RSAConference2017
San Francisco | February 13 – 17 | Moscone Center

POWER OF OPPORTUNITY

SESSION ID: AIR-R02F

One-hit Wonders: Dealing With Millions of Anomalies

Chris Larsen
Architect/Researcher, WebPulse Threat Research Lab
Symantec
@bc_malware_guy

Outline/Agenda

- Background
  - Dealing with Big Data (spoiler alert: you need AI/ML)
  - Machine Learning and Anomaly Detection
  - Problem: Lots of Anomalies!
  - The State of Threat Intelligence
- Research
  - Large-scale Anomalies: One-hit Wonders and One-day Wonders
  - Mining those Anomalies for Threats
- Recommendations

Nostalgia Time

- In my first-ever RSA talk (2010), I had a great slide to visually depict the Internet…
- …and I realized I could re-purpose it to show “Big Data”

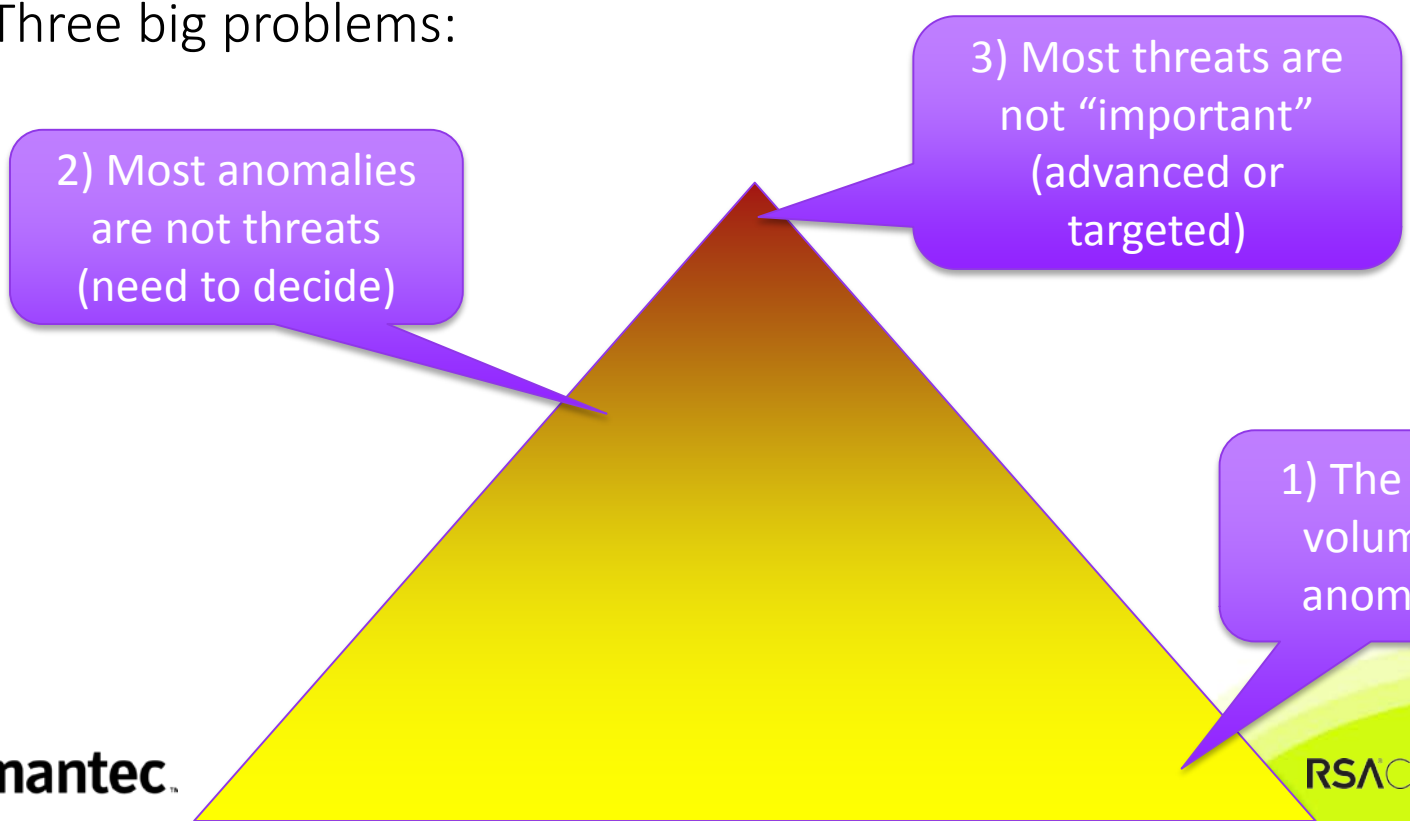

TOXIC
WASTE

Dealing with Big Data

- Security industry “discovered” Big Data several years ago
- Follow-up “discovery” was the need for AI/ML
  - (by definition, humans can’t deal with that much data)
