UNIVERSITY OF TWENTE.

Validation of Internet Census 2012

Bachelor assignment
B.Sc. Electrical Engineering

by

Dirk Maan
s1124420

Date:	February 12, 2014
Supervisors:	Dr. Ir. Pieter-Tjerk de Boer
	Dr. Anna Sperotto
	Jair Santanna M.Sc.
Institution:	University of Twente, Enschede, The Netherlands
Faculty:	Electrical Engineering, Mathematics, and Computer Science (EWI)
Chair:	Design and Analysis of Communication Systems (DACS)

Introduction

This document is the result of roughly six months of work on a bachelor assignment on the topic "Validation of Internet Census 2012". This assignment is the last requirement to fulfill before finishing the Bachelors curriculum and obtaining the Bachelor degree at the University of Twente, The Netherlands. In order to get more experienced in writing a scientific paper, the supervisors of this assignment approved that a paper was written to describe the obtained results.

At this place I would like to thank my supervisors for all advice and help they provided. I want to thank Jair Santanna for advising me during the entire process of the assignment and helping me with many practical issues that were encountered. He made sure that an optimistic attitude remained when troubles occurred in the process. Furthermore I am thankful to Anna Sperotto, for all the contributions and improvements to the paper. The suggestions she made during the research and the paper writing took the result to a higher level. Finally, I would like to thank Pieter-Tjerk de Boer for making me carefully consider the decisions made in my research and results. Besides this, I am also grateful that he provided the packet traces of his server to use in this study and gave me the opportunity to do this assignment.

An Approach to Validate the Internet Census 2012

H.C. Maan, J. Santanna, A. Sperotto, P.T. de Boer

University of Twente
Faculty of Electrical Engineering, Mathematics and Computer Science
Design and Analysis of Communication Systems (DACS)
P.O. Box 217, 7500AE
Enschede, The Netherlands
h.c.maan@student.utwente.nl,
{a.sperotto,j.j.santanna,ptdeboer}@utwente.nl

ABSTRACT

The Internet Census 2012 (IC) is an extensive dataset gathered from a scan that, according to the author, covers all IPv4 addresses. Although this dataset is very useful to provide the most recent picture of the entire Internet, the IC has a questionable reliability. The IC was gathered by compromising unprotected devices. In this research we provide a consistent approach to validate ICMP ping requests in the IC. To prove the suitability of our approach, we apply it in a case study, which compares the IC data with measurements of a /16 address block. Our results show that approximately 94% of host responsiveness records in the IC matches with local data. The achievements in our analysis indicates the correctness of our approach and the high reliability of the IC.

Keywords. Internet census 2012, Carna Botnet, validation, Internet scan, measurement, IPv4

1. INTRODUCTION

The term *census* means a statistical population sampling to study human or artificial populations. It is used to list all members of the population, and can also be used to focus on measuring end hosts of the Internet. In March 2013, an anonymous researcher published online the result of a project called 'Internet Census 2012', accompanied by a describing paper [1]. The author of the IC paper claims that a port scan was performed on the entire IPv4 Internet within one day, by using a distributed port scanner. This distributed port scanner is called the Carna Botnet, which consists of approximately 420 thousand compromised machines.

The dataset obtained from the port scan performed by this botnet is released on the Internet for further study. Because of its large amount of information, it may be a useful dataset for studies on device ports, firewalls, IP address allocation or device activity analysis. Besides this, it is of great importance when it comes to security awareness. In a blog post of the Cooperative Association for Internet Data Analysis (CAIDA) dating from May 2013, evidence is shown of the Carna Botnet scanning activity [2]. An increase of Nmap-hostprobes was observed in the CAIDA network telescope matching the IC data. This blog post shows the fact that the scan happened at the UCSD Network Telescope, which is a /8 address block without allocated IP addresses, also known as a darknet. Instead of using a darknet, we will apply our validation approach of IC data on a /16 address block which is partially allocated.

In order to use the IC data in future studies, it should be validated first. However, there is no dataset with the same characteristics (e.g. constructed at the same time, covering complete IPv4 address space) to validate it. Reproducing the dataset is not an option, due to the fact that the Internet has changed since the scan was performed, because of ethical issues, and because details of the Carna Botnet implementation are missing. So an important question arises, which will be treated in this study: *Can the data, as published on the website of Internet Census 2012, be trusted?*

In order to answer this question, we focus on validating this scan, i.e. showing if the scan happened and if the data in the IC are correct. Validation is started by pointing out that probes of the IC can be found in the network traffic of a single device. The device used in this study is a server using an IP address in the IP address block of the University of Twente (UT). If IC probes are found in the traces of incoming traffic on this server, we know that the scan was also performed on the UT address block. From that point on, the validation can be continued.

The next step in this validation is performed on a bigger scale. We compare the IC with a reference dataset, comprising a /16 IP address block. In order to validate the IC data concerning this /16 IP address block, the IC data of each IP address in this block is compared to the information of the same IP address in the reference dataset to see to what extent these datasets are similar. The more similar these datasets are, the more the IC is considered to be trustworthy.

The remainder of this paper is organized as follows: first, Section 2 summarizes related work, followed by an overview of the IC dataset in Section 3. Section 4 describes the proposed approach. Section 5 analyzes the results of our approach applied to a /16 IP address block. Section 6 summarizes our findings and proposes future work.

2. RELATED WORK

Since the beginning of the Internet, scans are performed to obtain information about the end hosts. What is known to be the first documented scan of the Internet is described in RFC-832 [3]. At that time, in 1982, 315 hosts were probed to see if they use the TCP protocol, which took less than one day. In 2003, Heidemann et al. performed a complete Internet scan by using two hosts [4]. The observations from this scan were validated by comparing them to scans of smaller address blocks (called surveys). This study only considers ICMP probes, as TCP caused thirty times more complaints than the use of ICMP.

Furthermore, Holz et al. conducted HTTPs scans of the top million popular hosts over a timespan of 1.5 year [5]. These scans were horizontal, as only port 443 was probed. The IC differs at this point, as the top 100 ports and several other random ports were scanned for about 660 million IPs in the IC. Moreover, the scan performed by Holz et al. scanned for certificates, opposed to the goal of estimating the IPv4 address usage, aimed for in the IC.

In 2011, a scan was performed on the entire IPv4 address space (a "/0" scan), described in Dainotti et al [6]. A botnet called Sality, which comprises approximately 3 million distinct IP addresses, took 12 days to complete this horizontal scan. This scan was observed and validated at a darknet.

