CS5331 Web Security – SmellsPhishy (Chrome Extension for Phishing Detection)

Samantha Wong, Lionel Teo, Leong Peng Kuen, Daryl Quek
School of Computing, National University of Singapore

NUS National University of Singapore

Introduction

- **Punycode** – Malicious parties replace certain characters in existing URLs with another Unicode character that looks exactly the same.
  - An attacker can host a site with the domain "google.com" using the Greek small letter (U+03BF) and redirect to his own malicious webpage.
- **Redirects** – Long URLs are difficult to distribute and remember. Malicious redirects can be hidden in shortened URLs and can be used to disguise the underlying address. Redirects can also be triggered by embedded JavaScript code.

Objective

- Our project aims to warn users before they visit malicious websites unknowingly to defeat homograph attacks and malicious redirects.
- Demonstrate SmellsPhishy performing an analysis on webpage that the user is browsing to. If the webpage is likely to be malicious, the user will be prompted with a warning and he will indicate his intent to proceed or cancel.

Detection Components

- **URL.** Makes an analysis on the URL being visited.
  - Domain Content
  - Domain Length
- **WHOIS.** Performs verification checks using data in the WHOIS record for the target URL's domain.
  - Domain Age: ~96% of phishing sites are taken down within 1 month
  - Domain Expiry: Legitimate domains are properly managed and renewed before they near expiry.
  - Registrant Verification: Registrant can be a parent/holdings company. They must be registrants for their own websites.
- **Site Statistics.** Perform comparison checks with possible identified original site provided with the following:
  - Page Ranking
  - Traffic Hits
- **Page Content.** Examines the target URL's web page's contents for suspicious elements.

Logical Overview
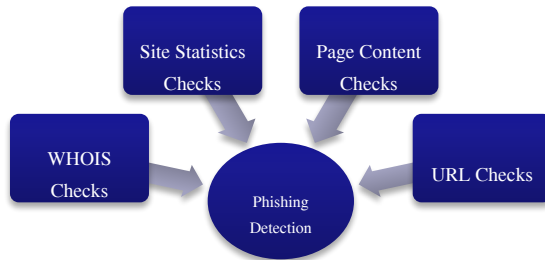
Breakdown of Detection Techniques

Technique	Recommended Threshold	Source
URL Checks		
Top Level Domains	Blacklist	Parsing
Length of Base Domain	< 12 characters	Parsing
WHOIS Checks		
Domain Age	> 1 month	WHOIS
Domain Expiry	< 3 months	WHOIS
Registrant Verification	Top Google search result for registrant must be a page registered by registrant	WHOIS
Site Statistics Checks		
URL Page Ranking	+/- 10 ranks from possible identified original site	API (SimilarWeb)
URL Traffic Hits	+/- 50% traffic hits from possible identified original site	API (SimilarWeb)
Page Content Checks		
Forms	Keyword Text	HTML

Implementation

1. SmellsPhishy uses the Chrome API to intercept HTTP requests containing Punycode and HTTP responses containing redirects.
2. A Google search is conducted for the target URL's page title. If URL of the top search result matches, this is considered to be a legitimate web page. Otherwise, it is put through the phishing detection engine.
3. The web page is passed through the various WHOIS checks. These make use of a WHOIS service provider and retrieves the required field(s) for verification.
4. Page statistics are examined for the target URL.
5. The target URL's web page is parsed for suspicious elements.
6. Text elements of target URL are examined.
7. Finally, the results of the phishing checks are displayed to the user. The user has to decide whether to proceed to visit the web page.

Challenges & Limitations

- WHOIS parsing is not trivial due to many different formats.
- Domain privacy strips availability of WHOIS data. This may soon become standard practice due to EU privacy regulations!
- Is using Google as a standard source of truth going to remain accurate going forward?
- May have limited text available on actual page
- In the case of a very viral phish - will site statistics lie?

Conclusion and Future Work

- There are many components involved in Phishing Detection
- Gather more experimental data on live phishing sites to fine-tune threshold parameters
- Add additional area for visual component checking
- Prompt user to input where they think they are going