Estimating IPv4 Address Space Usage with Capture-Recapture

Sebastian Zander, Lachlan L. H. Andrew, Grenville Armitage
CAIA, Swinburne University of Technology
Melbourne Australia
{szander, landrew, garmitage}@swin.edu.au

Geoff Huston
Asia Pacific Network Information Centre (APNIC)
Brisbane, Australia
gih@apnic.net

***Abstract*—As of April 2013 almost 95% of the IPv4 address space has been allocated. Yet, the transition to IPv6 is still relatively slow. One reason could be existing "IPv4 reserves" – allocated but unused IPv4 addresses. Knowing how many addresses are *actively used* is important to predict a potential IPv4 address market, predict the IPv6 deployment time frame, and measure progressive exhaustion after the IPv4 space is fully allocated. Unfortunately, only a fraction of hosts respond to active probes, such as "ping". We propose a capture-recapture method to estimate the actively used IPv4 addresses from multiple incomplete data sources, including "ping" censuses, network traces and server logs. We estimate that at least 950–1090 million IPv4 addresses are used, which is 36–41% of the publicly routed space. We analyse how the utilisation depends on various factors, such as region, country and allocation prefix length.**

***Index Terms*—Used IPv4 space, Capture-recapture.**

I. Introduction

As of April 2013 almost 95% of the usable IPv4 address space has been allocated and according to predictions, the Regional Internet Registrars (RIRs), except AfriNIC, will run out of IPv4 addresses by the end of 2014 [1]. However, the transition to IPv6 is still relatively slow. While most of the IPv4 address space has been *allocated*, it is unclear how many allocated addresses are *actively used* (from now on simply referred to as *used*). Knowing how many addresses are used is important to predict the value and costs of a potential IPv4 address market and the time frame of IPv6 deployment. Also, once the IPv4 space is fully allocated, its progressive exhaustion can only be measured through tracking the usage.

Surprisingly little work exists on identifying how much of the IPv4 space is used. To our knowledge the only existing studies are Pryadkin *et al*. [2], Heidemann *et al*. [3], Cai *et al*. [4] and a recent port scan census [5]. The previous studies were based mostly on active probing ("pinging") of the IPv4 address space. But pinging alone severely under-counts, since many hosts do not respond or their responses are filtered (e.g. firewalls). Apart from a simple correction factor in [3], previous work did not attempt to estimate the true population.

We propose an approach to estimate the number of used IPv4 addresses that 1) combines several sources of active and passive measurement data and 2) uses a statistical model to estimate the true population from the measurement data. We focus on 76% of the allocated space that is publicly routed (based on [6]), since only for this space we can (directly or indirectly) observe network traffic to detect used IPv4 addresses. Since our observation period extends over two years and many Internet addresses are allocated dynamically, the number of used addresses we measure is likely larger than the number of simultaneously used IPv4 addresses. However, we argue that often the difference are addresses that are de-facto used because they are on "stand-by", and we also provide separate estimates for dynamic and static space.

We "pinged" the whole allocated IPv4 space with ICMP echo requests and TCP port 80 SYNs. We also collected passive measurement data from different sources: logs of web servers that carry out IPv6 readiness measurements with random clients [7], logs from an email spam detector [8], Wikipedia edit logs, logs of Valve's Steam online gaming platform, logs from Measurement Lab [9], and NetFlow traffic logs from our university's access router.

The combined data from all sources detects significantly more used IPv4 addresses than the ping data alone. Still we assume there are many used addresses that for different reasons were not observed by any of our sources. We use the capture-recapture (CR) method [10]–[13] to estimate the total population of used addresses, including the *unobserved* used addresses. Simple CR methods, such as [10]–[11], make unrealistic assumptions, for example they assume the sample probability is the same for all IPv4 addresses. We use more complex log-linear models that are less restrictive and can deal with more realistic scenarios, for example when different types of addresses have different sample probabilities [12], [14].

From multiple ping censuses we detected 524 million used IPv4 addresses. When combined with the passive sources we detected 714 million used addresses. Based on log-linear CR models we estimate that there were at least 950–1090 million used IPv4 addresses, which means at least 36–41% of the publicly routed space was actually used. Due to various factors discussed later our results are likely to be underestimates. Besides a total estimate, we also provide some interesting insights into how the IPv4 address utilisation depends on RIR, country, allocation prefix size and allocation age.

The paper is organised as follows. Section II describes the basic concept of CR and log-linear CR models. Section III describes the IPv4 address dataset collection and processing. Section IV presents our results. Section V discusses related work. Section VI concludes and outlines future work.

Paper will be published at the 7th IEEE Workshop on Network Measurements (WNM) in conjunction with IEEE LCN 2013, Sydney, Australia, Oct. 2013.

II. Capture-Recapture

There are many techniques for estimating population sizes based on limited samples. Some use problem-specific approaches (e.g. [15]), but most use an approach called *capture-recapture* (CR). CR arose in ecology [10]–[11], but is also widely used in epidemiology [12]–[16], and was used to estimate missing links from observed autonomous system graphs [17]. We will first discuss the simplest possible CR technique, followed by the log-linear models that we use.

A. Two-sample method

The simplest CR model is the two-sample Lincoln-Petersen (L-P) method [10]–[11], which works as follows. Given a first sample, of M individuals, the size of the population would be known if we knew what *fraction* of the population had been sampled. To estimate this, L-P takes a second sample of C individuals, of which R individuals occur in both samples. If the fraction of "recaptured" individuals in the second sample equals the fraction of the total population captured in the first sample, then the population N can be estimated by [10]–[11]:[1]

$$R/C = M/N \ , \quad N = \frac{MC}{R} \ .$$

In our context, the samples or "sources" are different active and passive measurements (see Section III). For concreteness, consider one source to result from pinging the entire IPv4 space and another to be all addresses in a traffic trace.

