When threat hunting fails
Identifying malvertising domains using lexical clustering

Tucson, January 9th, 2018
Authors

Matt Foley
David Rodriguez
Dhia Mahjoub
Agenda

- Background
- Ad Network Profiling and Filtering
- Lexical Clustering
- Hosting space and top talkers
Background
Exploit Kits

Step 1.

Victim

Malvertising

Compromised Site

Ad Net. Publisher

Staged Site (Ad)

Gets lander (proxy)

EK Server
What is Malvertising

Visitors -> Ad Networks -> Ad Exchanges -> DSPs -> Ad Agencies

Publishers

Ad Servers
Ad Campaign Flow

User visits publisher site

Publisher site includes ad network javascript

Compromised Ad Net.

Examples:
- Tech support scam
- Rig Exploit Kit
- Fake flash/java update

Ad network fingerprints and sends user to malvertisement
<table>
<thead>
<tr>
<th>Date/Time</th>
<th>Dst</th>
<th>port</th>
<th>Host</th>
<th>Info</th>
</tr>
</thead>
<tbody>
<tr>
<td>2017-06-20 14:23:04</td>
<td>80.77.82.41</td>
<td>80</td>
<td>observice.info</td>
<td>GET /banners/uaps HTTP/1.1</td>
</tr>
<tr>
<td>2017-06-20 14:28:04</td>
<td>80.85.158.121</td>
<td>80</td>
<td>80.85.158.121</td>
<td>POST /Reg&amp;man=1617&amp;vvan=xHrQMrPybR3FFYbfKP_EUKFEMuWA0SKwYzHaGF5qxFTGpbH0s</td>
</tr>
<tr>
<td>2017-06-20 14:29:06</td>
<td>80.85.158.121</td>
<td>80</td>
<td>80.85.158.121</td>
<td>GET /olosfliet=4546&amp;man=4796&amp;vvan=xX3MvWybRXQCJ3EKv_cT6GMVHRCJL2YydmHCW</td>
</tr>
<tr>
<td>2017-06-20 14:29:07</td>
<td>80.85.158.121</td>
<td>80</td>
<td>80.85.158.121</td>
<td>GET /Mon&amp;flight=3607&amp;traveling=K3jx0ALQFozYdZBlxF9Pj7jELUYx-e1ZB-UPfYQ5Hr</td>
</tr>
<tr>
<td>2017-06-20 14:29:13</td>
<td>80.77.82.41</td>
<td>80</td>
<td>observice.info</td>
<td>GET /banners/uaps HTTP/1.1</td>
</tr>
<tr>
<td>2017-06-20 14:28:14</td>
<td>80.85.158.121</td>
<td>80</td>
<td>80.85.158.121</td>
<td>POST /vel&amp;man=2874&amp;traveling=SwZmnIw0AF5A9K2r2OKEnBTI1JW9RaPaALG-pbBE7J52</td>
</tr>
<tr>
<td>2017-06-20 14:28:16</td>
<td>80.85.158.121</td>
<td>80</td>
<td>80.85.158.121</td>
<td>GET /Reg&amp;man=163&amp;traveling=ThiRbSKAImdttfAw9H9P2qhkLXwBwfhcOF_heNYwLg-8CRR</td>
</tr>
<tr>
<td>2017-06-20 14:28:17</td>
<td>80.85.158.121</td>
<td>80</td>
<td>80.85.158.121</td>
<td>GET /Mon&amp;flight=3532&amp;man=3839&amp;traveling=ThiKbSKAZimdttfAw9H8_2qhkPXwBWYhcOF</td>
</tr>
<tr>
<td>2017-06-20 14:29:02</td>
<td>47.91.121.220</td>
<td>80</td>
<td>multifest.bit</td>
<td>POST / HTTP/1.0</td>
</tr>
<tr>
<td>2017-06-20 14:29:05</td>
<td>47.91.121.220</td>
<td>80</td>
<td>multifest.bit</td>
<td>POST /com/ HTTP/1.0</td>
</tr>
<tr>
<td>2017-06-20 14:29:07</td>
<td>47.91.121.220</td>
<td>80</td>
<td>multifest.bit</td>
<td>POST /com/ HTTP/1.0</td>
</tr>
<tr>
<td>2017-06-20 14:29:11</td>
<td>47.91.121.220</td>
<td>80</td>
<td>multifest.bit</td>
<td>POST /com/ HTTP/1.0</td>
</tr>
<tr>
<td>2017-06-20 14:29:14</td>
<td>47.91.121.220</td>
<td>80</td>
<td>multifest.bit</td>
<td>POST /com/ HTTP/1.0</td>
</tr>
<tr>
<td>2017-06-20 14:29:16</td>
<td>47.91.121.220</td>
<td>80</td>
<td>multifest.bit</td>
<td>POST /com/ HTTP/1.0</td>
</tr>
<tr>
<td>2017-06-20 14:29:20</td>
<td>47.91.121.220</td>
<td>80</td>
<td>multifest.bit</td>
<td>POST /com/ HTTP/1.0</td>
</tr>
<tr>
<td>2017-06-20 14:29:24</td>
<td>47.91.121.220</td>
<td>80</td>
<td>multifest.bit</td>
<td>POST /com/ HTTP/1.0</td>
</tr>
<tr>
<td>2017-06-20 14:29:27</td>
<td>47.91.121.220</td>
<td>80</td>
<td>multifest.bit</td>
<td>POST /com/ HTTP/1.0</td>
</tr>
<tr>
<td>2017-06-20 14:29:29</td>
<td>47.91.121.220</td>
<td>80</td>
<td>multifest.bit</td>
<td>POST /com/ HTTP/1.0</td>
</tr>
<tr>
<td>2017-06-20 14:29:40</td>
<td>47.91.121.220</td>
<td>80</td>
<td>multifest.bit</td>
<td>POST /com/ HTTP/1.0</td>
</tr>
<tr>
<td>2017-06-20 14:30:04</td>
<td>47.91.121.220</td>
<td>80</td>
<td>multifest.bit</td>
<td>POST / HTTP/1.0</td>
</tr>
</tbody>
</table>
Tech Support Scams
Fake Flash and Java Updates
Ad Network Profiling and Filtering
Proxy Network

