Identification of Malicious SSL Networks by Subgraph Anomaly Detection

Thomas Mathew, Dhia Mahjoub
Tucson, Jan 10\textsuperscript{th} 2018
Thomas Mathew, Senior Security Researcher, Cisco Umbrella (OpenDNS)
Data Science, Speaker at BruCon, Flocon, Kaspersky SAS, Black Hat, Defcon

Dhia Mahjoub, Head of Security Research, Cisco Umbrella (OpenDNS)
Graph theory, Speaker at Black Hat, Defcon, Flocon, RSA, Virus Bulletin
@DhiaLite
Actor(s) grow and maintain FF network

*FF service offered in underground forums

Victim

Crimeware consumer

Researcher

Zbot Fast Flux Proxy Network
Aka Fluxxy, Darkcloud

Botnet comprised of 30-40K compromised residential IPs, mainly in UA, RU

40-50 bot IPs provisioned per domain

SSL certs deployed by actor on customer’s bot IPs

Criminal customer’s site origin IP

Content delivered
Short lifetime: malware, ransomware
Medium lifetime: phishing
Long lifetime: carding, cybercrime forums

Introduced at Black Hat 2014, Botconf 2014, Defcon 2017
SSL CommonNames and Zbot FF

● Overall Goals:
  ○ Investigate the relationship between CommonNames found on x509 certificates and Zbot FF domains
  ○ Does Zbot FF’s use of SSL ‘leak information’?
SSL - The Bare Minimum

- x509 certificates can give information about ownership of IP space
- Help track how a domain is hosted over a range of IPs/ASNs
- Identify relationships between common names and IP space
Sonar Data

• 2x or 4x monthly scans of IPv4 space minus networks who opt out

• Sonar SSL data contains information that allows us to map an x509 certificate to an IP that hosts the certificate

• Track Sonar Data over a 5 month period
Sonar Data

- Table documents the number of SHAs and CommonNames per month
- Manually inspecting the domains is infeasible

<table>
<thead>
<tr>
<th></th>
<th>Unique SHAs</th>
<th>Unique Common Names</th>
</tr>
</thead>
<tbody>
<tr>
<td>JAN</td>
<td>1,068,402</td>
<td>850,236</td>
</tr>
<tr>
<td>FEB</td>
<td>692,542</td>
<td>589,609</td>
</tr>
<tr>
<td>MARCH</td>
<td>977,484</td>
<td>813,773</td>
</tr>
<tr>
<td>APRIL</td>
<td>249,252</td>
<td>233,834</td>
</tr>
<tr>
<td>MAY</td>
<td>1,098,914</td>
<td>958,321</td>
</tr>
</tbody>
</table>
Objectives

• Provide a series of simple statistical methods that allow a researcher to identify ZBot FF domains using SSL data

• More generally, provide a pipeline to analyze a large data set from a macro and micro view perspective to detect patterns

• All of the data discussed is open source and can be obtained at scans.io/study/sonar.ssl

• Thanks to rapid7 and the University of Michigan
Macro view analysis
Graphs as Representation

- SSL data naturally fits into a bipartite graph

- A bipartite graph is a graph whose vertices can be split into two disjoint sets

- Mapping could have been done differently:
  - IP range instead of ASN
  - ASN gave the best resolution to examine the data
Bipartite Graph - Investigation

- Anomalous behavior requires baseline metrics of normal
- Create a measure of popularity for each CommonName

Requirements:
- Have it based on topological features of the bipartite graph
- Calculate frequencies of what type of ASNs host each CommonName
- Type of ASN refers to the popularity of the ASN
- Popularity of an ASN is defined by how many CommonNames are hosted on it
Graph Analytics - ASNs

- Examine the degree distributions for both sets of the bipartite graph

**ASN degree**
- ASNs with degree 1: 2
- ASNs with degree 2: 3
- ASNs with degree 3: 2
Graph Analytics - ASNs

• Sampled histogram shows popularity frequency of the 22,682 ASNs in the January SSL dataset from a new sample set of 5k

