UNIVERSITY OF AMSTERDAM

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 users[1]
- Anti-spam filters include phishing detection solutions to combat phishing

[1] Aggarwal, S., Kumar, V., & Sudarsan, S. D. (2014, September). Identification and detection of phishing emails using natural language processing techniques. In Proceedings of the 7th International Conference on Security of Information and Networks (p. 217). ACM.

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 characteristics[2][3]:

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

[2] Fette, I., Sadeh, N., & Tomasic, A. (2007, May). Learning to detect phishing emails. In Proceedings of the 16th international conference on World Wide Web (pp. 649-656). ACM.
[3] Basnet, R., Mukkamala, S., & Sung, A. H. (2008). Detection of phishing attacks: A machine learning approach. In Soft Computing Applications in Industry (pp. 373-383). Springer, Berlin, Heidelberg.

Theoretical framework

Phishing email detection techniques[4]:

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

[4] Hajgude, J., & Ragha, L. (2012, October). Phish mail guard: Phishing mail detection technique by using textual and URL analysis. In Information and Communication Technologies (WICT), 2012 World Congress on (pp. 297-302). IEEE.

Theoretical framework

Example spam report:

```
$ 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)

 pts rule name              description
---- ---------------------- --------------------------------------------------
 1.6 SPOOF_COM2COM          URI: URI contains ".com" in middle and end
 0.0 HTML_MESSAGE           BODY: HTML included in message
 1.0 HTML_IMAGE_ONLY_16     BODY: HTML: images with 1200-1600 bytes of words
 1.5 TVD_PH_BODY_ACCOUNTS_PRE The body matches phrases such as "accounts
                            suspended", "account credited", "account
                            verification"
 0.0 T_DKIM_INVALID         DKIM-Signature header exists but is not valid
 0.1 MISSING_MID            Missing Message-Id: header
 1.0 ACCT_PHISHING          Possible phishing for account information
 0.0 T_REMOTE_IMAGE         Message contains an external image
```

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 ‘a’ (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

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

Results: analysis of phishing emails

Table 2: Rspamd - frequently triggered rules

Rule	Description
MIME_HTML_ONLY	Message has only HTML part
FROM_NEQ_ENVFROM	From address is different to the envelope
HAS_ATTACHMENT	Contains attachment
HAS_WP_URI	Contains WordPress URIs
FREEMAIL_REPLYTO	Freemail in Reply-To, but not From
PHISHING	Non matching URLs in HTML text and href
RSPAMD_URIBL	URL in URIBL.com blacklist
HFILTER_FROMHOST_NORES_A_OR_MX	From host no resolve to A or MX

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)

 pts rule name              description
---- ---------------------- --------------------------------------------------
1.6 SPOOF_COM2COM           URI: URI contains ".com" in middle and end
0.0 HTML_MESSAGE            BODY: HTML included in message
1.0 HTML_IMAGE_ONLY_16      BODY: HTML: images with 1200-1600 bytes of words
1.5 TVD_PH_BODY_ACCOUNTS_PRE The body matches phrases such as "accounts
                            suspended", "account credited", "account
                            verification"
0.0 T_DKIM_INVALID          DKIM-Signature header exists but is not valid
0.1 MISSING_MID             Missing Message-Id: header
1.0 ACCT_PHISHING           Possible phishing for account information
0.0 T_REMOTE_IMAGE          Message contains an external image
```

Results: applying obfuscation techniques

Spam report phishing email with fake HTML tag insertion:

Not effective

