Needle in a Haystack: Tracking Down Elite Phishing Domains in the Wild

Ke Tian, Steve T.K. Jan, Hang Hu, Danfeng Yao, Gang Wang
Computer Science, Virginia Tech
Phishing is a Big Threat

• Phishing: fraudulent attempt to obtain credentials (password)
• Big Threat: estimated $30M loss in 2017


Yahoo Data Breach in 2014
Affected 500 Million Yahoo! User Account

Ubiquiti Networks
Lost $46.7M dollar to scammers in 2015

• Exploiting human factor is easier than system vulnerabilities.
Some Phishing Websites are Easy to Tell

- Phishing is a long existing problem
- Good news: some phishing websites are easy to detect

URL not relate to Paypal: Phishing
Some Phishing Websites are Easy to Tell

- Phishing is a long existing problem
- Good news: some phishing websites are easy to detect

URL not include domain name: Phishing
More Sophisticated Phishing Example

• This is IDN (Internalized Domain Name) homograph attack
• Homograph domain squatting: Exploit the fact that many different characters look alike
More Sophisticated Phishing Example

- This is IDN (Internalized Domain Name) homograph attack
- Homograph domain squatting: Exploit the fact that many different characters look alike
How can we systematically capture these sophisticated phishing websites in practice?
This Study

• We focus on **squatting phishing domains**
  • Web contents: phishing content, mimicking real websites
  • Domain name: “squatting” domain that impersonates popular brands

• Research questions
  • How to systematically detect squatting phishing domains in practice?
  • What types of impersonation/evasion techniques do they use?
  • How effective are existing blacklists to detect them?

• Large-scale empirical measurements
  • Search over 224 million DNS records
  • 702 popular brands
Outline

• Introduction

• Detection methodology
  • Detect squatting domain
  • Detect phishing pages under squatting domain

• Measuring squatting-based phishing

• Conclusion
Detection Methodology

- Our detection methodology based on a series of filtering process

- Active DNS Project
  - DNS Records: 224,810,532
  - Popular brands: 702

- Squatting domain detection
  - Squatting Domains: 657,663

- Phishing classifier
  - Phishing: 1,741
    - Web: 857
    - Mobile: 908

- Manually check
  - Confirmed: Web 857 Mobile: 908
Detect Squatting Domain

• Goal: Detect squatting domain that impersonate brands
  • Given a brand, search squatting domains in DNS
• Capture five types of squatting domains
  1. **Homograph**: Look similar to target domain
     - facebook-stroty.com
  2. **Bits**: Flip a bit of target domain
     - faceb00k.com
  3. **Typo**: Mimic the incorrectly typed of target domain
     - fcaeb00k.com
  4. **Combo**: Connect target domain with other strings
     - facebook-stroty.com
  5. **WrongTLD**: Different TLD of target domain
     - facebook.audi

Detect Squatting Domain

- 224,810,532 DNS records → 657,663 squatting domains
  - Crawl web and mobile version of pages that are still alive
  - Dynamic crawler: It can load java scripts and process redirections
- 6,115 squatting domains (1.7%) are redirected to original brand
  - Some business purchase squatting domains to protect their own customers

Squatting Domain: pricelin.com
Original Brand: priceline.com
Re-direct
Phishing Classifier

• Goal: Classifying phishing pages under squatting domains

• Ground Truth Data:
  • 1,731 phishing pages from PhishTank (manually confirm)
  • 1,565 benign pages from squatting domain (manually confirm)

• Our classifier is motivated by observations on evasion techniques:
  1. Layout obfuscation
  2. String obfuscation
  3. Code obfuscation
Layout Obfuscation

- Change style/color/layout of target brand website
- Evade screenshot-similarity based detection method
String/Code Obfuscation

- Hide important text and keywords in the HTML source code
- Evade keyword-similarly based, or source code similarly based detection

```
<title> Log in to your PayPal </title>

String/Code Obfuscation

/<title> Log in to your PayPal </title>

<title> Log in to your PayPal </title>

CODE OBfuscation

String/Code Obfuscation

<script> String.fromCharCode(50) + "a" + ....

Target Brand HTML

Phishing HTML

Be detected by keyword-
similarly based methods
Our Design

• Intuition 1: *Phishing pages will be visually displayed to users*

• Extract keywords from their screenshots with OCR
  • Tesseract OCR: extract keywords from image

Keyword list:
- Payp[...]
- Email[...]
- [password]

Keyword list:
- Paypal[...]
- Email[...]
- [password]

Google OCR

Keyword list: Paypal Email [password]......

NLTK spell check

Keyword list: Paypal Email [password]......
Our Design Cont.

• Intuition 2: *Phishing pages contain forms to collect user credentials*

• Extract keywords from HTML forms

• Using text-based feature from the source code as compliment
Ground Truth Evaluation

- Feed features to machine learning classifiers
  - Image (OCR) features, form features, text-based features
- Naive Bayes, KNN and Random forest
- Results of 10-fold cross-validation:

<table>
<thead>
<tr>
<th>Classifier</th>
<th>False Positive</th>
<th>False Negative</th>
<th>AUC</th>
</tr>
</thead>
<tbody>
<tr>
<td>NaïveBayes</td>
<td>0.5</td>
<td>0.05</td>
<td>0.64</td>
</tr>
<tr>
<td>KNN</td>
<td>0.04</td>
<td>0.1</td>
<td>0.92</td>
</tr>
</tbody>
</table>

Random Forest is highly accurate
Outline

• Introduction

• Detection methodology
  • Detect squatting domain
  • Detect phishing pages under squatting domain

• Measuring squatting-based phishing

• Conclusion
Detection in Practice

DNS Records: 224,810,532, Popular brands: 702

Squatting domains: 657,663

Detected Phishing pages: 1,741

Confirmed phishing pages on both: 1175

Squatting phishing websites indeed exist
More phishing websites on mobile
Can Current Blacklists Detect Them?

- Run 70+ phishing blacklists, including PhishTank, eCrimeX, VirusTotal

Existing blacklists/tools are not capable to capture squatting phishing yet.
Squatting Domains Types

• Combo squatting domains contain the largest number of phishing pages
• Bits and homograph squatting domains: Hard to register
Example Study: Uber

• Attackers steal Uber truck driver’s account.
Example Study: Office 365

- Attackers compromises users’ office 365 account
Conclusion

• An extensive measurement of squatting phishing domain
  • From 224,810,532 DNS records and 700+ brands
  • Detect and identify 1,175 squatting phishing pages
  • Open-sourced our tool at: https://github.com/SquatPhish

• Future work
  • Adversarial attacks for OCR-based phishing detection
  • Deploy the system for long term measurement
Thank You
APPENDIX
Evasions in Squatting Phishing

- Layout obfuscation: average 28.5 hamming distance
- String obfuscation: 68% adopted
- Code obfuscation: 35% adopted

Obfuscation is common to squatting phishing.
IP Location

• Check geolocation of 1,021 IP addresses, hosted in 53 different countries.
• U.S. has most of the sites, then Germany
False Positive Prediction

http://paypal.me

Help others know it's you they're paying.