Internet-wide Scanning Taxonomy and Framework

David Myers[1] **Ernest Foo**[2] **Kenneth Radke**[3]

[1] Email: d1.myers@connect.qut.edu.au

[2]Email: e.foo@qut.edu.au [3]Email: k.radke@qut.edu.au

Abstract

Industrial control systems (ICS) have been moving from dedicated communications to switched and routed corporate networks, making it probable that these devices are being exposed to the Internet. Many ICS have been designed with poor or little security features, making them vulnerable to potential attack. Recently, several tools have been developed that can scan the internet, including ZMap, Masscan and Shodan. However, little in-depth analysis has been done to compare these Internet-wide scanning techniques, and few Internet-wide scans have been conducted targeting ICS and protocols.

In this paper we present a Taxonomy of Internet-wide scanning with a comparison of three popular network scanning tools, and a framework for conducting Internet-wide scans.

Keywords: Internet-wide scanning, Taxonomy, Framework, Industrial Control Systems, Critical Infrastructure, SCADA, ZMap, Masscan, Shodan.

1 Introduction

With the exhaustion of the IPv4 address pool, and the slow adoption of IPv6, researchers have the opportunity to conduct Internet-wide surveys for research. In the past few years, there have been several Internet-wide scans conducted by different organisations worldwide. With recent advances in Internet-wide scanning tools, computational power, and network bandwidth, the required time to scan the IPv4 address space has been dramatically reduced. It is now possible to scan the entire public IPv4 Internet in as little as three minutes (Graham, 2013*b*). Several tools currently exist which allow scans of the IPv4 internet, including ZMap, Masscan, Unicornscan and Shodan.

ICS have been moving from traditional serial communications to switched and routed corporate networks, either directly connected through ethernet or through devices to enable serial to ethernet conversion. These ethernet networks allow for easy access, management and operation of the devices, however connection to corporate networks can allow the devices to be directly accessible from the Internet (Hoover, 2013).

However, there is currently no framework for conducing Internet-wide scans, no comparison between commonly used Internet-wide scanning techniques has been made, and few Internet-wide scans have been conducted for Internet accessible ICS devices.

2 Background

The Electronic Frontier Foundation (EFF) EFF SSL Observatory project conducted an Internet-wide scan for study (Electronic Frontier Foundation, 2014). From this dataset, the EFF were able to ask key questions about the existing state of SSL Certificates on the internet, including the number of trusted Certificate Authorities (CAs), number of signers, and frequency of use. The EFF found there were a large amount of weak and vulnerable certificates.

The Internet Census (2012) was a distributed scan of the IPv4 Internet using the Carna Botnet, which infected over 400,000 embedded devices (Anonymous, 2012). Using the NMap Scripting Engine, the botnet was designed to initially scan random addresses, attempt a telnet login, and upload a small binary to infected devices which then was used to scan the Internet This distributed method of scanning the Internet dramatically reduced the time of the scan, from potentially months to hours.

Mining your Ps and Qs was a distributed scan of the IPv4 Internet using the NMap network mapping tool; the largest network survey of TLS and SSH servers at the time. The goal of the project was to search for TLS certificates with problems related to inadequate randomness upon generation (Heninger et al., 2012).

ZMap, created by a team from the University of Michigan (Durumeric et al., 2013), provides several improvements over traditional port-scanning programs such as NMap, used duringthe EFF SSL Observatory project, Internet Census, and the Mining your Ps and Qs scans (Electronic Frontier Foundation, 2014; Anonymous, 2012; Heninger et al., 2012). The ZMap tool dramatically reduces the time of scanning from days to as little as 45 minutes (Durumeric et al., 2013). In response to the release of ZMap, a new Internet-wide scanning tool Masscan was developed, which further reduces the time of scanning the Internet to a theoretical 3 minutes (Graham, 2013*b*). The development of these two tools has made conducting an Internet-wide scan easier, cheaper and more effective. Both ZMap and Masscan are tools designed specifically for conducting scans of the Internet, and provide significant performance improvements compared to NMap, a tool built for intensive local network scanning.

In comparison to conducting Internet-wide scans, the Shodan project allows users to bypass conduct-

ing scans themselves, and use use information gathered from conducting ports scans and banner grabbing to find information about specific target devices. Users search the Shodan database using an interactive web interface, and can search using queries designed to restrict searches to a type of device, port, or geographical location. Shodan captures information about many devices, including SCADA, ICS, IP cameras and routers (Shodan, 2014).

