Capturing Ghosts: Predicting the Used IPv4 Space by Inferring Unobserved Addresses

Sebastian Zander
CAIA, Swinburne University of Technology
Melbourne, Australia
szander@swin.edu.au

Lachlan L. H. Andrew
Faculty of IT,
Monash University
Melbourne, Australia
lachlan.andrew@monash.edu

Grenville Armitage
CAIA, Swinburne University of Technology
Melbourne, Australia
garmitage@swin.edu.au

ABSTRACT

The pool of unused routable IPv4 prefixes is dwindling, with less than 4% remaining for allocation at the end of March 2014. Yet, the IPv6 adoption remains slow. We demonstrate a new capture-recapture technique for improved estimation of "IPv4 reserves" (allocated yet unused IPv4 addresses or routable prefixes) from multiple incomplete data sources. A key contribution of our approach is the plausible estimation of both observed and unobserved-yet-active (ghost) IPv4 address space. This significantly improves our community's understanding of IPv4 address space exhaustion and likely pressure for IPv6 adoption. Using "ping scans", network traces and server logs we estimate that 6.3 million /24 subnets and 1.1 billion IPv4 addresses are currently in use (roughly 60% and 40% of the publicly routed space respectively). We also show how utilisation has changed over the last three years and provide an up-to-date estimate of potentially-usable remaining IPv4 space.

Categories and Subject Descriptors

C.2.3 [**Computer-Communication Networks**]: Network Operations—*Network Monitoring*; C.4 [**Performance of Systems**]: Measurement Techniques

General Terms

Measurement

Keywords

Used IPv4 space, capture-recapture

1. INTRODUCTION

At the end of March 2014 less than 4% of the IPv4 address space remained unallocated by Regional Internet Registrars (RIRs). RIPE and APNIC have exhausted their supply and the other RIRs (except AfriNIC) will run out of prefixes by the end of 2014 [1]. Understanding the pressures for IPv6 adoption, and the scope of possible IPv4 address markets, requires plausible estimates of actual IPv4 address use – particularly the efficiency with which allocated prefixes are filled with actively-used addresses. Ideally, our estimation techniques should also help the community track progressive exhaustion once all routable IPv4 prefixes are allocated.

Prior studies on IPv4 space growth [2–4] and a port scan census from 2012 [5] were based mainly on active probing ("pinging"). Yet pinging alone will under-count, as many hosts do not respond or their responses are filtered (e.g., by firewalls). Recently, Dainotti *et al.* [6] used IPv4 data from darknets to estimate the used /24 networks. Apart from a simple multiplier in [3], previous work did not attempt to correct for under-sampling.

Our key contribution in this work is a new method to estimate the true population of both observed and unobserved (yet still active) IPv4 addresses using a statistical *capture-recapture* (CR) [7–9] model applied over diverse sources of active and passive measurement data. We significantly extend our earlier workshop paper, with refined methodology, additional data sources and greatly extended analysis [10].

Our second contribution is a three-year study of address use using our CR method. We "pinged" the allocated space with ICMP echo requests and TCP port 80 probes, and also gathered IPv4 data from web server logs [11], email spam detector logs [12], Wikipedia edit logs, logs of Valve's Steam online game platform, logs from Measurement Lab [13], and university access router's NetFlow logs. Inevitably, our sources only detect used addresses from 80% of the allocated space that is publicly routed (based on [14]).

Although our sources provided diverse evidence of active IPv4 address use, there are likely many in-use addresses that we never see. We utilise our CR method to estimate a total population of used IPv4 addresses (and /24 networks) that *includes* these unobserved addresses (ghosts). As many sources obtain measurements over weeks or months, our estimates of the used IPv4 addresses (and /24 networks) are based on observation periods rather than points in time. By cross-validation with our datasets, and comparison with a few samples of ground truth, we show our CR method provides better estimates than prior techniques.

We analyse "demand" – growth in address use – over the last two years relative to factors such as the RIR, country, or prefix size, and estimate the remaining "supply" of unused

prefixes. Our combined sources observed 5.8 million used /24 subnets and 740 million used IPv4 addresses, yet our CR technique indicates significantly higher actual usage. We estimate 6.3 million /24 subnets and 1.1 billion IPv4 addresses were used by the end of March 2014 (approx. 60% and 40% of the publicly routed space respectively). From the end of 2011 to March 2014, the growth in used /24 subnets and IPv4 addresses was roughly linear, with an increase of 0.5 million /24 subnets and 160 million IPv4 addresses per year.

This trend means an estimated 4.3 million publicly routed but currently unused /24 subnets could supply us until 2022. If, for example, only 75% of all routed /24 subnets could ever be used, remaining supply will run out in 2017. However, unrouted unused space may provide more supply. Europe and Asia have the highest utilisation, while Africa and South America show the fastest growth.

The paper is organised as follows. Section 2 describes the concept of CR and our log-linear CR models. Section 3 describes our IPv4 address data collection and processing. Section 4 covers the validation of our CR model. In Section 5 we analyse the growth of used IPv4 space over time, and in Section 6 we estimate the space still unused. Section 7 discusses related work. Section 8 concludes and outlines future work.

2. CAPTURE-RECAPTURE

There are many techniques for estimating population sizes from limited samples. Some use problem-specific approaches, but many use CR methods. CR methods have been used in ecology [7, 8], epidemiology [9, 15], and to estimate missing links from observed AS-graphs [16]. To illustrate CR we will first discuss the simplest CR technique. Then we will discuss the log-linear models we use.

2.1 Two-sample method

The simplest CR model is the two-sample Lincoln-Petersen (L-P) method [7,8], which works as follows. Given a first sample, that observes M individuals, the size of the population would be known if we knew what *fraction* of the population had been observed. To estimate this, L-P takes a second sample. Say it contains 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, $R/C = M/N$, then the population N is [7, 8]:

$$N = \frac{MC}{R} .$$

In our context, the samples or "sources" are different active and passive measurements (see Section 3). For concreteness, consider Source 1 to result from pinging the entire IPv4 space and Source 2 to be all addresses in a server log. Based on the number of unique addresses observed by Source 1 and Source, and the number of unique addresses observed by both sources (Overlap) CR allows to estimate the number of unobserved addresses (Unseen), as illustrated in Figure 1.

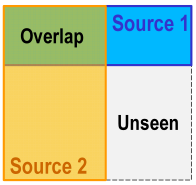

Figure 1: Two-source capture-recapture illustrated

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 (this may be some specialised devices, such as printers). Furthermore, the L-P estimate assumes that during measurement no individuals enter or leave the population (*closed population*).

Given our 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 logs. This gives rise to *apparent source dependence* that must be treated similarly. We must also avoid incorrectly believing an address has been sampled, due to address spoofing.

