Scanning 101

[Diagram showing a scanner sending SYN packets on port 22 to multiple targets and receiving SYN-ACK responses.]

SYN port 22
SYN port 22
SYN port 22
SYN port 22
SYN+ACK
Scanning 101

1. Identify vulnerable services/devices
2. Exploit vulnerabilities!
   • Attack vulnerable hosts/networks directly OR
   • Form botnets to carry out cyberattacks against other targets
Scanning 101

1. Identify vulnerable services/devices

2. Exploit vulnerabilities!
   - Attack vulnerable hosts/networks directly OR
   - Form botnets to carry out cyberattacks against other targets

Scanning hosts on the Internet:
Key component in many of today’s cyberattacks
Tracking Scanning Activity

Aggressive Scans by Hajime Botnet Targeting Port 8291 With a new Exploit

Immediately after the takedown, scan activity on ports 52869 and 37215 saw a huge spike, according to insight provided to Bleeping Computer by Netlab researchers. The Hajime Zeta is looking to scan and find bots for another Satori instance.

Tracking scanning activity: Assessment of ongoing and potentially upcoming cyberattacks
Detecting Scanning Activity: Darknets

X.0.0.0/8
Detecting Scanning Activity: Darknets

X.0.0.0/8

scanner
Internet-Wide Scans

Darknets only capture “naive” Internet-wide scanning

- Scanner targets the entire IPv4 space (e.g., ZMap)
- Scanner targets a random subset of the IPv4 space
Internet-Wide vs. Targeted Scans

X.0.0.0/8

Internet-Wide vs. Targeted Scans

X.0.0.0/8
Internet-Wide vs. Targeted Scans

X.0.0.0/8

scanner Internet-wide

scanner targeted
Internet-Wide vs. Targeted Scans

X.0.0.0/8

What about scans targeting specific prefixes or networks?
Undocumented and potentially much more dangerous!
Outline

Part 1
• What is a distributed network telescope?
• How much unsolicited traffic is there?

Part 2
• How much scanning traffic is there?
• How can we dissect scanning strategies?

Part 3
• How can an individual network detect targeted scans?
A Large CDN as a Network Telescope

- 89,000 CDN servers in 1,300+ ASes, 2,800+ BGP prefixes
  (a subset of the servers this CDN operates)
A Large CDN as a Network Telescope

- **89,000** CDN servers in **1,300+** ASes, **2,800+** BGP prefixes (a subset of the servers this CDN operates)

- Visibility into:
  - Internet-wide scans
  - Scans targeting the CDN itself
  - Scans targeting networks that host CDN caches
CDN Machine IPv4 Addresses

DNS

www.website.com ?

1.2.3.X1 (client-facing IP)

HTTP(S)

1.2.3.X1 (client-facing IP)

CDN cache
CDN Machine IPv4 Addresses

- **DNS**
  - www.website.com?
  - 1.2.3.X1 (client-facing IP)

- **HTTP(S)**
  - 1.2.3.X1 client-facing IP

- **CDN cache**
  - 1.2.3.X2 operations IP (cache refills, etc)
CDN Machine IPv4 Addresses

- Source IP hits only client-facing IPs of our servers:
  - CDN-targeted traffic

- Source IP hits client-facing and operations IPs 50/50:
  - CDN-agnostic traffic
CDN Firewall Log Dataset

- Traffic to active port numbers (TCP 80/443,..)
  - Handled by web server/app. layer firewall
  - Not captured in our dataset
CDN Firewall Log Dataset

• Traffic to active port numbers (TCP 80/443,..)
  ‣ Handled by web server/app. layer firewall
  ‣ Not captured in our dataset

• Traffic to inactive port numbers (any IP)
  ‣ Blocked by static firewall, logged
  ‣ Too much traffic (e.g., DDoS) ⇒ “burst” ⇒ sampling triggered
  ‣ 99.9% of time, a machine is NOT in “burst” state
CDN Firewall Log Dataset

- Traffic to active port numbers (TCP 80/443,..)
  - Handled by web server/app. layer firewall
  - Not captured in our dataset

- Traffic to inactive port numbers (any IP)
  - Blocked by static firewall, logged
  - Too much traffic (e.g., DDoS) ⇒ “burst” ⇒ sampling triggered
  - 99.9% of time, a machine is NOT in “burst” state

**Full visibility of “sporadic” unsolicited traffic**

<table>
<thead>
<tr>
<th></th>
<th>UDP</th>
<th>TCP</th>
<th>TCP SYN</th>
</tr>
</thead>
<tbody>
<tr>
<td>total pkts</td>
<td>19.4B</td>
<td>17.1B</td>
<td>16.8B</td>
</tr>
<tr>
<td></td>
<td>2.3B (12%)</td>
<td>17.1B (88%)</td>
<td>16.8B (86%)</td>
</tr>
</tbody>
</table>

