# Chapter 18
# Improving Auto-Detection of Phishing Websites using Fresh-Phish Framework

**Hossein Shirazi**
*Colorado State University, USA*

**Kyle Haefner**
*Colorado State University, USA*

**Indrakshi Ray**
*Colorado State University, USA*

## ABSTRACT

*Denizens of the Internet are under a barrage of phishing attacks of increasing frequency and sophistication. Emails accompanied by authentic looking websites are ensnaring users who, unwittingly, hand over their credentials compromising both their privacy and security. Methods such as the blacklisting of these phishing websites become untenable and cannot keep pace with the explosion of fake sites. Detection of nefarious websites must become automated and be able to adapt to this ever-evolving form of social engineering. There is an improved framework that was previously implemented called "Fresh-Phish", for creating current machine-learning data for phishing websites. The improved framework uses a total of 28 different website features that query using python, then a large labeled dataset is built and analyze over several machine learning classifiers against this dataset to determine which is the most accurate. This modified framework improves the accuracy of modeling those features by using integer rather than binary values where possible. This article analyzes not just the accuracy of the technique, but also how long it takes to train the model.*

DOI: 10.4018/978-1-7998-2460-2.ch018

## INTRODUCTION

The Internet has ushered in a new evolution of electronic deception called phishing, that involves the one-two punch of web and email that is very difficult for users to detect. In fact, according to Alsharnouby et al. only 53% of users successfully detect phishing websites (Alsharnouby et al., 2015).

Phishing, defined as, "the attempt to obtain sensitive information such as user-names, passwords, and credit card details, often for malicious reasons, by masquerading as a trustworthy entity in an electronic communication" (Wikipedia, 2016), is a problem that is as old as the Internet itself. Trying to get unsuspecting users to give up their money, credentials or privacy is a particularly insidious form of social engineering that can have disastrous effects on people's lives. Often this type of attack arrives in the form of an email containing the first part of what Chaudhry et al. describe as the lure, the hook and the catch (Chaudhry, Chaudhry, & Rittenhouse, 2016).

The lure is what entices the user to click on a link. It can be advertising a way to get easy money, obtain an illicit product, or a warning that a user's account has been compromised or blocked in some fashion. The hook is often a website that is designed to mimic a legitimate website of a reputable organization such as a bank or other financial institution. The hook is used to trick the user into entering and submitting their credentials such as user-name, password, credit card number, etc. The catch is when the user has submitted their private information and the malicious owner of the website collects and uses this information to exploit the user and his accounts.

Figure 1 shows the number of phishing attacks has been increasing year over year for the last decade. Anti-Phishing Working Group (APWG) reported an alarming 250% increase from the last quarter of 2015 to the first quarter of 2016 (APWG, 2016).

Not only have phishing attempts evolved and become more sophisticated, the motivation for implementing these attacks has changed as well. Attackers today have moved beyond simply probing the security of systems; now their primary goal has become financial gain. This commercialization of phishing is charted in Figure 2 showing the fourth quarter of 2016 where 41% of targeted industries are retail/services and 19% of them financial institutions. This wide diversity of targeted services, coupled with the trend of increasing attacks demonstrates that end-users are in more danger, from more sources, than ever before.

Phishing is a growing multi-vector problem that has real and devastating consequences for users. It is also a problem growing in sophistication, scope and reach. Automated detection techniques are critical to a safe and secure Internet. We use machine learning algorithms because they have been proven to have the capability to discover complex correlations among different data items of similar nature, however work to date leaves out one critical variable in this equation; we need an open and extensible framework capable of generating up-to-date data for researchers. We call this framework, Fresh- Phish.

There is no recent machine learning data that has been published on phishing websites. The data that does exist is several years out of date, a serious problem given the dynamic nature of the Internet. There is also no published framework, that we are aware of, for gathering new data.

In this paper, we introduce an open-source python-based framework called Fresh-Phish for generating up-to-date data of websites for training machine learning algorithms. The Fresh-Phish framework is intended to be an extensible building block that other researchers can modify, add, delete, or change what features are used to build datasets. We used our framework to crawl over 5,000 websites to generate a large labeled dataset with which we tested and analyzed several different machine learning techniques to accurately identify phishing websites.

## Related Content