Another Internet-wide scan which is similar to the IC was performed by Durumeric et al [7]. To the best of our knowledge, it is the most recent documented scan of the complete Internet. In their study a scanning tool like Nmap is used, but specifically designed to perform scans at a large scale like the complete IPv4 address space. This tool is called ZMap. The study of this scan is different from the IC because one device was used to perform the scan, instead of a coordinated network of computers used in the IC.

3. INTERNET CENSUS 2012

In this section, the Internet Census is described. Section 3.1 describes some general characteristics of the IC. Subsequently, Section 3.2 covers the time distribution of the IC scans. It is followed by Section 3.3, which explains the content of the IC traces and points to some examples and eventually Section 3.4 describes the rounding of timestamps in the IC.

3.1 IC Dataset

The dataset considered in this study, i.e. the IC, is a scan of the Internet gathered by the Carna Botnet [1]. The Carna Botnet was created by the anonymous author of the IC, to distribute the scanning process. In order to construct this botnet and perform the scan, the Nmap Scripting Engine (NSE) was used [8]. According to the paper that accompanied the IC, five scanning methods were used. These are called ICMP Ping, Reverse DNS, Nmap, Service probes and Traceroute. The outcome of these methods resulted in several datasets, called traces, which are combined in the IC. The data gathered by Nmap is split in traces '*hostprobes*', '*syncscan*', and '*TCP/IP fingerprint*'.

Table 1: Traces in IC

Trace	Content
ICMP Ping	Responsiveness and latencies
Reverse DNS (rDNS)	DNS records
Serviceprobes	Services behind open ports
Hostprobes	Responsiveness
Syncscan	State of ports
TCP/IP Fingerprint	Type of device and operating system identification
Traceroute	Path of data packet

All IC traces are listed in Table 1. In our research, the traces *'icmp_ping'*, *'hostprobes'* and *'syncscan'* are of special interest, because these traces indicate if a device was active or not at a certain timestamp. These three traces will be the only traces considered in the following sections. By matching these traces with a reference dataset containing host activity data, we validate the IC data.

The reference dataset used in this study is a /16 IP address block. Therefore, the analysis performed in the sequel are all related to a /16 address block.

3.2 Time distribution

As can be seen in Figure 1, the traces icmp_ping, hostprobes and syncscan have a different frequency distribution in the considered /16 address block.

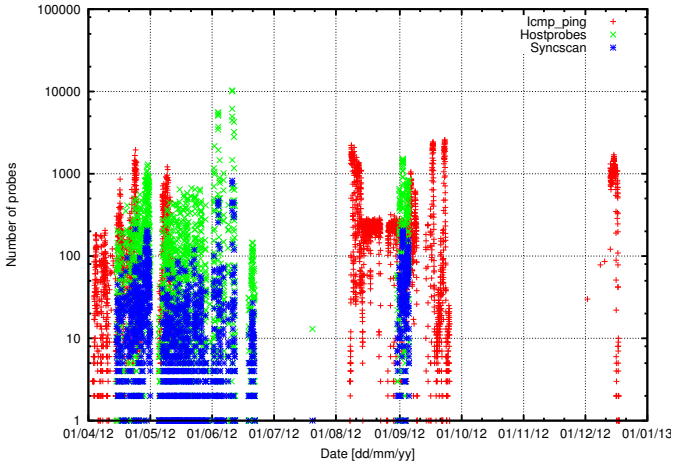

Figure 1: Number of probes per day for 130.89.0.0/16

Figure 1 shows the number of probes that were sent at a certain time. The red color indicates the number of icmp_ping probes, green refers to probes of the trace hostprobes and blue shows the number of syncscan probes. From the plot can be deduced that the maximum number of probes per timestamp varies per trace. For example, it is above 10000 in the hostprobes trace, just above 2500 in icmp_ping and

above 800 in syncscan. However, the syncscan probe frequency is not reflecting the actual amount of ports probed, because the probes of multiple ports were stored in just one entry. To see an example of this, refer to Table 4. The frequency plots also show that the probes of hostprobes and syncscan occur at the same time, as can be seen by the equal vertical axis points. This behaviour matches the description in the IC paper, stating that syncscan was preceded by hostprobes to determine if a device was alive.

3.3 IC trace contents

Each entry of the traces contains at least three data elements: an IP address of the probed device, a timestamp indicating the moment of probing and some data about the device. Which data is included in each trace depends on the scan method used. For an example of the possible entries, we refer the reader to Table 2, Table 3, and Table 4 in Appendix A. These tables show some entries of each studied trace.

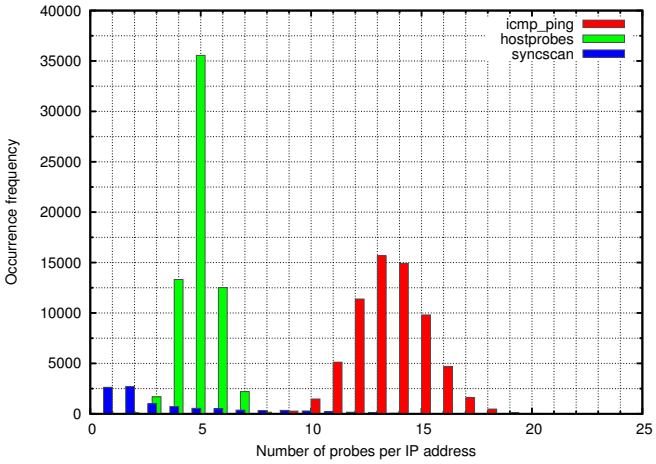

Figure 2: Occurence frequency of the number of probes per IP address

By inspecting the IC traces, it becomes clear that not every IP address is probed once. In some traces it is even more common that an IP address is probed multiple times, as shown in Figure 2. The x-axis of this figure is the number of probes that is sent to an IP address, while the y-axis shows how many IP addresses are probed. Icmp_ping is shown with red bars, hostprobes with green bars, and syncscan with blue bars. For example, in the hostprobes trace more than 35000 IP addresses were probed five times. From this figure can be deduced that the number of probes per IP address is really different for each of these traces. The most occurring probe frequency is reflected by the highest bar in the figure. For syncscan, the most occurring probe frequencies are one or two probes per IP address, but hostprobes has a clear top at five probes per IP address. Icmp_ping is centered around 13 probes per IP address.