Dataset: November 2018
Unsolicited Traffic at CDN Caches

Machine 1: Baseline radiation
Unsolicited Traffic at CDN Caches

Machine 2: CDN-targeted spikes
Unsolicited Traffic at CDN Caches

Machine 3: CDN-agnostic spikes
Unsolicited Traffic at CDN Caches

- **Baseline radiation**: Every IPv4 address: ~3000 pkts/day
- **Spikes can be grouped into**:
  - CDN-targeted $\Rightarrow$ sources target domain names
  - CDN-agnostic $\Rightarrow$ sources target prefixes/networks
Daily packets per DST IP in a /8 Darknet

DST IPs (normalized by max)

packets per destination IP address (1st of November 2018)

UCSD /8 darknet IPs (~15M IPs)
Daily packets per DST IP in a /8 Darknet

almost all darknet destination IP addresses:
baseline radiation - median: 3091 pkts/IP

heavily concentrated distribution
very few addresses receive more
Daily packets per DST IP: Darknet vs. CDN

[Graph showing the distribution of packets per destination IP address (1st of November 2018)]

- UCSD /8 darknet IPs (~15M IPs)
- CDN client-facing IPs (89K IPs)
- CDN operations IPs (89K IPs)
Daily packets per DST IP: Darknet vs. CDN

some CDN machines log slightly less. Network filtering!

[Graph showing packets per destination IP address (1st of November 2018)]

- UCSD /8 darknet IPs (~15M IPs)
- CDN client-facing IPs (89K IPs)
- CDN operations IPs (89K IPs)
Daily packets per DST IP: Darknet vs. CDN

Large share of CDN machines: only baseline radiation

UCSD /8 darknet IPs (~15M IPs)
CDN client–facing IPs (89K IPs)
CDN operations IPs (89K IPs)
Daily packets per DST IP: Darknet vs. CDN

- UCSD /8 darknet IPs (~15M IPs)
- CDN client-facing IPs (89K IPs)
- CDN operations IPs (89K IPs)

heavy tail: CDN machines log targeted traffic invisible in darknet
Daily packets per DST IP: Darknet vs. CDN

What is this additional unsolicited traffic?
Outline

Part 1
• What is a distributed network telescope?
• How much unsolicited traffic is there?

Part 2
• How much scanning traffic is there?
• How can we dissect scanning strategies?

Part 3
• How can an individual network detect targeted scans?
Identifying Scans

- Source IP contacts more than 100 destination IPs (IPv4)
- Maximum inter-arrival time 5400 seconds
- Effectively removes remaining DDoS attacks against CDN
Identifying Scans

- Source IP contacts more than 100 destination IPs (IPv4)
- Maximum inter-arrival time 5400 seconds
- Effectively removes remaining DDoS attacks against CDN

- 87% of all logged unsolicited packets are part of scans
- November ’18: 2.17M scans, 16.8B packets, 1.1M SRC IPs
Scan Target Selection Strategies

all scans

Internet-wide
(source targets entire IPv4 space)

targeted
(source targets specific IPs/ranges)
Scan Target Selection Strategies

all scans

Internet-wide
(source targets entire IPv4 space)

full IPv4 space
28% pkts

targeted
(source targets specific IPs/ranges)

full IPv4 space: scanner IP targets most/all of our IP addresses (e.g., ZMap)
Scan Target Selection Strategies

all scans

Internet-wide
(source targets entire IPv4 space)

full IPv4 space
28% pkts

targeted
(source targets specific IPs/ranges)
domain scans: targeted at DNS-exposed IPs
4.1% pkts

domain scans: scanner IP targets client-facing (DNS-exposed) IPs exclusively
E.g., scanning Alexa Top 1M etc.
Scan Target Selection Strategies

- all scans
  - Internet-wide (source targets entire IPv4 space)
    - full IPv4 space
      - 28% pkts
  - targeted (source targets specific IPs/ranges)
    - domain scans: targeted at DNS-exposed IPs
      - 4.1% pkts

Leaves us with 2/3s of scan packets that do not target the full space nor target our CDN DNS-exposed IPs:
Scan Target Selection Strategies

all scans

Internet-wide
(sources targets entire IPv4 space)

full IPv4 space
28% pkts

targeted
(sources targets specific IPs/ranges)
domain scans:
targeted at DNS-exposed IPs
4.1% pkts

Leaves us with 2/3s of scan packets that do not target the full space nor target our CDN DNS-exposed IPs:

Do they target a random subset of the space or target specific prefixes/networks?
Detecting if Scan Targets Random Subset