The L-P estimate assumes that the probability of an individual being captured in one source does not depend on the probability of being captured in a different source (*independent sources*). It also assumes that, within a sample, each individual has an equal chance of being sampled (*homogenous population*), specifically that the probability is not zero for any individual. Individuals with zero sample probability are not part of the estimated population (in our case this may be some specialised devices, see Section III-C). Furthermore, the L-P estimate assumes that during measurement no individuals enter or leave the population (*closed population*), but a violation of this assumption is simply another form of heterogeneity.

Given our current data sources, there is no significant causal relationship to introduce source dependence. However, the population is very heterogeneous; for example, servers are more likely to respond to pinging, while client machines may be more likely to appear in certain traffic traces. This gives rise to *apparent source dependence*, which must be treated similarly. We must also avoid incorrectly believing an address has been sampled, due to address spoofing.

We have multiple sources of data. In our results for L-P, we consider every possible way to split the data sources into two groups, and for each pair of groups, we calculated an L-P estimate, and report the overall minimum and maximum estimates. This is more accurate than simply taking pairwise estimates. We do not report L-P standard errors, since they severely underestimate the errors introduced by heterogeneity.

[1]The L-P estimator is biased. In our analysis we actually use Chapman's unbiased variant defined in [18].

Table I
THREE-SOURCE CONTINGENCY TABLE

		Source 1			
		yes		no	
		Source 2		Source2	
		yes	no	yes	no
Source 3	yes	Z_{111}	Z_{101}	Z_{011}	Z_{001}
	no	Z_{110}	Z_{100}	Z_{010}	Z_{000}=?

If there is (apparent) dependence such that two sources are positively correlated, the L-P estimator underestimates the true population size: $R/C > M/N$ and so $N > MC/R$. If the sign of the correlation is known, then L-P estimates can still be used to identify plausible lower or upper bounds [12]. In general, the correlation is not known. However, just as L-P uses a second sample to estimate the fraction of the population of the first sample, so a third sample can be used to estimate the correlation between the first two samples. This is the basis of the log-linear models discussed below.

B. Log-linear Models

Log-linear CR models (LLMs) [12], [14], [16] generalize L-P to model (apparent) source dependence among arbitrarily many sources.

1) Description: Let N be the unknown number of distinct individuals of the population. Let t denote the number of sources indexed by $1, 2, \ldots, t$. For each individual, let s_1 to s_t be defined such that $s_i = 1$ if the individual occurs in sample i and $s_i = 0$ otherwise. Then the string $s_1 s_2 \ldots s_t$ is called the "capture history" of the individual. The observed outcome of all measurements can then be represented by variables of the form z_s, which are the numbers of individuals with each capture history $s = s_1 s_2 \ldots s_t$. These are assumed to be instances of random variables Z_s. Note that individuals with the capture history $00\ldots0$ are unobserved, and our goal is to estimate $Z_{00\ldots0}$. This is illustrated in the form of an incomplete contingency table in Table I for $t = 3$.

For each history s, let $h(s)$ be the set of samples in which the individual occurs; for example, $h(101) = \{1, 3\}$. Define the indicator function $\mathbf{1}_A = 1$ if statement A is true and 0 otherwise. We can now write the following system of equations in 2^t variables $u, u_1, u_2, \ldots, u_{12}, \ldots, u_{23}, \ldots$ up to $u_{12\ldots t}$:

$$\log(\mathbb{E}\left(Z_s\right)) = \sum_{h \subseteq h(s)} u_h = \sum_h u_h \mathbf{1}_{h \subseteq h(s)} \ . \qquad (1)$$

For example, for $t = 3$, the system is

$$\begin{aligned} \log\left(\mathbb{E}\left(Z_{ijk}\right)\right) = & u + u_1 \mathbf{1}_{i=1} + u_2 \mathbf{1}_{j=1} + u_3 \mathbf{1}_{k=1} \\ & + u_{12} \mathbf{1}_{i=1 \wedge j=1} + u_{13} \mathbf{1}_{i=1 \wedge k=1} \\ & + u_{23} \mathbf{1}_{j=1 \wedge k=1} + u_{123} \mathbf{1}_{i=1 \wedge j=1 \wedge k=1} \ . \end{aligned}$$

The estimate of $Z_{00\ldots0}$ is then $\hat{Z}_{00\ldots0} = \exp(u)$. If we take $\mathbb{E}[Z_s] = z_s$ then this system has 2^t unknowns but only $2^t - 1$ equations, as $Z_{00\ldots0}$ is unknown. Hence it is customary to assume $u_{12\ldots t} = 0$ [12]. As the number of sources t increases, this t-way dependency becomes decreasingly important.

Paper will be published at the 7th IEEE Workshop on Network Measurements (WNM) in conjunction with IEEE LCN 2013, Sydney, Australia, Oct. 2013.

For large t, this model is sensitive to small values of Z_s; a zero count for some capture history may give $\hat{Z}_{00\ldots0} = 0$, regardless of the other Z_s [12]. This over-fitting is mitigated by "model selection" (see Section II-B2), in which some u_h are forced to 0, to reflect assumed independence between certain combinations of sources. For example, setting $u_{12} = 0$ indicates sources 1 and 2 are independent. With such incomplete models, the system of equations is overdetermined, and the maximum likelihood parameters u are typically used, based on the assumption that Z_s result from random sampling and are hence Poisson distributed. Our fitting is based on the glm() function of R [19] using iteratively reweighted least squares.

Even with appropriate model selection, it may be that some z_s are near zero. In our case this rarely occurs: only when we combine CR with stratification into many strata, such as stratification by country (see Section II-C). To mitigate this, we exclude strata where the number of samples is below a threshold and exclude estimates above the routed IPv4 space.

