Gefilte Phish: Stopping PII Phishing Campaigns with Believable Decoy

COLUMBIA UNIVERSITY DATA SCIENCE INSTITUTE

Dennis Roellke[1], George Litvinov[1], Mark Seiden[1], Shlomo Hershkop[2], Salvatore Stolfo[1]

Department of Computer Science
Columbia University[1], Allure Security[2]

Abstract

Human user behavior analytics is insufficient to prevent Account Takeover (ATO) attacks enabled by phishing attacks (via email, text and voice). We propose to focus on human adversary behavior analytics by developing methods to acquire data about attackers and thwart the PII theft ecosystem. We believe this can be accomplished by using deceptive methods against attackers. We propose to automate a) the generation of synthetic, highly believable decoy PII and b) the injection of this fake data into phishing websites to disrupt the adversary identity theft marketplace. Misuse of decoy PII is monitored by third-party collaborators to acquire detailed data to profile adversaries. Attacker profiles and injected decoy PII will improve risk-based access controls reducing ATO attacks. A key objective is the design of systems to scale the automated decoy injection to diminish world-wide PII theft.

Background and Motivation

Phishing is typically considered a credential-stealing attack. In our previous project [1], we attempted to test this statement and found that it was no longer correct. The overwhelming majority of phishing sites are now typically designed to steal digital identities, allowing attackers to impersonate victim users, rather than simply gaining access to their accounts or services. An analysis of 131,023 phishing sites from a 5 year time period indicated a rising number of over 31% of those sites pose an extreme danger to identity theft since they are designed to steal a victim's personally identifiable information (PII), while credential-theft focused phishing attacks are on the decline.

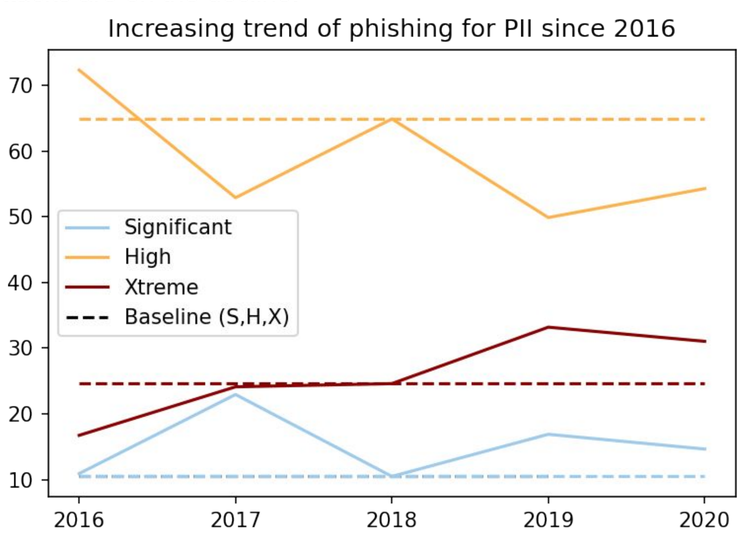

Figure 1. Phishing threat trend analysis

Project Overview

We introduce believable decoys to research modern phishing campaigns. For this goal we collaborate with multiple industry organizations including Yahoo! and The Internet Archive and we work closely with Columbia's IT department. We set up a dedicated system to filter phishing feeds for PII phishing and we automatically stuff them with automatically generated decoy information that we release from three different networks to mimic three different phishing victims. All our activity, including network traces are logged for forensic analysis and our system is built to scale using Docker, MongoDB, and Apache Kafka for event processing.