```
$ spamc -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)

 pts rule name              description
---- ---------------------- --------------------------------------------------
 1.6 SPOOF_COM2COM          URI: URI contains ".com" in middle and end
 0.0 HTML_OBFUSCATE_05_10   BODY: Message is 5 to 10 percent HTML obfuscation
 0.0 HTML_MESSAGE           BODY: HTML included in message
 1.0 HTML_IMAGE_ONLY_16     BODY: HTML: images with 1200-1600 bytes of words
 1.5 TVD_PH_BODY_ACCOUNTS_PRE The body matches phrases such as "accounts
                            suspended", "account credited", "account
                            verification"
 0.0 T_DKIM_INVALID         DKIM-Signature header exists but is not valid
 0.1 MISSING_MID            Missing Message-Id: header
 1.0 ACCT_PHISHING          Possible phishing for account information
 0.0 T_REMOTE_IMAGE         Message contains an external image
```

Results: applying obfuscation techniques

Spam report phishing email with Unicode obfuscation applied:

Effective

```
$ spamc -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)

 pts rule name              description
---- ---------------------- --------------------------------------------------
1.6 SPOOF_COM2COM           URI: URI contains ".com" in middle and end
0.0 HTML_MESSAGE            BODY: HTML included in message
1.0 HTML_IMAGE_ONLY_16      BODY: HTML: images with 1200-1600 bytes of words
0.0 T_DKIM_INVALID          DKIM-Signature header exists but is not valid
0.1 MISSING_MID             Missing Message-Id: header
0.0 T_REMOTE_IMAGE          Message contains an external image
```

Results: applying obfuscation techniques

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

Effective

```
$ spamc -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)

