PhishZoo: Detecting Phishing Websites By Looking at Them

Sadia Afroz
Rachel Greenstadt
The Phishing Problem

Alice uses online bank

Real bank

Fake bank

Alice thinks everything that looks like her bank is her bank!

URL

Browser Indicator

SSL
What is PhishZoo

Phishing Site

Fake site

Real site
Overview

• Motivation
• Related works
• Contribution
• PhishZoo: Approach
• Evaluation and results
• Conclusion and future works
Motivation

• Current state of phishing
  – According to the Anti-Phishing Working Group (APWG), there were at least 67,677 phishing attacks in the last six months of 2010.
• Importance of Appearance
  – 90.9% people trusts a site based on it’s appearance [Dhamija et al, 2006]

• Spear Phishing
  – Increase use of spear phishing or targeted attacks

• The majority of users provide sensitive credentials to a small set of sites (fewer than 20) [Cao et al, 2008]
Related works

- **Non-content based approaches:**
  - URL based phishing detection [Ma et al, 2009]
  - Blacklisting
  - Whitelisting

- **Content based approaches:**
  - CANTINA
  - Google’s anti-phishing filter

- **Visual similarity based phishing detection:**
  - Screenshots of websites [Chen et al, 2009]
  - Screen capture with Earth Mover’s Distance [Fu et al, 2006]
  - Layout and style similarity [Liu et al, 2006]
  - Optical character recognition of Screenshots of webpage [Dunlop et al, 2010]
1. We investigated vision techniques to detect phishing sites more robustly.

2. It can detect new phishing sites which are not yet blacklisted and targeted attacks against small brokerages and corporate intranets..
PhishZoo: Approach

Real site

Fake site

Extracts visual elements of the site

Images Visible text

Visual components match but the url, ssl don’t match

Profile Stored

Phishing Alert

Extracts visual elements of the site

Images Visible text

PhishZoo
PhishZoo: Site Profile

• Profile Making
  – A profile of a site is a combination of different metrics that uniquely identifies that site
  – A user chooses real sites that he wants to protect from phishing to be saved as profiles.
  – In a profile, PhishZoo stores SSL certificates, URL and contents related to a site’s appearance such as HTML files, extracted features of the logo.
  – SIFT is used to extract image features.
PhishZoo: Image matching

- Currently, only logo of a site is saved in the site profile.
- For image matching, we used SIFT
- SIFT extracts keypoints from an image
- These keypoints are invariant to affine transformation, scaling, rotation up to 30 degrees, and illumination.
- During matching, the keypoints of two images are matched.
- Two images are considered similar if the percentage of matched keypoints is greater than a threshold
Data Collection

• We used 20 stored sites.

• For true positive testing: 1000 verified phishing sites from Phishtank.

Evaluation and Results

• Profile Content Analysis:
  – HTML
  – Visible text in HTML
  – Keywords: top TF-IDF ranked words of a site and url
  – Images
  – Images and visible text
  – Screenshots
  – Images and keywords
Results: Texts

- HTML: True Positive - 21.5, False Positive - 1
- Visible text: True Positive - 70.3, False Positive - 0.5
- Keywords: True Positive - 97.6, False Positive - 18.7
Results: Images

- Images: True Positive 82.7, False Positive 2.5
- Images and visible texts: True Positive 96.4, False Positive 1.4
- Screenshots: True Positive 81.1, False Positive 30.3
- Images and keywords: True Positive 90.2, False Positive 0.5
• Performance Analysis:
  – Keyword and image matching has better accuracy than other profile contents.

  – This recommended approach takes about 7 to 17 seconds on average to compare a site against all the profiles.
Fetch site

Online profile matching → Select n keywords From profiles

Select most relevant profile → Search for the keywords in current site

Perform image matching with the most likely profile → Site with most keyword match is the most relevant profile

If matching found, then the site is phishing site

Offline profile matching → Perform image matching with all profiles in the background

If any matching found, then site is phishing site
Limitations

1. Site redirection
2. No or obfuscated use of the keywords,
3. Use of cropped, extended or rotated (more than 30 degrees) images
Conclusion and Future work

• We described a new approach of web-phishing detection using vision techniques.
  
• Use a faster image machine algorithm
• Scene analysis
References

