An Internet Census Taken by an Illegal Botnet – A Qualitative Assessment of Published Measurements

Thomas Krenc
TU Berlin
tkrenc@inet.tu-berlin.de

Oliver Hohlfeld
RWTH Aachen University
oliver@comsys.rwth-aachen.de

Anja Feldmann
TU Berlin
anja@inet.tu-berlin.de

This article is an editorial note submitted to CCR. It has NOT been peer reviewed.
The authors take full responsibility for this article's technical content. Comments can be posted through CCR Online.

ABSTRACT

On March 17, 2013, an Internet census data set and an accompanying report were released by an anonymous author or group of authors. It created an immediate media buzz, mainly because of the unorthodox and unethical data collection methodology (i.e., exploiting default passwords to form the Carna botnet), but also because of the alleged unprecedented large scale of this census (even though legitimate census studies of similar and even larger sizes have been performed in the past). Given the unknown source of this released data set, little is known about it. For example, can it be ruled out that the data is faked? Or if it is indeed real, what is the quality of the released data?

The purpose of this paper is to shed light on these and related questions and put the contributions of this anonymous Internet census study into perspective. Indeed, our findings suggest that the released data set is real and not faked, but that the measurements suffer from a number of methodological flaws and also lack adequate meta-data information. As a result, we have not been able to verify several claims that the anonymous author(s) made in the published report.

In the process, we use this study as an educational example for illustrating how to deal with a large data set of unknown quality, hint at pitfalls in Internet-scale measurement studies, and discuss ethical considerations concerning third-party use of this released data set for publications.

Categories and Subject Descriptors

C.2.1 [**Computer-Communication Networks**]: Network Architecture and Design—*Network topology*

Keywords

Carna Botnet, IPv4, Census

1. INTRODUCTION

Anonymous authors released an Internet census report on March 17, 2013, together with the underlying data set via a mailing list `http://seclists.org/fulldisclosure`, typically used for disclosure of security information [1]. The release contains a report anonymously hosted on BitBucket and GitHub, as well as 568GB of compressed data (9TB uncompressed) released via BitTorrent. In the Internet census report the authors claim to have conducted multiple scans of the entire IPv4 address space within 24 hours, using a large botnet which they call *Carna*. Primarily, these scans were directed at hosts via ICMP ping, at open ports and services, the reverse DNS tree, and some traceroutes. Part of these scans have been confirmed with CAIDA's Internet telescope which was scanned by the botnet as well [2]. Ironically enough, the anonymous authors build their botnet, supposedly consisting of 420k hosts, by exploiting default passwords. Note, using system resources without user permission is a violation of any reasonable terms of use. Thus, based on academic standards, their study is not only unorthodox but has to be considered *unethical*.

Although extensive ICMP censuses [3], port scans [4, 5, 6], and traceroutes [7, 8, 9] have been conducted before and even at a larger scale, the nature as well as the scale of the Internet census resulted in a media buzz [10, 11, 12, 13, 14, 15, 16, 17], an investigation by the Australian Computer Response Team [18], and in the creation of an Internet Census 2012 Search engine [19]. These responses, and the easy availability of the data set have attracted many hundred downloaders world-wide. By participating in the BitTorrent swarm during the days immediately after the release, we observed more than 470 peers located in 38 countries, predominantly in China, USA, and Germany. Further, by mapping peers to ASes and to reverse DNS hostnames, we identify among the downloaders more than 30 universities and research facilities, 20+ ISPs, several infrastructure providers, as well as governmental and security organizations.

Whether unethical or not, the interest is evident. However, in particular considering the level of attraction, there is almost no knowledge about the authenticity and the quality of the published data. This is further exacerbated by the fact that the authors are anonymous and the left-behind data set description is superficial or not existent at all. Without any kind of documentation or meta-data of the data *consumers* can easily misuse the data sets, as they do not know if the data quality is suitable for answering their questions in the first place. Often consumers simply assume that the data is of good enough quality for their purpose.

Due to many uncertainties that exist in and around the data set, we challenge the claims made by the authors of the Internet census. As contribution of this paper, we present the results of some authenticity checks in Section 3, after which we perform an analysis of the quality of the data in Section 4. In particular, we try to reverse engineer missing meta-data as best as possible, in order to verify the claims by the authors in Section 5. Our main findings are:

- Spot tests confirm that the data set is likely to be authentic.

- The analyzed data suffers from significant methodological problems, which results in a low data quality and lacks adequate meta-data.

- Several claims made by the authors cannot be verified using the data.

Indeed, contrary to the authors' claims, we find that the data set contains at most *one* census, as we will demonstrate in this paper. In a parallel effort to the analysis, this paper serves as example on how to validate measurement-based networking research, based on a methodology proposed by Krishnamurthy et al. [20]. We examine the data hygiene, i.e., how carefully the quality of the data sets was checked by the authors of the Internet census report, by analyzing the provided meta information. Furthermore, to drive our analysis we ask specific questions to ensure that certain requirements are fulfilled to reuse the data, e.g., questions that aim to uncover likely reasons for errors in the data. We conclude this paper with a discussion on whether the adequate rigor was used by the authors to estimate the size of the Internet, considering the available data. Finally, we elaborate on the novelty of the conducted measurements as well as the public reactions and ethical considerations.

