Iterative Publication of Phishing Sites

Qian Cui, Guy-Vincent Jourdan, Gregor v. Bochmann, Iosif Viorel Onut
School of Information Technology and Engineering - University of Ottawa

Phishing Attacks

Phishing attacks are one of the important threats to individuals and corporations in today's Internet.

Phishing sites have the following features:
- **Very Short Lifespan**: Slightly more than 10 hours on average
- **Rapid Update Cycle**: Phishers constantly "refresh" their attacks by publishing new phishing sites.

=> Our working assumption here is that attackers are not creating new attack sites, but instead are constantly modifying and re-publishing existing ones.

Methodology

We have created a method for the detection of phishing duplicates which is not sensitive to the textual content of the page and will resist to slight changes to the page structure.

Our method has 3 steps:
- Extract tag vector
- Calculate proportional distance
- Cluster similar sites

**Extracting Features of Pages: Tag Vectors**

**Corpus of HTML Tags**: We used the complete set of HTML elements provided by the World Wide Web Consortium, and removed some of the more common tags, such as body, head, html

Our **Tag Vector** is a vector of the size of the corpus, counting the number of times each tag in the corpus appearing in the page

<table>
<thead>
<tr>
<th>tag vector</th>
<th>tag corpus</th>
</tr>
</thead>
<tbody>
<tr>
<td>[1, 0, 2, 3, 1, 1, 2, 0, 4]</td>
<td>&lt;form&gt; &lt;br&gt; &lt;span&gt; &lt;input&gt; &lt;button&gt; &lt;video&gt; &lt;input&gt; &lt;iframe&gt;</td>
</tr>
</tbody>
</table>

**Challenges**

- The textual content of the phishing site may be updated substantially.
- The structure of the phishing site may be modified, for example because some parts are being swapped.

**Proportional Distance**

Our **Proportional Distance** is used to evaluate the similarity between two tag vectors (smaller = more similar).

Let \( t_1 \) and \( t_2 \) be two non-null tag vectors over the same corpus of size \( n \). The proportional distance \( PD(t_1, t_2) \) is defined by the following formulas:

\[
D(x, y) = \begin{cases} 
1 & \text{if } x \neq y \\
0 & \text{if } x = y 
\end{cases} \\
L(x, y) = \begin{cases} 
1 & \text{if } x \neq 0 \text{ OR } y \neq 0 \\
0 & \text{if } x = 0 \text{ AND } y = 0 
\end{cases} \\
PD(t_1, t_2) = \sum_{i=1}^{n} \frac{D(t_1[i], t_2[i])}{L(t_1[i], t_2[i])}
\]

**Clustering Algorithm**

Given a set of tag vectors and a threshold \( H \), our clustering algorithm groups together the vectors whose proportional distance is less than \( H \).

**Conclusion**

1) We have shown that Phishers repeatedly re-publish their attacks by minor modifications or seeking new host, which can keep the attack active for a long time.
2) Our clustering method is very efficient, with about 90% replicas reported, the more phishing instances caught, the higher chance to prevent further attacks.