"KATSE" was a scanning system designed to scan the nation of Finland to constantly search for exposed, Internet-connected ICS and analyse the systems for possible vulnerabilities. KATSE, a several component scanning system, scanned devices which were found using Shodan (Tiilikainen, 2014).

The ZMap team at the University of Michigan released in August, 2014, an analysis of traffic dataset received by a darknet over a 16 month period. Through the use of libpcap, the traffic was analysed and the team found that scans conducted targeting 10% or more of the IPv4 address space did not use ZMap or Masscan (Durumeric et al., 2014).

Existing research projects using Internet-wide scanning and surveying methods have been conducted in the past few years. Several new methods of scanning the internet have been developed, dramatically reducing the time required to conduct a full Internet-wide scan.

However, limited global Internet-wide scans have been conducted specifically targeting ICS. Furthermore, limited comparison between the new Internet-wide scanning methods have been conducted. The majority of Internet-wide scans described have used different methodologies and tools while conducting scans of the Internet, several developing their own tools to fulfil their research needs. Thus there is a need to analyse and compare techniques through the development of a Taxonomy and develop a Framework for conducting Internet-wide scans. Additionally, there exists the potential to scan the Internet to view the current landscape of publicly available ICS, through scanning the public IPv4 Internet for industrial control system protocols.

3 Taxonomy of Internet Scanning Methods

Through our investigation of Internet-scanning tools and previously conducted Internet-wide scans, we have seen Zmap, Masscan and Unicornscan share a large number of similarities. As such, we have distilled the following categorical breakdown of the Taxonomy of Scanning Methods. We have then compared the properties of ZMap, Masscan and Unicornscan using our taxonomy, as shown in Table 1.

3.1 Scanning Method

We define Scanning Method in Internet-wides canning as the method used by the scanners to connect, check port availability and disconnect from a target host. We categorise Scanning method into two categories: scanners which conduct SYN-scaning, and scanners which conduct complete 3-way handshakes.

Zmap utilises separate sending and receiving threads for packet transmission, and uses SYN-scanning for sending packets (Durumeric et al., 2013).

Masscan makes use of SYN-scanning, and like ZMap and Unicornscan, uses separate sending and receiving threads to transmit packets and receive responses.

Unicornscan conducts a full three-way handshake while conducting a scan, and breaks down the process of conducting scans into three processes. The main process "Unicornscan" is used to control the scan and keep track of packets,"unisend" which sends a SYN packet to the scan target, and "unilisten", which listens for the SYN-ACK response, and sends information back to the master process "Unicornscan".

3.2 Packet Transmission

We define Packet Transmission in Internet-wide scanning as the method used by the scanner to send and receive packets. We categorise Packet Transmission in to three categories: scanners which use the kernel TCP/IP stack, scanners which implement their own self-contained "user-mode" TCP/IP stack, and scanners which bypass the TCP/IP stack completely.

ZMap generates and sends packets using a raw socket at the Ethernet layer, which reduces kernel overhead and bypasses the TCP/IP stack. By generating and caching the Ethernet layer packet, ZMap prevents the Linux kernel from performing a routing lookup, arpcache lookup, and netfilter checks for each sent packet (Durumeric et al., 2013). Masscan uses a user self contained TCP stack, separate from the Linux kernel. In addition to this function, Masscan makes use of a kernel module "PF_RING" to improve packet transfer and capture speed. Unicornscan's method has similarities to Masscan, using a user TCP stack outside of the kernel.

3.3 Randomisation

We define Randomisation as the ability of the scanning tool to generate a random permutation of the IPv4 address pool, preventing iterative scanning of the IPv4 address space.

Traditional network scanning tools, such as NMap, iteratively scan through a list of IP addresses. Due to the methods new scanners use to generate packets, more traffic is generated and transmitted faster, reducing the time required to conduct an Internet-wide scan. However, this results in the possible overload of a destination network, potentially causing issues to the normal operation of that network(Durumeric et al., 2013).