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 [15]. 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.

2.2 Log-linear Models

Log-linear CR models (LLMs) [15, 17, 18] generalize L-P to model (apparent) source dependence among arbitrarily many sources.

2.2.1 *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

Table 1: 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}=?

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 1 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 $u_{12\ldots t}$:

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

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 $u_{12\ldots t} = 0$ [15]. As the number of sources t increases, this t-way dependency becomes decreasingly important.

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 [15]. This over-fitting is mitigated by "model selection" (see Section 2.2.2), 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 uniform random sampling and are hence Poisson distributed.

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 2.3). To mitigate this, we exclude strata with fewer than 1000 samples.

After model selection, we use the procedure in [19] 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.

Typical log-linear models used in CR assume the Z_s are Poisson distributed, which is appropriate if the upper limit for the Z_s is unknown. However, we can bound Z_s by the size of the publicly routed IPv4 space. Hence we use right-truncated Poisson distributions defined over $[0, L] \cap \mathbb{Z}$, where L is the upper limit. These improve estimates substantially for small strata, where the counters are relatively close to the limit (see Section 4.2), but otherwise make little difference.

2.2.2 *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 of the observed (and by assumption) unobserved individuals [17].

A common approach is to minimize an "Information Criterion" (IC). Two common ICs are [20]:

$$\text{AIC} = 2k - 2\log(L), \qquad \text{BIC} = \log(M)k - 2\log(L)$$

where L is the likelihood of the data given the assumed model, k is the number of free parameters of the model and M is the number of observed individuals. AIC is used more often, but each has merits [21]. Section 4 compares the BIC and the AIC for our data. We choose the simplest model m such that no other model n has $\text{IC}_n < \text{IC}_m - 7$ [20].

In our case, k is the number of non-zero u_h, but L is difficult to obtain. AIC and BIC assume that each source samples uniformly and so L is the likelihood of a Poisson model. If the number of samples is large, the central limit theorem indicates that substantial deviations from the mean have very low likelihood. In our case, as in [15, 18], the randomness comes largely from the choice of sources to monitor, which is hard to characterise but has substantially higher variance. Hence the Poisson assumption selects too complex a model.

We mitigate this overfitting using the simple heuristic of dividing all z_s by some integer u when calculating L. It remains to select u. If u is so large that any z_s gets rounded to zero, the LLM breaks down. The further heuristic of selecting u to be the largest power of 2 not less than $\min_s z_s$ appears to work well.

2.3 Stratification

We obtain insight and mitigate heterogeneity by stratifying the population. We use different stratifications. We classified IPv4 addresses as statically or dynamically assigned using the approach described in [10] and based on allocation/whois data we stratified by RIR (e.g. APNIC), country, prefix size, industry[1] 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.

[1] "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).

3. DATASETS AND PREPROCESSING

An IPv4 address is considered *used* if it responds to active probes or participates in connections. A *used* /24 subset contains one or more used addresses. This section describes our sources of used IPv4 address data, our data collection and processing, and our handling of both spoofed and dynamically assigned addresses.

3.1 Datasets

Our first two datasets involve actively probing the whole allocated IPv4 Internet using ICMP echo requests (**IPING**) and TCP SYN packets to port 80[2] (**TPING**). Since mid-2011 we probed each allocated IPv4 address (a census) once every 6 months. The first two used ICMP probing and the rest used both ICMP and TCP probing (with TCP probing seeing over 7% more observed IP addresses). We took care to avoid triggering intrusion detection systems (receiving only 10–20 queries per census). For the first half of 2011 we use ICMP ping data collected by USC/LANDER [22].

Passively observed IPv4 data includes addresses from Wikipedia's page edit histories[3] (**WIKI**), potential spam email senders from [12] (**SPAM**), addresses of clients tested by Measurement Lab [13] tools (**MLAB**), web clients participating in our IPv6 readiness test [11] (**APNIC**), server logs of game clients connecting to Valve's Steam online gaming platform (**GAME**), and NetFlow records of Swinburne University of Technology's access router[4] (**SWIN**) and Caltech's access router (**CALT**).

We utilise data gathered from 2011 onwards. We analyse the growth trends of the number of used IPv4 addresses and used /24 subnets. We generate datasets of unique /24 subnets by processing the IPv4 datasets and setting the last octet of each address to zero and then filtering out the duplicates. Table 2 shows the number of unique IPv4 addresses and /24 subnets in each dataset for each year. Note that the numbers in the table cannot be used as growth trends due to sample method variations.

Hosts using public IPv4 addresses are either routers, servers/proxies, clients (e.g. PCs, smart phones), or specialised devices (e.g. printers, cameras). ISP routers are sampled by IPING and TPING. Home routers are sampled by IPING and TPING (we confirmed that some responses were from Cable or DSL routers) and by all other sources (with NAT packets sent from home networks will appear to come from home routers). Servers/proxies are sampled by IPING, TPING, SWIN and CALT. They can also appear in WIKI, SPAM and APNIC. Clients are sampled by WIKI, SPAM, MLAB, APNIC, GAME, SWIN and CALT. NAT'ed clients also appear in IPING and TPING. Specialised devices are likely severely under-represented in our data, although IPING and TPING may sample a few of them.

[2]Initially we probed a sample of the Internet using different commonly used TCP ports and found port 80 to be the most responsive.
[3]Modification time and IPv4 address of edits by unregistered users.
[4]Excluding traffic from our own active probing.

3.2 Data collection and processing

For IPING we only counted IPv4 addresses that returned ICMP echo replies, "destination protocol unreachable" or "destination port unreachable" messages (ignoring addresses with other ICMP errors or "TTL exceeded" messages). For TPING we only counted addresses that returned SYN/ACKs. Lack of reply indicates an address was truly unused, a host ignored the probe, or the probe or response was filtered or lost. On average our prober sent one packet every two hours to individual /24 networks, to minimise congestion and stay below typical ICMP or TCP rate limit thresholds.

For the passive datasets we extracted the IPv4 addresses from log files. We filtered out multicast and private addresses (e.g., 10.0.0.0/8), and those in unallocated or unrouted space. For WIKI, SPAM, MLAB, APNIC and GAME (server logs), the addresses are only recorded for successful TCP sessions. The SWIN and CALT logs contain spoofed IPv4 addresses that do *not* represent used addresses. Our lack of packet data meant we needed a new heuristic to remove spoofed addresses (instead of the technique in [6]).

3.3 Removal of spoofed IPs

Our heuristic is based on the assumption that many spoofed IP addresses are uniformly distributed over the routed space. We observed that the unfiltered SWIN and CALT have uniformly randomly distributed IPv4 addresses in some /8 prefixes that are completely or almost completely unused by other sources (e.g. 53.0.0.0/8 or 55.0.0.0/8).[5] While the number of observed IPs from these 'empty' /8 subnets differs for SWIN and CALT, for a given dataset and time period the number of observed IPs is roughly identical for these /8 – consistent with the assumption that spoofed addresses are uniformly distributed over the IPv4 space.