After model selection, we use the procedure of [20]–[21] to compute a $100\,(1-\alpha)\,\%$ profile likelihood "confidence interval" (CI) for $\hat{N}$. Note that this is not a true confidence interval in our case, since it is based on the assumption that each sample is drawn randomly, resulting in a Poisson number of samples with each history. In contrast, our samples arise from different, not completely random sampling procedures. Hence we treat these "confidence intervals" as merely a useful heuristic indication of the sensitivity to modelling variations and we set $\alpha = 10^{-7}$ to obtain wide CIs.

LLMs strictly generalize the L-P estimator. We can rewrite the latter as $N - M - C + R = (M-R)(C-R)/R$, which in the LLM notation is $\mathbb{E}[Z_{00}] = z_{01}z_{10}/z_{11}$. Taking log of this and substituting (1) gives $\log(\mathbb{E}[Z_{00}]) = (u+u_1)+(u+u_2)-(u+u_1+u_2) = u$, which is the same as the LLM estimator.

2) Model selection: Model selection for an LLM consists of selecting which u_h will be assumed *a priori* to be 0. The goal is to select the least complex model with "adequate" fit. A key assumption is that the model that best describes the observed individuals, also describes the unobserved individuals [14].

Our approach is based on Akaike's Information Criterion (AIC), which is the most widely used IC when there is no prior information about the parameters, and which performs well when there is a "tapering effect" in the impact of the variables that is only revealed gradually as sample size increases [22]. The technique seeks to minimizes the AIC, defined as [23]

$$\mathrm{AIC} = 2k - 2\log(L)\,,$$

where L is the likelihood of the data given the assumed model and k is the number of free parameters of the model.

In our case, k is the number of non-zero u_h, but L is difficult to obtain. Like most information criteria, AIC assumes that each source samples randomly. As in epidemiology [12]–[16], the randomness in our case comes largely from the choice of which sources to monitor, which is hard to characterise.[2]

[2]For example, hosts will deterministically respond to pinging or not, but packet loss, dynamic addresses and network changes introduce randomness (overlap between two consecutive ping censuses is only 60–70%).

Randomly sampling typically gives Poisson distributed Z_s. However, given our large number of samples ($\sim 10^8$) the statistical fluctuations in a Poisson variable are negligible compared with the source inhomogeneity, and AIC would lead to models with excessively many parameters. However, for want of a better criterion, we use the following heuristic based on the AIC. We first divide all z_s by 10^3 (except for strata with few samples) and then calculate L for each model *as if* the samples were Poisson distributed. We then choose the simplest model m such that no other model n has $\mathrm{AIC}_n < \mathrm{AIC}_m - 7$ (see [23] chapter 4).

C. Stratification

We also try to mitigate heterogeneity by stratifying the population into more homogeneous sub-populations based on observable covariates (which also allows investigating the population of different sub groups). We use different stratifications. We classified IPv4 addresses as statically or dynamically assigned (see Section III-D) and based on the allocation/whois data we stratified by RIR (e.g. APNIC), country, prefix size, industry[3] and allocation age. For each stratification the estimated total number of used IPv4 addresses is the sum of the estimated used IPv4 addresses over all strata.

III. Datasets

Now, we describe the sources of observed used IPv4 address data and the data collection and processing. We also discuss the types of hosts sampled and how we deal with dynamically assigned addresses that cause potential bias.

A. Datasets

We obtained the first dataset (PING) by actively probing the whole allocated IPv4 Internet using ICMP echo requests and TCP SYN packets sent to port 80[4]. We probed each address four times (four censuses) with gaps of at least 3–4 months between censuses. Only ICMP probing was used for the first census, whereas ICMP and TCP probing were used for the other censuses. With TCP probing the number of observed IP addresses increased by over 7% compared to ICMP-only probing. We took care not to trigger intrusion detection systems, and on average we received only about 10 complaints per census.

We collected IPv4 addresses from Wikipedia's page edit histories (WIKI). All changes made by unregistered users are logged with the client's IPv4 address and the modification time. The next source was a list of potential spam email senders from [8] (SPAM), which includes open relays or clients that were used for spamming via botnets. We also collected the IPv4 addresses of clients tested by Measurement Lab [9] tools (MLAB) and of web clients that were tested by our IPv6 readiness test [7] or accessed the APNIC labs web

[3]"Industry" indicates whether address space is education, military, government, corporate, or ISP. We classified 88% of the allocated address space based on whois information (down to /17 networks), the rest remains unclassified.

[4]Initially we probed a sample of the Internet using different commonly used TCP ports and found port 80 to be the most responsive.

Table II
DIFFERENT OBSERVED IPV4 ADDRESS DATA SOURCES

Dataset	Time collected	Unique IPs [million]
PING	Sep 2011 – Mar 2013	524.7
GAME	Jan 2011 – Mar 2013	291.4
SWIN	Jan 2011 – Mar 2013	190.9
APNIC	Mar 2011 – Mar 2013	122.4
MLAB	Jan 2011 – Mar 2013	57.2
SPAM	May 2012 – Mar 2013	23.3
WIKI	Jan 2011 – Mar 2013	12.4

server (APNIC). Finally, we collected IPv4 addresses of users of Valve's Steam online gaming platform from their server logs (GAME) and from NetFlow records of Swinburne University's access router (SWIN).

