Threat Hunting Techniques at Scale

Dhia Mahjoub, PhD
Head of Security Research, Cisco Umbrella (OpenDNS)

Tuesday, June 26th, 2018
Agenda

- Intelligence cycle at scale
- Big data challenges
- Spike detection and classification
- Co-occurrences
- Tracking Malspam: combining techniques
- SSL Data mining
- Conclusion
Contributors

Dhia Mahjoub, Head of Security Research, @DhiaLite
PhD graph theory, network security, threat intel

Thomas Mathew, Senior Security Researcher
MS Computer Science, signal analysis, machine learning

Scott Sitar, Technical Leader
MS Applied Math, big data engineering, algorithms

David Rodriguez, Senior Security Researcher
MA Mathematics, statistics, machine learning
Day in the life of a SOC

**Internal Feed:**
- Security controls
  - Firewall, IDS/IPS
  - Other network security
  - Web security/proxy
  - Endpoint security
    (AV, EDR, VPN, etc.)
- Network Infrastructure
  - Routers/switches
  - Domain controllers
  - Wireless, Access pts

**External Feed:**
- Domain ownership
- Relationships with IPs and ASNs
- Passive DNS
- WHOIS record data
- Co-occurrences
- Reputation scores

**Threats** 

**YOU** 

**SIEM** 

**TIP** 

**Threat Intelligence**
Umbrella Investigate Intel Production Cycle

- **Feedback**: Protected customers, actionable use cases with Investigate
- **Requirements**: Detect and block domains/IPs to protect customers and provide insight and context around domains/IPs
- **Collection**: Retrieve raw DNS, IP, BGP, SSL, whois, hash, crawled web data, etc. at scale
- **Dissemination**: Domains, IPs into block list
  - Investigate UI and API
- **Analysis**: Threat detection and reputation scores using ML & graph models, human domain expertise
- **Processing**: Caching, indexing, enriching, summarizing data at scale
What makes us different

Lexical
- Live DGA prediction

Anomaly detection
- Newly seen domains
- Spike rank model

Predictive IP
- Predictive IP space monitoring

Graph-based
- Co-occurrence model

Umbrella
- Investigate

- Botnet
- Crimeware
- Exploit Kit
- Phishing
- Ransomware
- Spam
- Trojan
Malspam/Hancitor

Cybercrime sites

Criminal hosting space
Path of a malspam attack

1. Phishing email sent from delta@performanceair.com

2. Victims click on malicious URLs
   - myhearthstonehomes.org
   - ourrealtyguy.info
   - ourrealtyguy.org
   - ourrealtyguy.us
   - package2china.com

3. Malicious word doc drops Hancitor

4. Hancitor makes C2 call to domains for trojans
   - uneventrendi.com
   - ketofonerof.ru
   - thettertrefbab.ru

5. Trojans (Pony, Evil Pony, Zloader) make C2 call for extra malware or functionality
   - mebelucci.com.ua
   - uneventrendi.com
   - lycasofrep.com
   - rinbetarrab.com

6. Infection on device & positioned for data extraction
Совершенно другое представление о заработке

Im care

UNITED DUMPS PIN CVV ICOCVW 678351190

LadyBINS.com

MAKE DUMPS GREAT AGAIN TRUMP-DUMPS.SU

ProMarket.WS теневой рынок

Курсы товарного кардинга от команды форума ProCrd.

ProCrd.CO - Кардинг форум / Carding forum - Credit Cards - Dumps - Tracks - Bank Accounts
Zbot Fast Flux BPH operation

- Victim
- Crimeware consumer
- Researcher

Actor(s) grow and maintain FF network
*FF service offered in underground forums

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

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
Data Collection
Working with Big Data
Challenges and Uses

● Large datasets give researchers possibility to uncover widespread network threats

● When working with large data sets - traditional threat-hunting methods have to be modified

● We will explore:
  • Pivoting
  • Classification
Working with Big Data

Pivoting

● Pivoting is useful for analysts when given seeds of information

● Want to convert some seed information into further information about a malicious campaign
  • Difficult to connect multiple data sets together
  • Scale of data can make look-ups difficult
Working with Big Data

Classification

- Classification is the most self-explanatory challenge

- Revolves around sifting through a dataset and classifying threats

- Challenges arise when dealing with scale and class-imbalance problem
  - For example, at Cisco Umbrella we produce around 4 TB of hourly data that needs to be processed in near real-time
  - Class-imbalance refers to the percentage of benign to malicious domains that can be found
Case-Studies

- We will be examining some successful use-cases we worked on on dayjob involving classifier design, and designing a platform for pivoting.

