SipScan: the world scanning itself

A. Dainotti, A. King, K. Claffy, F. Papale*, A. Pescapè*

alberto@caida.org

CAIDA - University of California, San Diego

*University of Napoli Federico II, Italy
WHAT IS IT?

Feb 2011

• A “/0” scan from a botnet
• Observed by the UCSD telescope (a /8 darknet)
• Scanning SIP Servers with a specific query on UDP port 5060 and SYNs on TCP port 80
### OVERVIEW

**numbers for UDP**

<table>
<thead>
<tr>
<th>Metric</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td># of probes (1 probe = 1 UDP + multiple TCP pkts)</td>
<td>20,255,721</td>
</tr>
<tr>
<td># of source IP addresses</td>
<td>2,954,108</td>
</tr>
<tr>
<td># of destination IP addresses</td>
<td>14,534,793</td>
</tr>
<tr>
<td>% of telescope IP space covered</td>
<td>86.6%</td>
</tr>
<tr>
<td># of unique couples (source IP - destination IP)</td>
<td>20,241,109</td>
</tr>
<tr>
<td>max probes per second</td>
<td>78.3</td>
</tr>
<tr>
<td>max # of distinct source IPs in 1 hour</td>
<td>160,264</td>
</tr>
<tr>
<td>max # of distinct source IPs in 5 minutes</td>
<td>21,829</td>
</tr>
<tr>
<td>average # of probes received by a /24</td>
<td>309</td>
</tr>
<tr>
<td>max # of probes received by a /24</td>
<td>442</td>
</tr>
<tr>
<td>average # of sources targeting a destination</td>
<td>1.39</td>
</tr>
<tr>
<td>max # of sources targeting a destination</td>
<td>14</td>
</tr>
<tr>
<td>average # of destinations a source targets</td>
<td>6.85</td>
</tr>
<tr>
<td>max # of destination a source targets</td>
<td>17613</td>
</tr>
</tbody>
</table>
REL WORKS

• Analyses of botnet scans

• Coordinated scans
  - S. Staniford, V. Paxson, N. Weaver, “How to Own the Internet in Your Spare Time”, Usenix Sec. Symp. 2002
  - Carrie Gates, “Coordinated Scan Detection”, NDSS 2009

• Botnet code analysis
SIPSCAN

Anatomy of the scan

- Payload Signature
- Unspoofed
- Botnet
- /0 Scan
- Progression
- Bot Turnover
- Coverage vs Overlap
• Thanks to Saverio Niccolini @NEC (involved in IETF WGs on SIP) for brainstorming
• Thanks to Joe Stewart @SecureNetworks for finding the binary of the malware
• Matches a downloadable component of the Sality botnet documented by Symantec
SIPSCAN

isolating the “SipScan”

• Thanks to the unique payload fingerprint we could isolate it without inferences
UNSPOOFED

Because...

• Egyptian outage: we were actually not seeing “egyptian” IPs when the Egypt was isolated from the rest of the Internet
• It seems to be a scan (UDP requests + TCP SYN). No purpose in spoofing
• No IPs from our /8 or from unassigned space
• IPIDs and src ports from scanning hosts are consistent for the same host
UNSPPOOFED

The case of the Egyptian Killswitch (Feb 2011)

• No SipScan pkts are geolocated to Egypt during the Egyptian outage!

A BOTNET need of a Command & Control channel

• During the Egyptian blackout, some Conficker-infected networks were still able to send conficker scan traffic.
Animation created with an improved version of Cuttlefish, developed by Brad Huffaker
http://www.caida.org/tools/visualization/cuttlefish/
### /0 SCAN