B. Dataset collection and processing

For PING we grouped responses into positive acknowledgments (ACKs) and negative acknowledgements (NACKs). We only use the IPv4 addresses that responded with ACKs, because for NACKs it is often unclear if the probed addresses were actually used (e.g. routers or firewalls may have generated NACKs). For ICMP probing we treated ICMP echo replies, "destination protocol unreachable" and "destination port unreachable" messages as ACKs, if they were sent by the probed IPs. All other ICMP errors or "TTL exceeded" messages we counted as NACKs. For TCP probing we counted all SYN/ACKs as ACKs, and all RSTs as NACKs (25% of received RSTs cover nearly contiguous /25 or larger networks, which could mean they originated from firewalls).

A lack of reply could occur because an address was truly unused, a host ignored the probe, or the probe or response was filtered or lost. We assume that our prober did not induce congestion and the probe rate was below typical ICMP/TCP limiting thresholds (on average the prober sent one packet every two hours to a particular /24 network).

For the passive datasets we extracted the IPv4 addresses from log files. We filtered out private addresses (e.g. 10.0.0.0/8), multicast addresses, and unrouted addresses. For WIKI, SPAM, MLAB, APNIC and GAME server logs, the addresses are only recorded for established TCP sessions. However, SWIN is based on NetFlow data and can contain spoofed addresses: addresses that have not been actually used.

An analysis of SWIN revealed uniformly random distributed addresses in several routed /8 prefixes, some of which we know are not used or only lightly used. Many of the traffic flows from the presumably spoofed IPs showed a few characteristic average packets lengths, which we then used as heuristic to filter out potentially spoofed IP addresses. This removed most of the IP addresses in the space that we know was unused.

For datasets covering long time periods, such as WIKI, we only use addresses seen since January 2011. Table II shows the number of unique IPv4 addresses in each dataset after the processing.

C. Host types sampled

Different classes of hosts have different chances of appearing in each data set. We differentiate between the following host types: routers, servers/proxies, general-purpose clients (including PCs, mobile devices, PC-like devices such as TVs), and specialised devices (including printers, IP phones, cameras, controllers). We assume most networks are firewalled, and some networks use network address translation (NAT).

ISP routers are sampled by PING (and SWIN) only. Home routers are sampled by PING[5] and in most cases by all other sources because of NAT. Public servers/proxies are sampled by PING and SWIN. They can also appear in the WIKI, SPAM and APNIC datasets. Private servers/proxies are not sampled by any of our datasets and are effectively invisible.

General-purpose clients are sampled by WIKI, SPAM, MLAB, APNIC, GAME and SWIN. NAT'ed clients also appear in PING. Each of these datasets may be biased towards a specific set of clients. SPAM, MLAB, GAME and APNIC are likely biased towards home users or users at more "open" corporate or education sites allowing access to recreational web sites. However, WIKI, APNIC and SWIN may also contain client IPs of more restrictive sites. Clients that only communicate in private networks are effectively invisible.

Specialised devices not behind NATs are likely not sampled by any of our passive data sources; some may be sampled by PING, but probably most of them are effectively invisible.

D. Dynamic vs. static addresses

Many hosts have dynamically assigned IPv4 addresses that may change over time due to mobility or assignment by DHCP or PPPoE. Since measurement of the used IPv4 address space takes time, the number of addresses seen is likely to exceed the number of simultaneously used addresses. However, this "over-count" is counter-balanced, since dynamic addresses likely increase the correlation between sources, and so reduce the L-P estimates. Unfortunately, the resulting effect is very hard to quantify. We argue that over-count is in many cases addresses that are on "stand-by" and de-facto used.

We also divided the IPv4 space into dynamic and static /24 networks and provide separate estimates. We chose /24s subnets, as this is the smallest size advertised by BGP. We performed pairwise comparisons of the data from the four ping censuses and computed the overlapping used IPv4 addresses for each /24 and pair of censuses. Intuitively, for static /24s the majority of used addresses should be the same in different censuses. The time between two subsequent ping census was at least 3–4 months, which is long enough so that dynamic IPs would have changed.

We now define our approach more formally. For a particular subnet observed in two censuses i and j, let A_i and A_j be the sets of detected used IPv4 addresses, and let $A_b = A_i \cap A_j$ be the overlapping set. Let the "rate" of detected or overlapping

[5]We downloaded the web pages for devices responding to port 80 TCP SYN probes, and manual inspection confirmed that some responses were indeed from Cable or DSL routers.

addresses for a subnet i be $r_i = |A_i|/256$ and the minimum number of addresses be $\theta = \min(|A_i|, |A_j|)$. We classified a /24 as static if $\theta \geq 4$ and $r_b > r_i r_j$ and either $|A_b|/\theta \geq \delta$ or $\theta - |A_b| \leq 2$ (using $\delta = 75\%$). With four ping censuses, we obtained six classifications and used simple majority vote to make decisions, i.e. we classified a /24 as static if four out of six classifications indicated static.

Static subnets include subnets with static allocation, subnets where many addresses are effectively static, or subnets where most addresses are used (quasi-static). Dynamic subnets include all subnets where the used addresses changed significantly between censuses (presumably due to dynamic allocation), and (almost) empty subnets that we cannot classify. About 74% of /24s were almost empty in at least two censuses (for the stratification we treat them as dynamic), but 26% of /24s we could classify into static (9%) or dynamic (17%). The static IP address percentage is highest (16%) in PING, relatively low (5–8%) in WIKI, MLAB, APNIC, SWIN, and lowest (2–3%) in SPAM and GAME.

We used reverse DNS (rDNS) names to verify our classification. For roughly 870 000 non-empty classified /24s (23% of all non-empty classified) the rDNS names indicate a dynamic or static /24 subnet, e.g. by containing the words "dynamic" or "static" (or variations thereof). For 88% of these our classification and rDNS names are consistent. For 7% we classified the /24 as static but the name indicates dynamic (static networks with high fluctuations or misleading rDNS names). For 5% we classified the /24 as dynamic but the name indicates static (dynamically assigned but quasi-static /24s or misleading rDNS names).