- A classic (and important) basic use for AI/ML: anomaly detection
  - Machine Learning
    - System scans your log/traffic data, learns “normal”
    - Voila! It can then spot anomalies

Problem: Large-scale Anomalies

- Three big problems:



The State of Threat Intelligence

- Big push on this at conferences 3-4 years ago
  - "You need this! The more, the better!"
    - (Internal, OSINT, and Commercial)
  - Standard formats (e.g., STIX, TAXI) would make it easy
- Starting a couple years ago: conference feedback/pushback talks
  - Some TI was really good (and STIX was cool)
    - Good TI tended to be what I would call "boutique TI"
      - In-depth TI on specific attacks/groups/IOCs (but tended to lag…)
  - Most was large volume / low value (lacked context)
    - "Here's a bunch of bad IPs. Block them."

RSAConference2017

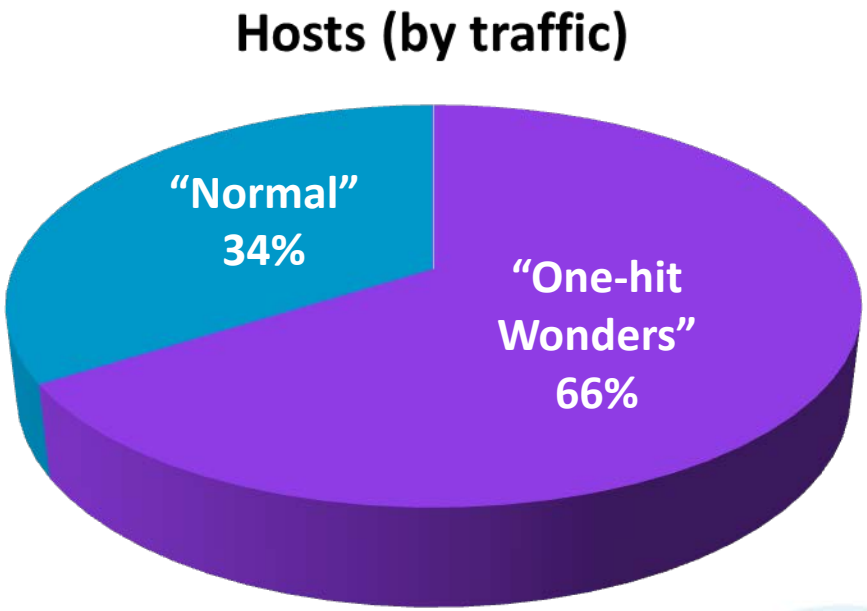
Background Completed

Time to see some research:

“One-hit Wonders”

