A Search Engine Backed by Internet-Wide Scanning

Zakir Durumeric
University of Michigan

David Adrian
University of Michigan

Ariana Mirian
University of Michigan

Michael Bailey
University of Illinois

J. Alex Halderman
University of Michigan
ZMap

2013 A **1200x performance improvement** over Nmap for an Internet-wide single port TCP scan

2014 Scan the Internet in **under 5 minutes**.

2015 Popular in industry and academia, used by over **104** academic studies.
ZMap Vision

Goals

Enable new and exciting research

Decrease the barriers to entry for Internet-wide surveys

Anyone can scan the entire Internet using a single host
# ZMap Vision

<table>
<thead>
<tr>
<th>Goals</th>
<th>Reality</th>
</tr>
</thead>
<tbody>
<tr>
<td>Enable new and exciting research</td>
<td>Not all researchers can run ZMap</td>
</tr>
<tr>
<td>Decrease the barriers to entry for Internet-wide surveys</td>
<td>Negotiate with network administrators for bandwidth and address space</td>
</tr>
<tr>
<td>Anyone can scan the entire Internet using a single host</td>
<td>Maintain an opt-out list and respond to complaints</td>
</tr>
</tbody>
</table>
A public archive of Internet-wide scan data

Data from University of Michigan, Rapid7, Fedora, and more

Over 35 TB downloaded in July 2015

https://scans.io
What is the impact of a particular vulnerability?

What types of cryptography are in use?

What version of software are most popular?
scans.io

Nearly 5TB of HTTPS data in the last year

What is the impact of a particular vulnerability?

Data needs to be processed and annotated

What types of cryptography are in use?

Analysis is time-consuming, and error-prone

What version of software are most popular?
What if?

…we could answer questions with a **single query**?

…we always knew the **current state** of the Internet?

…we built a **search engine** on top of Internet-wide scan data?
Search engine that allows researchers to ask questions about the devices and networks that compose the Internet.
Example

What is the impact of disabling support for SSLv3?
443.https.tls.version:SSLv3

Example

What is the impact of disabling support for SSLv3?
### IPv4 Hosts

<table>
<thead>
<tr>
<th>IPv4 Address</th>
<th>Details</th>
</tr>
</thead>
<tbody>
<tr>
<td>85.109.119.118</td>
<td>TTNET - Turk Telekomunikasyon Anonim Sis (9121), Turkey, 80/http, 443/https, 302 Document moved, Vigor Router, 443.https.tls.version: SSLv3</td>
</tr>
<tr>
<td>179.5.40.60</td>
<td>Telgua (14754), San Salvador, Departamento de San Salvador, El Salvador, 443/https, 7547/cwmp, brutus.neuronio.pt, 443.https.tls.version: SSLv3</td>
</tr>
<tr>
<td>82.163.47.0</td>
<td>MAILBOX - Mailbox Internet Ltd (8401), United Kingdom, 443/https, Vigor Router, 443.https.tls.version: SSLv3</td>
</tr>
</tbody>
</table>
Example

What is the impact of disabling support for SSLv3?
Example

What is the impact of disabling support for SSLv3?
This tool allows you to generate a report on the breakdown of a value present on the ipv4s returned by your query. For example, to generate a report on the cipher suites chosen by HTTPS servers in the United States, you could query for `location.country_code: US AND protocols:443/https` and then generate a report on the breakdown of the field `443.https.tls.cipher_suite.name`. A list of reportable fields is available here.

Many fields have both parsed and raw values available (e.g., `80.http.get.headers.server` and `80.http.get.headers.server.raw`). In these cases, the raw value will represent the exact string (e.g., Apache/2.2.22 (Debian)) and the parsed version will bucket on individual terms (e.g., Apache and Debian). Incidentally, in this case, you likely want to aggregate on a parsed out version of the web server, `metadata.web_server`.

