Compromised or Attacker-Owned: A Large Scale Classification and Study of Hosting Domains of Malicious URLs

Ravindu De Silva, Mohamed Nabeel, Charith Elvitigala, Issa Khalil, Ting Yu, Chamath Keppitiyagama
Malicious URLs from VirusTotal Feed

276K Malicious URLs/Week
How are these malicious sites hosted?
Compromised vs. Attacker-Owned

paypn1-son.shop (attack)

app-garden.com (bengin)

aaa.app-garden.com (compromised)
Why Distinguish Hosting Types?

- Different mitigation actions by different intermediaries
- Fine-grained blocking with minimal collateral damage
- Towards building better domain reputation systems
Current Status of Blacklists

VirusTotal

Google Safe Browsing

PhishTank
Malicious Hosting Types

- **Public Domain**
  - Subdomain.apex.tld/path
  - E.g.: 000webhostapp.com
  - Compromised subdomain or path
    - Attacker-owned subdomain or path
      - E.g.: fbook-png.000webapphost.com
      - Compromised Apex
        - E.g.: questionpro.com
        - Attacker-Owned Apex
          - E.g.: getbinance.org

- **Private Domain**
  - E.g.: nsa.gov
Machine Learning Pipeline Design

VT URL Feed -> VT Report Parser

Malicious URL Extractor -> Public/Private Domain Classifier

Compromised Domains
Attacker Owned Domains
Compromised Apexes
Attacker Owned Apexes

Compromised/Attack Classifier

VT NOD/NOH

Public/Private Classifier

Public
Private
Datasets – VT URLs

- Study Period: Aug. 2019 to Nov. 2019
- Total number of URLs: 800+ million

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Malicious URLs</th>
<th>Malicious Apexes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dataset1: Aug. 2019</td>
<td>3,434,226</td>
<td>373,238</td>
</tr>
<tr>
<td>Dataset2: Oct. 2019</td>
<td>4,398,584</td>
<td>358,762</td>
</tr>
</tbody>
</table>
Public/Private Classifier

- Key observations for feature engineering
  - Subdomain diversity
  - Variation and volume of scans

- Performance
  - Accuracy: 97.2%
  - Precision: 97.7%
  - Recall: 95.6%
Compromised/Attacker-Owned Classifier

- Key observations for feature engineering
  - Deviations in the hosting infrastructure
  - Lexical formation of URLs
  - Scan diversity

- Performance
  - Accuracy: 96.4%
  - Precision: 99.1%
  - Recall: 92.6%
Distribution of Attack Types – Malicious Apexes/URLs

- Malicious Apexes:
  - 1% Public
  - 99% Private

- Malicious Websites:
  - 46% Public
  - 54% Private
Distribution of Attack Types – Compromised/Attack

- **Private Malicious Websites**
  - Compromised: 66%
  - Attacker-Owned: 34%

- **Public Malicious Websites**
  - Compromised: 20%
  - Attacker-Owned: 80%
Properties of Compromised/Attack Websites

Popularity of Domain

Gap Between Registration and Attack Time
Comparing against GSB
Conclusions

- Content agnostic detection of various hosting types
  - Public vs. Private classifier
  - Compromised vs. Attacker-owned classifiers

- 81.7% malicious websites are hosted on apexes that attacks do not own
  - More needs to be done to secure benign websites and public domains

Future work

- Continuous evaluation of malicious URL types
- Integration with existing blacklists and reputations systems
Thank You!

Mohamed Nabeel
mnabeel@hbku.edu.qa