• Long tailed
  • There exists at least one ASN that hosts more than 50,000 unique CommonNames
  • Majority of mass concentrated in the range of ASNs hosting between 1 and 100 domains
Graph Analytics - Domains

• Examine the degree distributions for both sets of the bipartite graph

CN degree

• CNs with degree 1: 2
• CNs with degree 2: 0
• CNs with degree 3: 1
• CNs with degree 4: 1
• CNs with degree 5: 1
Graph Analytics - Domains

• First histogram displays a sampling from the set of 850,236 domains and their corresponding distribution

• The tick marks show the outliers at the tail

• Majority of the domains are hosted on (map to) 1-200 ASNs

• This is visually apparent by zooming into a sampled region of the range 1-140
Graph Analytics - Domains
Graph Analytics - Domains

• As we move towards the right of the graph we ‘gain’ more information about the domain
  • Information regarding how popular it is

• Goal is to identify structures within the graph.

• Structure requires some measure of information

![Graph Analytics Chart](chart.png)
Graph Analytics - Domains

• 831,704 domains map to only a single ASN
  • translates to around 97% of all domains fall in that band

• 99% of domains map to the range between 1-10

• 1-10 too sparse → not enough information
Graph Analytics - Domains

• Inspecting the histogram helps us determine an easy high pass filter
  • Use the mass density of the histogram in relation to information gain

• Examine domains that map to only 10 or more ASNs
  • Analogous to a document containing only one word
  • Allows us to remove around 99% of domains
Micro view analysis
Individual Domain Analysis

• Taken the macro approach so far
  • Examined domain and ASN distribution

New Goal:
• Examine the ASN distribution of two individual domains
  • Histogram shows frequency of different types of ASN that hosts a cert for the given domain
  • X-axis denotes the type of ASN
  • Y-Axis frequency
Individual Domain Analysis

• Noticeable differences between the two domains
• `naranyamarket.com` is found on some extremely populated ASNs
  • One ASN that hosts ~ 25k domains
• `meenyousecu.com` max ASN hosts ~ 1.8k domains
• For `meenyousecu.com` - majority of ASNs host less than 108 unique CNs
Filter Design

- You cannot directly compare histograms
  - Need to be compared on the same scale

- Devise a bucketing scheme that is sensitive to ‘low popularity’ ASNs
  - Increase the number of bands in the lower frequency spectrum
  - lower frequency spectrum refers to ASNs that map to <50 domains

- Intuition is based on the original distribution of ASN frequency counts
  - Interested in domains that are found on many ASNs where those ASNs are not highly popular
  - Majority of the domains hosted on 1-200 ASNs
Filter Design

- Bucket the frequencies into 9 different bands.
  - 1-5
  - 6-10
  - 11-20
  - 21-50
  - etc
    - Distance between the bands increase as we increase the ASN popularity
- Lower frequencies provide higher resolution
- The bucketing process transforms the distribution of ASN popularity counts into a 9-d vector
Filter Design - Histograms to Vectors

Different bands are: [1-5], [6-10], [11,20], [21,50], etc
Anomaly Detection per ASN Count Band

• Refocus on CNs and how many ASNs each domain maps to
Anomaly Detection per ASN Count Band

- Filter out the outliers (dlink, google, etc)

- Bucket domains that are mapped to roughly the same number of ASNs together
  - e.g. [160-150] → [itunes.apple.com, image-glb.qpyou.cn, asos-media.com, etc.]