**UCSD Telescope**

<table>
<thead>
<tr>
<th>Metric</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td># of probes (1 probe = 1 UDP + multiple TCP pkts)</td>
<td>20,255,721</td>
</tr>
<tr>
<td># of source IP addresses</td>
<td>2,954,108</td>
</tr>
<tr>
<td># of destination IP addresses</td>
<td>14,534,793</td>
</tr>
<tr>
<td>% of telescope IP space covered</td>
<td>86.6%</td>
</tr>
<tr>
<td># of unique couples (source IP - destination IP)</td>
<td>20,241,109</td>
</tr>
<tr>
<td>max probes per second</td>
<td>78.3</td>
</tr>
<tr>
<td>max # of distinct source IPs in 1 hour</td>
<td>160,264</td>
</tr>
<tr>
<td>max # of distinct source IPs in 5 minutes</td>
<td>21,829</td>
</tr>
<tr>
<td>average # of probes received by a /24</td>
<td>309</td>
</tr>
<tr>
<td>max # of probes received by a /24</td>
<td>442</td>
</tr>
<tr>
<td>average # of sources targeting a destination</td>
<td>1.39</td>
</tr>
<tr>
<td>max # of sources targeting a destination</td>
<td>14</td>
</tr>
<tr>
<td>average # of destinations a source targets</td>
<td>6.85</td>
</tr>
<tr>
<td>max # of destination a source targets</td>
<td>17613</td>
</tr>
</tbody>
</table>
/0 SCAN

DShield

DShield Observed Sources to UDP Port 5060

Date (UTC)

01/01 01/08 01/15 01/22 01/29 02/05 02/12 02/19 02/26

0 500 1000 1500 2000 2500 3000

Cooperative Association for Internet Data Analysis
University of California San Diego

http://www.dshield.org
We identified flow-level properties (e.g. 1 pkt + PS size) that allowed to spot the same traffic in MAWI/WIDE traces, which are anonymized.

- analysis of payload signature
- processing of MAWI traces to get flow-level logs
- sanitization (filtering) of MAWI logs
- plot

http://mawi.wide.ad.jp/mawi/
MAWI uses a specific configuration of Tcpdpriv for anonymization:

- **A50**: IP addresses are scrambled preserving matching prefixes.
- **C4**: IP classes (class A-D) are also preserved.
- **M99**: All multicast addresses are not scrambled.
- **P99**: TCP and UDP port numbers are not scrambled.

A few different /8 networks were found in the MAWI traffic associated with the SipScan.
/0 SCAN

Exploiting source port continuity

- Src_port++ in range 1025 - 5000
- ~512 average increments between 2 “visits” to the telescope
HILBERT CURVE

http://xkcd.com/195
The 1-dimensional IPv4 address space is mapped into a 2-dimensional image using a Hilbert curve. CIDR netblocks always appear as squares or rectangles in the image.

Software for hilbert-based IP heatmaps @ http://www.measurement-factory.com
BOT TURNOVER
new src IPs arrive constantly

Cumulative Sources (%)

Cum. Source IPs
Cum. Source /24s
Source IP Rate

Date (UTC)

Unique Source IPs (5 min bins)

Cooperative Association for Internet Data Analysis
University of California San Diego
BOT TURNOVER

most src IPs leave constantly

Figure 3.11 – Comparison of IPs arrivals with the number of packets per second of the scan received by the UCSD network telescope.

Figure 3.12 – Number of source addresses per number of packets sent.

Figure 3.13 – Number of senders for number of packets sent in range 1-10.

Figure 3.14 – Number of destination IPs that have been targeted by a certain number of source IPs.
# BOT TURNOVER

*few src IPs *stay* for a while*

<table>
<thead>
<tr>
<th>Metric</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td># of probes (1 probe = 1 UDP + multiple TCP pkts)</td>
<td>20,255,721</td>
</tr>
<tr>
<td># of source IP addresses</td>
<td>2,954,108</td>
</tr>
<tr>
<td># of destination IP addresses</td>
<td>14,534,793</td>
</tr>
<tr>
<td>% of telescope IP space covered</td>
<td>86.6%</td>
</tr>
<tr>
<td># of unique couples (source IP - destination IP)</td>
<td>20,241,109</td>
</tr>
<tr>
<td>max probes per second</td>
<td>78.3</td>
</tr>
<tr>
<td>max # of distinct source IPs in 1 hour</td>
<td>160,264</td>
</tr>
<tr>
<td>max # of distinct source IPs in 5 minutes</td>
<td>21,829</td>
</tr>
<tr>
<td>average # of probes received by a /24</td>
<td>309</td>
</tr>
<tr>
<td>max # of probes received by a /24</td>
<td>442</td>
</tr>
<tr>
<td>average # of sources targeting a destination</td>
<td>1.39</td>
</tr>
<tr>
<td>max # of sources targeting a destination</td>
<td>14</td>
</tr>
<tr>
<td>average # of destinations a source targets</td>
<td>6.85</td>
</tr>
<tr>
<td>max # of destination a source targets</td>
<td>17613</td>
</tr>
</tbody>
</table>
targets which we believe is likely a function of a parameter.