From Figure 2 we also conclude that each IP address may be marked both as alive or unreachable in the IC trace icmp_ping during separate time intervals.

3.4 Timestamp rounding

The IC paper does not explain anything about the probe timestamps in the IC data. It is unknown what instance created the timestamps and when. Possible options are the probed hosts, the probing bots or some central server collecting the data. The timestamp used in the IC is assumed to be the Unix time. The Unix time (or Unix timestamp) is a system for describing points in time, defined as the number of seconds elapsed since midnight proleptic Coordinated Universal Time (UTC) of January 1, 1970, not counting leap seconds.

During analysis of the IC traces, it became clear that any timestamp in the IC can be divided by 1800, which means that the time interval between two IC probes with a different timestamp is always a multiple of 1800. Besides this, dividing the timestamp by 1800 always returns a remainder of 900 seconds, which is equal to 15 minutes. Therefore, it is concluded that all timestamps can be translated to a time with 15 minutes before or after full hours, e.g. times of 10:15 or 08:45 are possible, but 07:14 is not.

Nothing is described either about this timestamp rounding in the IC paper. Timestamps could have been rounded in different ways, e.g. always up, always down or to the nearest possible timestamp. The introduced uncertainty is therefore approximately 3600 seconds, ranging from 1800 seconds before the IC timestamp to 1800 seconds after it.

4. METHODOLOGY

In order to validate the IC completely, there is one requisite: to check if each entry in the IC is either right or wrong, a dataset is needed that covers all records stored in the IC. For a dataset to be suitable to use in the IC comparison, it has to consist of at least three elements. First, this dataset has to contain all IPv4 addresses. The paper of the IC states that the entire IPv4 address space was probed. So to validate this, there should be at least one record of each address in the IPv4 address space to compare with.

As a second requirement, the dataset has to contain some state or info about each IP address. In the IC are traces that contain information about the included IP addresses, e.g. what ports are open or if it is alive according to ICMP. In a validation, this information should be compared to the information of another dataset to see if the information of the dataset and the IC corresponds.

Finally, the information that corresponds with each IP in the dataset has to be collected at the same time interval in which the IC was operating. Therefore, the dataset should include timestamps that are in the same operating interval as the IC. Using the validation approach presented in the following subsections, it is possible to perform a validation on the entire IC.

Unfortunately, there is not such a dataset of the entire IPv4 address space available. In this study, the validation is therefore carried out in two ways. At first, the IC is analysed at a small scale: for a server with one IP address, the traces of all incoming traffic are used to find an indication of IC activity. Using this small scale analysis, traffic patterns of the IC can be observed closely. Subsequently, a comparison

with a reference dataset of a /16 address block is used to validate the IC partially. Claims made in the IC are validated for a bigger range of IP addresses, namely the University of Twente /16 IP address block.

4.1 Single IP analysis

For a small scale analysis, the traces of all incoming traffic of a single machine are compared with the IC data. The best trace of the IC to use for this goal is the syncscan trace, because of its many probes in a short time interval. The hostprobes and icmp_ping IC traces are not inspected, because they are based on ICMP echo requests. ICMP echo requests are common in the traffic of the analysed machine. Therefore, it is hard to identify the IC probes in regular traffic.

From the syncscan trace the probed ports and the time of probing are known. This is enough data to filter the IC probes from other incoming traffic at the server. To validate each entry that corresponds to the analysed IP, two things are considered: first, it is checked if the machine was actually approached by the IC at the stated timestamp. Furthermore, the state that was reported in the IC has to be the same as the actual state of the machine.

The machine that is used in this research is a server with an IP address in the block of the University of Twente, i.e. 130.89.0.0/16. We analyze a dataset consisting of the packet traces of the traffic reaching the considered host, considering the time interval in which it was probed by the IC. Relevant parts of these packet traces are the timestamp of the packets, the source IP address and port, the destination IP address and port, and the state of the ports that was replied to the IC.

To see if comparison is possible, it is checked if the IP of our server is in the IC syncscan trace. If the entries of the syncscan trace that contain the server IP address are found, the corresponding timestamp is searched in the traces of incoming traffic of the server. The interval in which the probe was sent is determined by extending the timestamp, as described in Section 3. Subsequently, the traces with incoming traffic are searched within this interval for the probed ports according to the IC.

Furthermore, the state of the ports that is in the IC is validated with the reponses of the server gathered from the traces of incoming traffic. For instance, when the IC reports some port of the server to be open, the server should have replied to a probe in the indicated interval. No probe should be found when the IC reports a port to be filtered, because this points out that the probe packet was filtered before it received the host.

4.2 /16 address block analysis

4.2.1 *Comparison of two datasets*

In order to validate the IC, a comparison with a reference dataset is made. Both datasets consist of three key elements: an IP address, information about the host such as which ports are open, and a timestamp. The reference dataset consists of consolidated two-hour snapshots of the ARP table, maintained by the routers in the /16 block[1] that is assigned to the University of Twente (UT). This comparison can also be performed using other reference datasets, if the requirements for reference datasets stated in Section 4 are met.

The UT ARP data is gathered during the year 2012, which overlaps the interval of IC activity completely. For the considered /16 address block, this dataset contains all requirements that are stated in the beginning of Section 4: the ARP data contains timestamps with the interval in which an IP was active, it contains an IP address, and it implicitly shows an IP was active, because it is registered in the ARP data. Because these three elements are all in the dataset, it is suitable to use in a comparison for a partial validation of the IC. However, the ARP table contains small noise factors that influence the registration interval for several IP addresses. Several gaps in the dataset are present, due to SNMP timeouts that occurred or because the database table space was temporarily full. Although this reference dataset has some missing data, the comparison performed with it is still valid.

The first element to be compared is the IP address. Comparing timestamps or probe information makes no sense when they do not belong to the same IP. Comparing the IP addresses of each entry in two datasets, results in three different subsets, which can be visualized by a Venn diagram. An example of the IC compared to a reference dataset is shown in Figure 3. The three subsets resulting from the comparison are referred to as subsets A, B, and C. In this descriptions, we are assuming that the reference dataset is correct and precise. These subsets are described in the following paragraphs, starting with subsets A and C. Subset B is described last, because it contains the most thorough analysis.