Squid Proxy

Choice of region

Rotating IPs
Filtering on non-residential IP Address

Browsing with DigitalOcean proxy

GET

Ad Network

Ad Network Returns a 403

403
Attempts with other VPS providers
Advanced Proxy Check

The following lists several of the test results that we perform to attempt to detect a proxy server. Some tests may result in a false positive for situations where the IP being tested is a network sharing device. In some situations a proxy server is the normal circumstance (AOL users and users in some countries).

Thank you for participating in our test of detecting proxy servers. This proxy detector is constantly being updated. If you are using a proxy server and it was not detected please check back in a few days and see if we are able to detect the proxy server.

To test a different IP address please use the IP lookup tool.

VPN leaking your REAL IP address? Try our VPN Leak test.

Proxy server not detected.

<table>
<thead>
<tr>
<th>IP</th>
<th>104.238.137.26</th>
</tr>
</thead>
<tbody>
<tr>
<td>rDNS</td>
<td>FALSE</td>
</tr>
<tr>
<td>WIMIA Test</td>
<td>FALSE</td>
</tr>
<tr>
<td>Tor Test</td>
<td>FALSE</td>
</tr>
<tr>
<td>Loc Test</td>
<td>FALSE</td>
</tr>
<tr>
<td>Header Test</td>
<td>FALSE</td>
</tr>
<tr>
<td>DNSBL Test</td>
<td>FALSE</td>
</tr>
</tbody>
</table>
Anonymous Proxy detected, click here.
YOU ARE SEEING THIS PAGE, BECAUSE YOU ARE USING AN ANONYMOUS PROXY.

IN ORDER TO PROTECT OUR ADVERTISERS, WE DO NOT DISPLAY REAL ADS TO THE WEBSITES OWNERS. THE REAL ADS WILL BE DISPLAYED TO YOUR REGULAR TRAFFIC AND WILL BE PAID.
Lexical Clustering
Attention to Details
Fake Flash and Java Updates
<table>
<thead>
<tr>
<th>Domain Name</th>
<th>Security Category</th>
<th>First Seen</th>
</tr>
</thead>
<tbody>
<tr>
<td>contentfreeandsafe2updating.stream</td>
<td>Newly Seen</td>
<td>December 13, 2017, 3:17pm</td>
</tr>
<tr>
<td>contentfreeandsafetoupdating.review</td>
<td>Newly Seen</td>
<td>December 13, 2017, 3:09pm</td>
</tr>
<tr>
<td>contentfreeandsafe4updating.date</td>
<td>Newly Seen</td>
<td>December 13, 2017, 3:00pm</td>
</tr>
<tr>
<td>contentfreeandsafeupdatesgreat.win</td>
<td>Newly Seen</td>
<td>December 13, 2017, 2:18pm</td>
</tr>
<tr>
<td>contentfreeandsafeupdatingsnew.win</td>
<td>Newly Seen</td>
<td>December 13, 2017, 11:27am</td>
</tr>
<tr>
<td>contentfreeandsafetoupgrade.stream</td>
<td></td>
<td>December 13, 2017, 11:16am</td>
</tr>
<tr>
<td>contentfreeandsafe4upgrading.download</td>
<td></td>
<td>December 13, 2017, 10:39am</td>
</tr>
</tbody>
</table>
More or Less Traveled Roads
Consider the almighty RegeX Keywords

Known Keywords
- content
- free
- apple

UnKnown Keywords
- safe
- build
- click

Synonyms
- Typos
Consider the almighty RegeX

grep "*.fake.*"
Traffic Pattern of Fake Update Sites

contentfreeandsafe4update.bid

DNS queries

contentfreeandsafe2update.date

DNS queries
Traffic Pattern of Fake Update Sites

Look for burst in traffic
For one word, many
Shingling Fake Flash and Java Update