We discovered a correlation between coverage and overlap in the rate of the scan or specific subnets being scanned. However, the overall coverage and overlap of target addresses are independent of the number of bots active at any given time.

These properties—reverse representation of the address bytes, making detection highly unlikely—are a non-negligible overlap in terms of bots hitting the same probed IP addresses. Whether probed zero, one, or multiple times, more than one bot, and on average a targeted IP is probed by 850 packets.

About 5 million IP addresses were probed by more than one bot. The high bot turnover rate makes the scan impressive, as bots would most likely arrive from a different IP address on a 69% network and see only a tiny fraction of packets to the same port. Not only would an automated intrusion detection system be ineffective, but they would most likely be overwhelmed by the volume.

The scan fails to cover the entire darknet's address space, making detection highly unlikely. But they would most likely be assigned to a CxC channel with a smaller pool of bots. In such a situation, bots could not reach the head of the queue twice.

The CxC channel may assign a list of target IP ranges to a queue, and the probed IP addresses are scattered all over the address space without clusters or holes. In both the standard and reverse order, the probed IP addresses are scattered all over the address space. Distinguish bots whether probed zero, one, or multiple times, more than one bot, and on average a targeted IP is probed by 850 packets.

In combination with the reverse-byte order property of the address space, the probed source IP addresses are scattered without clusters or holes. These properties—reverse representation of the address bytes, making detection highly unlikely—are a non-negligible overlap in terms of bots hitting the same probed IP addresses. Whether probed zero, one, or multiple times, more than one bot, and on average a targeted IP is probed by 850 packets.

COVERAGE & OVERLAP

different phases w/ different parameters?

Figure shows the distribution of the number of packets sent by each bot. The diagram on the left uses a log-log scale to zoom in to the left side of the distribution, where we use the Hilbert curve map to highlight the overlap of the scan. The same regions are visible in Figure 8. The corresponding address space shows the same phenomenon in terms of overlap, the brighter areas indicating a greater coverage of the target space.

Figure 8 shows the distribution of the number of packets sent by each bot. The diagram on the left uses a log-log scale to zoom in to the left side of the distribution, whereas the diagram on the right uses a linear scale to zoom in to the left side of the distribution to show all the data. The right side of the figure shows the overlap of the scan, the left side shows the coverage of the scan, and the middle shows the overlap of the scan.

The Hilbert curve map in Figure 8 shows three different phases of the scan: Phase A, Phase B, and Phase C. The different densities of areas on the map represent target addresses probed by more than one bot. The different phases correspond to three different phases of the scan as indicated in Table 8. The figure shows statistics calculated separately for the three different phases with different parameters.
COVERED & OVERLAP

different phases w/ different parameters?
The strong coordination of bot activity is also visible in the distributions of the number of packets sent in each reverse 350 subnet, with a consistent and distinctive behavior in each phase. The graph shows for each phase the distribution of the number of packets, reflecting the varying coverage and overlap of the target address space. Table 6 summarizes the characteristics of the three phases of the scan, with different coverage and overlap of the target address space, showing a trade-off between the number of IP spaces covered and the overlap. The curves represent for each phase the distributions of packets sent, with the average values centered around different values and mostly non-overlapping, indicating a probabilistic approach in the choice of target IPs and a large overlap in the transition from one phase to another, likely happening at the level of the subnet.
SIPSCAN FEATURES

some are unique

• Operated by a botnet
• Global vs Global
• Observed by a /8
• No inferences on pkts: unique payload “signature”
• Lasting 12 days
• Sequential progression in reverse byte order
• Continuous use of new bots
• Stealth: IP progression, speed, use of new bots
• Coordination between sources (global sequential progression and small redundancy)
• Targeting SIP
THANKS