CS259D: Data Mining for Cybersecurity
Phishing

Goal:
- Account information
- Logon credentials
- Identity information

Attack vectors:
- Legitimate-looking emails
- Legitimate-looking websites
Scale of the problem
Detection

- Toolbars
  - Spoofguard
  - Netcraft

- Email filtering
  - Examples:
    - SpamAssassin
    - Spamato
  - Advantages:
    - More complete context (content, headers, etc.)
    - Completely shield user from decision-making process
Phishing classification: Features

- IP-based URLs
  - Example:
  - Compromised PCs with no DNS entries
  - Binary feature

- Age of linked-to domain names
  - Registered legitimate-sounding domain names
    - Example: playpal.com, paypal-update.com
  - Typically short life-span
    - Registered using stolen credit cards, canceled by registrar
    - Domain caught by anti-phishing monitors
    - Often lasting only ~ 48 hours
  - Obtained using a WHOIS query
  - Binary feature: Lifetime < 60 days
Phishing classification: Features

- Non-matching URLs
  - Example: `<a href="badsite.com">paypal.com</a>`
  - Binary feature: URL text different from HREF

- “Here” links to non-modal domain
  - Example: Click here to restore your account
  - Modal domain: domain most frequently linked to
  - Binary feature: link with text “link”, “click”, “here” that links to a domain other than modal domain
Phishing classification: Features

- **HTML emails**
  - Binary feature: email section with MIME type text/html

- **Number of links**
  - Numeric feature: # links in HTML part(s) of email
  - Link defined by an `<a>` tag with `href` attribute
    - Including mailto: links
Phishing classification: Features

• **Number of domains**
  ◦ Domain names for URLs starting with http/https
  ◦ Only the main part of the domain name
    • What registrar gets paid for
      • Not necessarily same as combination of top- & 2nd-level domain
    • Example:
      • university.edu for [www.cs.university.edu](http://www.cs.university.edu)
      • company.co.jp for [www.company.co.jp](http://www.company.co.jp)
        • Top-level: .jp, second-level: .co
  ◦ Numeric feature: #distinct domains
Phishing classification: Features

- Number of dots
  - Subdomains: http://www.my-bank.update.data.com
    - Looks to naïve user to be from google.com
    - Redirects browser to badsite.com
  - Numeric feature: Maximum number of dots in any of the links in the email
Phishing classification: Features

- Contains javascript
  - Binary feature: string “javascript” appears in email

- Spam filter output
  - Binary feature: class assigned to email by SpamAssassin
Evaluation

- 10-fold cross validation
- Classifier: Random forest
  - 10 decision trees
  - Each decision made on a random attribute
  - Trees pruned
Evaluation

- SpamAssassin ham corpora
  - ~6950 non-phishing non-spam

- Publicly available phishing corpus
  - ~ 860 phishing messages
  - Challenge with WHOIS queries
    - Only 505 domains out of 870 domains
    - Increases false negative rate
## Evaluation

<table>
<thead>
<tr>
<th>Feature</th>
<th>Non-Phishing Matched</th>
<th>Phishing Matched</th>
</tr>
</thead>
<tbody>
<tr>
<td>Has IP link</td>
<td>0.06%</td>
<td>45.04%</td>
</tr>
<tr>
<td>Has “fresh” link</td>
<td>0.98%</td>
<td>12.49%</td>
</tr>
<tr>
<td>Has “nonmatching” URL</td>
<td>0.14%</td>
<td>50.64%</td>
</tr>
<tr>
<td>Has non-modal here link</td>
<td>0.82%</td>
<td>18.20%</td>
</tr>
<tr>
<td>Is HTML email</td>
<td>5.55%</td>
<td>93.47%</td>
</tr>
<tr>
<td>Contains JavaScript</td>
<td>2.30%</td>
<td>10.15%</td>
</tr>
<tr>
<td>SpamAssassin Output</td>
<td>0.12%</td>
<td>87.05%</td>
</tr>
</tbody>
</table>
## Evaluation

<table>
<thead>
<tr>
<th>Feature</th>
<th>$\mu_{\text{phishing}}$</th>
<th>$\sigma_{\text{phishing}}$</th>
<th>$\mu_{\text{non-phishing}}$</th>
<th>$\sigma_{\text{non-phishing}}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of links</td>
<td>3.87</td>
<td>4.97</td>
<td>2.36</td>
<td>12.00</td>
</tr>
<tr>
<td>Number of domains</td>
<td>1.49</td>
<td>1.42</td>
<td>0.43</td>
<td>3.32</td>
</tr>
<tr>
<td>Number of dots</td>
<td>3.78</td>
<td>1.94</td>
<td>0.19</td>
<td>0.87</td>
</tr>
</tbody>
</table>
## Evaluation