Figure 1 plots the number of static and dynamic /24 subnets against the corresponding allocation prefix sizes. The number of static and dynamic subnets is roughly equal for /8 and /18 or larger prefixes, but the number of dynamic subnets is higher for prefixes /9 to /17. This result is consistent with expectations. Smaller prefixes (larger blocks) are typically owned by ISPs, which commonly use dynamic addressing for many of their customers; however, until 1998 the pre-CIDR /8 and /16 prefixes were mainly assigned to non-ISP corporations, universities, government or military.

IV. Results

We first use cross-validation to show that LLMs provide reasonable accurate estimates and outperform L-P. Then we estimate the total number of used IPv4 addresses depending on different stratifications, and finally we investigate how the estimated used IPv4 space depends on RIR, country, allocation prefix size and allocation age.

A. Cross-validation

We cannot directly validate the accuracy of estimates of used IPv4 addresses, since the actual number (the ground truth) is unknown. However, we can perform cross-validation with our t data sources. Instead of estimating the addresses not seen by any of the sources, we treat the number of IPv4 addresses seen *only* by one particular source i as unknown and

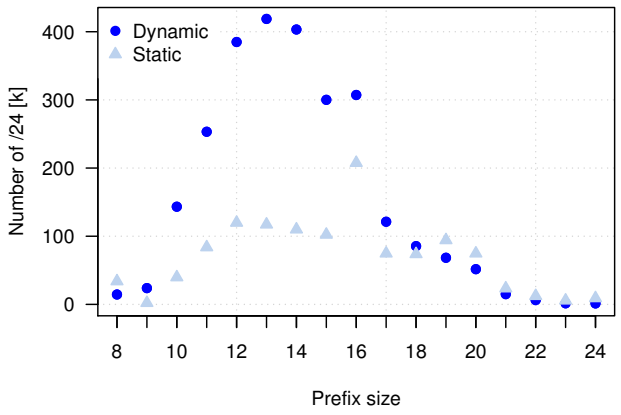

Figure 1. Number of /24 subnets classified as dynamic or static depending on the prefix size of the allocation a /24 is part of

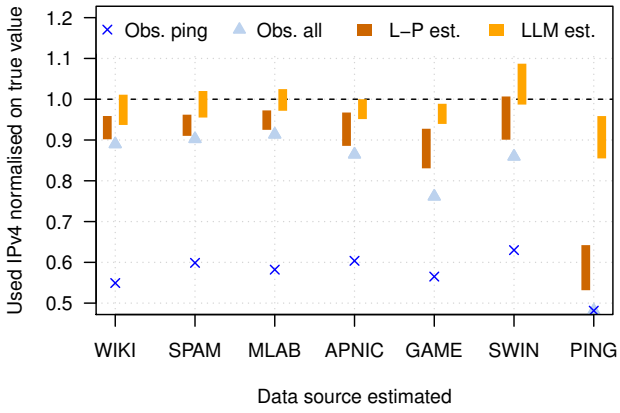

Figure 2. IPv4s observed by ping, IPv4s observed by all other sources, and the estimated ranges for L-P and LLM normalised on the true number of unseen IPv4s for each data source

estimate this number using CR with the remaining data (the overlap of source i with all other $t-1$ sources). We do this for all sources. In this situation we know the ground truth and can evaluate the accuracy of the CR estimates. Furthermore, we can identify whether there is positive or negative bias.

If the cross-validation shows low accuracy, it suggests that the accuracy of an estimate of the IPv4 addresses not seen by any of the sources is also low. If the cross-validation shows high accuracy, an estimate of the unseen IPv4 addresses may be reasonably accurate, but we do not know with certainty.

Figure 2 shows the number of IPs in each source also observed by PING, the number of addresses of a source also observed by any other sources, and the ranges of the L-P estimates (minimum to maximum) and LLM estimates (confidence interval based on profile likelihood). Since the sources have very different sizes we normalised the number of addresses based on the total number of addresses observed by each source (the ground truth). All sources other than PING have relatively high overlap (50–60% is due to PING), but 10–25% of addresses are unique to each source. In contrast almost 50% of addresses in PING were only seen by PING.

Table III
OBSERVED AND ESTIMATED USED IPV4 ADDRESSES FOR AGGREGATED, STATIC AND DYNAMIC SPACE (AND SUM OF STATIC AND DYNAMIC ESTIMATES)

	Observed [M]	Unstratified estimates [M]	Stratified estimates [M]					Range over all stratifications [M]
			RIR	Country	Age	Prefix size	Industry	
Aggregated	714	980–1040	958–995	1049–1136	985–1070	978–1073	999–1045	958–1136
Static	91	122–133	121–129	121–132	121–132	121–131	119–126	119–132
Dynamic	623	815–827	849–898	894–962	842–901	837–907	867–912	837–962
Static+Dynamic	714	937–960	970–1027	1015–1094	963–1033	948–1038	986–1036	**948–1094**

We treat an L-P or LLM estimate as correct if the ground truth is in the estimated range. Figure 2 shows that L-P produces a correct estimate only for SWIN but otherwise underestimates, and the estimate for PING is very poor. Still in all cases the L-P estimates are an improvement over the number of observed addresses. LLMs, despite producing slightly smaller ranges, estimate WIKI, SPAM, MLAB, APNIC and SWIN correctly and GAME almost correctly. LLMs still underestimate PING, but the estimate is relatively close to the ground truth and a huge improvement over L-P.

Since LLM outperforms L-P, we only present LLM estimates in the following sub sections.