2. PUBLISHED DATA SETS

In this section we introduce the data sets, their file organization, along with the data structure within the files. The Internet Census 2012 announcement [1] points to a Web site [21] containing the report as well as the data sets, available to everyone for download and analysis—in principle a great service to the community. The data spans several archives, 568GB compressed/9TB uncompressed, and is offered via BitTorrent, which is how we obtained the data. It includes the following data sets: *i) ICMP ping* reachability and latency information, *ii) nmap* port scans for open ports and per-port service information (host and service probes), *iii) nmap* TCP/IP fingerprints and IP ID sequence information, *iv) reverse DNS* records, and *v) traceroute* records. Each data set is subdivided into smaller files, grouping all the probes into /8 blocks, based on the respective destination IP. In those /8 block files, each tab-separated probe record includes the destination IP, timestamp, and the probing result (e.g., ICMP ping result, list of open ports, etc.). Regarding the service probes, separate /8 block files are provided for each probed service, e.g., port 80/http. Finally, the downloaded data includes some (wallpaper) images, some data for the website, along with the source code of the website, a modified nmap tool, and a Hilbert graphic generator.

Having compiled the most basic description of the published data, we first want to understand the properties of the measurements. First, we find that the data collection started in April 3rd, 2012 and lasted until December 18th, 2012[1]. For this measurement period, in Table 1, we summarize the number of total ICMP probes, host probes, reverse DNS queries, and traceroutes records available in the data as well as the number of total probes that were stated in the report. Surprisingly, there are various mismatches in what is claimed in the report to what is in the actual data set. For example the report states that there are 2.5B (5%) ICMP ping probes more than we count. We elaborate more on the inconsistencies in Section 4.1.

[1]Two timestamps date back to 1978 which obviously is outside the range of the Internet census.

data set	total probes		probed hosts
	data	report	
ICMP Ping	49.5B	52B	3,706,585,088
Host probes	19.7B	19.5B	3,705,342,574
Reverse DNS	10.5B	10.5B	3,700,481,860
Traceroute	68.7M	68M	64,666,758

Table 1: High-level statistics for some of the data sets.

Further, we analyze the targeted address space and count the number of unique hosts that were probed. Since this information was not provided by the *producers* of the data, we see it as our responsibility to fill the missing information. Except for the traceroute records, the number of unique hosts is more than 3.7B which corresponds to the currently allocatable address space [22]. Indeed, we did not see any probes for IP blocks that are listed as reserved by IANA, private address space, as well as multicast space. So in total, the data sets comprise of probes launched towards 221 /8 blocks. Notably, in the case of ICMP, considering the overall and the unique probed hosts, the data allows for at most 13 censuses.

We note that all data sets but the traceroute records do not include the source IP address, i.e., the IP of the probing host. This is problematic and challenges the use of the data as some results depend on the location of both the source as well as the destination, e.g., the ICMP latencies. But there are also more subtle problems that are not addressed by the authors, e.g., is the destination IP behind a firewall or a proxy that may alter the reachability results? We note that including information about the operating conditions of any involved network during the measurement periods can be crucial to properly interpret the data.

Finally, we observe that the measurement periods of the individual data sets are of varying lengths, irregular, and only partially overlap. In fact, we cannot recognize any reasonable kind of measurement schedule in the data as well as in the documentation.

3. AUTHENTICITY

We continue our analysis by asking the most fundamental question: is the data authentic or manufactured—an April fool hoax? We aim to answer this question by reproducing some of the measurements. This is non-trivial, as network conditions in the Internet are subject to constant changes and therefore decrease the reproducibility over time. However, if we are able to reproduce some of the measurements, it can be a strong indication of authentic data. We note that CAIDA has confirmed that the scanning took place, using a combination of telescope data and the census data [2]. In this section, we add to this by taking a closer look at those parts of the data sets that are less time dependent, e.g., the reverse DNS records and the server IP addresses—for which we happen to have comparable data sets.

3.1 Reverse DNS

We start with the reverse DNS data set which we compare to a separate, external data set of reverse-resolved IP addresses captured in November 2012. Note, November 2012

is just shortly after the reverse DNS data collection of the Internet census ended. Our data set contains 70.6M IP addresses across 177 /8s for comparison, while the Internet census data contains 3.7B in 221 /8s.

We check the data sets for consistency via string comparison of the hostnames from both data sets using the external data set as basis. We find exact hostname matches for 95.2% of the tested IPs. For 3.1% of the IPs the external data set finds a reverse name, which is not reported in the census data. A closer look at these 3.1% shows that the unsuccessful lookups are due to DNS lookup errors (86.5%), nonexisting reverse DNS entries (8.2%), and timeouts (5.2%). For the remaining 1.7% of the IPs, we do not find exact matches in the hostnames. The reasons for this are different hostnames (61.6%), even though the domain name and top level domain match, differences in capitalization (3.5%), or multiple reverse entries with different reverse entries in each of the data sets. The latter requires manual checking.

Overall, our test finds that almost all entries match in principle (>96%) for a data set that was unknown to the authors of the Internet census. This indicates that the data is unlikely to be artificially manufactured.

3.2 Akamai IPs

The Internet census report states that 5% (4M) of all web servers on port 80 return the *AkamaiGHost* user agent string [21]. This user agent string is announced by Akamai CDN caches when requesting content that is not hosted by Akamai, e.g., as seen by nmap service probe scans during the Internet census. Similar to reverse DNS, we want to reproduce the measurements to verify the authenticity of the data.

Therefore, we collect another data set by probing all 4M IPs from the Internet census a single time from a single IP address from our local university network in July 2013. For our probes we use an in depth understanding of Akamai's caching infrastructure. It was shown that any CDNized object is accessible from any Akamai cache [23, 24]. We exploit this by downloading two image objects hosted by Akamai (one from a major social network, another from a major car manufacturer), and one non-Akamai object. The latter download lets us distinguish open proxies from Akamai caches. Our script validates the SHA-1 hash and HTTP return code of all retrieved objects. We consider an IP address to belong to an Akamai cache, iff all three tests passed, i.e., if the hash and HTTP return code for the two CDNized objects match, and the non-CDNized object cannot be retrieved.