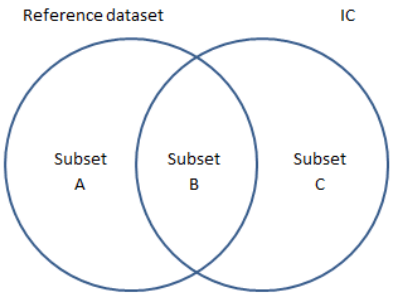

Figure 3: Venn Diagram of the IC compared to a reference dataset

Subset A. This subset contains all entries of the ARP table with an IP address that does not occur in the IC. By checking if any IP addresses of the /16 address block of the UT are present in this subset A, indications of errors in the IC are pointed out. The IC paper claims that the entire IPv4 address space was scanned. When IP addresses of the

[1]The description 'block' is preferred instead of 'subnet' because we refer to the address block, not the router configuration, in accordance with the terminology used by Heidemann et al. [4]

analysed address range in the IC occur in subset A after this comparison, there is an error in the IC, because the IP should have been in the IC data, i.e. subset B or C. On the contrary, if this subset A is empty, it implies that none of the IP addresses registered in the ARP table were skipped by the scan. This is one indication of the validity of the IC.

Subset C. By analyzing subset C, the IC entries with an IP address that is not present in the ARP table are found. This means that an IP address is probed by the IC when it was not registered, i.e. when it was not active, according to the ARP table. No problem occurs when the IC marks these hosts as being unreachable. However, when the IC states that a host is alive though its IP address is not in the ARP table, this is considered an error. Eventually, an IP address that is a public IP address, should be registered in some ARP table. Otherwise the IP destination will never be resolved to a physical address (MAC address) and packets that are destinated to this IP addresses will not arrive at their destination.

Subset B. This subset is the intersection of the both datasets, i.e. the entries of the IC and the ARP table that have matching IP addresses. Hosts that were active according to the ARP table and where also probed by the IC will be in this subset. This last subset will be highlighted in this study, because it requires a more thorough analysis. Entries that have matching IP addresses and are therefore in this subset, can be compared again. The timestamps are the next subject of comparison.

4.2.2 Subset B timestamp analysis

The following paragraphs describe the comparison based on time performed on the probes in subset B. First, the expansion of the IC probe timestamp is explained, followed by the four overlaptypes considered in the time based comparison.

Timestamp expansion. The complete interval in which the IC probe can be sent, should be considered. Recall from Section 3.4 that the exact moment of probing is unknown, due to an uncertainty of 1800 seconds around each IC timestamp. We want to eliminate this uncertainty to be sure that a probe was sent in a certain time interval. To realise this, every timestamp is expanded to create an interval. The start of the interval is determined by subtracting 1800 seconds, the end is determined by adding 1800 seconds. By using these intervals instead of the IC timestamps, the uncertainty caused by the rounding is eliminated.

The time interval in which an IP address was considered active is described in the ARP table by two timestamps. The first timestamps indicates the start of the interval and the last timestamp indicates the end. Comparing the interval of the ARP table with the interval of the IC will result in four possible outcomes with four different conclusions. Refer to Figure 4 for the different cases, which will be called overlaptypes further on, that could apply. The interval of the IC is referred to with the dashed line, while the interval of the ARP table is shown with the solid line.

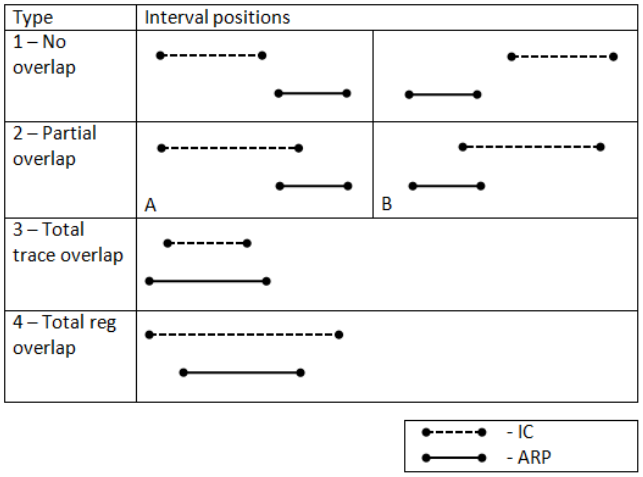

Figure 4: Different overlap possibilities.

When using this approach, it should be kept in mind that an IP could be registered at multiple time intervals. For simplicity it is assumed that these time intervals do not overlap. Because multiple ARP intervals for the same IP could therefore exist, it is possible that more than one of the cases mentioned in Figure 4 applies, e.g. *Partial overlap* and *Total reg overlap*. In this case, only one of the overlaptypes is matched to the IC record. The order of these overlaptypes is equal to their occurence when sliding two intervals over each other from left to right: *Partial overlap A*, *Total trace overlap*, *Partial overlap B*, *Total reg overlap*, *No overlap*.

No overlap. When a comparison is performed on two intervals, it is possible that no overlap between the IC trace and the reference trace exists. This situation occurs when an IP address was registered according to the ARP table, but the IC probed the IP address later or earlier than it was registered in this ARP interval. Therefore, the IC should not state this IP address as being alive, because the IP was only considered alive during the ARP interval. However, if it is marked as alive, it is assumed that there's an error in the IC.

Partial overlap. A second case that can occur is partial overlap. Assume that an IP address was marked as active by the ARP table in a certain registration interval. Now assume that the IC interval of this IP overlaps this registration, but not completely. Because of the uncertainty in the trace interval, it is not clear if the probe was sent during the overlapping part of the trace and registration or not. Therefore, a conclusion about the validity of this record can not be drawn from this data alone.

Total IC overlap. In this case the IC interval is entirely within the time boundaries of the ARP time interval. No uncertainty is left for this situation: it makes no difference in what moment the actual probe was sent, because the IP address was considered active according to the registration interval during the complete trace interval. The information

in the IC should therefore contain an indication of the IP address being alive. If this is not the case, the IC might contain an error.

Total ARP overlap. The last case can only occur when the trace interval is longer than the registration interval. It is characterized by the registration interval being completely overlapped by the trace interval. The start of the trace interval begins earlier than the start of the registration interval and the trace interval ends later than the registration interval. Similarities exist between this case and the case of partial overlap because of the uncertainty in the exact moment of the IC probe. If the IC probe was sent somewhere during the overlap of the trace and the registration, the IC data and ARP data should correspond with each other, i.e. the IC should report some kind of information that indicates that the IP was alive during the interval. But if the IC probe was sent at a moment near the borders of the trace interval when no overlap with the registration interval exists, the IC should report an unreachable state. When the IC data deviates from these assumptions, an error in the IC is assumed.