- For a given count band, e.g. [160-150] calculate the pairwise Euclidean distance between two domains using the domain’s histogram vector
Anomaly Detection per Frequency Band

- Hypothetical distance matrix for a band containing 3 domains

- Each column gives us the distance between a domain and every other domain in that band
  - i.e. red $\rightarrow$ all the distances from d1

$$d(p, q) = d(q, p) = \sqrt{(q_1 - p_1)^2 + (q_2 - p_2)^2 + \cdots + (q_n - p_n)^2}$$

$$= \sqrt{\sum_{i=1}^{n} (q_i - p_i)^2}.$$
Anomaly Detection per Frequency Band

• Calculate the Euclidean norm (EN) per column to determine the similarity of each domain to others

• Ran this algorithm over the Jan dataset.
  • Returned one interesting case in the band 100-110
  • e.g. $\text{EN(tangerine-secure.com)} = 567.003$
Anomaly Detection per Frequency Band

• The histogram clearly shows a significant outlier (more than 2 std away)

The outlier in the high band was ‘tangerine-secure.com’ which we verified as a ZBot Fast flux domain
Anomaly Detection per Frequency Band

- Another interesting band is the 30-40 ASN popularity range
- The distance spectrum is tighter than the 100-110 range
- The tail end of the histogram contains interesting domains
Anomaly Detection - Results

Out of these 5 domains - three are ZBot fast flux (meenyousecu.com, securedatassl.net, secure.tangerineaccess.com)

- Confirmed TPs via PDNS and active probing (of SSL certs)
- Removed FPs with additional signals

This anomaly detection method was able to reduce ~800k domains down to a manageable list of 8 domains
Validating TPs and Removing FPs

- Further signals were incorporated to help delineate between outliers
- Examine the number of unique SHAs associated with a particular commonName
- Examine ASN count over IP count ratio
  - ZBot FF is bounded by the number of IPs the actor provisions to client domains
  - For ZBot FF, on average a client domain gets 1 IP per ASN
- Gave us actionable intelligence regarding which ASNs to monitor more closely
ASNs of ZBot FF vs Popular Domains

• CNs on the ASN popularity band [30,40]
  • secure.tangerineaccess.com
    • AS50161|PE Vasylchyshyn Iurii Oleksandrovyych
    • AS15895|Kyivstar PJSC
    • AS42975|Chiliy Valery Mykhaylovych PE

• blueapron.com
  • AS20940|Akamai International B.V.
  • AS16625|Akamai Technologies, Inc.
  • AS5511|Orange
  • AS14618|Amazon.com, Inc.
Takeaways

• Global structure of SSL graph can inform local behavior of SSL graph
  • We can use the macro knowledge to help us create similarity measures for each micro layer

• We can use this statistical method as a mechanism to cluster anomalous domains in the graph
  • We can reduce the set of domains of interest from ~800k → 8

• Unsupervised method to detect suspicious domains on IP/SSL space

• This method revealed malicious domains in other monthly scans:
  • mobilebanking1.scotiabankcan.com (Feb)
  • datasslsecure.net (Feb)
Our other related work

- Defcon 2017 [https://www.youtube.com/watch?v=AbJCOVLQbjs](https://www.youtube.com/watch?v=AbJCOVLQbjs)
- Black Hat 2017
- Usenix Enigma 2017 [https://www.youtube.com/watch?v=ep2gHQgjYTs&t=818s](https://www.youtube.com/watch?v=ep2gHQgjYTs&t=818s)
- Black Hat 2016 [https://www.youtube.com/watch?v=m9yqnwuqdSk](https://www.youtube.com/watch?v=m9yqnwuqdSk)
- RSA 2016 [https://www.rsaconference.com/events/us16/agenda/sessions/2336/using-large-scale-data-to-provide-attacker](https://www.rsaconference.com/events/us16/agenda/sessions/2336/using-large-scale-data-to-provide-attacker)
- BruCon 2015 [https://www.youtube.com/watch?v=8edBgoHXnwg](https://www.youtube.com/watch?v=8edBgoHXnwg)
- Black Hat 2014 [https://www.youtube.com/watch?v=UG4ZUaWDXSs](https://www.youtube.com/watch?v=UG4ZUaWDXSs)
Thank you

Questions?

We are hiring

Thomas Mathew, thomathe@cisco.com

Dhia Mahjoub, dmahjoub@cisco.com