Out of the 4M IPs, 84.2% pass all tests and thus are consistent with Akamai caches. The remaining 15.8% fall into two categories. 10.5% of the IPs were unreachable, e.g., because of firewalled hosts that cannot be reached from the public Internet. The remaining 5.3% did not pass one or two of our tests, e.g., due to timeouts. However, 84% served at least one Akamai hosted object correctly and thus *appear* to be valid Akamai caches. Overall, the large number of validated Akamai caches again shows that presumably the data is not manufactured.

4. LOOKING BEHIND THE CURTAIN

Given that the data appears to be authentic, we want to validate the claims of the authors, e.g., claims that they have performed several censuses, among which are fast scans

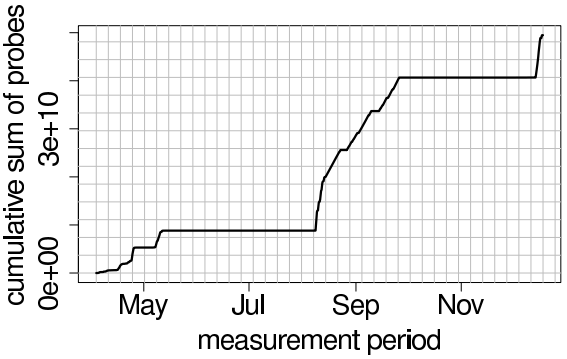

Figure 4: Measurement periods: Cumulative sum of all ICMP probes over entire measurement period (horizontal lines: 3.7B target address space).

sweeping the complete Internet address space within 24 hours. However, we note that there is no meta-data in the report that relates to censuses, except for the description of two ICMP ping scanning methods, i.e., *"long term scan [...] for 6 weeks on a rate of [a complete scan] every few days"* or *"fast scans [...] probed the IPv4 address space within a day"*. In order to verify those claims, the ultimate goal of this section is to reverse engineer as much meta-data as possible, starting from what we are able to uncover in Section 2, to identify the censuses in the ICMP data set.

4.1 Meta-data? Wrong!

As a first step, we examine as much information as possible from the meta-data reported by the authors, in order to check the reusability of the data sets. Regarding the Internet census report, one would expect a detailed description of the measurement tool (botnet) and measurement data. However, while the Internet census report contains some rather superficial information about the measurement methodology using a large botnet, the data set documentation itself lacks detailed information about the measured data. It gets even worse, when checking the consistency between the report and the data. For instance, the report mentions the ICMP measurement period spans *"from June 2012 to October 2012"*. However, in June and July no probes are reported in the data set. This is confirmed by Figure 4 which shows the cumulative number of ICMP probes over the whole measurement period. The horizontal support lines correspond to 3.7B IP addresses, supposedly the base line of the probed address space, while the vertical support lines correspond to weeks. (The plot appears to be consistent with the interactive plot included in the report.)

Further inconsistencies between the data and the report concern the probed address space, and the number of samples. For instance, while the report states 52B ICMP probes, the data set only contains 49.5B ICMP probes, see Table 1. Also, while the report refers to scans of *"all 3.6 billion IP addresses of the Internet"* or *"240k sub-jobs, each [...] scanning approximately 15 thousand IP addresses"*, the data set reports roughly 3.7B probed IP addresses, see Table 1. When evaluating the completeness of the censuses, we therefore assume that the *target address space includes at least 3.7B IP addresses*.

From a hygiene perspective, a well maintained documentation of the data, e.g., meta-data, includes as much infor-

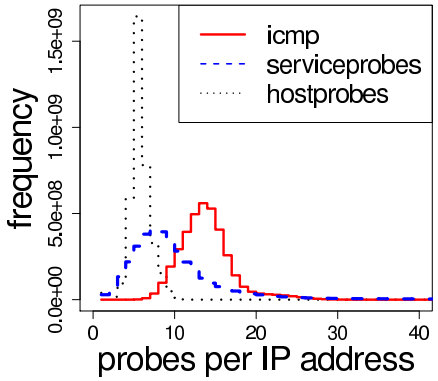

Figure 1: Probing Frequency: Distribution of how often each IP probed was probed for three kinds of probes.

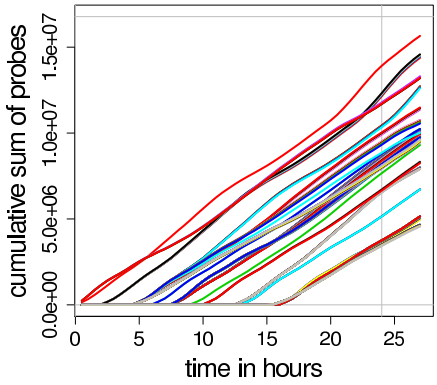

Figure 2: Misalignment of timestamps within 24 hours: Evolution of cumulative sum of ICMP probes for each /8 prefix in $icmp_3$.

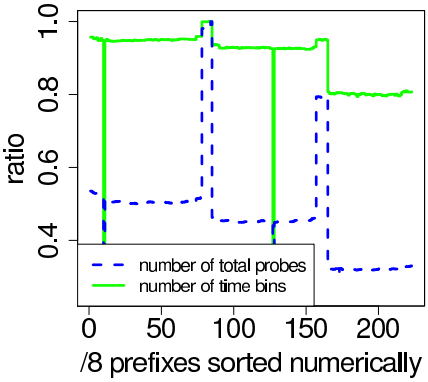

Figure 3: Scan diversities: ordered /8 prefixes plotted against the #probes/total (dashed) and #timebins/total (solid) ratios in $icmp_2$.

mation as possible in order to allow any consumer to reuse the data adequately. The low level of documentation we find here, however, is problematic: While we cannot rely on the documentation as it is at best superficial and inconsistent with the published measurements, it is the only source of information given by the anonymous authors.