IPv4 /8 prefix

machines per /8

0 5K 10K 15K

0.0.0.0/8 64.0.0.0/8 128.0.0.0/8 192.0.0.0/8 255.0.0/8

IPv4 /8 prefix
Internet-wide partial (random) scan

- Packets in scan total: 148.
- Pearson correlation DST IPs / CDN IPs = 0.96.
- This scan targets a random subset of CDN IPs.
  - Internet-wide Partial scan.
Localized Scan

- Packets in scan total: 827.
- Pearson correlation DST IPs / CDN IPs = 0.38.
- This scan targets a non-random subset of CDN IPs.
  - Localized scan
Internet-wide vs. Localized Scans

localized scans: low pearson correlation

Internet-wide scans: high pearson correlation
Scan Target Selection Strategies

- **all scans**
  - **Internet-wide** (source targets entire IPv4 space)
    - full IPv4 space: 28% pkts
    - partial: random subset of IPv4: 39.9% pkts
  - **targeted** (source targets specific IPs/ranges)
    - domain scans: targeted at DNS-exposed IPs: 4.1% pkts
    - localized: targets prefixes/ASes/?: 29% pkts
Scan Target Selection Strategies

all scans

Internet-wide
  (source targets entire IPv4 space)
  - full IPv4 space: 28% pkts
  - partial: random subset of IPv4: 39.9% pkts

targeted
  (source targets specific IPs/ranges)
  - domain scans: targeted at DNS-exposed IPs: 4.1% pkts
  - localized: targets prefixes/ASes/?: 29% pkts

~68% scan packets: Internet-wide scans
Scan Target Selection Strategies

**Internet-wide**

- Full IPv4 space: 28% pkts
- Partial: random subset of IPv4: 39.9% pkts

**Targeted**

- Domain scans: targeted at DNS-exposed IPs: 4.1% pkts
- Localized: targets prefixes/ASes/?: 29% pkts

~68% scan packets: Internet-wide scans

~29% scan packets: target prefixes/ASes/?
Some 30% of all scan traffic localized!

- **Do not** target a random subset of the space
- **Do not** target the CDN’s customers

Vastly different characteristics:

- Services scanned, Scanner origins, repeated scans, etc.
- Internet-wide scans
- Full IPv4 space: 28% pkts
- Partial: random subset of IPv4: 39.9% pkts
- Domain scans: targeted at DNS-exposed IPs: 4.1% pkts
- 4.1% scan packets targeting the CDN’s customers (e.g., Alexa Top 1K): 29% pkts
Top 5 Target Port Numbers

![](Top 5 Target Port Numbers.png)

For all scans (CDN): 23% on port 23, 445% on port 445, 1433% on port 1433, 81% on port 81, and 22% on port 22.
Table 3: Top origin countries of scan traffic.

<table>
<thead>
<tr>
<th>Country</th>
<th>Internet-wide full scans</th>
<th>Internet-wide partial scans</th>
<th>Localized multi-port scans</th>
</tr>
</thead>
<tbody>
<tr>
<td>U.S.</td>
<td>38.2%</td>
<td>11.3%</td>
<td>30.5%</td>
</tr>
<tr>
<td>Russia</td>
<td>11.5%</td>
<td>7.4%</td>
<td>11.6%</td>
</tr>
<tr>
<td>China</td>
<td>20.3%</td>
<td>8.3%</td>
<td>10.0%</td>
</tr>
<tr>
<td>Netherlands</td>
<td>9.0%</td>
<td>11.5%</td>
<td>8.8%</td>
</tr>
<tr>
<td>Brazil</td>
<td>16.2%</td>
<td>7.8%</td>
<td>10.0%</td>
</tr>
</tbody>
</table>

We note that CDN-targeted scans show different behavior, with the two port numbers being 500/UDP and 137/UDP (NetBIOS name resolution). These two port combinations can be attributed to incarnations of the Mirai botnet [1], whereas the 4-tuple of 23, 8080, 7547, 8291 (TCP) accounts for only 19% of all scans.

While Internet-wide full scans are highly concentrated, with the presence of 445/tcp the most likely Web, we believe that the majority of these packets do not resemble actual scans or exploits of these servers. Cleaning these packets will improve our ability to forecastation and/or NetBIOS name resolution upon establishing a TCP connection in order to perform localized scans to countries.

We do not restrict our definitions of scans to countries. We do not consider scans to countries. We do not consider scans to countries.

(b) Scan source IP addresses.

(c) We note that when looking at aggregate port statistics, percentage packets from countries.

Table 3 shows the top-5 origin countries of scan traffic.