B. Overall used IPv4 space

Table III shows the observed addresses and the LLM-estimated used IPv4 addresses for the aggregated, static and dynamic space, both without stratification and depending on different stratifications. It also shows the sum of the static and dynamic estimates and the estimated ranges for each stratification. We observed 714 million addresses, but we estimate that 950–1090 million addresses were actually used (estimates are largely consistent across different stratifications). This means we observed 27% of the routed addresses, and we estimate that 36–41% of the routed addresses were used (based on [6]). Our estimated range covers the number of IPv4 addresses with rDNS entry, which is 1050 million [5]. However, whether the number of rDNS entries is a useful indicator for the number of used IPv4 addresses or not requires further study.

The estimates are plausible for all strata including the 209 estimated countries[6], since the minimum estimates are always below the number of publicly routed addresses. With the exception of four countries, the maximum estimates are also below the limit. As discussed in Sections II and III our sources do not adequately capture certain types of IPv4 addresses, such as specialised devices. Hence, the minimum of our estimated range is probably an underestimate of the total number of IPv4 addresses used, but the number of actually used IPv4s may also exceed the maximum of our estimated range.

C. Used IPv4 space by RIR

Figure 3 shows the LLM-estimated number of used IPv4 addresses and the percentage of used publicly routed space (utilisation) by RIR. We differentiate between addresses observed through ping (Obs. ping), all observed addresses (Obs. all), and the estimated minimum (Used min.) and maximum (Used max.) from the LLM estimates. APNIC (Asia), RIPENCC (Europe) and ARIN (North America) have the highest number of used addresses, but the utilisation is highest for LACNIC (South America), followed by RIPENCC and APNIC.

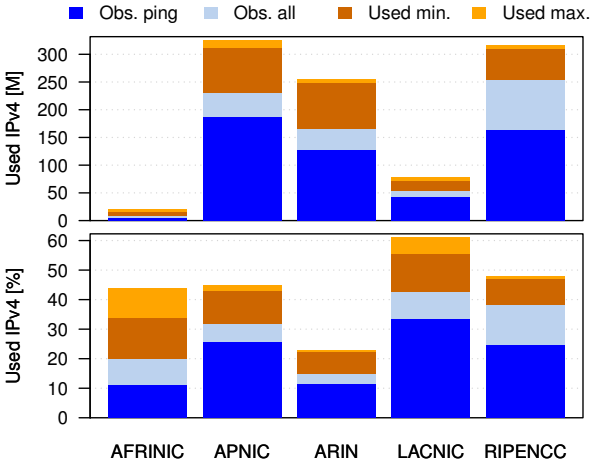

Figure 3. Number and percentage of observed and estimated used IPv4 addresses by RIR

D. Used IPv4 space by allocation age

Figure 4 shows the LLM-estimated number of used IPv4 addresses and the utilisation based on the allocation year. Most of the used address space is in allocations made since the 1999 .com boom heights, and these later allocations also have higher utilisations. Allocations until 1987 were mostly small with a relatively high percentage of universities (explaining the moderate utilisation). From 1988 to 1998 large allocations were given to corporations, military and government directly, and based on the estimates they only use small fractions of their space (partly this may be due to underestimation given our data sources). The utilisation decrease in recent years suggests that it takes up to 2–3 years to fill the address space.

E. Used IPv4 space by prefix size

Figure 5 shows the LLM-estimated number of used IPv4 addresses and the utilisation depending on the size of the allocated prefix. Prefix size has a relationship with allocation age, which is reflected in the results. The pre-CIDR prefixes /8, /16 and /24 that were allocated mainly until 1998, with

[6] For another 25 country codes of small countries/territories we did not compute estimates due to the very small number of samples.

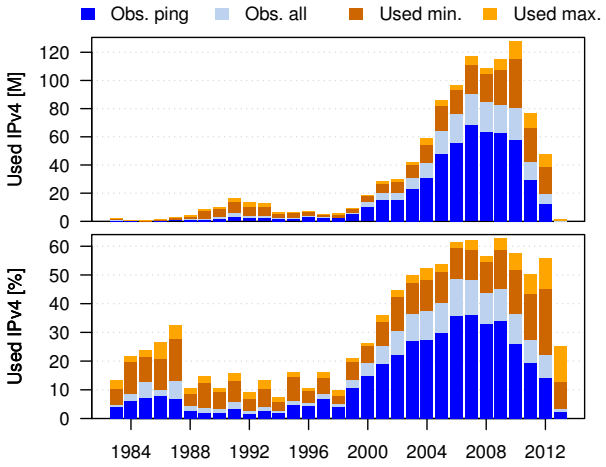

Figure 4. Number and percentage of observed and estimated used IPv4 addresses by allocation year

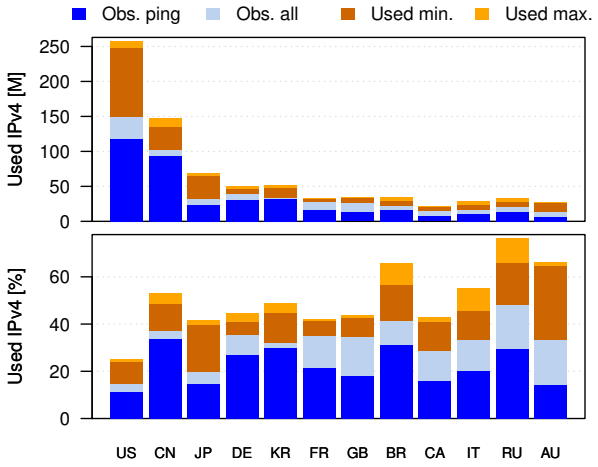

Figure 6. Number and percentage of observed and estimated used IPv4 addresses of the 12 countries with the largest allocations