One-hit Wonders

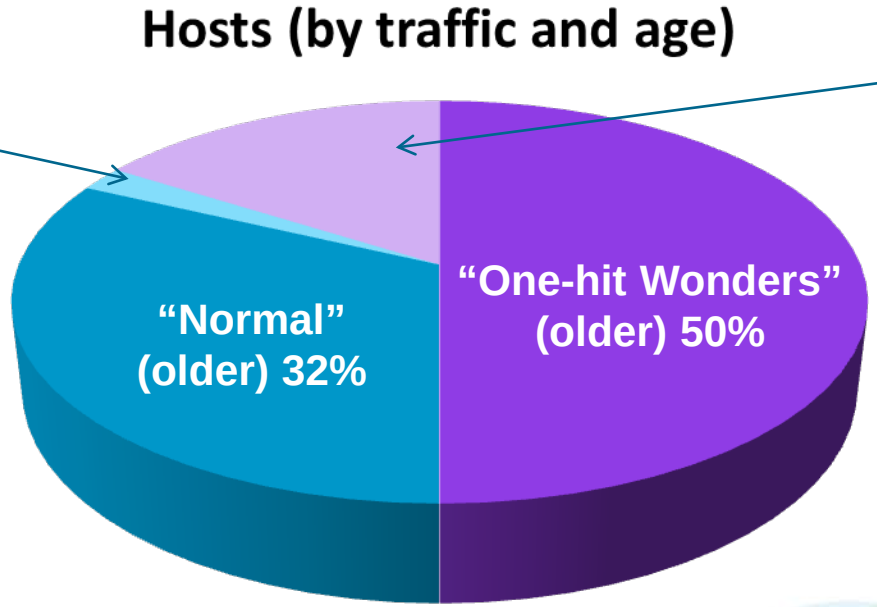
- Start with the “most interesting” part of the Web…
  - …in a 24 hour period: over 6 million active hosts*



* Host: unique Domain, Subdomain, or IP address (the base of the URL)

One-hit Wonders: A Closer Look

- **Next step: look at the age of the hosts…**
  - (the light color parts are the New ones in that day's traffic)

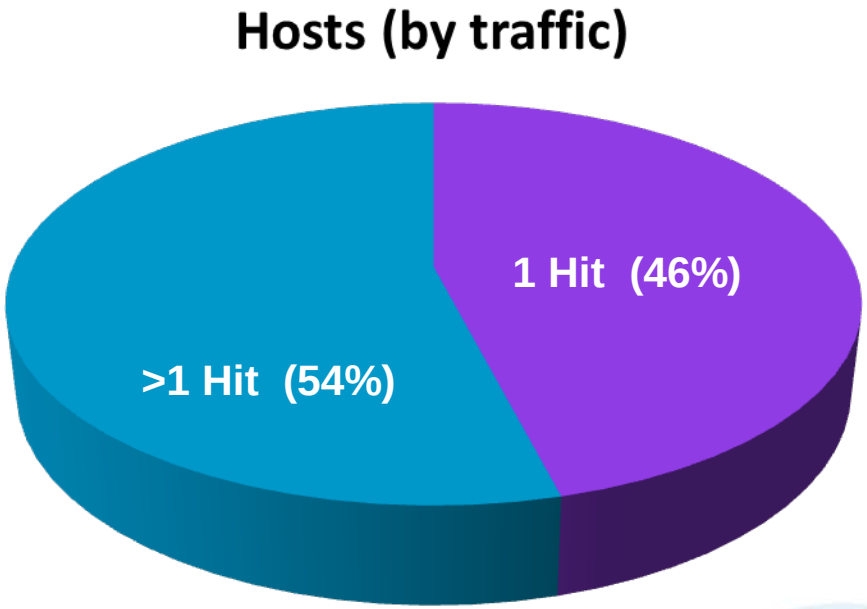


Very few normal-traffic hosts were New… (2% overall)

…but 1/4 of one-hit wonders were New! (16% overall)

Once that big AI system got built…

- Recent look: Out of almost 25M “interesting things”
  - (if 2-hitters are counted as OHWs, it’s close to original %)



So what are they?