4.2.3 Validation

In this study, we focus in particular on the erroneous entries in the IC, because it is an indication of the reliability of the IC. An erroneous entry is defined as an entry that contradicts the data in the ARP tables of the UT, e.g. when an IP is reported as 'alive' in the IC while it was not according to the ARP tables.

Basically, the comparison consists of the following steps, which will be elaborated further on:

- Split IC trace in unreachable and alive subtraces;
- Determine the appropriate subset for each entry;
- Count probes per subset;
- Find errors.

In order to clarify the comparison in a more visual way, we refer the reader to the flowchart in Figure 5. This figure shows how an IC probe is categorized in a certain subset. Furthermore, it shows what type of IC probes are erroneous, as explained in Section 4.2.2.

Split IC trace. As stated in Section 4.2.2, the conclusions that can be drawn from the comparison are dependent of the state that is reported in the IC. By splitting the IC trace in a subtrace that contains all 'alive' entries and another subtrace containing all 'unreachable' entries, the analysis is simplified. The icmp_ping and hostprobes traces directly report an alive or unreachable state of a machine.

Besides this, the syncscan trace can also be used to determine if a device was active during the probing. The information of these probes comprise the state of some ports, of which the most common are 'closed', 'filtered', or 'open'. A 'closed' state means that this port on the probed device receives and responds to probe packets but does not have an active application listening on it. The 'filtered' state indicates that it is not possible to determine if the port was open or closed, since packet filtering prevents the probe packet to reach the host. The 'open' state reveals that a port responds to probes and an application behind the port is actively accepting packets. The IP addresses with 'closed' and 'open' port states thus indicate that a device is alive at the moment of probing. A subtrace is created with probes comprising these two states, which is similar to an 'alive' subtrace. The 'filtered' state however is not considered, because from this state can not be deduced if a device was alive or unreachable. Therefore no 'unreachable' subtrace is created for the syncscan trace.

By analyzing the IC syncscan trace considering the /16 address block of the UT we observe that 3,5% of the probes is obtained using the UDP protocol. Refer to Table 4 for an example of a probe sent using the UDP protocol. These probes are not considered in this study, because they cover such a small percentage of all probes. The remaining probes of this trace are sent via the TCP protocol. These probes are used to determine whether a probed host was alive or unreachable.

Determine the appropriate subset. In order to find the subset that each entry of a subtrace belongs to, the IP addresses are inspected and compared to the IP addresses in the ARP table, as shown in Figure 5. The entries that appear to be in subset B, have to be analysed even further. The timestamp analysis of 4.2.2 is performed to categorize each entry in this subset.

Count probes per subset. By counting the number of elements in each subset, statistics about the comparison are provided. Each entry per subset is counted to indicate the distribution of the probes over the subsets. The entries that are part of subset B are counted separately for each overlaptype, because each overlaptype is associated with a different type of error. For example, the subtrace of 'alive' entries that belong to the overlaptype *Total trace overlap* in Figure 4 are considered valid, but the 'alive' entries that are counted in the overlaptype *No overlap* are considered erroneous.

Find errors. Checking the entries that were assigned to subset A is required once more to validate the assumption that was made in Section 4.2.2. It was assumed that no records of the considered address block can end up in this subset. When subset A does not contain any public IP addresses, this is an indication of the completeness of the IC. If no IP addresses are in subset A, it can be stated that no IP addresses of the total /16 address block of the University of Twente were skipped in the scan. Furthermore, the percentage of erroneous probes per trace are described, to get an impression of the IC validity.

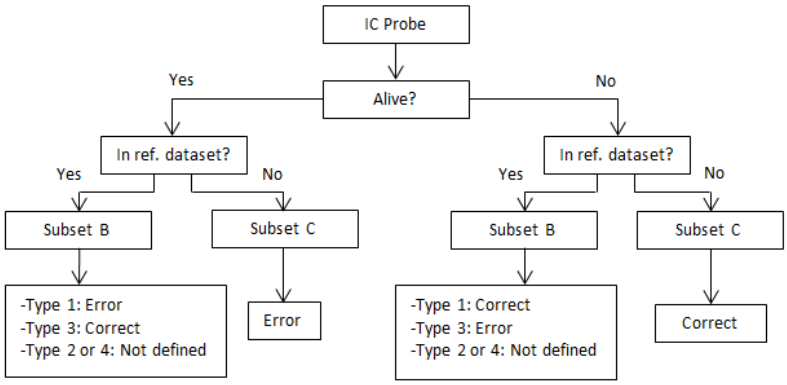

Figure 5: Flowchart of the comparison process

5. RESULTS

Similar to Section 4, the results are described in two parts and ordered in the same way, first the single IP analysis is discussed, followed by the results of the /16 address block analysis.

5.1 Single IP analysis results

Performing the method described in Section 4.1 on the given traces, results in the conclusion that the IC records concerning the server correspond to the traffic traces of the server. At first, the IP of the machine appears to be present in the IC records of the syncscan trace. Furthermore, packets have been sent to the machine during the IC probe intervals, according to the traces of incoming traffic of the server. By inspecting the packets in these intervals, it becomes clear that the packets sent to the ports described in the IC, come from one IP address. This IP address probably belongs to a compromised device of the Carna Botnet. Most of the probes were replied with a packet having the RST and ACK flags set, revealing that the probed host exists but has these ports closed. The ports of the server IP were marked as 'closed' in the IC, which is as expected. Refer to Table 6 (in Appendix C) for an overview of all probed ports with the according server responses.

By comparing the timestamps of the packet traces of the single host to the probe timestamps in the IC, the rounding of IC timestamps was observed. From this observation is concluded that for the syncscan trace and this host, the IC probe timestamps were rounded down with respect to the probing timestamps observed in the packet traces. Although we only have this data of one IP and one trace, it is assumed that all timestamps in the IC are rounded down. Even though this assumption reduces the timestamp uncertainty, it is not applied in the timestamp expansion as described in Section 4.2.2 because of a lack of time to implement it. However, it is mentioned for further research.

By inspecting the packet traces of the host considered in this study, we observe the following:

- When the probed host replies by sending a packet with flags [RST,ACK] set, the IC reports a closed state accordingly.
- When the IC probe does not reach the probed host, the IC reports a filtered state accordingly.
- When the IC probe reaches the probed host and the host does not reply but drops the probe, the IC reports a filtered state accordingly.
- Some ports are probed multiple times. For example, consider port 80 in Table 6, which is reported as 'closed' by the first probe, but reported as 'filtered' in the seventh probe.