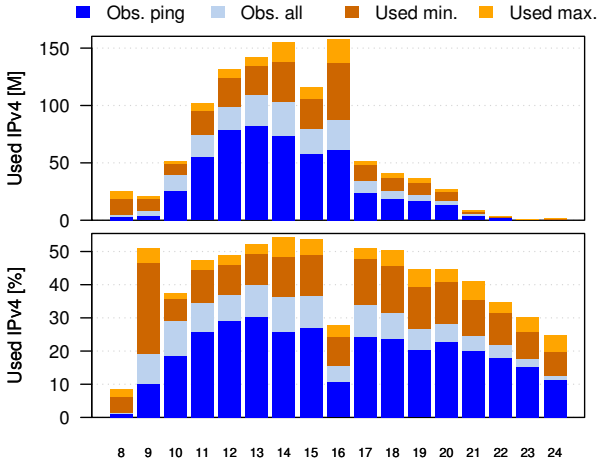

Figure 5. Number and percentage of observed and estimated used IPv4 addresses by prefix size

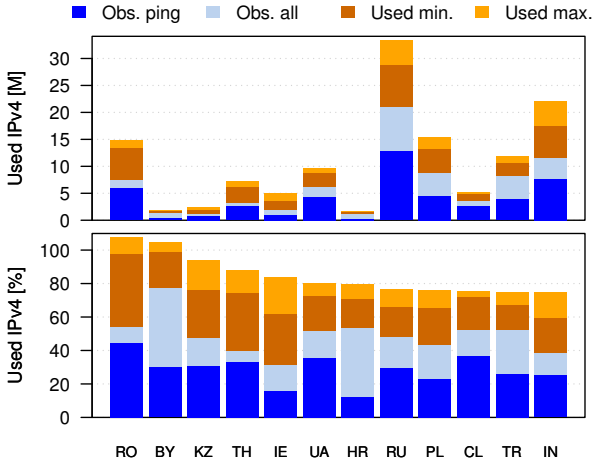

Figure 7. Number and percentage of observed and estimated used IPv4 addresses of the 12 countries with the largest estimated used percentages

a large part of the space belonging to non-ISP corporations and military/government, have lower utilisation. The /22 and /23 prefixes also have slightly lower utilisation, likely because their number has greatly increased over the last 1–2 years due to the address space shortage of RIRs, and the new space is still filling. The remaining prefixes, allocated mainly after 1998 largely by ISPs, have relatively similar utilisations is in the range of 35–50%.

F. Used IPv4 space by country

Figure 6 shows the LLM-estimated number of used IPv4 addresses and the utilisation for the 12 countries with the largest routed space (ordered by size of routed space from left to right). The USA has the most used IPv4 addresses and the lowest utilisation. The other countries (CN:China, JP:Japan, DE:Germany, KR:Korea, FR:France, GB:Britain, BR:Brazil, CA:Canada, IT:Italy, RU:Russian Federation, AU:Australia) have varying allocation sizes and used addresses, but the estimated utilisation is mostly 40–60%.

Figure 7 shows the top-12 countries (with at least one million observed addresses) with the highest percentages of estimated used addresses (ordered by maximum estimated used percentage from left to right): RO:Romania, BY:Belarus, KZ:Kazakhstan, TH:Thailand, IE:Ireland, UA:Ukraine, HR:Croatia, RU:Russian Federation, PL:Poland, CL:Chile, TR:Turkey, and IN:India. The estimated utilisation of these countries is at least 60–80% – significantly higher than for the 12 countries in Figure 6.

V. Related Work

Some prior research tried to infer address usage based on prefixes advertised by BGP [24], [25]. However, *actively used* addresses differ from *routed* addresses.

Pryadkin *et al.* [2] used ICMP echo and TCP SYN probing to probe the allocated Internet. They discovered 62 million used IPv4 addresses in 2003/2004. Pryadkin *et al.* also analysed the distribution of used IPv4 addresses within routable and allocated prefixes. They observed that only a small number

of prefixes appeared to be heavily used, while a large part of the space appeared either unused or underutilized.

Heidemann *et al*. [3] probed all allocated IPv4 addresses every few months (census) and more frequently probed selected address samples (survey) with ICMP echo pinging to study the usage, availability and up-time of addresses. Their last census in 2007 accounted for 112 million used IPv4 addresses. Heidemann *et al*. compared ICMP probing with TCP port 80 probing and passive measurements based on small samples and proposed a correction factor of 1.86, thus estimating the total number of used IPv4s in mid 2007 was 200–210 million.

Cai *et al*. [4] used ping survey data from [3] and conducted more surveys in 2009–2010 to analyse whether contiguous addresses are consistent, the typical block sizes used, how many addresses are dynamically assigned, and the edge-link bit-rates. They did not estimate the used IPv4 address space, but made the observation that "most addresses in about one-fifth of /24 blocks are in use less than 10% of the time".

From June to October 2012, anonymous researchers used hacked low-performance routers to perform a port scan of the IPv4 Internet [5]. They detected 420 million addresses that responded to ICMP echo, and another 36 million addresses that responded to TCP SYN probes on several hundred ports. The number of addresses responsive to ICMP is broadly consistent with our findings. It is larger than the number we observed in a single seven-week census (300 million) but smaller than the number we observed in all four censuses combined. In our censuses 15–20 million addresses only reacted to port 80 TCP SYNs but not to ICMP; [5] indicates that probing hundreds of ports would only double this number.

VI. Conclusions and Future Work

We estimated the total number of *actively used* IPv4 addresses based on multiple data sources and capture-recapture (CR) models. From multiple ping censuses of the Internet we detected 524 million used IPv4 addresses. When combined with the other data sources we detected 714 million used IPv4 addresses. Based on log-linear CR models we estimate at least 950–1090 million used IPv4 addresses, which equates to 36–41% of the publicly routed space. Our results are likely to be underestimates; however we think they are closer to the truth than previous estimates, and they are more up to date.