4.2 Data Quality

In the previous section, we find that the Internet census report only includes limited and partially inconsistent information about the data and how it was collected. In order to verify the claims of the authors, we need to reverse engineer as much meta-data as possible. Throughout this analysis, we focus on the ICMP data set only, since from the report we assume it contains the censuses in question.

4.2.1 Probing Distribution

We attempt to reverse engineer the meta-data step by step to find the missing information, e.g., when does a census start and when does it end? Part of filling in the missing information requires knowing how each IP address was probed. This can help us to understand 1) how many censuses we can expect, and 2) what other data, except for clean censuses, are included in the data set. For example, are there scans of different types? Do they overlap in time? Do we miss probes or see reprobes, e.g., due to bot failures?

In this section, we investigate the distribution of probes by counting the number of probes per IP address. Ideally, assuming our hypothesis in Section 2 is correct, we should find that each IP address was probed 13 times, resulting in 13 censuses. Figure 1 shows a histogram of the number of probes per IP address for three different kinds of probes: ICMP pings, host probes, and service probes[2]. We find that all are highly skewed. Regarding ICMP pings, while most IP addresses are probed between 6 and 25 times, some IP addresses are probed more than 600 times and others only once. Indeed, the latter is highly problematic, as strictly speaking the data can thus only contain a *single complete census*, contrary to what the authors claim.

[2]Due to resource constraints, we only focus on service probes directed at well-known ports.

name	period	total probes	probed hosts	days
$icmp_1$	Apr.-May	8.8B	3,682,182,938	40.0
$icmp_2$	Aug.-Oct.	31.8B	3,706,583,819	49.5
$icmp_3$	December	8.8B	3,704,509,119	4.3

Table 2: ICMP measurement periods overview.

Due to the skewed distribution we cannot assume that the data consists of clean censuses. Indeed, there are many possible explanations for these different probing frequencies. One explanation is that beside the censuses, there is additional data included in the ICMP ping data set. Since the meta-data description is poor, we cannot reject the hypothesis that census data is mixed with other unreported data. Another explanation can be problems with the bots. Failures by a subset of the botnet, which is a worldwide distributed set of workers and aggregation nodes, can severely impact the measurement data.

4.2.2 Probing Activities

Our first attempt to determine the number of censuses from the number of probes per IP has failed. Instead of 13 clean censuses we find a skewed distribution, which indicates that the alleged censuses may be mixed with other data collected for different purposes. In this section, we classify the measurement activity periods, to identify potentially separate experiments. This may enable us to not mix the results from incoherent scans and determine their individual purposes. Therefore, we analyze the ICMP probing activities, i.e., determining when and for how long the IPv4 based Internet was probed. Recall, Figure 4 shows the probing activity for the overall measurement period which spans more than eight months. Even a cursory glance at the plot indicates that the probing intensity varies significantly over time, thus it make sense to separate these periods. Initially, there appears to be some initialization period, then some scans, then a break, another scanning period, another break, and a final scanning period. Although not documented, we find three major activity periods separated by longer inactivity periods. Thus, we split the data into three subsets: $icmp_{1-3}$. Table 2 reports the number of the total

number of probed IPs, unique IPs, as well as the measurement duration.

However, the purpose of those activity periods is not immediately apparent. The relatively slow scanning rate of $icmp_1$ (220M probes/day on average) and its irregular scanning behavior (short activity burst and a two-week break) suggest that it contains test runs while gathering experience with using the botnet as a measurement tool. Further, Table 2 shows that $icmp_1$ contains less than 3.7B uniquely probed hosts, contrary to the other periods. Test runs may explain the skewed distribution from Section 4.2.1. While $icmp_2$ (642Mp/d) consists of several stable scans separated by small breaks, including two 5-days breaks, the probing behavior in the very beginning is rather steep. Together with $icmp_3$ (2047Mp/d), a steep, short and stable period, these two periods appear to be potential candidates for fast scans.

Due to the surprisingly different and thus noteworthy characteristics that $icmp_{1-3}$ expose, we, in the remainder of this paper, report our findings using these measurement periods. Note, that there is no related description available in the Internet census report.

4.2.3 Botnet Architecture

As discussed in Section 4.2.1, the architecture of the botnet can be one reason for the skewed distribution of probes per IP address. For example, the challenges to be addressed by the data collection are handling failures both at the worker level as well as at the aggregation node level. Does the controller start the job from scratch at the same or another intermediate node? What happens with the results of the workers? Can these be stopped or reintegrated into the process? We assume that, due to the distributed nature of botnets, failures that relate to the orchestration of aggregation nodes and workers are reflected in the probing frequency of particular IP groups. For example, if one aggregation node fails, it can miss all the data that was measured and transmitted by the workers, while other aggregation nodes keep collecting data, leading to a skewed distribution. Thus, we need to understand at which granularity the jobs on the aggregator level are delegated and collected. This view enables us to see whether the botnet infrastructure causes some IP groups to be probed differently than others.

Since the measurement data is organized in /8 block files, we begin with checking if /8 is also the granularity for which an aggregation node is responsible. Accordingly, for the first day of $icmp_3$, Figure 2 plots the cumulative number of probes for each individual /8 prefix against the timestamps. Note the horizontal support line, indicating the /8 address space of around 16.7M IPs, and the vertical support line boxing the first day. We choose $icmp_3$, because it is a short and stable measurement period, thus our observations will most probably not be biased by different probing activities, as argued in the previous section. Surprisingly, with regard to the timestamps, the plot highlights a temporal misalignment of the start of the probing for all /8s which spans a time period of more than 15 hours. Similar observations also hold for the other measurement periods, i.e., $icmp_1$ and $icmp_2$. However, for smaller aggregation levels, e.g., /16 within the /8s, this misalignment is *not* present. In addition, in Figure 2 we observe that several prefix groups show similar characteristics. We therefore conjecture that *the scans are organized by /8 prefixes*. Finally, we note that from the first started prefixes not a single one received more than 16.7M probes within the first 24 hours.

