case: Web)*

URLs resembling a URL from a known benign source.

- Given: Inducive $\leftrightarrow$ Content of *messages*
- Predictiveness using this strategy signal bias
- Incomplete coverage of axioms

Axioms

($\checkmark$) Impersonating
$\div$ Inducive
$\checkmark$ *Scalable*

Statistical Modeling

Strategy

Given a dataset containing information related to *messages*, and the presence of attacks within them, approximate a function $f(x)$ that can detect attacks.

Axioms

($\checkmark$) Impersonating
($\checkmark$) Inducive
$\checkmark$ *Scalable*

- Highly dynamic strategy, delimited by the information of the used dataset
- Selecting a model is often a trade-off between complexity and interpretability
- Parameters are selected using empirical performance
    - Assuming generalization to out-of-distribution inputs
- Complex functions can be in the magnitude of millions of parameters
    - WhiteNet (Abdelnabi et al., 2019): $\geq 100M$

WhiteNet (Abdelnabi et al., 2019)

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

- Model: *Siamese* Deep Neural Network (DNN) with $\geq 100M$ parameters.
- Given two visual representations of web sites yield a similarity measure
- Adversarial examples (AE) are a known vulnerability to DNNs
- Found that stated robustness towards AE to be inaccurate
  - Likely due to under-sampling during the creation of attacks

Design Guidelines

We introduce a set of design guidelines for future detection solutions to follow:

Accessible

- Provide a widely available implementation
  - Statistical Models: Weights and/or dataset.
- Benefit: Allow for continuous evaluations (both empirical and adaptive)

Explicit Attributes

- Clarify how information from the input space is used to infer attacks
  - (Complex) Statistical Models: Attribution Methods

Align with Axioms

- Focus on using functional properties of attacks for detection
- Absence: Predictiveness stemming from bias (symptoms not cause)

Thank you!

Thanks for listening!

Thomas Kobber Panum
tkp@es.aau.dk

PhD Student
Department of Electronic Systems
Aalborg University, Denmark