Bypassing Phishing Filters

Shahrukh Zaidi
MSc System and Network Engineering
(University of Amsterdam)
Supervisors:
Alex Stavroulakis, Rick van Galen
(KPMG)
Phishing emails

- Special type of spam message
- Fraudulent social engineering techniques to elicit sensitive information from unsuspected users¹
- Anti-spam filters include phishing detection solutions to combat phishing

Research question

*Which aspects of a phishing email can be modified in order to bypass common phishing filters?*
Research question

Sub-questions:

- What are common characteristics of phishing emails?
- What detection techniques are commonly utilised by phishing filters?
- What methods can be deployed to bypass these detection techniques?
Theoretical framework

Phishing email characteristics²³:

- 'Fresh' linked-to domains
- Disparity between domain names in message body and sender’s domain
- Non-matching URLs
  - <a href="badsite.com"> paypal.com </a>
- Frequently repeated keywords
  - 'update', 'confirm', 'suspend', 'verify', 'account'

Theoretical framework

Phishing email detection techniques⁴:

- Blacklists
- Whitelists
- Heuristics
  - Content-based filtering
  - Machine learning (e.g. Bayesian classification)

Theoretical framework

Example spam report:

```
$ spmc -R < tests/capitalone.txt

Content preview: Dear Capital One Customer. Sincerely, Capital One Security Department www.capitalone.com Dear Capital One Customer. [...] 
```

Content analysis details: (5.4 points, 5.0 required)

<table>
<thead>
<tr>
<th>pts</th>
<th>rule name</th>
<th>description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.6</td>
<td>SPOOF_COM2COM</td>
<td>URI: URI contains &quot;.com&quot; in middle and end</td>
</tr>
<tr>
<td>0.0</td>
<td>HTML_MESSAGE</td>
<td>BODY: HTML included in message</td>
</tr>
<tr>
<td>1.0</td>
<td>HTML_IMAGE_ONLY_16</td>
<td>BODY: HTML: images with 1200-1600 bytes of words</td>
</tr>
<tr>
<td>1.5</td>
<td>TVD_PH_BODYACCOUNTS_PRE</td>
<td>The body matches phrases such as &quot;accounts suspended&quot;, &quot;account credited&quot;, &quot;account verification&quot;</td>
</tr>
<tr>
<td>0.0</td>
<td>T_DKIM_INVALID</td>
<td>DKIM-Signature header exists but is not valid</td>
</tr>
<tr>
<td>0.1</td>
<td>MISSING_MID</td>
<td>Missing Message-Id: header</td>
</tr>
<tr>
<td>1.0</td>
<td>ACCT_PHISHING</td>
<td>Possible phishing for account information</td>
</tr>
<tr>
<td>0.0</td>
<td>T_REMOTE_IMAGE</td>
<td>Message contains an external image</td>
</tr>
</tbody>
</table>
Related work

Detection evasion techniques:

- Statistical evasion
- Tokenization
  - HTML tricks:
    - acc<i></i>ount vs. account
    - acc<font size="0"> </font>ount
- Obfuscation
  - Unicode transliteration:
    - latin ‘a’ (U+0061) vs. cyrillic ‘а’ (U+0430)
  - Scrambling
  - Misspelling
  - URL obfuscation
    - URL shorteners
Methodology

Analysis of phishing emails:

● Test data set containing ~300 phishing emails
● Analyse output of spam reports
  ○ SpamAssassin
  ○ Rspamd
● Determine frequently triggered rules
● Apply obfuscation techniques and observe effect
  ○ ProtonMail
  ○ Office 365 (/KPMG)
  ○ G Suite Gmail
  ○ Amazon WorkMail
  ○ RackSpace Email
## Results: analysis of phishing emails