- **Shady subdomains (spam networks, exploit kits, etc.)**
- **Ultra-low-traffic blogs (e.g., Tumblr, Blogspot, Wordpress)**
- **A lot of IP addresses**
  - (more at the new/OHW end of the spectrum, fewer at the older/Normal)
- **Botnet/DGA names…**
- **Errors (invalid domains, other non-resolving junk)**
- **And many normal-looking domains…**
  - ...there are just a lot of low-traffic legitimate sites out there

RSAConference2017

More Interesting Research…

“One-day Wonders”

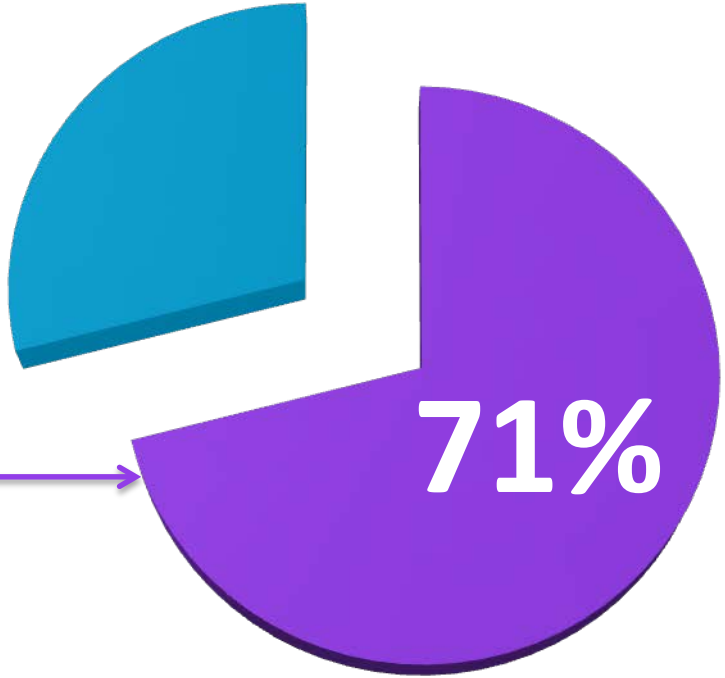
Deciding what “normal” looks like

- Researching traffic levels for sites…
  - How much daily traffic for “big” sites?
    - Google, Facebook, Twitter, Youtube, Baidu, etc. will have a **lot!** (every day!)
  - But how much daily traffic for other sites?
    - (it’s like our version of Alexa…)
- We looked at 90 days of all our traffic
  - (whether already in our main database or not)
- Over 660 million unique hosts* showed up at least once…

* Host: unique Domain, Subdomain, or IP address (the base of the URL)