From these observations is concluded that the probes are interpreted correctly in the IC, considering the probes that were received and replies sent by our host. Besides this, we conclude that when a probe is lost, it will result in a 'filtered' state of the probed IP address.

5.2 /16 address block analysis results

In order to validate a /16 subset of the IC, an ARP table with activity data of the entire UT address range is compared to the part of the IC concerning the same IP address range. In conformity with the division in subsets of Section 4.2.1, the results of the /16 address block analysis are split in three parts, because each of these parts requires a different analysis. The result of the interval comparison in subset B is an overview of the occurrence of different overlaptypes described in Figure 4.

5.2.1 Subset A

Subset A contains all entries of the ARP table that do not have a matching IP address within the IC traces. Comparing the ARP table with the icmp_ping traces resulted in an empty subset A. This indicates that no IP addresses registered in the ARP table were skipped in the icmp_ping scan. Furthermore, the ARP table is compared to the hostprobes trace, resulting in an empty subset A. So for the hostprobes scan also holds that no IP addresses registered in the ARP table were skipped.

Last, the IP addresses in the ARP table are compared to the syncscan trace. Differing from the comparison with hostprobes and icmp_ping traces, 71.4% of the ARP table IP addresses appeared to be not matching with the IP addresses in the IC traces. This does not necessarily reflects an error in the IC. The author of the IC paper actually states

that a syncscan was limited to only about 660 million IP addresses [1]. So it just reflects that nearly all IP addresses in the ARP table were not probed during the syncscan.

5.2.2 Subset C

By collecting all IC entries with an IP address that is not matching any IP address in the ARP table, subset C is formed. The percentage of these entries with respect to the total probes in each subtrace (e.g. hostprobes_alive or icmp_unreachable) is shown in Figure 6. The first overlaptype on the x-axis (noipmatch) shows the percentage of entries that are categorized in subset C. Refer to the column of Subset C in Table 5 of Appendix B for more numerical results of nonmatching IP addresses.

According to the characteristics described in Section 4.2.2, no IC entries marked as alive should be in this subset. As shown in Figure 6, the alive subtraces of hostprobes, icmp_ping and syncscan contain 5.31%, 2.82%, and 3.94% respectively nonmatching IP addresses. These are considered erroneous.

However, further analysis shows that several alive IP addresses of trace icmp_ping that are categorized in this subset are actually broadcast addresses or network addresses. These addresses are categorized in subset C, because they are not included in the ARP table of the UT. ICMP echo requests sent to these addresses are generally replied by routers of the UT. The IC reported these addresses as alive, due to the reply of some UT router that was received. Therefore, these probes are not incorrect, since the probed IP addresses were allocated, but not registered in the ARP table. Since we do not have access to a list of all broadcast and network addresses, we can not determine all IP addresses in subset C that are broadcast and network addresses.

In addition, several hosts reported as alive in this subset were probed close in time to moments were ARP table errors occurred. This can be another reason why some IP addresses in the /16 address block that were alive according to the IC are not present in the ARP table. From this observation is concluded that alive entries in this subset are not necessarily errors in the IC.

Figure 6 shows that more than 50% of the probes in the unreachable subtraces of hostprobes and icmp_ping do not match the IP addresses of the ARP table. These probes are consistent with the ARP data, due to the fact that IP addresses can not be alive without them being registered in the ARP table. Hence these probes are considered correct.

5.2.3 Subset B

By categorizing the records of subset B into separate overlaptypes, the validity of each subtrace can be determined. The result is a table with the number of probes that is counted for each overlaptype, as shown in Table 5.

Refer to Figure 6 for the cumulative distribution function plots of each subtrace. In this Figure, the line indicates the percentage of counted probes. On the x-axis are the possible overlaptypes (i.e. no_overlap, partial_overlap, totaltraceoverlap and totalregoverlap), the y-axis displays the CDF. From this graph, it becomes clear which part of the total probes in a trace is categorized in each overlaptype.

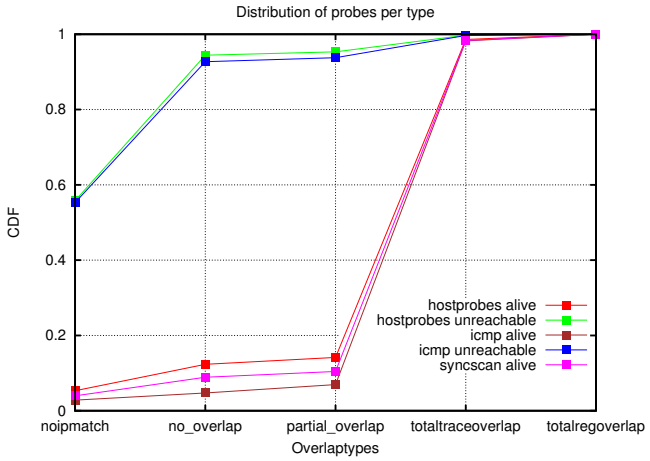

Figure 6: Cumulative Distribution Function of overlaptypes in subtraces

The following two paragraphs summarize the observed distribution of the subtraces. In Section 5.2.4 and Section 5.2.5 the analysis continues with a determination of correct and incorrect probes.

Unreachable. As seen in Figure 6, the largest increase of the unreachable subtraces occurs in between overlaptypes noipmatch and no_overlap. Because these traces add up to more than 90% of all probes in these traces, we can conclude that most of these probes did indeed not reach active hosts. From these probed IP addresses can be said that their unreachable state is noted correctly in the IC.

Alive. Another thing standing out in Figure 6 is the large percentage of the totaltraceoverlap in traces hostprobes_alive, icmp_alive, shown by a clear increase from overlaptype partial_overlap to totaltraceoverlap. The share of total probes in these subtraces in overlaptype totaltraceoverlap of subset B is 84.46% and 91.30% respectively. The probes of these traces in this category are correct, i.e. these probes in the IC data correspond with the entries in the ARP table. A similar percentage of 87.90% totaltraceoverlap is observed in the syncscan_alive subtrace.

5.2.4 Errors per trace

By counting the occurences of overlaptype no_overlap, the amount of errors for subset B in the alive subtraces is obtained. For the hostprobes_alive subtrace, this is equal to 7.01% of probes. Regarding the icmp_alive subtrace, the share is 1.90% of all probes in this subtrace. Subtrace syncscan_alive is just in between these two percentages, since 4.94% of its probes has no overlap with the ARP table.