A likely reason for the temporal misalignment seems to be the use of the local time at the aggregation nodes, in order to specify when a probe is launched (or a response is received). As the timestamps are not addressed in the report, we do not know how to interpret them. In which time zone are timestamps reported? How is time normalized (e.g., to UTC)? The logical presumption is that the experiment uses a single reference timezone, and that all timestamps are accordingly normalized. Otherwise, the data set should contain timezone information which is not the case.

4.2.4 Probing Characteristics

Now that we can assume that the jobs are organized per /8s, we wonder whether each /8 was probed equally or not, in order to eventually find the explanation for the skewed probing distribution, as shown in Section 4.2.1. Throughout this section, we address this task by contrasting the individual /8 prefixes from $icmp_2$, but note that similar behavior holds for /16 and /24 prefixes as well as $icmp_1$ and $icmp_3$. Figure 3 plots for each /8 (in sorted order) (a) the ratio (dashed) of probes towards this /8 vs. the maximum number of probes any /8 got and (b) the ratio (solid) of time bins with probes towards this /8 vs. the maximum number of time bins seen for any /8. This way of plotting ensures that at least one prefix will have value 1 for both metrics. We note that there are no probes for 10/8 and 127/8. For time bin granularity we choose 30 minutes, as it is the maximum accuracy available in the data.

From the plot we can identify five different prefix groups when focusing on the ratio of probes at 100%, 80%, 50%, 45% and 35%. This implies that prefixes from one group are probed in a similar fashion. Moreover, prefixes from one group are IP-wise adjacent when sorted numerically. We point out the significant differences in probing frequency across the groups. Some /8s are probed at least three times as often as other /8s. This hints at some problems with the control flow of the experiment, or the distributed nature of the botnet. Thus, we conclude that how /8 prefixes are probed differs across /8s.

Also, for the second ratio which is normalized by the number of time bins, we again notice at least five prefix groups. This underlines the above observations that the probing is not done in a uniform manner across the /8s. Moreover, if we consider the relationships of the two ratios for the same /8 the different probing rates for the /8s, i.e., how fast the probing is done, become apparent. For example, we observe prefixes 1/8 to 165/8 to have a similar time bin ratio between 90 and 100%. However, as shown before, the ratio of probes in the same prefix range shows drastic differences, in particular for the prefixes 75/8 to 80/8, or 160/8 to 165/8. We conclude that the /8 prefixes are probed with different probing rates. More specifically, some prefixes are probed up to twice as fast as others.

Unfortunately, we find that the probing characteristics per /8 (/16 as well as /24) differ significantly both in terms of number of probes, as well as in terms of probing rate. We attribute these probing diversities to failures in the botnet architecture, which eventually seem to be responsible for the bad data quality. Therefore, as conclusion of our analysis, we point out that significant meta information remains un-

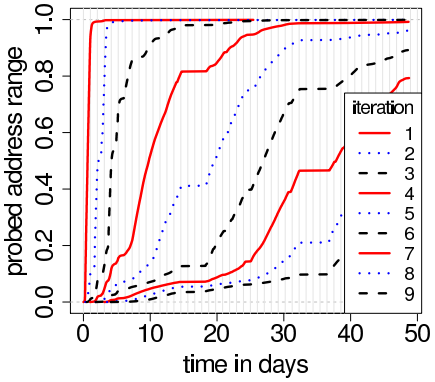

Figure 5: Overlapping iterations: First nine iterations over the probed address range in $icmp_2$.

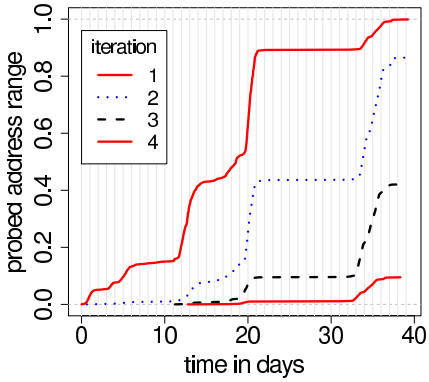

Figure 6: Overlapping iterations: First four iterations over the probed address range in $icmp_1$.

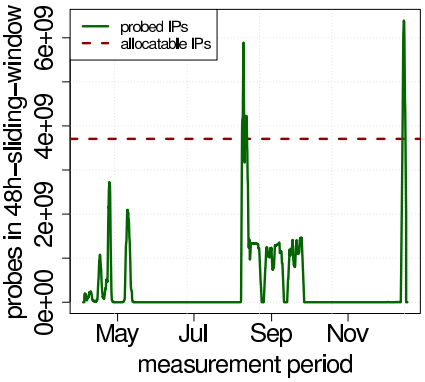

Figure 7: Finding *fast scans*: Sum of probes in 48h-sliding window over entire measurement period.

known and the gained insights are not sufficient to verify the claims of the authors.

5. CLAIMS OF THE AUTHORS

After having pointed out some inconsistencies between the Internet census report and data, and commented on the lack of meta information, as well as the data quality, we now turn our attention to two central claims of the authors related to the ICMP ping data set. Concretely, the authors claim to have conducted several censuses. Some of them—the fast scans—supposedly are done within 24 hours, while the long term scan spans a period of six weeks. Thus, in this section we try to find all the censuses, and identify the fast scans.