Our approach works in two stages. First, we estimate which /24 subnets should be removed entirely, and then we remove potentially spoof addresses from used /24s.

From SWIN and CALT we removed all /24 subnets that:

1. have fewer than m observed IPs, *and*
2. have no overlapping IPs that are also in the spoof-free WIKI, APNIC, MLAB and GAME datasets.

We choose m as follows. Treating spoofed IPs as uniformly sampled from a space of s IPs with probability p, the number X of spoofed IPs in the space follows a Binomial distribution. Specifically:

$$\Pr(X > k) = 1 - \sum_{i}^{k} \binom{n}{i} p^i (1-p)^{n-i} \ .$$

[5]The number of addresses from these /8 in our non-spoofed sources is negligible (no more than a few tens of addresses) and in some cases we know from the network administrators that these /8 are hardly used. However, for SWIN and CALT we see a much larger number of addresses in these /8.

Table 2: Data sources and observed unique IPv4 addresses and /24 subnets per year (SWIN and CALT after spoofed IP filtering)

			2011		2012		2013	
Dataset	**Description**	**Time collected**	**IPs [M]**	**/24 [M]**	**IPs [M]**	**/24 [M]**	**IPs [M]**	**/24 [M]**
WIKI	Wikipedia's page edit histories	Jan 2011 – Mar 2014	5.5	1.70	5.9	1.97	6.8	2.16
SPAM	Potential spam email senders	May 2012 – Mar 2014	-	-	19.2	1.56	17.5	1.73
MLAB	Clients tested by Measurement Lab	Jan 2011 – Mar 2014	30.0	2.66	27.6	2.69	21.5	2.50
APNIC	Web clients tested for IPv6	Mar 2011 – Mar 2014	22.0	2.92	88.0	3.90	108.7	4.13
GAME	Game clients logged into Valve's Steam	Jan 2011 – Mar 2014	89.7	3.10	120.1	3.62	340.0	4.33
SWIN	Swinburne access router NetFlow records	Jan 2011 – Mar 2014	150.6	3.13	142.4	3.38	112.9	3.36
CALT	Caltech access router NetFlow records	Jun 2013 – Mar 2014	-	-	-	-	356.8	3.92
IPING	TCP port 80 census of IPv4 Internet	Mar 2011 – Mar 2014	320.3	4.25	358.3	4.55	411.1	4.82
TPING	ICMP ping census of IPv4 Internet	Mar 2012 – Mar 2014	-	-	70.0	3.38	92.7	3.71

In our case of /24 subnets, $s = 256$ and we estimate p based on the number of spoofed IPs S in each 'empty' /8 prefix, so $p = S/2^{24}$. We then choose $m = k$ where $\Pr(X > k) < 10^{-8}$. Note that for SWIN, S is relatively constant across all time periods (10,000–15,000), but for CALT it increases from 15,000–20,000 until December 2013 to almost 250,000 in March 2014.

Spoofed IP addresses will also fall into /24 subnets that have actually used IP addresses. The second phase is to filter out potentially spoofed IPs in used /24 as follows. Since we assume the spoofed IPs to be uniformly random distributed, the number of spoofed IPs is S for used /8 prefixes as well. Subtracting the number of already removed IPs in spoofed /24 subnets we have S'_i spoofed IPs left in /8 prefix i. Given the observed number of IPs T_i in /8 prefix i in SWIN or CALT the expected number of not-spoofed addresses per /8 prefix (out of 2^{24} addresses) is

$$2^{24} \cdot \frac{T_i - S'_i}{2^{24} - S'_i} .$$

On average the probability that an IP in i is valid (V) is

$$\Pr(V) = \left(\frac{T_i - S'_i}{T_i}\right)\left(\frac{2^{24}}{2^{24} - S'_i}\right) \approx (T_i - S'_i)/T_i .$$

This tells us how many IPs to keep in each /8 prefix, but to use capture-recapture, we must also determine which IPs to keep. To do this we use the fact that the distribution of the final byte B of used addresses is not uniform. We estimate the probability $P(B|V)$ from the IPs observed by all sources except SWIN and CALT. Then assuming that $P(B|\text{not } V) = \frac{1}{256}$ (uniform distribution), Bayes' rule gives that an IP is not spoofed in SWIN or CALT with probability

$$\Pr(V|B) = \frac{\Pr(V)\Pr(B|V)}{\Pr(V)\Pr(B|V) + (1 - \Pr(V))/256} .$$

We then filter SWIN and CALT by independently removing addresses ending with B with probability $1 - P(V|B)$.

We cannot evaluate the true accuracy of our approach, but the following circumstantial evidence shows that it is effective. With filtering, randomly distributed IPs in the 'empty' /8 networks are removed. With filtering, the number of used /24 subnets gradually increases over time and does not show large abrupt increases and decreases anymore. Without filtering the number of /24 subnets in SWIN or CALT is much higher than in any other dataset, e.g. it is up to 30% higher than for our largest dataset (IPING) and up to 60% higher compared to APNIC, GAME. After filtering the number of used /24 subnets in SWIN and CALT is lower or similar to that in APNIC and GAME.

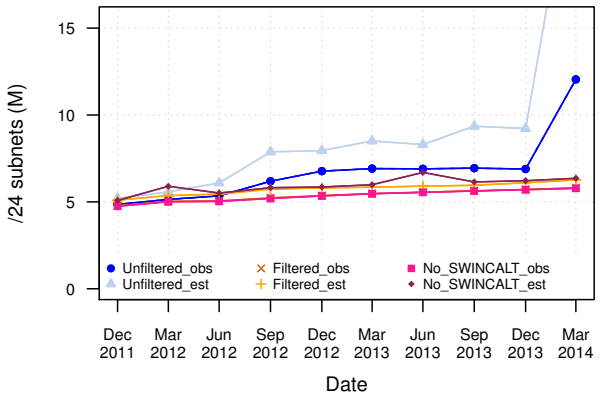

Figure 2: Observed (obs) and estimated (est) /24 subnets with and without spoof filtering compared to observed and estimated /24 subnets without SWIN and CALT

Figure 2 shows the benefit of filtering spoofed addresses. LLM estimates that include filtered SWIN and CALT are quite consistent with LLM estimates made without SWIN and CALT. LLM estimates using unfiltered SWIN and CALT are much higher (exceeding the possible maximum for March 2014). To save space, we only show this comparison for /24 subnets, as spoofed IPs have less negative impact on the observation and estimation of used IPv4 addresses (due to the uniform random nature and low – 10% or less – estimated percentage of spoofed IPs).