 pts rule name              description
---- ---------------------- --------------------------------------------------
 0.0 HTML_MESSAGE           BODY: HTML included in message
 1.0 HTML_IMAGE_ONLY_16     BODY: HTML: images with 1200-1600 bytes of words
 0.0 T_DKIM_INVALID         DKIM-Signature header exists but is not valid
 0.1 MISSING_MID            Missing Message-Id: header
 0.0 T_REMOTE_IMAGE         Message contains an external image
```

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

Sample phishing mail: original

<HTML><head><meta http-equiv="Content-Type" content="text/html; charset=iso-8859-1"/></head><BODY><P align=right><IMG src="https://s.graphiq.com/sites/default/files/765/media/images/t2/Capital_One_827157.png" width=210 align=left height=40></P>
<P>
</P>
<P></P>
<P><B><FONT size=-1 face="Verdana, Arial, Helvetica, sans-serif">Dear </FONT><FONT size=-1 face=Arial><SPAN id=lw_1336748011_0 class=yshortcuts>Capital One</SPAN></FONT><FONT size=-1 face="Verdana, Arial, Helvetica, sans-serif"><SPAN></SPAN>
Customer.</FONT></B></P>
<FONT style="FONT-SIZE: 12px; LINE-HEIGHT: 18px" color=#000000 size=2 face=verdana,arial,helvetica,sans-serif>Your Capital One Internet **Banking account** has been temporary **suspended**.

We **require** you to **Unlock** your **account** <B><A **href="http://www.christianmccannauctions.com.au/cp/images/images/Cap1/Capitalone/OnlineBanking.htm"** rel=nofollow target=_blank><SPAN id=lw_1336748011_1 class=yshortcuts>Unlock Access</SPAN></A></B>.

Sincerely,
Capital One **Security** Department</FONT><FONT size=-1 face="Verdana, Arial, Helvetica, sans-serif">
</FONT>
<P><FONT size=2 face=Verdana><A **href="http://capitalone360.com.alsheheri.com/capital360/index.html"** rel=nofollow target=_blank><B><SPAN id=lw_1336748011_2 class=yshortcuts>www.capitalone.com</SPAN></B></A></FONT></P></TD></BODY></HTML>

Sample phishing mail: obfuscated

<HTML><head><meta http-equiv="Content-Type" content="text/html; charset=utf-8"/></head><BODY><P align=right><IMG src="https://s.graphiq.com/sites/default/files/765/media/images/t2/Capital_One_827157.png" width=210 align=left height=40></P>
 <P>
</P> <P></P> <P><B><FONT size=-1 face="Verdana, Arial, Helvetica, sans-serif">Dear </FONT><FONT size=-1 face=Arial><SPAN id=lw_1336748011_0 class=yshortcuts>Capital One</SPAN></FONT><FONT size=-1 face="Verdana, Arial, Helvetica, sans-serif"><SPAN></SPAN> Customer.</FONT></B></P>
<FONT style="FONT-SIZE: 12px; LINE-HEIGHT: 18px" color=#000000 size=2 face=verdana,arial,helvetica,sans-serif>Your Capital One Internet **Bаnkⅰng аccοunt** has been temporary **suspеndеd**.

We **rеquⅰrе** you to **ᑌnlοck** your **аccοunt** <B><A **href="http://bit.ly/2JWtONR"** rel=nofollow target=_blank><SPAN id=lw_1336748011_1 class=yshortcuts>Unlock Access</SPAN></A></B>.

Sincerely,
Capital One **Sеcurⅰty** Department</FONT><FONT size=-1 face="Verdana, Arial, Helvetica, sans-serif">
</FONT> <P><FONT size=2 face=Verdana><A **href="http://bit.ly/2K9bltl"** rel=nofollow target=_blank><B><SPAN id=lw_1336748011_2 class=yshortcuts>Go to bank</SPAN></B></A></FONT></P></TD></BODY></HTML>

Results: effectiveness of obfuscation techniques (ProtonMail)

Sample phishing email	*Phishing related rules triggered using original email*	*Phishing related rules triggered after obfuscation techniques applied*
bitstamp	URI_WPADMIN (Spam score: 3.0)	~~URI_WPADMIN~~ (Spam score: 0.2)
capitalone	SPOOF_COM2COM TVD_PH_BODY_ACCOUNTS_PRE (Spam score: 3.5)	~~SPOOF_COM2COM~~ ~~TVD_PH_BODY_ACCOUNTS_PRE~~ (Spam score: 1.5)
dhl	URIBL_PH_SURBL_PQS RAZOR2_CHECK (Spam score: 9.8)	~~URIBL_PH_SURBL_PQS~~ ~~RAZOR2_CHECK~~ (Spam score: -0.1)
fedex	URI_WPADMIN TVD_PH_BODY_ACCOUNTS_PRE (Spam score: 4.6	~~URI_WPADMIN~~ ~~TVD_PH_BODY_ACCOUNTS_PRE~~ (Spam score: 1.8)

Results: effectiveness of obfuscation techniques (Office 365)

Sample phishing email	*Short URL*	*Unicode Obfuscation*	*Short URL + Unicode Obfuscation*
bitstamp	✗	✓	✓
capitalone	✗	✗	✓
dhl	✓	✗	✓
fedex	✗	✗	✓
dropbox	✗	✗	✗

Results: effectiveness of obfuscation techniques (Office 365 KPMG)

Sample phishing email	*Short URL*	*Unicode Obfuscation*	*Short URL + Unicode Obfuscation*
`dhl`	✗	✗	✗
`fedex`	✗	✗	✓
`docusign`	✓	✗	✓
`netflix`	✗	✗	✓
`security_alert`	✓	✗	✓

Results: effectiveness of obfuscation techniques (G Suite Gmail)

Sample phishing email	*Short URL*	*Unicode Obfuscation*	*Short URL + Unicode Obfuscation*
`bitstamp`	✗	✗	✓
`acc_terminate`	✓	✗	✓
`docusign`	✓	✗	✓
`dropbox`	✗	✗	✗
`bank_of_america`	✓	✗	✓

Results: effectiveness of obfuscation techniques (Amazon WorkMail)

Sample phishing email	*Short URL*	*Unicode Obfuscation*	*Short URL + Unicode Obfuscation*
bitstamp	✓	✓	✓
capitalone	✗	✓	✓
dhl	✗	✗	✓
fedex	✗	✗	✓
dropbox	✗	✗	✗

Results: effectiveness of obfuscation techniques (Rackspace email)

Sample phishing email	*Short URL*	*Unicode Obfuscation*	*Short URL + Unicode Obfuscation*
`acc_terminate`	✓	✗	✓
`blacklist`	✗	✗	✗
`alibaba`	✓	✗	✓

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?