- Two datasets:
  - Recursive Layer DNS Traffic
  - IPv4 SSL scans
Cisco Umbrella Datasets
Recursive DNS Data

- 28 data centers worldwide
- ~150 billion queries a day
- Translates to around 24 TB a day
- Valuable client query information
Cisco Umbrella data center locations
Open source Datasets

SSL Data

- SSL data collected from internet wide scan of IPv4 space

- Store and retrieve over 2 TB of total data
  - Scalability
  - Speed
  - Flexibility

- Primary source of data (scans.io)
  - Secondary – active scanning
Processing
Challenges Working at Cisco Umbrella Scale

1. Logs are big:
   - We've peaked at over 150 billion user queries per day and growing

2. Interesting algorithms are slow:
   - Always worse than linear, and sometimes far worse

3. Event horizons move:
   - The data you need to make your new idea work is always one day past the current retention window
Lessons Learned

- Currently somewhere between the 4th and 5th iteration of our systems
- Algorithms are always improving and systems are always getting faster, but...

The most cost effective way to improve your search performance is almost always to reduce the size of your search space
The Internet Is...

Noisy

- People doing scans and scraping expeditions
- Misconfigured search domains
- Infrastructure chattering away
- Low value/inconsequential entities fighting each other
The Internet Is...

Repetitive

- Content distribution networks - everyone needs to download that latest javascript framework
- Low TTLs to aid fault tolerance mean clients need to constantly ask the same questions
- The incessant push to move everything to "the cloud"
The Internet Is...

**Boring** (for the most part)

- Typically, only 5-10% of our raw logs are useful for threat hunting purposes
- Need to strike a balance between cost effective scaling and losing a small amount of signal
Analysis
Spike Detection
Datasets

Goals - Recursive DNS Data

- Can we identify exploit kit and ransomware domains from DNS client traffic?
- Examine traffic logs for possible signals
Datasets
Recursive DNS data

<table>
<thead>
<tr>
<th>Domain</th>
<th>QTYPE</th>
<th>RCODE</th>
<th>Resolvers</th>
<th># of unique IPs</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>0 Resolving</td>
<td>List of resolvers</td>
<td></td>
</tr>
<tr>
<td>1 – A</td>
<td>15 – MX</td>
<td>0 Resolving</td>
<td>List of resolvers</td>
<td></td>
</tr>
</tbody>
</table>

- **QTYPE**
  - 1 – A
  - 15 – MX
  - 28 – AAAA
  - 16 – TXT
  - 99 – SPF
  - 255 – ANY

- **RCODE**
  - 0 Resolving

- **Resolvers**
  - List of resolvers

- **# of unique IPs**
  - 0 Resolving
DNS Features Taxonomy

**Assigned**
- Lexical
- DGA setup
- Hosting
- Registration

**Inherent**
- DNS query trends
- Diversity of clients across geography and IP space
- DNS query volume
- Query types
- Number of querying IPs
- Distribution of queries across resolvers

Harder to obfuscate and change by actors at global scale
Classify domains based on two sets of features:
  • Spike DNS data
  • Historical query volume patterns

Spike DNS data
  • Qtype distributions
  • Resolver distributions

Historical query data:
  • Volatility
  • Sparsity
Classification
Recursive DNS Data
Spike Detection pipeline

Recursive DNS Data

- Spike detection
- Domain History Filter
- Domain Records Filter
- Exploit kits
  - Phishing
  - DGA
  - Fast flux CnCs
  - Fake software
- Random Forest

Training sets

Extra malware domains and IP ranges

Pivot around IPs, prefixes, ASNs, hoster, registrant to enrich Technical Intelligence
Classification
Random Forests

- **Use random forest for classification step**
  - Random forests parallelize easily
  - Non-parametric model
  - Handle non-linear boundaries