3.4 Time windows

To analyse the growth of use of IPv4 addresses, we split our data into overlapping 12 month windows. Windows started every three months from 1 Jan 2011 until 1 April 2013 (with the last window ending 31 March 2014). This is a suitable trade-off between temporal resolution and noisy estimates. For some datasets we cannot really make the time window smaller, e.g. we only conducted IPING/TPING censuses every six months and the GAME data was only collected every 3+ months.

Overlapping windows smooth out quick changes, but we believe fast transients in the number of used IPv4 addresses are unlikely. In the rest of the paper we associate statistics with the end of time windows. For example, for the 2011 calendar year, the observed and estimated used space is associated with 31 December, 2011.

3.5 Dynamic and static addresses

Many IPv4 addresses are (re)assigned dynamically (such as with DHCP or PPPoE). Hence, long passive measurements may observe multiple addresses for a single host, and over-count the number of simultaneously used addresses.

If each assignment uses the lowest/highest unused address of a pool, then the total number of addresses used from the pool is the maximum simultaneous pool utilisation and the LLM estimate would indeed estimate the maximum number of simultaneously used addresses. However, if addresses are drawn uniformly, as measurements suggest, then all pool addresses could be observed even if at most one address is in use at a time. Similarly, a single host moving between multiple statically assigned addresses may report multiple addresses, even if at most one is in use at a time.

However, addresses assigned to pools cannot be used elsewhere. So we argue that any over-count captures addresses (or /24 subnets) that are on "stand-by" and de facto 'in use' at the time of our measurement. (In the future under-utilised pools may be reduced in size and the freed addresses may be used for other purposes. However, this is the same as re-purposing addresses of de facto unused hosts. We cannot quantify such future optimisations.)

We also study /24 subnets, which are less affected by dynamic addressing [6]. While some address reassignments (e.g., host mobility) may cross different /24 subnets, a large fraction of them will be within the same /24 subnet(s).

4. VALIDATION

We now validate the heuristics and assumptions used in deriving our model. First, we pick a specific model-selection algorithm from among those described in Section 2.2.2, based on test data. Next, we compare estimated use of addresses and /24s against ground truth for a handful of networks, and show that CR gives better estimates than simply summing the observed addresses. Since we have no ground truth for most networks, we use cross-validation demonstrate that this also applies to the whole address space

Table 3: Cross-validation errors depending on different parameter settings

	IP addresses		/24 subnets	
Setting	**RMSE [M]**	**MAE [M]**	**RMSE [k]**	**MAE [k]**
fixed1	32.2	13.1	105.3	47.0
fixed1bic	37.2	15.1	115.4	52.0
fixed10	21.8	9.8	114.7	51.2
fixed100	16.6	8.1	123.4	54.9
fixed1000	18.1	9.1	122.6	60.8
adapt1000	19.7	9.2	105.2	47.8
adaptbic1000bic	17.8	8.7	108.4	49.0

To perform cross-validation with our $k = 9$ data sources we consider a particular source i as the "universe" of possible IPv4 addresses. We apply CR to the addresses/subnets in i that are also in the other $k - 1$ sources, to estimate the number of individuals unique to source i. Since we know the true number of individuals unique to i, we can evaluate the effectiveness of CR. We do this for each source, to obtain the mean error and mean-square error. We then assume that the CR estimator based on the full k sets is equally accurate at estimating the true number of ghosts, although we do not have confidence intervals for this.

4.1 Model selection

First, we performed cross-validation for each time window for both used IPv4 addresses and used /24 subnets for different parameter combinations. We vary the IC used (AIC, BIC), and the count pre-processing (adaptive, fixed to different values). For each setting we performed cross-validation. For each time window we computed the Root Mean Square Error (RMSE) and Mean Absolute Error (MAE) between the estimates and the ground-truth, averaged over all sources. Table 3 shows the different parameters settings we investigated and the error averaged over all sources and time windows.

From the results we see that using the actual counters (fixed1) results in the highest errors for IPs but provides the lowest error for /24 subnets. We think this is because a) there is much more noise in the observed IPs than in the observed /24 subnets and b) the number of observed /24 subnets is much lower than the number of IPs, and hence even for a small divisor of 10 we already start losing information for /24 subnets which leads to much reduced accuracy. Choosing a divisor around 100 results in the smallest error for IPs but the largest error for /24 subnets. Effectively, the choice depends on the type of data and it is unclear what choice would be the best for estimating the IP addresses and /24 subnets unseen overall.

In contrast, our adaptive approach (especially with a maximum divisor of 1000) works quite well for both IPs and /24 subnets, with errors only slightly larger than the minimum

Table 4: Pingable, observed (obs) and estimated (Poisson, right-truncated Poisson) vs. peak usage (ground truth)

Network	Ping [%]	Obs. [%]	Estimated(Error) [%]		Truth [%]
			Poisson	TruncPoisson	
A	0.4	5.7	23.4(-2.5)	26.7(+0.8)	25.9
B	6.7	8.5	13.9(+2.5)	12.3(+0.9)	11.4
C	12.0	13.7	32.5(–)	36.1(–)	30–35
D	24.0	31.8	41.3(-6.3)	51.6(+4.0)	47.6
E	9.4	17.3	52.1(-6.2)	60.5(+2.2)	58.3

errors. With the adaptive approach, using the BIC instead of the AIC lowers the error for IPs but increases the error for /24 subnets, but the increase for /24 subnets is small and even for /24 subnets the estimates obtained with the BIC are smoother.[6] Hence, in the rest of the paper the estimates presented are based on our adaptive approach with a maximum divisor of 1000 and we use the BIC.

4.2 Comparison with ground truth

We compared our estimations with the ground truth for several networks where we obtained information on how many IPv4 addresses were actively used at peak times (during March to June 2013). Note that the ground truth here is rough estimates of the number of actively used IPv4 addresses. Based on the time window ending 30 September 2013, for each network Table 4 shows the number of addresses that responded to ping, the number of addresses observed, the number of addresses estimated (for both Poisson and right-truncated Poisson), and the actual number of used addresses as percentages of the sizes of the networks. For privacy reasons we cannot reveal the identity of the networks or their sizes (the largest network covered is a two /16 and the smallest network is roughly one /20 combined from multiple allocations).[7]

The results show that the percentage of pingable and observed addresses is much smaller than the actual peak usage for networks A, C, D and E, whereas for the more "open" network B the percentage of observed addresses is relatively close to the actual peak usage. However, the CR estimates are always much closer to the truth. Using right-truncated Poisson distributions gives better estimates than using Poisson distributions. The right-truncated Poisson estimates are always too high, but since we use 12-month windows, the estimated number of used IPv4 addresses is likely above short-term peak numbers due to dynamic addresses.