Trigram host name

contentfreeandsafe4update

Shingling Fake Flash and Java Update

Trigram host name

contentfreeandsafe4update


MinHash

LSH
Locality Sensitive Hashing Fake Flash

3 Domains with a lot of shingles in common

contentfreeandsafe4update
contentfreeandforupdate
content4freeandsafeupdate

c on  t en t  f re a nd s af  d at
On to production
Clustering Pipeline Realtime/Batch

goodnewcontentssafe.download

pipeline

hasher

Count min-sketch

Out pipeline

Cluster DB

Analyst Dashboard
Payday
Fake Flash and Java Update Lexical Clustering

cluster_1:
goodnewcontentssafe.download
goodnewfreecontentsload.date
goodnewfreecontentall.trade
...

cluster_3:
artificialintelligencesweden.se
artificialintelligencechip.com
artificialintelligence.net.cm
...

cluster_2:
call-microsofntnw-err81711102.win
call-microsofntnw-err99817109.win
call-microsofntnw-err81711101.win
...

cluster_4:
mkto-sj220048.com
mkto-sj220146.com
mkto-sj220162.com
...
We need help
for j, domain in enumerate(entry['domains']):
    entry['domains'][j] = {'domain': domain, 'timestamp': entry['timestamp']}
    entry.pop('timestamps')

for i, idx in enumerate(date_changes):
    n = date_changes[idx]  # i.e., n = [1, 3, 5]
    if len(date_changes) == 1 else None
    r[idx:n] = sorted(r[idx:n], key=lambda x: x['c_num'])

@app.route("/clusters/attribute", methods=['POST'])
def attribution():
    if not request.json:
        return "Error!"
        res = {}
    for cluster_id in request.json:
        attr = request.json[cluster_id]
        ret = add_attribute(cluster_id, attr)
        resp[cluster_id] = ret
    if ret == 'success' and BLOCKING:
        domains = m.get_cluster_domains(cluster_id)["domains"]
        block_description = "Domain showed similarities to {0} malware"
        print "Blocking domains: {0}".format("", ','.join(domains))
        block(domains, block_description=block_description)
    return jsonify(resp)

@app.route("/clusters/attribute/<string:cluster_id>")
def get_attribute(cluster_id):
    return jsonify(m.get_attribute(cluster_id))

@app.route("/clusters/uncategorized")
def get_uncategorized():
    r = [entry for entry in m.get_uncategorized()]
    if not r or len(r) == 1:
        return jsonify(results=r)
Hosting space and top talkers
Where are these hosted? Any patterns?

- Take 1 week’s worth of detections and their hosting space; Jan 1-7

- Some hosters are consistently abused
  
  AS12876, FR  
  AS14618 Amazon AWS and more  
  Some IPs are actively hosting thousands of domains for months

- Some hosters are highly infested with shady, toxic content; dedicated?
  
  AS202023, LLHOST, RO; phishing, tech support scams, fake updates, porn
Who is querying these domains?

- Take 1 week’s worth of detections; Jan 1-7 and user IPs

- 10 busiest hours

  20000+ user IPs querying 2000+ malvertising domains

- Some top talker clusters emerge

  Security companies owned ranges querying hundreds of domains
  Some rogue networks querying hundreds of domains
grep "*.fake.*"

Look for burst in traffic

<table>
<thead>
<tr>
<th>Domain Name</th>
<th>Security Category</th>
<th>First Seen</th>
</tr>
</thead>
<tbody>
<tr>
<td>contentfreeandsafe5_updating.stream</td>
<td>Newly Seen Co...</td>
<td>December 13, 2017, 3:17pm</td>
</tr>
<tr>
<td>contentfreeandsafe6_updating.mview</td>
<td>Newly Seen Co...</td>
<td>December 13, 2017, 3:09pm</td>
</tr>
<tr>
<td>contentfreeandsafe4_updating.date</td>
<td>Newly Seen Co...</td>
<td>December 13, 2017, 3:00pm</td>
</tr>
<tr>
<td>contentfreeandsafe5_updating.user.win</td>
<td>Newly Seen Co...</td>
<td>December 13, 2017, 2:19pm</td>
</tr>
<tr>
<td>contentfreeandsafe6_updating.user.exe</td>
<td></td>
<td>December 13, 2017, 11:27am</td>
</tr>
<tr>
<td>contentfreeandsafe5_updating.exe.stream</td>
<td></td>
<td>December 13, 2017, 11:16am</td>
</tr>
<tr>
<td>contentfreeandsafe6_updating.download</td>
<td></td>
<td>December 13, 2017, 10:28am</td>
</tr>
</tbody>
</table>

user IPs

hosting IPs
Current and Future Work

- NLP on misspellings and common typos
- Models to categorize clusters
- Identifying malicious file hosts using belief propagation
Thank you

Questions?

We are hiring

Matt Foley, matfoley@cisco.com

David Rodriguez, davrodr3@cisco.com

Dhia Mahjoub, dmahjoub@cisco.com