When the number of overlaptypes no_overlap (part of subset B) and noipmatch (part of subset C) are now summed, the total amount of errors in the alive subtraces is obtained. This is equal to 12.33% erroneous probes in subtrace hostprobes_alive, 4.73% in subtrace icmp_alive, and 8.89% in subtrace syncscan_alive.

In order to give an indication of the amount of incorrect probes in the unreachable subtraces, the probes that definitely should have been marked as alive, are counted. Recall from Section 4.2.2 when the trace interval of the probe completely overlaps the registration interval of the according IP in the ARP table, it should be marked as alive. Thus any probes containing the unreachable state that are classified by this overlaptype totaltraceoverlap are therefore erroneous. In this assumption, we do not take into account that an IP could remain registered in the ARP table for a certain period after disconnecting from the Internet. This is equal to 4.41% of probes in hostprobes_unreachable and 5.91% of probes in icmp_unreachable. The reader is referred to Table 5 for these results.

At last, the alive and unreachable subtraces are combined again to determine the portion of errors for each trace. Combining these two subtraces of hostprobes results in a percentage of 4.95% incorrect probes in this trace. By merging the count of errors in the subtraces of trace icmp_ping, 5.84% of probes in the trace icmp_ping appears to be wrong. For the syncscan trace, this is equal to 4.11%.

5.2.5 *Trace correctness*

In order to give some correctness indication for each IC trace, considering the analysed block, the number of correct probes are summed. Probes are considered to be correct if the information about an IP agrees with the data known by local resources like an ARP table. In this analysis, correct probes are the probes in an unreachable subtrace, with no matching IP in the local data, i.e. probes of an unreachable subtrace in subset C. Moreover, in subset B probes are correct if they belong to an alive subtrace and overlaptype totaltraceoverlap. Correct probes can also be found in subset B if probes belong to an unreachable subtrace and are categorized in overlaptype no_overlap.

Using these strategy, the minimum of correct probes in each trace is calculated. By adding up these numbers, at least 93.75% of probes in the hostprobes trace appears to be correct. Moreover, icmp_ping contains at least 92.63% correct probes. The share of correct probes in syncscan is at least 40.69%.

The maximum correctness percentage is obtained by subtracting the percentage of errors from the total amount of probes in each trace. This comes down to maximum correctness percentages of 95.05%, 94.16%, and 95.89% for the traces hostprobes, icmp_ping and syncscan respectively.

Finally, this results in the correctness ranges of each IC trace. Regarding hostprobes, the correctness is in between 93.75%-95.05%. For icmp_ping this correctness is equal to 92.63%-94.16% and in syncscan, 40.69%-95.89% is correct. IC trace syncscan has a considerable smaller lower bound. This is caused by the fact that only the alive subtrace of syncscan was considered to determine correct probes.

6. CONCLUSIONS

Summarizing, our work shows the successful validation of more than 90% of the IC probes in the traces icmp_ping and hostprobes, considering a /16 address block. As a result of this validation, the correctness of our proposed approach and the partial validity of Internet Census 2012 dataset is attested.

Using the incoming traffic traces of a single host, it was validated that the IC scan included devices in the /16 address block of the UT. We were able to identify the syncscan probes of the IC in the normal server traffic, which is a strong indication that the IC was also performed on the 130.89.0.0/16 address block of the UT. We showed that the syncscan probe timestamps were rounded down with respect to the probe timestamp as observed on our server. Although this study assumes the timestamps to be correct, it is very difficult to prove this, because the origin of the timestamps is unknown. It is unknown where (e.g. at the probing machine or a collecting machine) and when (e.g. during the probing or afterwards) the timestamps were determined.

Many error sources could have affected the IC scan when it was performed. Some of the possible error causes might be packet loss or bot misconfiguration. Missing information could for example lead to an incorrect unreachable state of an IP address in the IC. Different from the hostprobes and icmp_ping traces, the syncscan trace is only validated for about 40%. Many probes of this trace were neglected in the process, e.g. probes sent by the udp protocol and probes that have the state 'filtered'. Therefore, we consider this result is not really accurate and should not be used as a measure for correctness of the entire syncscan trace.

The IC is a snapshot of the Internet, which is obtained during a total time interval of roughly eight months [1]. The Internet is changing a lot because of the large increase of connected devices, according to Cisco's major global mobile data traffic projections and growth trends [9]. So any conclusions drawn from this dataset can not be translated directly to the current state of the Internet.

Several opportunities for future work exist. By using traces of incoming traffic of various servers, the rounding method of the IC can be identified. Furthermore, other address blocks of the IPv4 address space can be validated using the proposed method. When more address blocks in the IC are validated, a better conclusion about the validity of the entire IC can be drawn. Another possibility for future research is the validation of IC traces that were skipped in our research. The serviceprobes and tcp_ip_fingerprint traces for example contain Nmap data about the devices that were scanned. This Nmap data offers other information about the scanned devices than the data observed in this study. A few examples of this are services run on the scanned ports (e.g. HTTP, TCP, SSH), the device type (e.g. printer, router), version number, OS family (e.g. Windows, Linux), etc. Moreover, the IC would be an interesting starting point for any statistical analysis of the Internet.

7. ACKNOWLEDGEMENTS

Special thanks goes to Jeroen van Ingen at ICTS for providing the dataset of the UT considering ARP data of the /16 address block.

8. REFERENCES

[1] Anonymous. Internet census 2012. `http://internetcensus2012.bitbucket.org/paper.html`, Mar 2013.

[2] Alistair and Alberto. Carna botnet scans confirmed. `http://blog.caida.org/best_available_data/2013/05/13/carna-botnet-scans/`, May 2013.

[3] D. Smallberg. RFC 832: Who talks TCP?, Dec 1982.

[4] John Heidemann, Yuri Pradkin, Ramesh Govindan, Christos Papadopoulos, Genevieve Bartlett, and Joseph Bannister. Census and survey of the visible internet. In *Proceedings of the ACM Internet Measurement Conference*, pages 169–182, Vouliagmeni, Greece, Oct 2008. ACM.

[5] Ralph Holz, Lothar Braun, Nils Kammenhuber, and Georg Carle. The ssl landscape: A thorough analysis of the x.509 pki using active and passive measurements. In *Proceedings of the 2011 ACM SIGCOMM Conference on Internet Measurement Conference*, IMC '11, Nov.

