# SmellsPhishy

## Introduction

## Background

### Punycode Detection

#### Homograph Attacks

Since 2003, Unicode characters have been allowed to be used in domain names through the initial IDNA protocol (“IDNA2003”). The idea behind the introduction of unicode characters was to give international internet users the ability to follow links in their own languages. Unfortunately, unicode domains are constantly being exploited in phishing attacks.

Malicious parties can replace certain characters in existing URLs with another unicode character that looks exactly the same. For example, the “o”s in “google.com” are normally represented with the Latin small letter “o” (U+006F), whereas an attacker can host a site with the domain “google.com” using the Greek small letter Omnicron(U+03BF). The attacker can then give the user a “google.com” link which actually directs to his own malicious webpage. This is known as a homograph attack. A more drastic example can be seen in Figure 1 below for “apple.com”.

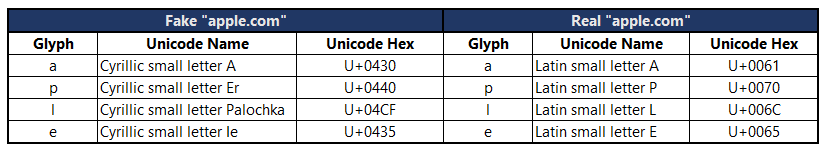


Figure 1. Taken from https://blog.malwarebytes.com/101/2017/10/out-of-character-homograph-attacks-explained/

### Redirect Detection

#### Malicious Redirects

One problem with URLs is in the form of shortened URLs. In 2002, the first notable URL shortening service, TinyURL started and has gathered public interests. Long URLs are difficult to distribute and remember. The motivation behind shortened URLs was to allow instant messaging services, websites or hard copy publications which have characters limitation or to “beautify” a link. Unfortunately, URL shortening can be used to disguise the underlying address.

Redirects can also be triggered by embedded JavaScript code.

#### Understanding Redirects

The HTTP/1.1 RFC lists the different types of redirects:

1. Redirects that indicate the resource might be available at a different URI, as provided by the Location field, as in the status codes 301 (Moved Permanently), 302 (Found), and 307 (Temporary Redirect).
2. Redirection that offers a choice of matching resources, each capable of representing the original request target, as in the 300 (Multiple Choices) status code.
3. Redirection to a different resource, identified by the Location field, that can present an indirect response to the request, as in the 303 (See Other) status code.
4. Redirection to a previously cached result, as in the 304 (Not Modified) status code.

#### Intercepting Redirects

Chrome

## Related Work

## Design

*SmellsPhishy* is a Chrome browser extension that aims to warn users before they visit malicious websites unknowingly. It aims to defeat homograph attacks and malicious redirects. These are common techniques used by malicious parties to lead users to browse to phishing websites that masquerade as the legitimate ones.

*SmellsPhishy* consists of 3 components:

1. Punycode Detection
2. Redirect Detection
3. Phishing Page Analysis

These components are explained in further detail in subsequent sections. *SmellsPhishy* looks out for two common triggers: a URL consisting of non-alphanumeric Unicode characters and page redirects. When one of these is present, *SmellsPhishy* will perform an analysis on the webpage that the user is browsing to. If it is determined that the webpage is likely to be malicious, the user will be prompted with a warning, and he will have to indicate his intent to proceed to the page or to cancel the request.

## Implementation

### URL-based Detection Techniques

### Domain-Based Detection Techniques

Domain-based detection techniques examine the WHOIS records of the website that is being visited. WHOIS data have been crucial components in anti-phishing efforts (The Anti-Phishing Working Group DNS Policy Committee, 2007). Fette et al. used WHOIS queries to detect phishing links in emails (Fette, Sadeh, & Tomasic, 2007).

We obtained WHOIS records through an API call to a WHOIS service provider. We examine the following fields in the records:

1. Created Date

Li et al. found that 96.67% of phishing domain names are detected within 1 month from registration (Li, Geng, Yan, Chen, & Lee, 2016). In our extension, we define 1 month as 31 days. We ensure that the domain to be visited has been registered more than 31 days ago. Otherwise, we attribute it with a 4% probability of being a phishing site.

1. Expiry Date

Registrants of phishing sites will not register the sites for too long a period of time as they are normally detected and taken down quickly. Legitimate organizations would ensure that their domains are not near expiry as they risk losing it. We check that the domain will not expire in the next 3 months.

1. Registrant

**Limitations**

Website owners can opt to use something called domain privacy to obscure their identity, which means that the information cannot be viewed through WHOIS. Domain privacy is a legitimate service that many domain registrars and web hosts offer, but sometimes it’s abused by scammers.

### Site Statistics Detection Techniques

### Page Content Detection Techniques

## Evaluation

## Conclusion

# References

Dhanalakshmi, R., C., P., & C., C. (2011). Detection Of Phishing Websites And Secure Transactions. *International Journal Communication & Network Security (IJCNS)*, 7.

Fette, I., Sadeh, N., & Tomasic, A. (2007). Learning to Detect Phishing Emails. *WWW.* Banff, Alberta, Canada.

Li, X., Geng, G., Yan, Z., Chen, Y., & Lee, X. (2016). Phishing Detection Based on Newly Registered Domains. *IEEE International Conference on Big Data (Big Data)* (p. 8). Washington, DC, USA: IEEE.

The Anti-Phishing Working Group DNS Policy Committee. (2007). *Issues in Using DNS Whois Data for Phishing Site Take Down .*