### Table 1: SpamAssassin - frequently triggered rules

<table>
<thead>
<tr>
<th>Rule</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>MIME_HTML_ONLY</td>
<td>Message has only HTML part</td>
</tr>
<tr>
<td>ACCT_PHISHING</td>
<td>Possible phishing for account information</td>
</tr>
<tr>
<td>TVD_PH_BODY_ACCOUNTS_PRE</td>
<td>Body matches phrases such as 'accounts'</td>
</tr>
<tr>
<td>FREEMAIL_FORGED_REPLYTO</td>
<td>Freemail in Reply-To, but not From</td>
</tr>
<tr>
<td>SUBJ_ALL_CAPS</td>
<td>All capital letters in subject</td>
</tr>
<tr>
<td>HEADER_FROM_DIFFERENT_DOMAINS</td>
<td>From and EnvelopeFrom different</td>
</tr>
<tr>
<td>URI_WPADMIN</td>
<td>WordPress login/admin URI</td>
</tr>
<tr>
<td>RDNS_NONE</td>
<td>Delivered by host with no rDNS</td>
</tr>
</tbody>
</table>
Results: analysis of phishing emails

Table 2: Rspamd - frequently triggered rules

<table>
<thead>
<tr>
<th>Rule</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>MIME_HTML_ONLY</td>
<td>Message has only HTML part</td>
</tr>
<tr>
<td>FROM_NEQ_ENVFROM</td>
<td>From address is different to the envelope</td>
</tr>
<tr>
<td>HAS_ATTACHMENT</td>
<td>Contains attachment</td>
</tr>
<tr>
<td>HAS_WP_URI</td>
<td>Contains WordPress URIs</td>
</tr>
<tr>
<td>FREEMAIL_REPLYTO</td>
<td>Freemail in Reply-To, but not From</td>
</tr>
<tr>
<td>PHISHING</td>
<td>Non matching URLs in HTML text and href</td>
</tr>
<tr>
<td>RSPAMD_URIBL</td>
<td>URL in URIBL.com blacklist</td>
</tr>
<tr>
<td>HFILTER_FROMHOST_NORES_A_OR_MX</td>
<td>From host no resolve to A or MX</td>
</tr>
</tbody>
</table>
Results: applying obfuscation techniques

Example phishing email:

Dear Capital One Customer.

Your Capital One Internet Banking account has been temporary suspended.

We require you to Unlock your account Unlock Access.

Sincerely,
Capital One Security Department

www.capitalone.com
Results: applying obfuscation techniques

Spam report original phishing email:

$ spamc -R < tests/capitalone.txt

Content preview: Dear Capital One Customer. Sincerely, Capital One Security Department www.capitalone.com Dear Capital One Customer. [...] 

Content analysis details: (5.4 points, 5.0 required)

<table>
<thead>
<tr>
<th>pts</th>
<th>rule name</th>
<th>description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.6</td>
<td>SPOOF_COM2COM</td>
<td>URI: URI contains &quot;.com&quot; in middle and end</td>
</tr>
<tr>
<td>0.0</td>
<td>HTML_MESSAGE</td>
<td>BODY: HTML included in message</td>
</tr>
<tr>
<td>1.0</td>
<td>HTML_IMAGE_ONLY_16</td>
<td>BODY: HTML: images with 1200-1600 bytes of words</td>
</tr>
<tr>
<td>1.5</td>
<td>TVD_PH_BODY_ACCOUNTS_PRE</td>
<td>The body matches phrases such as &quot;accounts suspended&quot;, &quot;account credited&quot;, &quot;account verification&quot;</td>
</tr>
<tr>
<td>0.0</td>
<td>T_DKIM_INVALID</td>
<td>DKIM-Signature header exists but is not valid</td>
</tr>
<tr>
<td>0.1</td>
<td>MISSING MID</td>
<td>Missing Message-Id: header</td>
</tr>
<tr>
<td>1.0</td>
<td>ACCT_PHISHING</td>
<td>Possible phishing for account information</td>
</tr>
<tr>
<td>0.0</td>
<td>T_REMOTE_IMAGE</td>
<td>Message contains an external image</td>
</tr>
</tbody>
</table>
Results: applying obfuscation techniques

Spam report phishing email with fake HTML tag insertion:

Not effective

```bash
$ spmc -R < tests/capitalone_obf_html.txt
```

Content preview: Dear Capital One Customer. Sincerely, Capital One Security Department www.capitalone.com Dear Capital One Customer. [...] 