[6]The BIC selects fewer parameters representing interactions of many sources, which have a much lower number of samples than interactions between fewer sources and hence are noisier.

[7]One network is that of Swinburne University. For this estimate we omitted SWIN and replaced PING by a dataset collected by pinging Swinburne's network from an external vantage point.

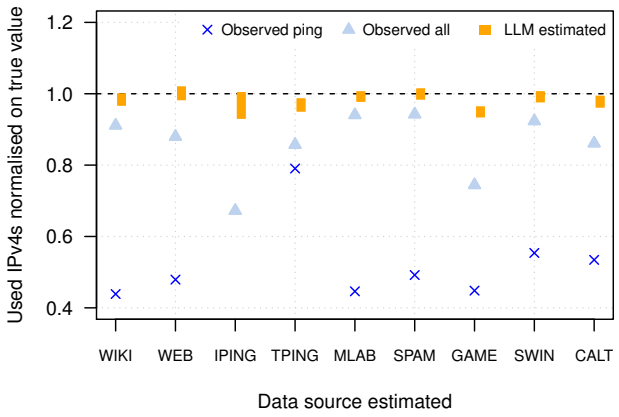

Figure 3: IP addresses observed with ping, observed by any source, and estimated ranges for LLM normalised on the true number of unseen IPs for each data source

4.3 Cross-validation results illustrated

Figure 3 illustrates the results of the cross-validation for addresses and subnets for time window 7 (results for other time windows are largely consistent). The figure shows the number of IPs in each source also observed by IPING (Obs ping), the total number of addresses of a source also observed by any other sources (Obs. all), and the ranges of the CR estimates (confidence interval based on profile likelihood with $\alpha = 10^{-7}$ to get wide intervals). Since the sources are of different sizes we normalised the number of addresses on the total number of addresses observed by each source (the ground truth). A CR estimate is good, if the normalised range includes 1 and the range is not too large.

Figure 3 shows that for IPv4 addresses all sources other than IPING and GAME have relatively high overlap, but between 10% and 15% of addresses appear only in one source. Only 50–60% of addresses of each source (except TPING) is in IPING, showing that ICMP pinging undercounts significantly. The CR estimates for WIKI, SPAM, MLAB, APNIC, SWIN, TPING and CALT are quite good. The estimated range for IPING almost covers the true value but the uncertainty is higher, and the estimate for GAME is slightly too low. Nevertheless, overall the CR estimates for our LLM model are a substantial improvement over just using the number of observed IPs or using the simpler L-P model (see [10]).

For brevity we do not show the figure for /24 subnets. For /24 subnets there is a very high overlap between all data sources. However, for most sources only 90% or less of the /24 subnets appear in IPING, so just using ICMP pinging significantly undercounts even the used /24 subnets. While the difference between CR estimates and observed addresses is much smaller for /24 subnets (in most cases the difference is only 1–2%, except for IPING), our CR estimates are still an improvement.

Table 5: Observed and estimated used IPv4 addresses and /24 subnets at the end of March 2014 based on different stratifications

	Estimated total [M]							Observed [M]	Est. unseen [M]	Routed [M]
Stratification	**None**	**RIR**	**Country**	**Age**	**Prefix size**	**Industry**	**Stat/Dyn**			
IP addresses	1132.0	1161.6	1141.7	1108.5	1062.9	1071.4	1105.7	740.9	320–420	2706.6
/24 subnets	6.26	6.34	6.32	6.30	6.30	6.28	6.22	5.81	0.4–0.5	10.57

5. USED SPACE ANALYSIS

Now, we present the results for the estimated used IPv4 addresses and /24 subnets. We present both total estimates as well as estimates for different RIRs, countries, allocation ages, and allocation prefix sizes. We also investigate whether our estimates are sensible given the growth of Internet users.

5.1 Used IPv4 space totals

Table 5 shows the estimated used IPv4 addresses and /24 subnets depending on different stratifications, as well as the observed and the estimated unseen addresses and /24 subnets at the end of March 2014. The totals for different stratifications are always the sum of the estimates over all strata. Our estimates are fairly consistent across stratifications: roughly 1050–1150 million used IPv4 addresses and 6.2–6.3 million used /24 subnets. Based on Routeviews [14] this means we observed roughly 27% of the routed IPv4 addresses and 55% of the routed /24 subnets, and we estimate that roughly 40% of the routed IPv4 addresses and 60% of the routed /24 subnets were used.

For all stratifications our estimates are always plausible (below the number of routed addresses). The quotient of estimated used addresses divided by the addresses detected only with ICMP echo pinging is 2.5–2.6, which is larger than the correction factor of 1.86 used in [3].

5.2 Used IPv4 space over time

Figure 4 shows the number of estimated used /24 subnets against the number of observed and routed /24 subnets both as absolute numbers and normalised. The dashed line is the actual estimates and the solid line is the estimates smoothed. The total number of observed /24 increased from 4.8 million to 5.8 million, but we estimate that the number of /24 subnets actually increased from 5.1 million to 6.3 million (an increase of 0.5 million subnets per year). Whilst the routed space only increased by 8% in two years, the number of observed and estimated used /24 subnets increased by 22% over the same time.

Figure 5 shows the number of estimated used IPv4 addresses against the number of observed and routed IPv4 addresses both as absolute numbers and normalised. The number of observed IPv4 addresses increased from 450 million to 740 million, but we estimate that the number of addresses actually increased from 730 million to 1.1 billion (an average increase of about 160 million IPv4 addresses per year). As for /24 subnets, the observed and estimated number of IPv4 addresses increased faster than the routed addresses. The difference between estimated and observed relative growth may be in part because of earlier undercounting due to fewer sources and a hole in the GAME data.

5.3 Used IPv4 space by RIR

Figure 6 shows the estimated number of IPv4 addresses over time depending on the RIR responsible for their allocation both as absolute numbers and normalised. For brevity we omitted the broadly similar statistics for /24 subnets. APNIC has the largest number of used addresses followed by RIPE and ARIN. Looking at relative growth, AfriNIC is growing at the fastest rate, followed by LACNIC. Of the three RIRs with the most allocated space, relatively APNIC and ARIN are growing faster than RIPE.

5.4 Used IPv4 space by prefix size

Figure 7 shows the average yearly growth rate for addresses estimated for different prefix sizes. For brevity we do not show the estimates for /24 networks here, as the trends are broadly similar. Absolute growth is strongest in the large prefixes /10 to /16 (/8 and /9 have not grown much). However, if we look at relative growth, growth has been more equally across many prefixes. Exceptions are the old /8 allocations which haven't grown and /9, /21 and /22 allocations which show the strongest growth (/9 is driven up by a few ISPs since there are less than ten /9 allocations overall, and /22 is the largest allocation handed out by APNIC since 15 April, 2011, and by RIPE since 14 Sep, 2012).