- Feed in spike domain feature vectors hourly to classify spiked domains

- Out-of-bag error at 3%
Classification

Results

- Two types of domains caught:
  - Dedicated
  - Compromised

- Have different time series for the two classes
- Compromised domains more difficult to detect due to the presence of additional noise

- ‘nowupdate4free.thelinkersgoodfreeforcontentall.date’
- ‘vancouverwashingtonpersonaltraining.com’
Classification

Results

- Layering of signals post initial classification
- Good method for identifying compromised domains
Classification

Results

- Use available WHOIS data to pivot
  - Search through other registrant domains
  - Look for similar signal patterns
- Pivoting done through a combination of manual and automated work
- A whole set of compromised domains are found
  - Able to block these compromised domains ahead of going live
Detected Threats

- Exploit kits
- DGA
- Phishing
- Malware C2s
- Ransomware
Track Malspam; Combining techniques
Path of a malspam attack

1. Phishing email sent from delta@performanceair.com
2. Victims click on malicious URLs
   - myhearthstonehomes.org
   - ourrealtyguy.info
   - ourrealtyguy.org
   - ourrealtyguy.us
   - package2china.com
3. Malicious word doc drops Hancitor
4. Hancitor makes C2 call to domains for trojans
   - uneventrendi.com
   - ketofonerof.ru
   - thetettertrefbab.ru
5. Trojans (Pony, Evil Pony, Zloader) make C2 call for extra malware or functionality
   - mebelucci.com.ua
   - uneventrendi.com
   - lycasofrep.com
   - rinbetarrab.com
6. Infection on device & positioned for data extraction
Malicious malspam campaign

Dear client,

Your order has been processed and your credit card has been charged. Please download and print your ticket by clicking here.

Please find your order details below.

FLIGHT NUMBER : DT3547136446U3
ORDER# : DELTA64377537
CARD NUMBER : 4XXX-XXXX-XXXX-5741
CARD TYPE : VISA
AMOUNT CHARGED : 950.50

For more information regarding your order, contact us by visiting http://www.delta.com.

Thank you for flying with us
Delta Airlines

Reference hazmalware.wordpress.com
performanceair.com
Spoofed email used in mailspam attack
August 30: Peak of malicious redirect
Duration: 7 hour period
Attack took place between 14:00-21:00 UTC
## Insight into the IP network

**myhearthstonehomes.org**

### IP Addresses

<table>
<thead>
<tr>
<th>First seen</th>
<th>Last seen</th>
<th>IPs</th>
<th>TTL</th>
</tr>
</thead>
<tbody>
<tr>
<td>9/14/17</td>
<td>9/14/17</td>
<td>184.168.221.49</td>
<td></td>
</tr>
<tr>
<td>8/31/17</td>
<td>9/13/17</td>
<td>184.168.221.49</td>
<td>600</td>
</tr>
<tr>
<td><strong>8/30/17</strong></td>
<td><strong>8/30/17</strong></td>
<td><strong>52.14.244.225</strong></td>
<td><strong>600</strong></td>
</tr>
</tbody>
</table>
Known malicious domains on the same IP

Known domains hosted by 52.14.244.225

agentssellingtips.info antoineandmuse.com apadriana.com brookestonehousevalue.info centralfhousevalue.info

**heymamaradio.com** imap.antoineandmuse.com imap.centralfhousevalue.info imap.vetstuff.com myoutdoorchild.com

rexahunter.com susannahope.com thechristianblog.com verumpharmaceuticals.com whymovenow.info writerbloggers.com

greathomesellingtips.info newwesorangehomes.info package2china.com realestatetruth.info vetstuff.com

wgopodcastbooking.com writerblogger.com www.agentssellingtips.info zasbiopharmaceuticals.com zasproperties.com

zasbiopharm.com zashealthsystems.com zasholdings.com zashealth.com lovelyfirealestate.com ourrealtyguy.org


myhearthstonehomes.info myhearthstonehomes.net myhearthstonehomes.org ourrealtyguy.info ourrealtyguy.net

ourrealtyguy.us www.myhearthstonehomes.info www.ourrealtyguy.org
This domain is associated with the following attack: Hancitor Dropper