### Host Report

<table>
<thead>
<tr>
<th>Cipher Suite</th>
<th>Count</th>
<th>Total Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>TLSv1.2</td>
<td>12,329,034</td>
<td></td>
</tr>
<tr>
<td>TLSv1.0</td>
<td>3,843,276</td>
<td></td>
</tr>
<tr>
<td>TLSv1.1</td>
<td>33,193</td>
<td></td>
</tr>
<tr>
<td>SSLv3</td>
<td>3,936</td>
<td></td>
</tr>
</tbody>
</table>
A Search Engine Backed by Internet-Wide Scanning

- Full-text search
  - SQL
- Current and historical data
  - API
Motivation

Architecture

Looking Forward
Data Collection

1. Identify listening hosts
2. Gather application-layer data
3. Annotate with additional metadata
4. Aggregate by host
Data Collection

1. Identify listening hosts
2. Gather application-layer data
3. Annotate with additional metadata
4. Aggregate by host

1.2.3.4
23.196.166.175
141.211.243.44
198.41.209.140
...

ZMap
ZGrab
ZTag
ZDb
Data Collection

1. Identify listening hosts
2. Gather application-layer data
3. Annotate with additional metadata
4. Aggregate by host

```json
{
  host: "1.2.3.4",
  cipher: "DHE_AES",
  version: "SSLv3",
  certificate: ...
}
```
Data Collection

1. Identify listening hosts
2. Gather application-layer data

3. **Annotate with additional metadata**

4. Aggregate by host

```json
{
  host: "1.2.3.4",
  cipher: "DHE_AES",
  version: "SSLv3",
  certificate: ...
  tags: ["POODLE"],
  server: "nginx"
}
```
Annotations are simple Python functions

1. Identify listening hosts
2. Gather application-layer data
3. Annotate with additional metadata
4. Aggregate by host

https://github.com/zmap/ztag

```python
class CiscoServer(Annotation):
    protocol = protocols.HTTP

    def process(self, obj, meta):
        server = obj['headers']['server']
        if 'cisco' in server.lower():
            meta.manufacturer = "Cisco"
        return meta
```
Data Collection

1. Identify listening hosts
2. Gather application-layer data
3. Annotate with additional metadata
4. Aggregate by host

ZMap

ZGrab

ZTag

ZDb

1.2.3.4,22,SSH
1.2.3.4,80,HTTP
1.2.3.4,443,TLS
...

A Search Engine Backed by Internet-Wide Scanning • David Adrian
Querying
A Search Engine Backed by Internet-Wide Scanning  •  David Adrian

Querying

Streaming Deltas

Elastic Search

ZDb
A Search Engine Backed by Internet-Wide Scanning • David Adrian

Querying

Full-text search
Queries over current data

Elastic Search

Streaming Deltas

ZDb
Querying

- Full-text search
- Queries over current data

ZDb

Streaming Deltas

Elastic Search

Google

Big Query

Daily Snapshots

Streaming Deltas
Querying

- Full-text search
- Queries over current data
- SQL queries
- Queries over historical data

ZDb → Elastic Search

ZDb → Daily Snapshots

ZDb → Streaming Deltas

ZDb → Google Big Query
Querying

ZDb

Streaming Deltas

Elastic Search

Google Big Query

Daily Snapshots

Web Interface

API

A Search Engine Backed by Internet-Wide Scanning • David Adrian
Motivation

Architecture

Looking Forward
Censys aims to be open and community-driven
Contributing

Are you extending ZMap, ZGrab, or another scanner with a new protocol?

Do you have annotations to add to our framework?

We’ll work with researchers to add new scan modules to Censys

https://github.com/zmap/zmap

https://github.com/zmap/zgrab

https://github.com/zmap/ztag
Future Research

Censys strives to be **research enabling more research**

**Contribute back** scanners and annotations — we do the heavy lifting

Bring **measurement-driven security** to a wider audience
Acknowledgements

Google, for providing much of the infrastructure that runs Censys

Ben Burgess, Alishah Chator, Harsha Gotur, Drew Springall

Elie Bursztein, Brad Campbell, Aleksander Durumeric, James Kasten, Kyle Lady, Adam Langley, HD Moore, Pat Pannuto, Paul Pearce, Niels Provos, Mark Schloesser, Eric Wustrow

The many contributors to the ZMap and ZGrab open source projects
https://www.censys.io

team@censys.io

censysio davidcadrian

A Search-Engine Backed by Internet-Wide Scanning.
Zakir Durumeric, David Adrian, Ariana Mirian, Michael Bailey, J. Alex Halderman