[6] Alberto Dainotti, Alistair King, kc Claffy, Ferdinando Papale, and Antonio Pescapè. Analysis of a "/0" stealth scan from a botnet. In *Proceedings of the 2012 ACM Conference on Internet Measurement Conference*, IMC '12, Nov.

[7] Zakir Durumeric, Eric Wustrow, and J. Alex Halderman. Zmap: Fast internet-wide scanning and its security applications. In *Proceedings of the 22Nd USENIX Conference on Security*, SEC '13, Aug.

[8] Gordon Lyon. Nmap scripting engine (NSE). `http://nmap.org/book/man-nse.html`, Dec 2012.

[9] Cisco. Cisco visual networking index: Global mobile data traffic forecast update, 2012–2017. Technical report.

Appendices

Table 2: Content example of IC trace icmp_ping

IP Address	Timestamp	Result
130.89.0.0	1335224700	unreachable
130.89.0.2	1335244500	alive, 4088
130.89.0.2	1345272300	alive, 29589
130.89.0.8	1344442500	alive from 145.145.4.2, 36027
130.89.0.194	1347902100	Icmp Error: 1,ICMP Host Unreachable, from 130.237.140.109

Table 3: Content example of IC trace hostprobes

IP Address	Timestamp	State	Reason
130.89.0.2	1335825900	up	unknown
130.89.0.2	1346825700	up	echo-reply (0.064s latency).
130.89.0.4	1335739500	down	no-response

Table 4: Content example of IC trace syncscan

IP Address	Timestamp	State	Reason	Tcp/Udp	Ports
130.89.0.2	1335825900	closed	reset	tcp	20,21,22,23,53,80,110,111,143,443,993,995, 1723,3306,3389,5900,8080
130.89.0.2	1346825700	closed	reset	tcp	21,22,23,25,26,80,81,110,111,514,515,1025, 1026,1027,2000,2001,10000,113,143,1433,1720, 1723,179,199,32768,3306,3389,443,465,49152, 49154,5060,53,548,554,5666,587,5900,6001, 646,8000,8008,8080,8443,8888,993,995
130.89.0.9	1335012300	filtered	no-response	tcp	25,135,139
130.89.0.9	1337501700	open|filtered	no-response	udp	109,124,135,202,249,262,291,320,322,338, 448,465,607,781,841,900,908,950,1017,1568, 5267,5492,9507,10853,19024,32974,35134, 37870,39793,42853,43157,46687,47938,50267, 50929,51417,55976,62003,62122
130.89.0.21	1335581100	open	syn-ack	tcp	22

Table 5: Result of comparison per subtrace of IC with ARP table

	Subset C	*Subset B*				*Total*
Subtrace	noipmatch	no_overlap	partial_overlap	totaltraceoverlap	totalregoverlap	
hostprobes_alive	1192 [5.31%]	1573 [7.01%]	411 [1.83%]	18945 [84.46%]	311 [1.39%]	22432
hostprobes_unreachable	170565 [55.80%]	118078 [38.63%]	2775 [0.91%]	13481 [4.41%]	769 [0.25%]	305668
icmp_alive	1497 [2.82%]	1010 [1.90%]	1176 [2.22%]	48406 [91.30%]	930 [1.75%]	53019
icmp_unreachable	458833 [55.27%]	310824 [37.44%]	8688 [1.05%]	49048 [5.91%]	2721 [0.33%]	830114
syncscan_alive	670 [3.94%]	839 [4.94%]	264 [1.55%]	14930 [87.90%]	282 [1.66%]	16985

Table 6: Closed port comparison single IP

IC Timestamp (extended range)	Server trace begin-end time, (duration in s)	IC state	IC probed ports	Ports replied by server [Reply with RST,ACK flags sent]
1335811500 (1335809700-1335813300)	1335812318-1335812324, (6)	closed	20,21,23,53,80,110,111,143,443,993, 995,1723,3306,3389,5900,8080	20,21,23,53,80,110,111,111,143,443,993, 995,1723,3306,3389,5900,8080
1335811500 (1335809700-1335813300)	-	filtered	25,135,139	-
1335811500 (1335809700-1335813300)	-	open	22*	-
1337334300 (1337333500-1337336100)	1337334855-1337334872, (17)	closed	29,54,119,179,260,317,343,450,518, 529,573,621,654,710,741,808,878,882, 929,943,1795,5364,6547,6845,9217, 11292,16696,19052,23931,24159,26274, 30871,34307,40752,44361,45584,49983, 50575,52211,55908	29,54,119,179,260,317,343,450,518, 529,573,621,654,710,741,808,878,882, 929,943,1795,5364,6547,6845,9217, 11292,16696,19052,23931,24159,26274, 30871,34307,40752,44361,45584,49983, 50575,52211,55908
1338596100 (1338594300-1338597900)	1338596829-1338596912, (83)	closed	72,133,180,274,345,385,487,547, 615,646,663,683,751,760,763,785, 848,918,950,974,4289,8864,11269, 11831,13585,14096,22134,24822, 29555,31080,34663,48366,49567, 49891,55268,56236,56514,58325, 61394,63542	72,133,180,274,345,385,487,547, 615,646,663,683,751,760,763,785, 848,918,950,974,4289,8864,11269, 11831,13585,14096,22134,24822, 29555,31080,34663,48366,49567, 49891,55268,56236,56514,58325, 61394,63542
1339375500 (1339373700-1339377300)	1339375724-1339375796, (72)	closed	57,64,174,176,188,205,235,447,456, 466,511,515,518,524,583,655,727, 739,772,805,853,2825,3919,4538, 6331,20806,22319,28796,30682,30836, 31558,33030,33286,35202,42812, 43604,45345,48301,58258,61052	57,64,174,176,188,205,235,447,456, 466,511,515,518,524,583,655,727, 739,772,805,853,2825,3919,4538, 6331,20806,22319,28796,30682,30836, 31558,33030,33286,35202,42812, 43604,45345,48301,58258,61052
1346868900 (1346867100-1346870700)	-	filtered	21,23,26,53,80,81,110,111,113,143, 179,199,443,465,514,515,548,554, 587,646,993,995,1025,1026,1027, 1433,1720,1723,2000,2001,3306, 3389,5060,5666,5900,6001,8000, 8008,8080,8443,8888,10000,32768, 49152,49154	-

**No traces of incoming traffic on port 22 were logged.*