This domain has a suspicious prefix score

This domain has a suspicious RIP score

Classifier prediction: suspicious

Umbrella risk score: -83

DNS queries

2k

1k

DNS Queries/hour

WHOIS information of myhearthstonehomes.org

<table>
<thead>
<tr>
<th>Email Address</th>
<th>Associated Domains</th>
<th>Email Type</th>
<th>Last Observed</th>
</tr>
</thead>
<tbody>
<tr>
<td><a href="mailto:john@liveingarnetvalley.net">john@liveingarnetvalley.net</a></td>
<td>17 Total - 7 malicious</td>
<td>Administrative, Registrant, Technical</td>
<td>Current</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Nameserver</th>
<th>Associated Domains</th>
<th>Last Observed</th>
</tr>
</thead>
<tbody>
<tr>
<td>ns70.domaincontrol.com</td>
<td>Greater than 500 Total</td>
<td>Current</td>
</tr>
<tr>
<td>ns69.domaincontrol.com</td>
<td>Greater than 500 Total</td>
<td>Current</td>
</tr>
</tbody>
</table>
## Domains Associated with john@liveingarnetvalley.net

<table>
<thead>
<tr>
<th>Domain Name</th>
<th>Security Categories</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>myhearthstonehomes.info</td>
<td>Malware</td>
<td></td>
<td></td>
</tr>
<tr>
<td>myhearthstonehomes.net</td>
<td>Malware</td>
<td></td>
<td></td>
</tr>
<tr>
<td>myhearthstonehomes.org</td>
<td>Malware</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ourrealtyguy.info</td>
<td>Malware</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ourrealtyguy.net</td>
<td>Malware</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ourrealtyguy.org</td>
<td>Malware</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ourrealtyguy.us</td>
<td>Malware</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Details for ourrealtyguy.info

This domain is currently in the Umbrella block list

This domain is associated with the following attack: Locky Ransomware

DNS queries

<table>
<thead>
<tr>
<th>DNS queries/hour</th>
</tr>
</thead>
<tbody>
<tr>
<td>1,500</td>
</tr>
<tr>
<td>1,000</td>
</tr>
<tr>
<td>500</td>
</tr>
</tbody>
</table>

Co-occurring domains tied to the same malspam campaign

myhearthstonehomes.org

Co-occurrences

www.delta.com (18)  a1.verisigndns.com (14)  performanceair.com (11)  a3.verisigndns.com (10)
 a.dnspod.com (10)  a2.verisigndns.com (10)  b.dnspod.com (9)  c.dnspod.com (9)  mx00.1and1.com (4)
 mx01.1and1.com (4)  myhearthstonehomes.net (4)  ourrealtyguy.net (4)  ourrealtyguy.org (3)
Co-occurrences

- Find domains queried by clients in close temporal proximity
- Data is one hour of querylog traffic - 2TB of raw data
- Identify domains looked up by same clients within one minute window of one another
- Output {domain: [List of Domains]}
- Example:
  
  100luimg.361lu.com. -> {"ucsec1.ucweb.com":3.0,"d2.avgc.us":3.0,"home.1100lu.info":4.0}
Co-occurrences

- The closer in time, the higher the co-occurrence score
- The more clients exhibiting this behavior, the higher the score
Co-occurrences

- Domains having similar topic, e.g. security sites, hacking, carding sites
  - Visited by users with related interest
- Example: first.org

Co-occurrences

nakedsecurity.sophos.com (92.14) www.bleepingcomputer.com (7.86)

- Botnet CnC domains, e.g. DGAs
- Infection chains: compromised sites -> Exploit kit landing domains
Scaling Up co-occurrence detection algorithms

<table>
<thead>
<tr>
<th>Before</th>
<th>After</th>
</tr>
</thead>
<tbody>
<tr>
<td>job ran daily</td>
<td>job runs hourly</td>
</tr>
<tr>
<td>job used heavily sampled logs</td>
<td>no sampling apart from initial &quot;algorithm relevant&quot; cleaning</td>
</tr>
<tr>
<td>heuristics used to further cut down data size to help catch initial compromise/infection</td>
<td>no further data size reduction necessary</td>
</tr>
</tbody>
</table>
Path of malspam attack

1. Phishing email sent from delta@performanceair.com
2. Victims click on malicious URLs
3. Malicious word doc drops Hancitor
4. Hancitor makes C2 call to domains for trojans
   - uneventrendi.com
   - ketofonerof.ru
   - thettertrefbab.ru
5. Trojans (Pony, Evil Pony, Zloader) make C2 call for extra malware or functionality
   - mebelucci.com.ua
   - uneventrendi.com
   - lycasofrep.com
   - rinbetarrab.com
6. Infection on device & positioned for data extraction