One-day Wonders

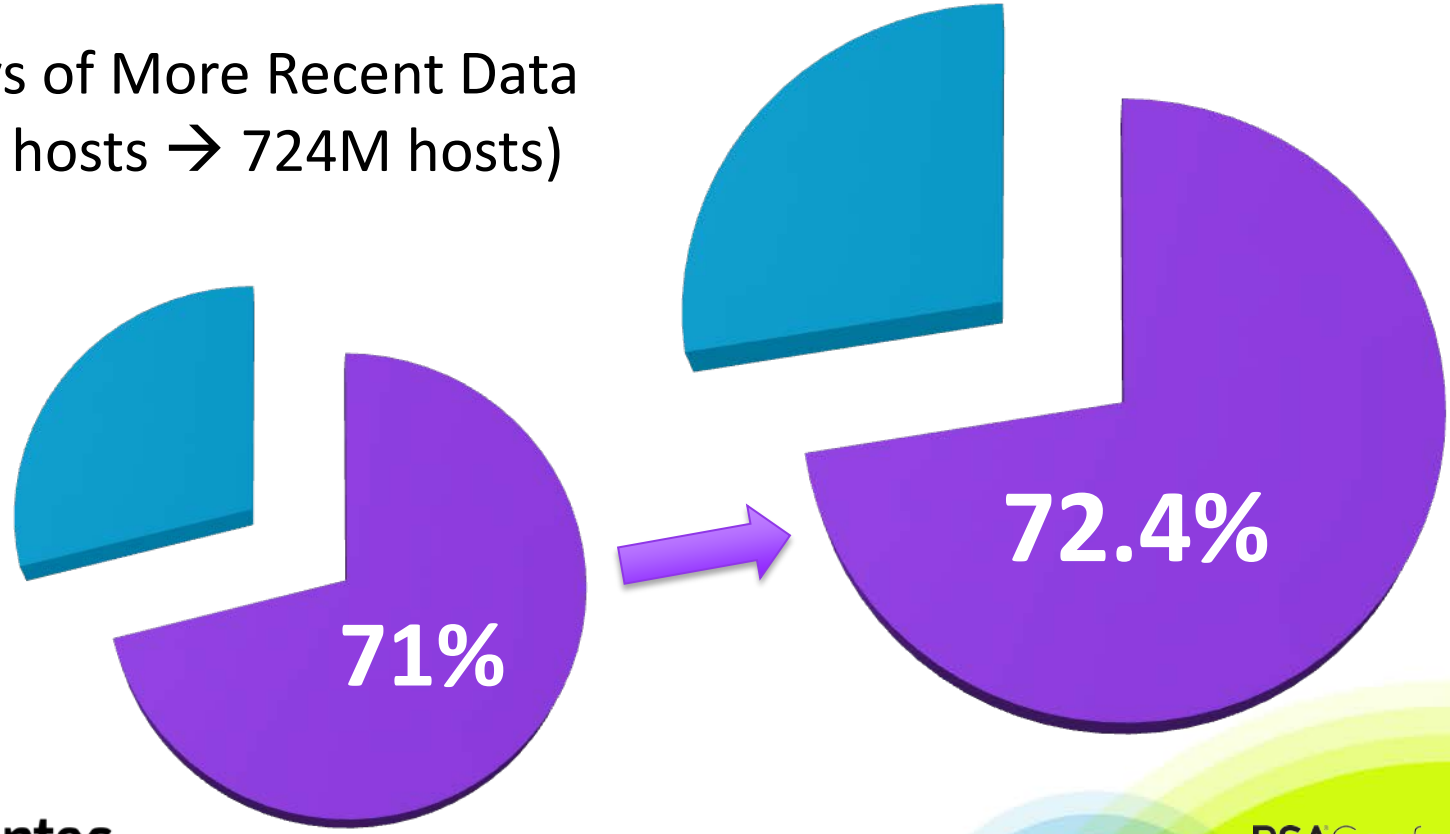
Of 660 Million Unique Hosts…



…470 Million existed 24 hours or less

One-day Wonders: The Pattern Continues

90 Days of More Recent Data
(660M hosts → 724M hosts)



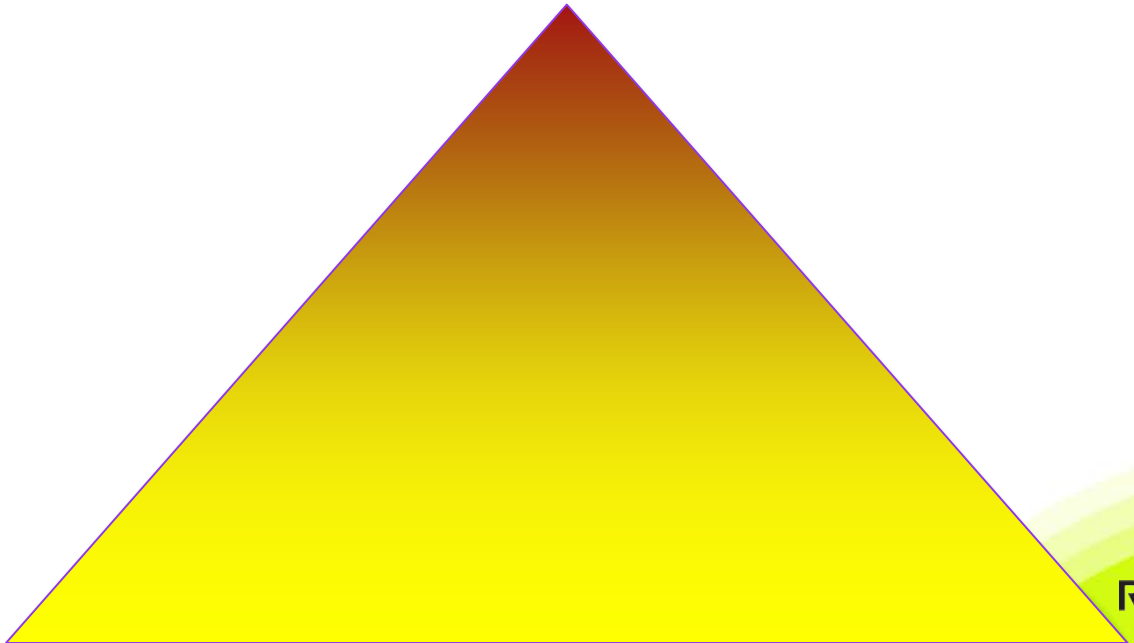
BTW,
~ 80%
of
ODWs
are
OHWs

So what are they?

- **CDNs using unique subdomain codes**
- **Ultra-low-traffic blogs (e.g., Tumblr, Blogspot, Wordpress)**
- **Beyond that, yes, there's Bad Stuff:**
  - **Spam**
  - **Exkits**
  - **Some botnets (e.g., C&C via DNS)**
  - **SEO**

OHWs / ODWs show the scope of the problem

- **Big Data -> Anomalies -> Threats -> Interesting Threats**



RSAConference2017

Mining Anomalies

Examples of what you find, and what kinds of additional Threat Intelligence are helpful

Mining Anomalies
Example: Korean IP Range

OHWs – Example (Korea)

- IP block 123.111.0.0/16
  - Mostly home/mobile, some domains
  - Thousands of requests a day
  - Some IPs with hundreds of hits a week (not OHWs)
- Notice some direct-to-IP traffic with high ports
  - *123.111.11.149:65004, 123.111.214.51:32343, …*
  - They seem to be low traffic (check: yes, **all** are OHWs)

OHWs – Example (Korea)

- So now we have a nice pile of Anomalies
  - KEY: finding good factor to filter/cluster on
- Next: are they Threats? Try a license/geo breakdown:
  - Some S.Korean licenses (assume normal?)
  - But Lebanon? Russia? Ukraine? Malaysia? UAE? US? …
    - (and these are Enterprise class licenses)
- So that's looking like possible botnet traffic...

KEY: finding a useful bit of outside-source data (like Geo)

Interlude: Confession Time…

- Original idea for this talk:
  - Gather a bunch of OHW examples (like the “Korean IP Range” set)
  - Illustrate awesome Big Data techniques to deal with massive anomalies! ☺
- Realization: This is fundamentally a hard problem
  - (often, there’s just not enough info in your available data)
  - Most useful: good Threat Intelligence (either detailed/specific, or bigger picture)
- New Goal:
  - Highlight the types of TI I found most useful; suggest ways to find/use it
  - Humbly ask audience for additional suggestions on helpful TI

RSAConference2017

Mining Anomalies
Example: Spamhaus DROP List

“Do not Route Or Peer”

OHWs – Example (Spamhaus DROP list)

- This could be considered an example of “Level 0 TI”
  - (“here’s a list of Bad IPs to block”)
  - Or, maybe “Level 0.5 TI” (since we do know it’s a Spamhaus list)
  - But, we can’t assume that ALL the IPs in the block are spamming…
  - …and we may be re-purposing the list (i.e., not for spam blocking)
- Grab a random block from the DROP list (4.0.0.0/17)
  - Check for traffic…