The regions with the highest utilisation are South America and Europe, closely followed by Asia. North America has the lowest utilisation. Most of the used IPv4 address space is in allocations made since 1999, and these allocations also have higher utilisations than older allocations. The 12 countries with the largest allocations have utilisations of 40–60%, except for the USA that has 25–30% utilisation. In contrast, the 12 countries for which we observed the highest fractions of used addresses, have utilisations of 60–80%.

We plan to improve our estimates with more data sources and a refined technique, to evaluate our estimates against ground truth for selected networks where we know peak usage, and to estimate the number and size of unused address blocks.

Acknowledgements

This research was supported under ARC's Linkage Projects funding scheme (project LP110100240) in conjunction with APNIC Pty Ltd and by ARC grant FT0991594. We thank Valve Corporation and Swinburne ITS for providing IPv4 address data. We thank D. Buttigieg and C. Tassios for their help in getting the Swinburne data.

References

[1] G. Huston, "IPv4 Address Report." http://www.potaroo.net/tools/ipv4/index.html.
[2] Y. Pryadkin, R. Lindell, J. Bannister, R. Govindan, "An Empirical Evaluation of IP Address Space Occupancy," Technical Report ISI-TR 598, USC/ISI, 2004.
[3] J. Heidemann, Y. Pradkin, R. Govindan, C. Papadopoulos, G. Bartlett, J. Bannister, "Census and Survey of the Visible Internet," in *ACM Conference on Internet measurement (IMC)*, pp. 169–182, 2008.
[4] X. Cai, J. Heidemann, "Understanding Block-level Address Usage in the Visible Internet," in *ACM SIGCOMM Conference*, pp. 99–110, 2010.
[5] Anonymous, "Internet Census 2012 – Port scanning /0 using insecure embedded devices." http://internetcensus2012.bitbucket.org/paper.html.
[6] University of Oregon Route Views Project. http://www.routeviews.org/.
[7] S. Zander, L. L. H. Andrew, G. Armitage, G. Huston, G. Michaelson, "Mitigating Sampling Error when Measuring Internet Client IPv6 Capabilities," in *ACM Internet Measurement Conference (IMC)*, Nov. 2012.
[8] DNS-based Blacklist of NiX Spam. http://www.dnsbl.manitu.net/.
[9] Measurement Lab. http://www.measurementlab.net/.
[10] C. G. J. Petersen, "The Yearly Immigration of Young Plaice into the Limfjord from the German Sea," *Rept. Danish Biol. Sta.*, vol. 6, pp. 1–77, 1895.
[11] F. C. Lincoln, "Calculating Waterfowl Abundance on the Basis of Banding Returns," *U.S. Dept . Agric. Circ.*, vol. 118, pp. 1–4, 1930.
[12] E. B. Hook, R. R. Regal, "Capture-Recapture Methods in Epidemiology: Methods and Limitations," *Epidemiologic Reviews*, vol. 17, no. 2, pp. 243–264, 1995.
[13] A. Chao, "An Overview of Closed Capture-Recapture Models," *Journal of Agricultural, Biological, and Environmental Statistics*, vol. 6, no. 2, pp. 158–175, 2001.
[14] S. E. Fienberg, "The Multiple Recapture Census for Closed Populations and Incomplete 2k Contingency Tables," *Biometrika*, vol. 59, pp. 591–603, Dec. 1972.
[15] J. Zhou, Y. Li, V. K. Adhikari, Z.-L. Zhang, "Counting YouTube Videos via Random Prefix Sampling," in *ACM Internet Measurement Conference (IMC)*, pp. 371–380, 2011.
[16] A. Chao, P. K. Tsay, S. H. Lin, W. Y. Shau, D. Y. Chao, "The Applications of Capture-Recapture Models to Epidemiological Data," *Statistics in Medicine*, vol. 20, pp. 3123–3157, October 2001.
[17] M. Roughan, J. Tuke, O. Maennel, "Bigfoot, Sasquatch, the Yeti and other missing links: what we don't know about the AS graph," in *ACM Internet Measurement Conference (IMC)*, pp. 325–330, October 2008.
[18] D. G. Chapman, "Some Properties of the Hypergeometric Distribution with Applications to Zoological Censuses," *Univ. California Publ. Stat.*, vol. 1, pp. 131–160, 1951.
[19] R Core Team, *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2012.
[20] R. M. Cormack, "Interval Estimation for Mark-Recapture Studies of Closed Populations," *Biometrics*, vol. 48, pp. 567–576, 1992.
[21] S. Baillargeon, L.-P. Rivest, "Rcapture: Loglinear Models for Capture-Recapture in R," *Journal of Statistical Software*, vol. 19, pp. 1–31, April 2007.
[22] K. P. Burnham, D. R. Anderson, "Multimodel Inference - Understanding AIC and BIC in Model Selection," *Sociological Methods & Research*, vol. 33, pp. 261–304, 2004.
[23] E. Cooch, G. C. White, *Program MARK: A Gentle Introduction*. Cornell University, 2009.
[24] X. Meng, Z. Xu, B. Zhang, G. Huston, S. Lu, L. Zhang, "IPv4 Address Allocation and the BGP Routing Table Evolution," *ACM Computer Communication Review (CCR)*, vol. 35, no. 1, pp. 71–80, 2005.
[25] A. Sriraman, K. R. B. Butler, P. D. McDaniel, P. Raghavan, "Analysis of the IPv4 Address Space Delegation Structure," in *IEEE Symposium on Computers and Communications (ISCC)*, pp. 501–508, July 2007.