Content analysis details: (5.4 points, 5.0 required)

<table>
<thead>
<tr>
<th>pts</th>
<th>rule name</th>
<th>description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.6</td>
<td>SPOOF COM2COM</td>
<td>URI: URI contains &quot;.com&quot; in middle and end</td>
</tr>
<tr>
<td>0.0</td>
<td>HTML_OBFUSCATE_05_10</td>
<td>BODY: Message is 5 to 10 percent HTML obfuscation</td>
</tr>
<tr>
<td>0.0</td>
<td>HTML_MESSAGE</td>
<td>BODY: HTML included in message</td>
</tr>
<tr>
<td>1.0</td>
<td>HTML_IMAGE_ONLY_16</td>
<td>BODY: HTML: images with 1200-1600 bytes of words</td>
</tr>
<tr>
<td>1.5</td>
<td>TVD_PH_BODY_ACCOUNTS_PRE</td>
<td>The body matches phrases such as &quot;accounts suspended&quot;, &quot;account credited&quot;, &quot;account verification&quot;</td>
</tr>
<tr>
<td>0.0</td>
<td>T_DKIM_INVALID</td>
<td>DKIM-Signature header exists but is not valid</td>
</tr>
<tr>
<td>0.1</td>
<td>MISSING MID</td>
<td>Missing Message-Id: header</td>
</tr>
<tr>
<td>1.0</td>
<td>ACCT_PHISHING</td>
<td>Possible phishing for account information</td>
</tr>
<tr>
<td>0.0</td>
<td>T_REMOTE_IMAGE</td>
<td>Message contains an external image</td>
</tr>
</tbody>
</table>
Results: applying obfuscation techniques

Spam report phishing email with Unicode obfuscation applied:

Effective

```
$ spmc -R < tests/capitalone_obf_unicode.txt

Content preview: Dear Capital One Customer. Sincerely, Capital One Security Department www.capitalone.com Dear Capital One Customer. [...] [...]

Content analysis details: (2.8 points, 5.0 required)

<table>
<thead>
<tr>
<th>pts rule name</th>
<th>description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.6 SPOOF_COM2COM</td>
<td>URI: URI contains &quot;.com&quot; in middle and end</td>
</tr>
<tr>
<td>0.0 HTML_MESSAGE</td>
<td>BODY: HTML included in message</td>
</tr>
<tr>
<td>1.0 HTML_IMAGE_ONLY_16</td>
<td>BODY: HTML: images with 1200-1600 bytes of words</td>
</tr>
<tr>
<td>0.0 T_DKIM_INVALID</td>
<td>DKIM-Signature header exists but is not valid</td>
</tr>
<tr>
<td>0.1 MISSING_MID</td>
<td>Missing Message-Id: header</td>
</tr>
<tr>
<td>0.0 T_REMOTE_IMAGE</td>
<td>Message contains an external image</td>
</tr>
</tbody>
</table>
Results: applying obfuscation techniques

Spam report phishing email with Unicode obfuscation applied and URL replaced with bit.ly short URL: Effective

$ spmc -R < tests/capitalone_obf_unicode_url.txt

Content preview: Dear Capital One Customer. Sincerely, Capital One Security Department www.capitalone.com Dear Capital One Customer. [...] 

Content analysis details: (1.1 points, 5.0 required)

<table>
<thead>
<tr>
<th>pts</th>
<th>rule name</th>
<th>description</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.0</td>
<td>HTML_MESSAGE</td>
<td>BODY: HTML included in message</td>
</tr>
<tr>
<td>1.0</td>
<td>HTML_IMAGE_ONLY_16</td>
<td>BODY: HTML: images with 1200-1600 bytes of words</td>
</tr>
<tr>
<td>0.0</td>
<td>T_DKIM_INVALID</td>
<td>DKIM-Signature header exists but is not valid</td>
</tr>
<tr>
<td>0.1</td>
<td>MISSING_MID</td>
<td>Missing Message-Id: header</td>
</tr>
<tr>
<td>0.0</td>
<td>T_REMOTE_IMAGE</td>
<td>Message contains an external image</td>
</tr>
</tbody>
</table>
Proof of Concept

- Python script
  - Input: HTML email
  - Input: common phishing words
  - Iterate through HTML contents:
    - Apply Unicode obfuscation to common phishing words
      - replace vowels with Unicode visually identical character
    - Replace all href links with short URL
  - Save new HTML
Dear Capital One Customer,

Your Capital One Internet Banking account has been temporarily suspended. We require you to Unlock your account. 

Sincerely,
Capital One Security Department

Unlock Access

[Website Link]

www.capitalone.com
Sample phishing mail: obfuscated

Dear Capital One Customer,

Your Capital One Internet Banking account has been temporarily suspended. We require you to unlock your account. Sincerely,

Capital One Security Department

[B#1072;nk#8560;ng &#1072;cc#959;unt has been temporary susp&#1077;nd#1077;d. We r&#1077;qu#8560;r&#1077; you to &#5196;nl#959;ck your &#1072;cc#959;unt]<A href="http://bit.ly/2JWtONR" rel=nofollow target=_blank><SPAN id=lw_1336748011_1 class=yshortcuts>Unlock Access</SPAN></A>.

Sincerely,

Capital One Security Department

### Results: effectiveness of obfuscation techniques (ProtonMail)

<table>
<thead>
<tr>
<th>Sample phishing email</th>
<th>Phishing related rules triggered using original email</th>
<th>Phishing related rules triggered after obfuscation techniques applied</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>bitstamp</strong></td>
<td>URI_WPADMIN (Spam score: 3.0)</td>
<td>URI_WPADMIN (Spam score: 0.2)</td>
</tr>
<tr>
<td></td>
<td>SPOOF_COM2COM TVD_PH_BODY_ACCOUNTS_PRE (Spam score: 3.5)</td>
<td>SPOOF_COM2COM TVD_PH_BODY_ACCOUNTS_PRE (Spam score: 1.5)</td>
</tr>
<tr>
<td><strong>capitalone</strong></td>
<td>URI_WPADMIN (Spam score: 4.6)</td>
<td>URI_WPADMIN TVD_PH_BODY_ACCOUNTS_PRE (Spam score: 1.8)</td>