Newly seen domains:
- myhearthstonehomes.org
- ourrealtyguy.info
- ourrealtyguy.org
- ourrealtyguy.us
- package2china.com
Newly SeenDomains: A real-time Stream

Known Domains

Unknown Domains
Overview of Newly Seen Domains System
Clear Patterns Emerge
Cybercrime sites
For every 1 hour of traffic, we define:

- Chattiness: # unique domains a machine queries
- Popularity: # unique machines that queried the domain

- amplify dom domain chattiness popularity nbhourspast level
- Pivot through domains and machines by keeping a threshold of chattiness and popularity
Amplify through signals using seeds

Pivot from procrd.co -> other crimeware sites
amplify.sh dom domain chattiness popularity nbhourspast level
amplify.sh dom procrd.co 10 10 120 2

Carding/dump shops: carder007.org, carder.site, cardx.biz, mastercvv.in, trump-dumps.ru
Crimeware forums: fuckav.ru
Jabber/chat servers used by criminals: jabber.ru, blah.im
Anonymous, vpn, proxy, socks: doublevpn.com, hidemevpn.de, vpmmonster.ru, hidevpn.me
Stolen accounts, shell, RDP: dedicrdp.ru
Criminal hosting space
Zbot Fast Flux BPH operation

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

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 Data mining
SSL
Goals

- What questions will we ask?
  - How can we connect domain, IP data with SSL

- IPv4 SSL scans: 2TB of data - a few million IPs and a few million SHAs
SSL Backend Architecture - Data Format

- Two separate data formats
  - x509 certificate
  - IP $\leftrightarrow$ SSL SHA mappings

- Require different forms of indexing
  - Document store
  - Key/Value store

- Communication between data stores
SSL Backend Architecture

Data Challenges

- Different data stores for different data types
  - Documents don’t store well in traditional RDBMS
  - X509 certs are sparse documents

- Combination of big data technologies
  - HBASE
  - ElasticSearch
SSL Backend Architecture

API Endpoint

HBase

Elastic Search

x.509

Doc
SSL Backend Architecture
HBase and Elastic Search Table Design

- 4 HBase major components:
  - Rowkey
  - Column Family
  - Column Qualifier
  - Cell

- Design of the rowkey most important
  - Use RK to match SHAs to IP spaces

- Elasticsearch stores parsed x509 docs
  - Indexed on fields; e.g. search on CNs
Reveal origin IP of domains hiding behind reverse proxies

- darkmoney.cc, a cybercrime forum hides behind Cloudflare
  
  darkmoney.cc. 299 IN A 104.31.1.166  
  darkmoney.cc. 299 IN A 104.31.0.166

- Search CN=darkmoney.cc in SSL data base
  
  33b64e11a6e8529d9b719bf9e91bf8b9fd0ad6fa,darkmoney.cc

- Search the sha in SSL data base
  
  33b64e11a6e8529d9b719bf9e91bf8b9fd0ad6fa 2016-06-27 181.174.164.101

- Confirm content is hosted on the hidden IP
  
Dealing with large scale threat intel problems, you need to:

- Know your requirements: what are you looking for?
- Know what to collect
- Know how to store and process the data at scale
- Know what analysis to apply: human or machine based at scale or a combination
- What is your final product: discrete IOCs, or trends and TTPs
Some of our related work

- Hack in the Box 2018 [https://youtu.be/co2cvi_5Flc](https://youtu.be/co2cvi_5Flc)
- Virus Bulletin 2017 [https://www.youtube.com/watch?v=sbzvZ8ChTiU](https://www.youtube.com/watch?v=sbzvZ8ChTiU)
- Defcon 2017 [https://www.youtube.com/watch?v=AbJCOVLQbjs](https://www.youtube.com/watch?v=AbJCOVLQbjs)
- Usenix Enigma 2017 [https://www.youtube.com/watch?v=ep2gHQqiYTs&t=818s](https://www.youtube.com/watch?v=ep2gHQqiYTs&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
dhia@opendns.com

@DhiaLite