5.5 Used IPv4 space by allocation age

Figure 8 shows the average yearly growth rate of IPv4 addresses for different allocation ages (we omitted the year 2013 because the estimates are unreliable due to few data points). For brevity we omitted the results for /24 subnets as the trend is very similar. In absolute numbers the more recent allocations made since 2005 are growing the most, with a clear positive correlation between recentness and growth. In relative terms growth is strongest for allocations made in the last three years, but we can also see 20% or higher growth in some old allocations.

5.6 Used IPv4 space by country

Figure 9 shows the absolute and relative growth for IPv4 addresses for the countries with the largest number of observed used IPv4 addresses (at least 1.5 million addresses). Again, we don't show results for /24 subnets as the trends

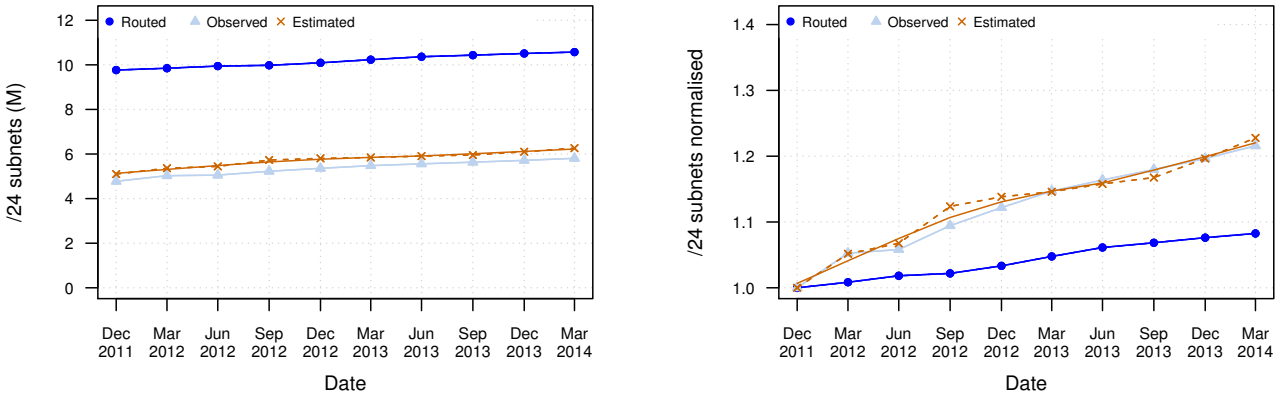

Figure 4: Absolute and relative growth of estimated, observed and routed /24 subnets

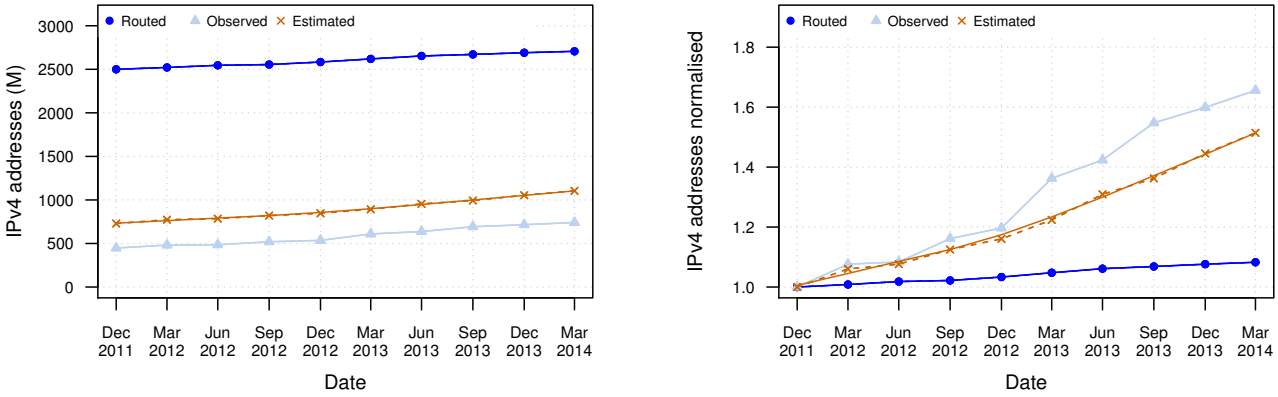

Figure 5: Absolute and relative growth of estimated, observed and routed IPv4 addresses

are broadly similar. Absolute growth is strongest in the two nations with the largest allocations (USA, China) followed by Brazil and South Korea. Relative growth is between 10% and 40% for many countries, but Romania and several Asian and South American countries (Indonesia, Brazil, Columbia, India, Argentina, Taiwan, Thailand, and Vietnam) have grown much faster.

5.7 Comparison with Internet user growth

According to data from the ITU [23] the number of Internet users has grown from 16 million in December 1995 to 2.75 billion (roughly 39% of the world's population) in December 2013 (see Figure 10). The growth rate of Internet users looks exponential at the beginning, however since 2006/2007 the growth appears roughly *linear*.

We think the growth of the number of used IPv4 addresses is primarily driven by an Internet population increase, irrespective of the number of devices per user. All home devices are behind NATs and mobile devices are also largely behind NATs. Similarly, if we look at increasingly complex commercial networks, these are also mainly behind NATs (or not even connected to the Internet). Then, given the linear growth of estimated Internet users, it is logical that the used IPv4 addresses also grow linearly (shown in Section 5.2).

Between the start of 2007 and mid 2012 the number of Internet users grew by roughly 250 million per year (c.f. Figure 10). For private use typically a household shares one public IP address. In industrial nations the household size is 2–3, but in developing countries it can be higher, for example it is over 5 in India [24]. We assume the average household size of new Internet users is between 2 and 5. In addition a fraction of people will have a public IPv4 address at work. We assume an average employment ratio of 65% [25]. As upper limit we assume one IPv4 address per two employees, as in reality many employees (especially in developing countries) have no Internet at work, work at home, or share computers with other workers. As a lower bound we assume on average there is one public address per ten workers.

With these assumptions, as a lower bound we would expect the number of used IPv4 addresses to grow between 65 million and 205 million per year (plus additional addresses for service and infrastructure growth). Our current growth estimate is 160 million IPv4 addresses per year, which fits well within these bounds.