5.1 Finding Censuses

One of our main problems in validating these claims is the lack of meta information. With regards to finding censuses, this is particularly relevant, as the data set description does not state when a census starts, or when it ends. Therefore, in this section we try to uncover that missing information, and give an estimation of the number of censuses.

5.1.1 Scan Iterations

In the Internet census report, the authors claim to have scanned the IPv4 address space multiple times. When performing multiple scans in parallel, it is crucial to ensure that they do not overlap in time or, if they do overlap, to be able to separate the resulting data sets. To distinguish different scans of the address space, regardless of their duration, we use the concept of iterations: The first time an address is scanned belongs to the first iteration. The next scan belongs to the second iteration, etc. Should the data contain two full censuses then the first and the second iteration would cover the full IP address range. Should the data contain two full censuses, and some reprobing in order to do error recovery, the data should contain a full first and second iteration of the IP address range, a partial third iteration, and maybe even partial fourth and fifth iterations. We note, that reprobing is almost a necessity in order to recover from failures among the measurement bots. If two censuses are properly separated in time, namely non-overlapping, then the concept of iterations over limited time periods lets us separate censuses, since we can distinguish reprobings from the next census.

Figure 5 plots the CDF for the first nine iterations for time period $icmp_2$. The probed address range (y-axis) is premised on the respective number of probed hosts in Table 2. We notice that even the first iteration takes more than 6 weeks in order to cover the full address range. The other eight iterations also start within the first few hours, but do not reach 100%, indicating that there is reprobing and that there are no non-overlapping censuses in $icmp_2$. The second through fifth iterations reach 96-99%, while the sixth iteration does not even reach 90% of the IP address range. Moreover, it highlights the different probing frequencies of the different IP address ranges. Similar observations hold for the other measurement periods, e.g., see Figure 6 for $icmp_1$, or smaller time periods. We conclude that we cannot distinguish the scans, and therefore cannot count the number of individual and clean censuses in the ICMP data set.

5.1.2 Censuses vs. Test Runs

When comparing Figure 6 for $icmp_1$ to Figure 5 for $icmp_2$, we notice a significant difference. While in $icmp_2$ the CDF for iteration one is very steep in the beginning, the first iteration in $icmp_1$ is rather flat and irregular for more than five weeks. We conjecture that the latter observation is due to test runs. Next, we take a closer look at Figure 6: We observe reoccurring probing activities at days 0-5, 11-13, and 17-21 for $icmp_1$. We note that the probing activities are not comprehensive, i.e., they only contain partial probing of the address space, which is 3.6B IPs, according to Table 2. Moreover, except for the first, each of those probing activities include restarts or reprobes of previous activities which is reflected in the emergence of new iterations. Overall, 87% of the IPs are probed twice (∼42% are probed three times, and ∼10% four times). Thus, from the irregular and incomprehensive scans, as well as the restarts, we conclude that $icmp_1$ includes test runs.

5.1.3 How many Censuses are there?

Since the iterations overlap throughout the entire measurement period, we are not able to count how many censuses the data set contains. Thus, we cannot verify the claims of the authors to have conducted several scans. Therefore, we elaborate on the number of censuses based

on our findings so far.

If we are strict with regard to the findings in Section 4.2.1, i.e., if we require that each census contains records for all IP addresses that were probed throughout the entire measurement period, then the data can contain at most a single census. The same holds for the individual measurement periods, as we have shown in Section 5.1.1 that only the first iterations are the most complete. Thus, the number of censuses would be three, one for each measurement period. If we are satisfied with partial censuses in the sense of scanning up to 99.5% of the IP address space, then $icmp_1$ and $icmp_3$ contain one census each, while $icmp_2$ contains three censuses. Finally, if we look at each time period, and the respective numbers in Table 2 separately, there is room for two censuses in $icmp_1$ and $icmp_3$, as well as eight censuses in $icmp_2$. However, we need to keep in mind that $icmp_1$ probably includes test runs, and we are not able to separate and count censuses in the data set.

5.2 Where are the Fast Scans?

In the previous sections we were not able to identify the censuses, or to find out how many there are in total. Given that there is at least one census in the ICMP data, in this section we try to identify the fast scans that were advertised by the authors, i.e., scans of the entire IPv4 address space completed within 24 hours. For this we resort to the "typical" approach of using a sliding window to count the number of unique IPs within 24 hours. In principle, a sliding window of 24 hours length should suffice for this analysis. However, given that the timestamps in the data set are misaligned and scattered over almost an entire day (see Section 4.2.3), we use a 48-hour sliding window as conservative approach.

Figure 7 shows the number of probes per 48-hours sliding window across time. We added a supporting dotted line at 3.7B IPs. This corresponds to the number of currently allocatable IP addresses. We see that only for two measurement periods the number of probes exceeds the required number of probes, whereby the extraordinary high number indicates overlapping iterations. Within these periods, we find that each candidate window contains probes of all 221 probed /8 prefixes. However, none of the /8s was probed completely. The number of missing IPs per /8 ranges from 2,522 IPs for the most frequently probed prefix to 1,060,415 IPs for the least frequently probed prefix. We thus conclude that we are unable to find any complete fast scans, even when we use a 48-hour sliding window. Thus, we are unable to verify the claims of the authors.

6. DISCUSSION

In this section, we examine the size of the Internet, as it was determined by the authors of the Internet census report and highlight related, typical pitfalls. Also, we comment on the novelty of the census as well as the public reactions that followed the release of the data sets. We close by discussing ethical considerations and concerns.

6.1 Robustness of the Data

Part of the Internet census report deals with estimating the size of the Internet. As part of our validation efforts, we ask whether the proper rigor was used to estimate the size of the Internet, considering the bad data quality, e.g., due to different probing rates.