ZMap uses a mathematic method for generating a random permutation of the IPv4 address pool. Using modular mathematics, ZMap iterates over a multiplicative group of integers, ensuring the scanner will reach all IPv4 addresses, with exception to the address 0.0.0.0, an IANA reserved address (Durumeric et al., 2013). ZMap has recently been improved by including parallelised generation of IP addresses over multiple cores, allowing faster address generation (Adrian et al., 2014).

Masscan creates random permutations of the IPv4 address pool using a custom cryptographic algorithm "Blackrock", based on a Feistel network to encrypt an index. The Blackrock encryption function is based on Data Encryption Standard (DES) (Graham, 2013*a*).

ZMap and Masscan both have the ability to "seed" the randomisation element of the scans, allowing the random permutation of IP addresses to be repeatable.

From using Unicornscan to perform restricted local network scans, we found Unicornscan does not have a randomisation function, and scans iteratively through addresses in the specified network range.

3.4 Scan Distribution

We define Scan Distribution in Internet-wide scanning as the ability for the scanner to conduct dis-

tributed scans from multiple source hosts. We categorise Scan Distribution into two categories: scanners which can conduct distributed scans, and scanners which cannot conduct distributed scans.

Both ZMap and Masscan have the ability to conduct distributed scans of the Internet, and use the term "Shard" to describe the distributed hosts. ZMap and Masscan have similar methods of conducting distributed scans; first a "seed" is set to specify the same randomised address permutation over all hosts, then assign multiple IP addresses to scan from. Unicornscan does not have the ability to conduct a distributed scan from multiple hosts.

3.5 Blacklisting and Whitelisting

We define Blacklisting in Internet-wide scanning as a user created or edited list used to exclude IP addresses from scans, resulting with any address listed in a blacklist not being scanned at any point. We define Whitelisting in Internet-wide scanning as a user created or edited list used to specify a network range to scan, resulting in only that address or range of addresses being scanned.

We categorise Blacklisting and Whitelisting in to four categories: scanners which can use blacklisting, scanners which can use whitelisting, scanners which can use both blacklisting and whitelisting, and finally scanners which can use neither blacklisting or whitelisting.

ZMap can use both blacklisting and whitelisting for Internet-wide scans (Durumeric et al., 2013). Masscan can use blacklisting in the form of an "exclude file", but not whitelisting. Like ZMap, blacklisting is configured through a configuration file, accepting the same format as ZMap (Graham, 2013*b*). Unlike ZMap and Masscan, Unicornscan does not implement either blacklisting or whitelisting.

3.6 Modularity

We define modularity in Internet-wide scanning as the scanner being extensible with internal or external modules, to increase the functionality of the scanner or provide some additional benefit. We categorise modularity in to two categories: scanners which are modular, and scanners which are not modular.

ZMap is a modular scanner, internally having a series of extensible probe modules which can be customised for different types of probes and payloads, such as the UDP probe module (Durumeric et al., 2013). In addition to the internal probe modules, ZMap has output handlers which allow the scan results to be pushed into external modules to provide additional processing. Neither Masscan or Unicornscan have the ability to be extended with modules.

3.7 Scanning Speed

We define scanning speed in Internet-wide scanning as the speed it is theoretically possible to conduct an Internet-wide scan using a scanning tool. We categorise scanning speed in to three categories: 1gigE, scanners which can theoretically scan up the limit of 1gbps Ethernet Cards, and 10gigE, scanners which can theoretically scan up to the limit of 10gbps Ethernet Cards. These differences are determined on networking

The receiving component of ZMap utilises libpcap, a library for capturing network traffic and filtering results. ZMap can send packets close to the theoretical limit of a 1gbps ethernet card, approximately 1.5 million packets-per-second (Mpps) (Durumeric et al., 2013). Recently, ZMap has been further developed and optimised, improving the performance of address generation and "PF_RING" resulting in the ability to scan using a 10gbps network card, reaching similar speeds to Masscan at 15Mpps to 25Mpps (Adrian et al., 2014).

Through the PF_RING module, Masscan can use up to 10gbps Ethernet card to send packets at a maximum of 15Mpps, or up to 25Mpps using a dual-port 10gbps Ethernet card (Graham, 2013*b*).

Unicornscan uses a similar method of sending packets compared with Masscan, and additionally uses the libpcap library for receiving network traffic. (Lee and Louis, 2005). Using the same library, Unicornscan would be able to send packets at the same rate as ZMap, at approximately 1.5Mpps over a 1gbps Ethernet card.