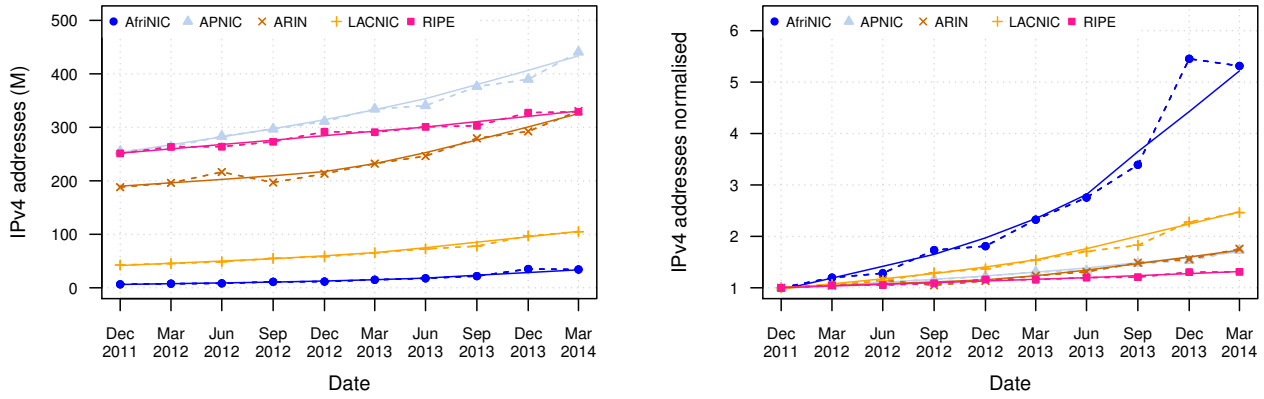

Figure 6: Absolute and relative growth of estimated IPv4 addresses for different RIRs

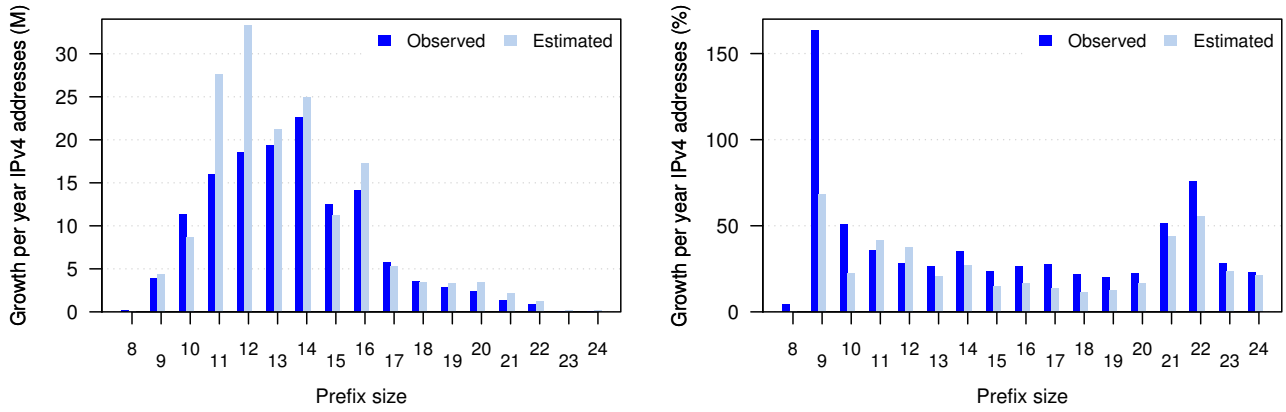

Figure 7: Average absolute and relative yearly growth of observed and estimated IPv4 addresses for different allocation prefixes

6. UNUSED SPACE PREDICTION

Our CR technique tells us how many unobserved IPv4 addresses to expect, but says nothing about the distribution of free blocks/prefixes. This is a challenging issue – recipients of newly assigned IPv4 address blocks typically prefer usably-sized contiguous allocations, and forwarding information base (FIB) tables in routers are more efficiently packed if address blocks are allocated hierarchically.

Some information is given by the CR estimate of the used but unobserved /24 networks (in Figure 4). However, this again does not tell us whether these small blocks are isolated or parts of unused larger blocks. In this section, we try to understand how the unseen addresses are distributed among seemingly empty subnets, by observing what happens when data sources are combined sequentially; each new source brings addresses that were unseen by the previous sources, and we can model how those addresses fill the previously empty space.

6.1 Model

Let x_i be the number of observed free /i blocks. Let $Z_{00\ldots0}$ be the total number of new addresses to allocate (given by CR). Let N_i be the number of new addresses assumed to be allocated to vacant /i blocks, assuming sequential allocation. Specifically, if two addresses are added to the same vacant /i, then only the first of these contributes to N_i , since the block is no longer vacant when the second one is added.

Similarly, let x_i^S be the number of free /i blocks in a set S of addresses, $Z_{00\ldots0}^{S,\Delta}$ be the number of new addresses when a new set Δ is merged with S, and $n_i^{S,\Delta}$ be the number of addresses added to vacant /i blocks in the process. Without the subscript i, the variables x and n denote vectors.

Note that adding an address to a vacant /i will reduce the number of vacant /i blocks by 1, but increase by one the number of /j blocks for each $j > i$, regardless of where within the /i the address is added. That is,

$$x^{S\cup\Delta} - x^S = An^{S,\Delta}, \tag{2}$$

where

$$A = \begin{pmatrix} -1 & 1 & 1 & \cdots & 1 \\ 0 & -1 & 1 & \cdots & 1 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & -1 \end{pmatrix}. \tag{3}$$

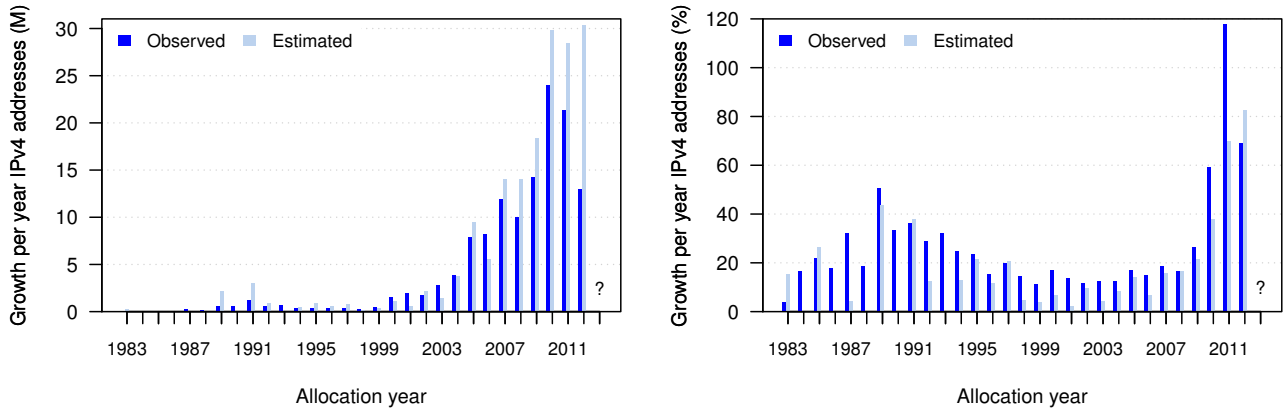

Figure 8: Average absolute and relative yearly growth of observed and estimated IPv4 addresses for different allocation ages