6.1.1 Size of the Internet

The authors of the Internet census report to have used data *"from June 2012 to October 2012"* to estimate the size of the Internet. Note that in Section 4.1 we have already reported that there is no data collected in June and July. Still, using such a long time range can significantly bias the results. Consider the following thought experiment: A customer uses the Internet once a day and is assigned a new IP address by its ISP every time it connects. Then, considering the five months of measurements, this single customer is responsible for 150 IP addresses. Thus, mixing incoherent measurement periods together may exaggerate the size of the Internet. However, measurement failures and probe drops may underestimate the size.

In their final remarks the authors add up numbers from all the different data sets. They assume if they have any indication that an IP address might have had any activity, then it needs to be included in the calculation of the size of the Internet. This may be problematic, as they consider an IP used, if it has a reverse DNS entry in the reverse DNS tree. However, reverse DNS entries are not mandatory to be assigned to the IP addresses by their owners. Furthermore, network operators sometimes automatically prepopulate entire (large) address ranges. This way a reverse DNS entry can be assigned to an otherwise unused IP address.

6.1.2 Typical Pitfalls

Getting a good grip on the size of the overall Internet is definitely an interesting research challenge. However, whether the number of IP addresses "in use" is a good proxy for the size of the Internet is debatable, since today a single IP address is rarely assigned to a single real person or machine. Rather it is an any-to-any relationship. Among the culprits for this are NAT gateways, which allow many hosts to share a single IP, Proxy servers, load-balancers, etc. Moreover, many hosts have multiple network interfaces and may, therefore, have multiple IPs. In addition, services such as anycast and multicast are used frequently in the Internet.

Still, given the discussion of IPv6 deployment and IPv4 address exhaustion [22], knowing which IP addresses are currently in use is of interest. However, this is a highly dynamic process. One example is dynamic address assignment in residential access networks and companies. Customers are assigned an IP address when they use the Internet; once they are offline their address can be reused. Indeed, many ISPs assign a different IP address to their customers whenever they connect. Also, allocations of complete IP blocks, and their usage can change drastically, e.g., infrastructure providers can renumber each host when allocations are changed.

Moreover, IP addresses that do not respond to probes are not necessarily "unused". Rather, they may have been configured to ignore any kind of probe. In addition, some probes can be dropped along the way, e.g., due to ICMP rate limiting. Furthermore, there are some devices that are configured to respond to any probe, even if the destination IP address is actually not in use. For example, part of the nmap probes, i.e., TCP acknowledgments on port 80, did not reach the CAIDA telescope [2]. They were intercepted and replied by proxies in networks where some of the bots were located, such that some IPs from the telescope address space were reported active.

6.2 What's the News?

The anonymous authors announced the availability of the Internet census in March 2013 with the statement: *"This project is, to our knowledge, the largest and most comprehensive IPv4 census ever"* [1]. The media picked up these statements and claimed: *"[t]he Most Detailed, GIF-Based Map Of The Internet"* [11], *"the Most Detailed Picture of the Internet Ever"* [25], *"one of the most comprehensive surveys ever"* [14], *"remarkable academic paper"* [13], etc. What is behind this buzz? Is the Internet census 2012 really unique?

In our view what makes this census so unique is not necessarily the data itself, but rather the unethical measurement methodology. ICMP censuses have been captured by Heidemann et al. since 2003 [3]. Moreover, extensive and Internet-wide *nmap*-based port scans and service probes have been conducted in the past [4, 5, 6]. With regards to traceroutes, this Internet census data set provides 68M records. More extensive studies have been conducted, e.g., by Shavitt et al. with 230M [7], Chen et al. with 541M [8], and Claffy et al. with 2.1B [9] sample records. Thus, what remains unique and novel is the combined data set that was allegedly captured using a large distributed network of 420k bots. The technical contribution, however, appears to be overrated by the press.

6.3 Ethical Considerations

One of the most fundamental questions for researchers given the availability of the Internet census data is whether it can/should be used for publications. While answering this question is beyond the scope of this paper, we raise concerns on the ethical validity of the data. While using traceroute, ping, and nmap for measurement purposes is in principle legitimate, using resource of end-users without permission is not only unorthodox but a violation of the terms of use. Thus, based on academic standards, the study as well as the data has to be considered *unethical*.

When discussing the data set with members of the community, we observed a diverse set of opinions ranging from *never touch such data* to *why not?* Such controversy raises the question of whether we as community need better ethical guidelines for networking and security research. While ethical standards in medical research are well defined (see e.g., the Belmont Report [26]), similar standards for Internet research are still not clearly defined. However, first steps exist: For example, the Menlo Report [27] as equivalent to the Belmont Report, or the Internet Measurement Conference. The IMC enforces adherence to its ethical standards as specified in the call for paper [28]. However, the interpretation is up to the individual PC members.

7. CONCLUSIONS

The novel, but unethical way the Internet census 2012 was performed attracted many different reactions, e.g., the technical contribution that was overrated by the press, or ethical discussions that came up in the community about using the data for publications. On the other hand, the interest in the provided measurement data is high, as we have observed downloaders from ISPs, universities, governmental institutions, etc., from all over the world. Therefore, and because many relevant questions about the data remained unanswered, in this paper we try to put the scope of the Internet census into perspective.

We show that the provided measurement data seems to be authentic, based on some spot tests. However, analyzing the quality of the data reveals a rather chaotic picture. While the data suffers from qualitative problems that are caused by methodological flaws, we note the significant lack of meta information, which cause us problems to verify statistics and claims made by the authors of the Internet census.

In particular, we find that the address space is not probed evenly. Also, due to the misalignment of timestamps, as well as the overlapping scans of the address space we are unable to single out a clean census, let alone find the promised fast scans. Our conjecture about the number of censuses ranges somewhere between one and twelve. We also note that the Internet census report is not always in sync with the provided data. These problems render the data unusable for many further analyses and conclusions drawn from the measurements, e.g., the size of the Internet estimated by the authors and recited in the press.