</tr>
<tr>
<td><strong>dhl</strong></td>
<td>URIBL_PH_SURBL_PQS RAZOR2_CHECK (Spam score: 9.8)</td>
<td>URIBL_PH_SURBL_PQS RAZOR2_CHECK (Spam score: -0.1)</td>
</tr>
<tr>
<td><strong>fedex</strong></td>
<td>URI_WPADMIN TVD_PH_BODY_ACCOUNTS_PRE (Spam score: 4.6)</td>
<td>URI_WPADMIN TVD_PH_BODY_ACCOUNTS_PRE (Spam score: 1.8)</td>
</tr>
</tbody>
</table>
Results: effectiveness of obfuscation techniques (Office 365)

<table>
<thead>
<tr>
<th>Sample phishing email</th>
<th>Short URL</th>
<th>Unicode Obfuscation</th>
<th>Short URL + Unicode Obfuscation</th>
</tr>
</thead>
<tbody>
<tr>
<td>bitstamp</td>
<td>✗</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>capitalone</td>
<td>✗</td>
<td>✗</td>
<td>✓</td>
</tr>
<tr>
<td>dh1</td>
<td>✓</td>
<td>✗</td>
<td>✓</td>
</tr>
<tr>
<td>fedex</td>
<td>✗</td>
<td>✗</td>
<td>✓</td>
</tr>
<tr>
<td>dropbox</td>
<td>✗</td>
<td>✗</td>
<td>✗</td>
</tr>
</tbody>
</table>
Results: effectiveness of obfuscation techniques (Office 365 KPMG)

<table>
<thead>
<tr>
<th>Sample phishing email</th>
<th>Short URL</th>
<th>Unicode Obfuscation</th>
<th>Short URL + Unicode Obfuscation</th>
</tr>
</thead>
<tbody>
<tr>
<td>dhl</td>
<td>✗</td>
<td>✗</td>
<td>✗</td>
</tr>
<tr>
<td>fedex</td>
<td>✗</td>
<td>✗</td>
<td>✓</td>
</tr>
<tr>
<td>docusign</td>
<td>✓</td>
<td>✗</td>
<td>✓</td>
</tr>
<tr>
<td>netflix</td>
<td>✗</td>
<td>✗</td>
<td>✓</td>
</tr>
<tr>
<td>security_alert</td>
<td>✓</td>
<td>✗</td>
<td>✓</td>
</tr>
</tbody>
</table>
Results: effectiveness of obfuscation techniques (G Suite Gmail)

<table>
<thead>
<tr>
<th>Sample phishing email</th>
<th>Short URL</th>
<th>Unicode Obfuscation</th>
<th>Short URL + Unicode Obfuscation</th>
</tr>
</thead>
<tbody>
<tr>
<td>bitstamp</td>
<td>✗</td>
<td>✗</td>
<td>✓</td>
</tr>
<tr>
<td>acc_terminate</td>
<td>✓</td>
<td>✗</td>
<td>✓</td>
</tr>
<tr>
<td>docusign</td>
<td>✓</td>
<td>✗</td>
<td>✓</td>
</tr>
<tr>
<td>dropbox</td>
<td>✗</td>
<td>✗</td>
<td>✗</td>
</tr>
<tr>
<td>bank_of_america</td>
<td>✓</td>
<td>✗</td>
<td>✓</td>
</tr>
</tbody>
</table>
Results: effectiveness of obfuscation techniques (Amazon WorkMail)

<table>
<thead>
<tr>
<th>Sample phishing email</th>
<th>Short URL</th>
<th>Unicode Obfuscation</th>
<th>Short URL + Unicode Obfuscation</th>
</tr>
</thead>
<tbody>
<tr>
<td>bitstamp</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>capitalone</td>
<td>✗</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>dhl</td>
<td>✗</td>
<td>✗</td>
<td>✓</td>
</tr>
<tr>
<td>fedex</td>
<td>✗</td>
<td>✗</td>
<td>✓</td>
</tr>
<tr>
<td>dropbox</td>
<td>✗</td>
<td>✗</td>
<td>✗</td>
</tr>
</tbody>
</table>
Results: effectiveness of obfuscation techniques (Rackspace email)

<table>
<thead>
<tr>
<th>Sample phishing email</th>
<th>Short URL</th>
<th>Unicode Obfuscation</th>
<th>Short URL + Unicode Obfuscation</th>
</tr>
</thead>
<tbody>
<tr>
<td>acc_terminate</td>
<td>✓</td>
<td>✗</td>
<td>✓</td>
</tr>
<tr>
<td>blacklist</td>
<td>✗</td>
<td>✗</td>
<td>✗</td>
</tr>
<tr>
<td>alibaba</td>
<td>✓</td>
<td>✗</td>
<td>✓</td>
</tr>
</tbody>
</table>
Discussion

- Unicode obfuscation not triggered as being suspicious by any of the tested spam filters

- URL shortening obfuscation undetected

- Mitigation can be fairly simple
  - Set up list containing identical clones of suspicious word
  - Flag any character not common in English language
  - Short URL detection may be trickier
Conclusion

- Phishing filters commonly apply blacklisting and heuristic techniques to identify phishing emails

- Obfuscation of certain words and URLs can be sufficient to fool these filters
Future work

- Consider additional aspects other than the contents only
- Determine effect of phishing emails sent in bulk
Questions?