<table>
<thead>
<tr>
<th>Classifier</th>
<th>False Positive Rate $fp$</th>
<th>False Negative Rate $fn$</th>
</tr>
</thead>
<tbody>
<tr>
<td>PILFER, with S.A. feature</td>
<td>0.0013</td>
<td>0.036</td>
</tr>
<tr>
<td>PILFER, without S.A. feature</td>
<td>0.0022</td>
<td>0.085</td>
</tr>
<tr>
<td>SpamAssassin (Untrained)</td>
<td>0.0014</td>
<td>0.376</td>
</tr>
<tr>
<td>SpamAssassin (Trained)</td>
<td>0.0012</td>
<td>0.130</td>
</tr>
</tbody>
</table>
Review of TF-IDF

- Measure importance of word in document
- TF = frequency of word in document
- IDF = measure popularity of word in corpus
  - $\log(N/\#\{\text{documents having the term}\})$
- $\text{tf-idf}(t, d, D) = \text{tf}(t, d) \times \text{idf}(t, D)$
Robust hyperlinks

- Lexical signatures for identifying URLs
- Signature words chosen using TF-IDF
- Experiments: 5 terms enough for unique page identification
Observation

- Minimal changes to original page detectable via robust hyperlinks
- Phishing sites often include brand names
  - Common on brand’s webpages
  - Rare on the web
Algorithm

- Compute term TF-IDFs
- Find top 5 terms
- Submit terms as query to Google
- Check if domain is among top-N results
- Assumption: phishing pages have low pagerank
Lowering false positives

- Include domain name in lexical signature
- Heuristic: Zero results Means Phishing
Example

Sign In

New to eBay? or

Already an eBay user?

If you want to sign in, you'll need to register first.

Registration is fast and free.

eBay members, sign in to save time for bidding, selling, and other activities.

eBay User ID

Forgot your User ID?

Password

Forgot your password?

Sign In Securely

Keep me signed in on this computer unless I sign out.
Example
Example

- Top terms: eBay, user, sign, help, forgot
Other features

- Age of domain
- Known images
  - Presence of inconsistent well-known logos
  - Top-10 identified targets: eBay, PayPal, Citibank, Bank of America, Fifth Third Bank, Barclays Bank, ANZ Bank, Chase Bank, and Wells Fargo Bank
- Suspicious URL
  - Contains @ or – in domain name
Other features

- Suspicious links
  - Same as suspicious URLs
- IP address as domain
- Dots in URL
  - Binary: \#Dots > 5
- Forms
  - HTML `<input>` tag, with text such as “credit card”, “password”
Evaluation

The diagram shows the true positive and false positive rates for different methods:

- Basic TF-IDF: 94% true positive, 30% false positive
- Basic TF-IDF + domain: 67% true positive, 10% false positive
- Basic TF-IDF + ZMP: 94% true positive, 10% false positive
- Basic TF-IDF + ZMP + domain: 97% true positive, 10% false positive
Evaluation

![Bar chart showing true positive and false positive rates for different rankings.]

- True positive:
  - Top 1: 97%
  - Top 10: 97%
  - Top 30: 97%
  - Top 50: 97%

- False positive:
  - Top 1: 30%
  - Top 10: 15%
  - Top 30: 10%
  - Top 50: 10%
Evaluation

![Evaluation Table]

<table>
<thead>
<tr>
<th>Method</th>
<th>True Positive</th>
<th>False Positive</th>
</tr>
</thead>
<tbody>
<tr>
<td>Final-TF-IDF</td>
<td>97%</td>
<td>6%</td>
</tr>
<tr>
<td>Final-TF-IDF+heuristics</td>
<td>89%</td>
<td>1%</td>
</tr>
<tr>
<td>SpoofGuard</td>
<td>91%</td>
<td>0%</td>
</tr>
<tr>
<td>Netcraft</td>
<td>97%</td>
<td>0%</td>
</tr>
</tbody>
</table>
References

• “Learning to Detect Phishing Emails”, Fette et al, 2007

• “Cantina: A content-based approach to detecting phishing websites”, Zhang et al, 2007