OHWs – Example (Spamhaus DROP list)

- "Weird Port/Protocol" logs:
  - Some heavy hitters (VPNs? P2P?)…
  - …but about 80% were OHWs
- One day, some interesting "ftp:" traffic…
  - From Philippines (a Government license)
  - Randomly checking the 4.0.0.0/17 range for FTP servers?
    - (ftp://4.0.114.115:21/, ftp://4.0.82.149:21/, ftp://4.0.17.200:21/ …)
- What else?

OHWs – Example (Spamhaus DROP list)

- Switch to Web Traffic logs:
  - About 90% were OHWs; licenses from US, Chile, Japan, Ireland, France, Brazil, India, Saudi Arabia, Canada, China, UK…
  - Wait – China?
    - A US-based APT attack on China **could** hide here…
    - …but this probably **isn't,** based on context…
    - …**still**, if you were that Chinese customer, and saw this traffic, you'd worry.
    - But, if you had this context, you **might** relax a bit…
    - …or would you? You'd never really be **sure**, right?

KEY: given the low traffic volume, certainty is hard to come by…

RSAConference2017

Mining Anomalies
Example: Kelihos Botnet Traffic

(As a jumping-off / comparison point.)

OHWs – Example (Kelihos)

- Long life, wide distribution
- Choose a block in Japan: 101.96.32.0/19
- 5 IPs here had Kelihos traffic in a two-week period (Aug-Sep)
  - 1, 2, 6, 1, 7 requests, respectively (so they will look like OHWs)
- Of course, with TI about Kelihos traffic, you could identify it
  - (we might call this “Level 1” TI)

KEY: Level 1 TI, identifying indicators of specific malware, is always nice…

OHWs – Example (Kelihos)

- …but the Kelihos traffic is just the beginning of the story here
- 98 other IPs had “weird port/protocol” traffic
  - And 92 of those were OHWs (<7 hits, in 7 days of logs)
  - So, if you were one of them, you’re not alone (maybe relax a bit)
    - (unless you can find something unique to you?)

KEY: this is a case where “outside TI” is useful (different perspective/area/device/etc.)

RSAConference2017

Mining Anomalies Example: Exploit Kits

Magnitude and Angler, among others…

OHW/ODW Example: Magnitude Exkit

- Magnitude Exploit Kit (September 2016)
  - Rapidly generated subdomains on changing domains

Root Domain	Subdomains Seen
peakturn.bid	59
whyalias.bid	88
fixwill.bid	123

(Any would have looked like a OHW, or at least a ODW, in your logs…)

- *33bp5f87deg.peakturn.bid, f3d683tf75u7e.peakturn.bid, c2a867dcv.peakturn.bid, …*
- *37vaqbc731751h.whyalias.bid, ef360k151.whyalias.bid, 3214e5b4e.whyalias.bid, …*
- *71efdbfj87.fixwill.bid, f78ccccbh6.fixwill.bid, 889p0818k6m.fixwill.bid, …*

KEY: either a larger external TI source (subdomain count/list), or dossier on this exkit's traffic and host naming scheme.

OHW/ODW Example: Angler Exkit

- Angler is another exploit kit that made heavy use of OHWs/ODWs
  - (this batch from August 2015)
    - *abcadrcrbea.publi.sahvalestoilesheroiques.net*
    - *abbdpdcapcg.tugso.bievuixawiki.net*
    - *abbbaarpzrz.asahi.oonoyvestidadenoiva.com*
    - *azrprqrpgz.ilove.foosifilmycz.com*
    - *azqzegapar.frien.bievuixawiki.net*
    - *azqzebgrrd.brazz.iemooentypo.com*

KEY: again, either a larger external TI source (parent domain count/list), or a dossier on this exkit's traffic and host naming scheme.