<table>
<thead>
<tr>
<th>Country</th>
<th>Internet-wide full scans</th>
<th>Internet-wide partial scans</th>
<th>Localized multi-port scans</th>
</tr>
</thead>
<tbody>
<tr>
<td>U.S.</td>
<td>38.2%</td>
<td>11.3%</td>
<td>30.5%</td>
</tr>
<tr>
<td>Russia</td>
<td>11.5%</td>
<td>7.4%</td>
<td>11.6%</td>
</tr>
<tr>
<td>China</td>
<td>20.3%</td>
<td>8.3%</td>
<td>10.0%</td>
</tr>
<tr>
<td>Netherlands</td>
<td>9.0%</td>
<td>11.5%</td>
<td>8.8%</td>
</tr>
<tr>
<td>Brazil</td>
<td>16.2%</td>
<td>7.8%</td>
<td>10.0%</td>
</tr>
</tbody>
</table>

We leverage the CDN's proprietary geolocation database to map scanner IP addresses and assess the effect of sharded scans on our inferences.

(a) Scan packets.

(b) Scan source IP addresses.

(c) We note that when looking at aggregate port statistics, percentage packets from countries.

Table 3 shows the top-5 origin countries of scan traffic.

<table>
<thead>
<tr>
<th>Country</th>
<th>Internet-wide full scans</th>
<th>Internet-wide partial scans</th>
<th>Localized multi-port scans</th>
</tr>
</thead>
<tbody>
<tr>
<td>U.S.</td>
<td>38.2%</td>
<td>11.3%</td>
<td>30.5%</td>
</tr>
<tr>
<td>Russia</td>
<td>11.5%</td>
<td>7.4%</td>
<td>11.6%</td>
</tr>
<tr>
<td>China</td>
<td>20.3%</td>
<td>8.3%</td>
<td>10.0%</td>
</tr>
<tr>
<td>Netherlands</td>
<td>9.0%</td>
<td>11.5%</td>
<td>8.8%</td>
</tr>
<tr>
<td>Brazil</td>
<td>16.2%</td>
<td>7.8%</td>
<td>10.0%</td>
</tr>
</tbody>
</table>
The figure shows the top 5 target port numbers for different types of scans:

1. **All scans (CDN)**: The port numbers 23, 445, 1433, 81, and 22 are observed. Among these, port 23 is the most common, accounting for about 23% of all scans.
2. **UCSD darknet**: Similarly, port 23 is the most common, followed by ports 80, 445, 22, and 81.
3. **Localized scans**: For localized scans, port 8291 is the most common, followed by ports 445, 22, and 7547.

The figure also indicates the percentage of packets for each port, with port 23 being the most commonly scanned in all categories. The data suggests that port 23 is particularly targeted across different scan types and locations.
Critical vulnerabilities targeted in localized scans! But barely scanned by Internet-wide scanners -> Darknets severly underestimate scanning activity
Outline

Part 1
- What is a distributed network telescope?
- How much unsolicited traffic is there?

Part 2
- How much scanning traffic is there?
- How can we dissect scanning strategies?

Part 3
- How can an individual network detect targeted scans?
Baseline Composition

logged packets per day

day [November '18]

- client-facing IP
- operations IP
Baseline Composition

remaining unsolicited traffic after removal of Internet-wide (full and random subset) scans
Baseline Composition

remaining unsolicited traffic after removal of Internet-wide (full and random subset) scans
Baseline Composition

- 90% of baseline radiation is the result of Internet-wide scans
  - Full scans of the + random subsets of the IPv4 space
Baseline: Detect Targeted Scans

- 90% of baseline radiation is the result of Internet-wide scans
  - Full scans of the + random subsets of the IPv4 space
- Threat detection / detection of targeted activity
  - Determine current baseline level (darknets, etc.)
  - See if unsolicited scan traffic exceeds baseline
Key Takeaways

• Distributed telescope illuminates undocumented scan activity

• Baseline: Every routed IPv4 address receives ~3000 pkts/day
  ▸ Deviation from the baseline indicates targeted activity

• Methodology to dissect scan strategies
  ▸ 30% of scan traffic are result of localized scans
  ▸ Different characteristics (services targeted, origins, etc.)
Baseline Evolution

Number of packets per IP address over time:

- 2016-01-01 to 2017-01-01
- 2017-01-01 to 2018-01-01
- 2018-01-01 to 2019-01-01
CDN vs. UCSD Darknet: DSTs per Source IP

- destination IPs hit in CDN
- destination IPs hit in UCSD-NT

source IPs

- 10M
- 1M
- 100K
- 10K
- 1K
- 100
- 10
- 1

destination IPs hit in CDN

- 170K
- 10K
- 100
- 0

destination IPs hit in UCSD-NT

- 10M
- 100K
- 10K
- 1K
- 100
- 10
- 1