3.8 Speed Limiting

We define Speed Limiting in Internet-wide scanning as the ability to slow or limit a scan's speed, in order to conduct the scan slower if necessary. We categorise Speed Limiting in to four categories: limiting speed by rate of scan in packets per second (pps) or bandwidth (G,M,Kbps), limiting speed by duration of scan (seconds), limiting speed by number of results, and limiting speed through a combination of methods.

ZMap has the ability to limit the speed of a scan by rate in packets per second (pps), by bandwidth in bits per second (G,M,Kbps), limit number of hosts and results, and by total time. Both Masscan and Unicornscan can limit the rate of the scan in packets per second, however neither Masscan or Unicornscan can limit by amount of hosts, results, or time.

4 Internet-wide Scanning Framework

In this section, we present our framework for conducting Internet-wide scans. We present our framework in four sections; a scan policy, a primary scan, secondary scan, and scan analysis.

4.1 Scanning Policy

While developing the Internet-wide scanning tool ZMap and conducting internet-wide scans as part of research, the team at the University of Michigan developed a list of seven recommended practices for future researchers to use as as guidelines for "Good Internet Citizenship" (Durumeric et al., 2013). We followed this list of recommended practices where it was feasible while developing our policies. The ZMap team's guidelines for "Good Internet Citizenship" were used as a base for implementing our two policies for ensuring all required parties are aware of any Internet-wide scans.

We have defined a clear policy to be used for communicating with internal groups at our ISP, ensuring all required groups are informed of the Internet-wide scans.

1. Request Ethics approval where necessary.
2. Inform and Discuss the nature and extent of the scans with the ISP.
3. Coordinate network usage with the ISP to prevent any disruption to normal network operation.
4. Coordinate with the ISP to ensure any emails will be received and processed by the scanning team.

Properties	ZMap	Masscan	Unicornscan
Scan Method	SYN-scanning	SYN-scanning	3-Way Handshake
Packet Transmission	Bypass Kernel	User-mode TCP/IP	User-mode TCP/IP
Randomisation	Uses Randomisation	Uses Randomisation	No Randomisation
Distributed Scanning	Can conduct	Can conduct	Can not conduct
Black/Whitelisting	Both Black & Whitelisting	Blacklisting	Neither
Scanning Speed	1gigE (10gigE as of August, 2014)	10gigE	1gigE
Speed Limiting	Combination (Rate, Duration, Results)	Rate of Scan (pps)	Rate of Scan (pps)
Modularity	Is modular	Is not modular	Is not modular

Table 1: Comparison of ZMap, Masscan and Unicornscan using our Taxonomy of Internet Scanning Methods.

In addition to this policy for working with internal groups, we have defined a clear policy for working with scan-traffic recipients to receive, and process any requests for information or requests to opt-out of any future scanning activities.

1. Maintain a constant, clear contact point for receiving any information or opt-out requests, through use of web pages, reverse-DNS and contact email.
2. Respond to information or opt-out requests promptly after receiving the request, ensuring responses are taken seriously.
3. Immediately add opt-out requests to an IP address blacklist for future scans, and update blacklist as soon as possible.
4. Refine the address blacklist as needed if a repeat request is received.

4.2 Primary Scan

The Internet-wide scanning framework has two scanning stages, a primary scan and a secondary scan. The primary scan is conducted to find a broader range of hosts to be narrowed down by the secondary scan. In the primary scan, an Internet-wide scan is conducted using ZMap, Masscan, or Unicornscan, against a port or ports necessary to obtain a range of IP addresses for research. The main outcome of the primary scan is a list of IP addresses to be used for the secondary scans.

4.3 Secondary Scan

The Internet-wide scanning framework uses a secondary scanning stage to further identify the initial hosts, in order to identify or gather more required information from the hosts. The secondary scan is conducted using the outcomes from the primary scan on a second port or ports which are used by common services, that have the potential to identify the devices. These services, when queried, can provide information or banners containing software versions and device information such as device name and type. These ports include the web server port TCP/80, Simple Network Management Protocol (SNMP) port TCP/161, Telnet port TCP/23 and File Transfer Protocol (FTP) port TCP/21. These protocols are commonly used to interact with ICS devices, for managing, accessign and uploading and downloading of files to the devices. Using this list of ports, scanning a list of IP addresses can return banners and status of the device.