Finally, we believe that our analysis provides educationally useful hints about pitfalls in Internet scale measurements and large data analysis. Moreover, we question the scientific contribution of the Internet census and point to related work.

8. REFERENCES

[1] internet census, "Port scanning /0 using insecure embedded devices." `http://seclists.org/fulldisclosure/2013/Mar/166`, 2013.

[2] A. King and A. Dainotti, "Carna botnet scans confirmed." `http://blog.caida.org/best_available_data/2013/05/13/carna-botnet-scans/`, 2013.

[3] J. Heidemann, Y. Pradkin, R. Govindan, C. Papadopoulos, G. Bartlett, and J. Bannister, "Census and survey of the visible internet," in *ACM IMC*, 2008.

[4] P. Eckersley and J. Burns, "An observatory for the SSLiverse." Talk at Defcon 18, 2010. `https://www.eff.org/files/DefconSSLiverse.pdf`.

[5] D. Leonard and D. Loguinov, "Demystifying service discovery: implementing an internet-wide scanner," in *ACM IMC*, 2010.

[6] N. Heninger, Z. Durumeric, E. Wustrow, and J. A. Halderman, "Mining your Ps and Qs: detection of widespread weak keys in network devices," in *USENIX Conference on Security*, 2012.

[7] Y. Shavitt and E. Shir, "DIMES: let the internet measure itself," *ACM CCR*, vol. 35, pp. 71–74, Oct. 2005.

[8] K. Chen, D. R. Choffnes, R. Potharaju, Y. Chen, F. E. Bustamante, D. Pei, and Y. Zhao, "Where the sidewalk ends: extending the internet as graph using traceroutes from P2P users," in *ACM CoNEXT*, 2009.

[9] k. Claffy, Y. Hyun, K. Keys, M. Fomenkov, and D. Krioukov, "Internet mapping: from art to science," in *IEEE DHS Cybersecurity Applications and Technologies Conference for Homeland Security*, 2009.

[10] The Register, "Researcher sets up illegal 420,000 node botnet for IPv4 internet map." `http://www.theregister.co.uk/2013/03/19/carna_botnet_ipv4_internet_map/`, 2013.

[11] Huffington Post, "The most detailed, GIF-based map of the internet was made by hacking 420,000 computers." http://www.huffingtonpost.com/2013/03/22/internet-map_n_2926934.html, 2013.

[12] Spiegel Online International, "Mapping the internet: A hacker's secret internet census." http://www.spiegel.de/international/world/hacker-measures-the-internet-illegally-with-carna-botnet-a-890413.html, 2013.

[13] wired.co.uk, "Botnet-generated map of internet gathered data 'unethically'." http://www.wired.co.uk/news/archive/2013-05/16/internet-census, 2013.

[14] arstechnica.com, "Guerilla researcher created epic botnet to scan billions of IP addresses. with 9TB of data, survey is one of the most exhaustive—and illicit—ever done." http://arstechnica.com/security/2013/03/guerilla-researcher-created-epic-botnet-to-scan-billions-of-ip-addresses/, 2013.

[15] Yahoo News, "Watch 24 hours of internet activity around the world in 8 seconds." http://news.yahoo.com/watch-24-hours-internet-activity-around-world-8-113700364.html, 2013.

[16] Daily Mail, "This is what the internet looks like: Spectacular image created by computer hacker captures every iota of online data for 'day in the life' of the web." http://www.dailymail.co.uk/sciencetech/article-2299936/This-internet-looks-like-Spectacular-image-created-hacker-captures-iota-online-data-day-life-web.html, 2013.

[17] ZDNet, "What 420,000 insecure devices reveal about web security." http://news.cnet.com/8301-1009_3-57574919-83/what-420000-insecure-devices-reveal-about-web-security/, 2013.

[18] Anonymous, "Carna botnet scanning of all IPv4 addresses." http://www.auscert.org.au/render.html?it=17258, 2013.

[19] "Internet census 2012 search." http://www.exfiltrated.com/querystart.php, 2013.

[20] B. Krishnamurthy, W. Willinger, P. Gill, and M. Arlitt, "A socratic method for validation of measurement-based networking research," *Computer Communications*, vol. 34, no. 1, pp. 43 – 53, 2011.

[21] Anonymous, "Internet census 2012: Port scanning /0 using insecure embedded devices." http://internetcensus2012.bitbucket.org/paper.html, 2013.

[22] IANA, "The allocation of internet protocol version 4 (IPv4) address space." http://www.iana.org/assignments/ipv4-address-space/ipv4-address-space.xml, 2013.

[23] C. Huang, A. Wang, J. Li, and K. W. Ross, "Measuring and evaluating large-scale cdns," in *ACM IMC*, 2008. Paper withdrawn.

[24] S. Triukose, Z. Al-Qudah, and M. Rabinovich, "Content delivery networks: protection or threat?," in *European Conference on Research in Computer Security*, 2009.

[25] Motherboard, "This is the most detailed picture of the internet ever (and making it was very illegal)." http://motherboard.vice.com/blog/this-is-most-detailed-picture-internet-ever, 2013.

[26] "Ethical principles and guidelines for the protection of human subjects of research." http://www.hhs.gov/ohrp/humansubjects/guidance/belmont.html, 1979.

[27] M. Bailey, D. Dittrich, E. Kenneally, and D. Maughan, "The menlo report," *IEEE Security & Privacy*, vol. 10, pp. 71–75, Mar. 2012.

[28] "IMC 2013 call for papers." http://conferences.sigcomm.org/imc/2013/cfp.html, 2013.