RSAConference2017

Mining Anomalies Example: WebAds

An interesting strategy for anti-adblocking…

OHW/ODW Example: WebAds

- Interesting “anti-adblocker” technique:
- Wild-and-crazy domains (anon. registration, of course)…
  - *orangepekoeanorexianervosa.com*
- …with junk/random looking subdomains
  - In 5 days of logs, over 93,526 unique subdomains
    - *izhfo59tzb7r1eazom6z7bq57eb3g1.orangepekoeanorexianervosa.com*
    - *2tfc9wqd1mbjd15l5oxpooftxerevm.orangepekoeanorexianervosa.com*
    - *rvmog16astpughuac9cqe57rz438bp.orangepekoeanorexianervosa.com*
    - …

KEYs: group by parent domain; recognize patterns over time; “bigger picture” check

Time for a Brief Editorial

Adblocking / Malvertising Editorial

- The company behind the previous example published a whitepaper on adblocking…
  - Focus: entirely on the economics and morality
  - No mention of **Malvertising**
    - (which is a major attack vector)
  - No mention of detailed visitor **tracking**
    - (which is a big user privacy concern)

Conveniently ignoring security and privacy is NOT helpful...

RSAConference2017

Mining Anomalies Example: Botnets

Some are “noisy” and some are “quiet”…

OHW/ODW Example: Botnets

- Fast-flux botnet: IPs hosting suspicious porn-spam network sites
- Queries seen for multiple subdomains, on several active domains:
  - 200.68.64.189 (AR)
    - 15 hits in 7 days (mostly OHWs) – 10 subdomains on 3 root domains
  - 178.150.4.211 (UA)
    - 7 hits in 7 days (OHWs) – 7 subdomains on 3 root domains
  - 91.144.134.148 (RU)
    - 15 hits in 7 days (mostly OHWs) – 11 subdomains on 3 root domains

Key: probably not enough traffic just in your logs to be able to make the connections; need larger/external TI to see the patterns…

OHW/ODW Example: Botnets

- Necurs botnet is much more concentrated (and noisier)
  - 185.118.66.196 had 1176 hits in 7 days of traffic,
    - Across 340 licenses
    - (it was a OHW, on average, for >300 of them)
  - 185.127.24.189 had 941 hits in 7 days
    - Across 321 licenses
    - (it was a OHW, on average, for almost 300 of them)
  - Etc.

Key: again, for this class of OHW, you need some simple outside TI (a bigger picture), to see that it's a mass-market bot of some kind...

RSAConference2017

Summary / Action Slides

Ideas from me to you and you to me…

Summary (Actions for You)

- Basic questions for when you're diving into your Big Data:
  - Can you identify one or more common characteristics to filter/cluster?
    - (like the Korea example)
  - Can you find Threat Intelligence to identify specific threat traffic?
    - (like the Botnet and Exkit examples)
  - Can you find TI to cluster threat traffic as part of a bigger picture?
    - (like most of the examples)
- You may not be able to classify all of your OHWs/ODWs
  - (many times, there's just not enough data, and you can't find good TI)
  - Can you at least track those machines, going forward?

Key: **time** gives you another dimension/view on your data

Summary (Actions for Me)

- This talk started out with idea of suggesting a bunch of creative ideas for clustering and filtering your Anomalies…
- …but the most useful thing, most of the time, was better TI
  - (often internal: pulled from different logs/sources)
- So what kinds of TI would **you** like to see more of?
  - Better tying things to malware/attack "families"?
  - Industry/Vertical data? Geo data?
  - Overall traffic levels? (e.g., grouping by parent domain)
  - Other? (chris_larsen [at] symantec.com)

RSAConference2017

??? Questions ???

Acknowledgements:

Dr. Jonathan Dinerstein (One-hit Wonders)

Dr. Timothy van der Horst (One-day Wonders)