4.4 Scan Analysis

The Scan Analysis section of the framework is for gathering insight from the results gained from the primary and secondary scans, such as statistical information, and geolocation information. Statistical information can be gathered from using standard UNIX tools. ZMap, Masscan and Unicornscan output files to multiple formats, and by default use a human readable format for viewing files. Extensions to these tools, such as banner grabbing modules, have the ability to output to the human readable to ascii format. Using unix tools such as *grep, wc, diff,* and *comm*, information can be observed from the results; such as how many IP addresses are in a range, and the number of times a server appears in a banner grab.

Our scan analysis stage uses geographical IP address information, retrieved from regional internet registries (RIR). Geolocation software uses databases of IP address data gathered from the RIR's to allow users of the software to search for geographic information related to an IP address, such as the approximate geographical location (Fiori, 2014). The results obtained through conducting primary and secondary scans can be processed through a GeoIP Server, and used as input to generate geographic maps of results to visually display scan results.

5 Discussion

While designing the Framework for Primary and Secondary scans, we initially considered sending messages to industrial control system protocols. Based on the responses we receive from the messages, we could quickly eliminate what devices were not ICS. However, this method of scanning would require us to craft packets to send commands as a payload specifically designed for the the destination protocol. A scan using this method of crafting packets could be construed as an attempted attack on the destination system. In addition to the construed nature of the scan, it is possible that commands sent to a destination industrial control system could potentially interrupt the function of the device. Due to possible misinterpretation of the messages, and the potential of interrupting the function of the devices, we eliminated this method of scanning as a possible way of identifying ICS. Instead, we designed the Secondary Scan as a method for identifying ICS without the use of specialised or crafted payloads.

Acknowledgements

This work was supported in part by Australian Research Council Linkage Grant LP120200246, Practical Cyber Security for Next Generation Power Transmission Networks.

References

Adrian, D., Durumeric, Z., Singh, G. and Halderman, J. A. (2014), Zipper ZMap: Internet-Wide Scanning at 10Gbps, *in* 'Proceedings of the 8th USENIX Workshop on Offensive Technologies'.

Anonymous (2012), 'Internet census 2012: Port scanning/0 using insecure embedded devices.'.
URL: *http://internetcensus2012.bitbucket.org/paper.html*

Durumeric, Z., Bailey, M. and Halderman, J. A. (2014), An Internet-Wide View of Internet-Wide Scanning, *in* 'Proceedings of the 23rd USENIX Security Symposium'.

Durumeric, Z., Wustrow, E. and Halderman, J. A. (2013), ZMap: Fast Internet-wide scanning and its security applications, *in* 'Proceedings of the 22nd USENIX Security Symposium'.

Electronic Frontier Foundation (2014), 'The EFF SSL Observatory'.
URL: *https://www.eff.org/observatory*

Fiori, A. (2014), 'freegeoip.net'.
URL: *http://freegeoip.net/*

Graham, R. (2013*a*), 'Masscan: designing my own crypto'.
URL: *http://blog.erratasec.com/2013/12/masscan-designing-my-own-crypto.html*

Graham, R. (2013*b*), 'Masscan: the entire internet in 3 minutes'.
URL: *http://blog.erratasec.com/2013/09/masscan-entire-internet-in-3-minutes.html*

Heninger, N., Durumeric, Z., Wustrow, E. and Halderman, J. A. (2012), Mining Your Ps and Qs: Detection of Widespread Weak Keys in Network Devices, *in* 'Proceedings of the 21st USENIX Security Symposium'.

Hoover, J. N. (2013), 'Thousands of industrial control systems at risk: Dhs study'.
URL: *http://www.darkreading.com/risk-management/thousands-of-industrial-control-systems–at-risk-dhs-study/d/d-id/1108149?*

Lee, R. and Louis, J. (2005), 'Unicornscan - History, Background and Technical Details', Presentation.

Shodan (2014), 'man shodan'.
URL: *http://www.shodanhq.com/help*

Tiilikainen, S. (2014), Improving the National Cybersecurity by Finding Vulnerable Industrial Control Systems from the Internet.