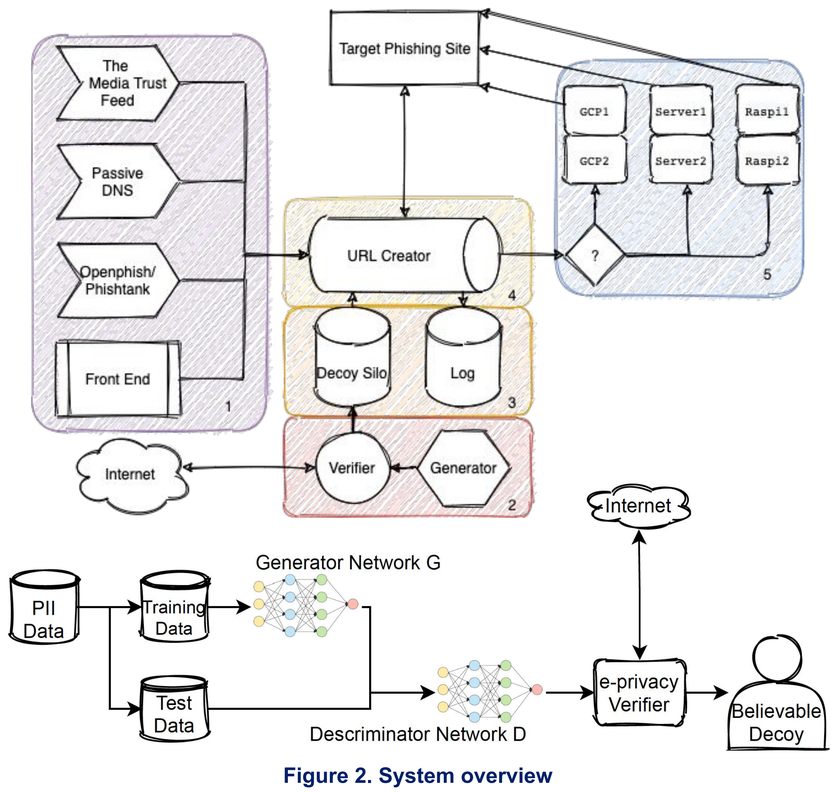

Figure 2. System overview

Decoy Generation

The generation of believable phishing decoys that pass data verification checks and include all required PIIs for advanced stuffing requires a large amount of time, thus the automation of the creation of such decoys would highly improve the sample size of the experiment. We have compiled a list of details about common PII tokens, such as the Luhn checksum for credit card numbers and the birthdate/geo assignment of SSN prior to 2011. We aim to pair these static information with a Generative Adversarial Network (GAN) [2] generating combinations of name, gender, religion and ethnicity which are believable and appropriate for the phishing target. As pointed out by the iconic movie "Superbad", the combination of the world's most popular first name and the world's most popular last name is not a popular name itself.

A GAN learns to avoid such cases by simultaneously training two models: a generative model G that captures the distribution of the ground-truth data (e.g., real world PII distributions), and a discriminative model *D* that classifies whether the sample came from the training data or from the generator model *G*. In other words, *D* and *G* play the following two-player minimax game with value function *V(G, D)*:

$$\min_G \max_D V(D,G) = \mathbb{E}_{\boldsymbol{x}\sim p_{\text{data}}(\boldsymbol{x})}[\log D(\boldsymbol{x})] + \mathbb{E}_{\boldsymbol{z}\sim p_{\boldsymbol{z}}(\boldsymbol{z})}[\log(1 - D(G(\boldsymbol{z})))]$$

Equation 1. Definition of a Generative Adversarial Network [2]

Further, we constrain our model to strictly avoid collision with real persons' identities (including deceased individuals). This morally motivated goal can be achieved by utilizing differential privacy [3] to verify the generator's creations. Our verifier scrapes the internet for information about the newly created believable decoy. It then ensures that the effect of making an arbitrary single substitution in the results (creating D_2 from D_1) is small enough, that the query result cannot be used to infer much about any single individual. Formally, we specify:

$$\Pr[\mathcal{A}(D_1) \in S] \leq \exp(\epsilon) \cdot \Pr[\mathcal{A}(D_2) \in S],$$

Equation 2. Definition of Differential Privacy [3]

We note that we can configure the verifier's privacy budget ϵ to only confirm generations that are believable but do not collide with a real person at all.

Conclusion

Phishing is an ever-evolving area of cyber-crime. With trends arising quickly and becoming ever more sophisticated, it is important to be on the forefront of research in this field. Financial losses of victims are in the billions of dollars yearly, while prosecution is minimal. Over the years, there was a large shift within the area of phishing where campaigns went from tricking users for login credentials to asking them for sensitive PII, thus allowing the threat actors to commit identity theft. Our project is the first of its kind, in that it focuses on the study of this ongoing trend by generating believable and verifiable decoy identities and stuffs them into newly started phishing campaigns.

References

[1] "Ranking PII Threat and Fingerprinting Modern Phishing Websites", Roellke et al, tbd
[2] "Generative Adversarial Networks", Goodfellow et al., NIPS, 2014
[3] "Calibrating Noise to Sensitivity in Private Data Analysis", Dwork et al., TCC, 2006