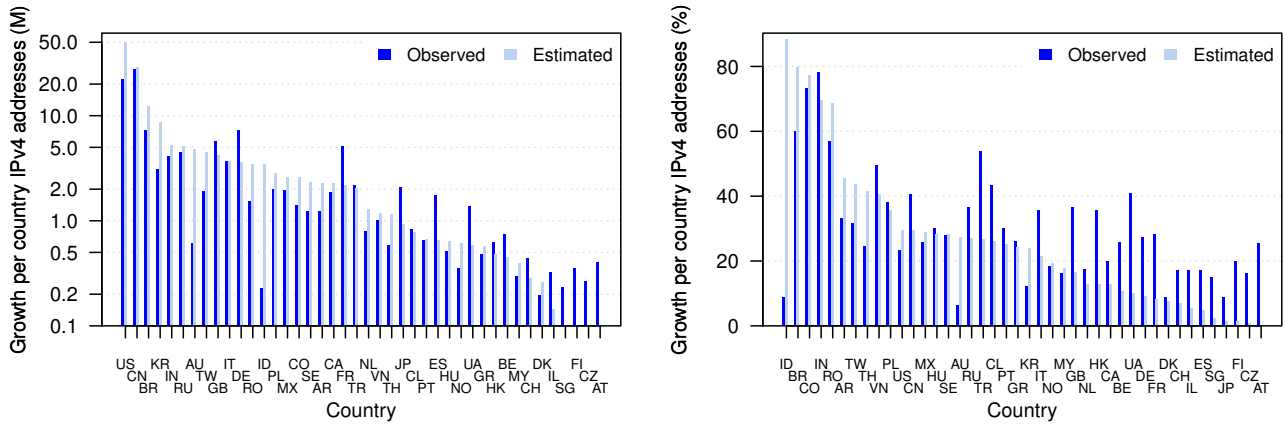

Figure 9: Average absolute and relative yearly growth of observed and estimated IPv4 addresses for different countries sorted by estimated growth (only the largest 42 countries). Note, the absolute numbers are plotted in log scale.

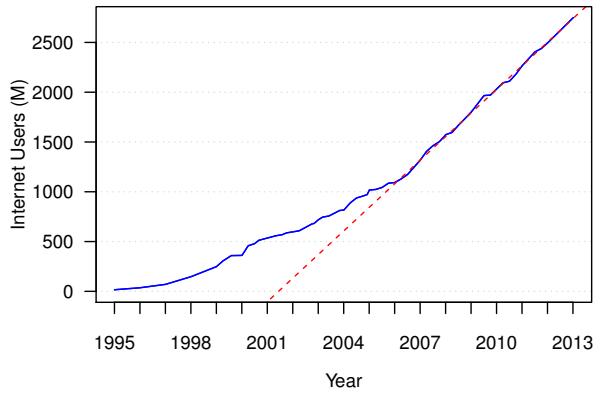

Figure 10: Number of Internet users based on data from ITU

A natural approach is to estimate the previous fraction of addresses revealed by each new source that have been allocated to free blocks of a given size, and assume the new $Z_{0,0,\ldots,0}$ addresses will be distributed in the same way. This is not sufficient, because the allocation process changes the number of available blocks. Instead, our model uses the observation that the probability that a new address is allocated to a free /i block is proportional to x_i, the number of such free blocks. In particular, it assumes that there are $f_1, \ldots, f_{32}$ such that the ratio

$$\frac{N_1}{x_1} : \frac{N_2}{x_2 + N_1} : \cdots : \frac{N_{32}}{x_{32} + \sum_{j=1}^{31} N_j} = f_1 : f_2 : \cdots : f_{32} \quad (4)$$

remains approximately constant as more batches of addresses are discovered.

The model includes subnets larger than /8, even though blocks larger than /8 have not been allocated. Similarly, we consider all vacant subnets down to vacant /32s, even though subnets smaller than a /24 are not routed on the public Internet. However, before computing the remaining unused prefixes we split a few /7 into /8, and we also exclude all private, multicast, experimental and reserved prefixes, such as 224.0.0.0/3 or 10.0.0.0/8. Note that we do not exclude non-publicly routed prefixes.

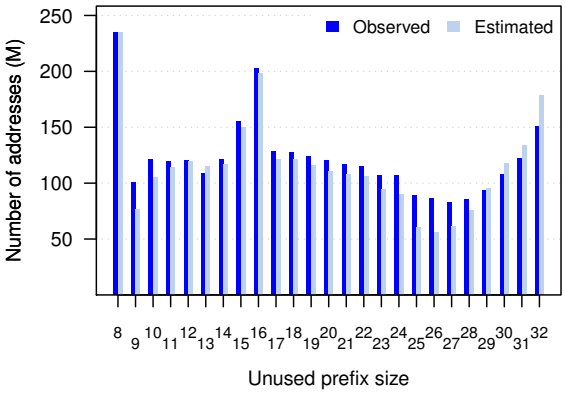

Figure 11: Number of addresses in observed and estimated unused prefixes for different prefix sizes

To determine how the unobserved addresses predicted by CR will affect the distribution of free blocks, it remains to determine f_i. To do this, we observe the change in x when a new data set Δ is added to an existing list S of used IPv4 addresses, and from that calculate n. Since A in (3) is invertible, (2) gives

$$n^{S,\Delta} = A^{-1}(x^{S\cup\Delta} - x^S).$$

The f_i are then found by (4), normalized so that $f_{32} = 1$. Since few large subnets become newly used for each data set, estimates of f_i for $i \lesssim 12$ are noisy. This is unfortunate, since these are the blocks of greatest interest. To reduce this noise, estimates were averaged over four cases: $\Delta =$ IPING, GAME, APNIC, WIKI; in each case, S is the union of all remaining datasets, except SWIN and CALT.

One concern with this model is that, as the address space fills up, the values of f_i may vary. To check this, we performed tests where datasets were added to S one at a time, in both increasing and decreasing order of the dataset size. The values were reasonably consistent in each case.

6.2 Results

Figure 11 shows the number of addresses in unused prefixes in the last time period, based on all sources except SWIN and CALT. Results are for both direct observation and CR. The majority of empty prefixes are longer than /20 (fewer than 2^{12} addresses), but the unused space is roughly uniformly distributed among prefixes of lengths /9 to /24 (except /15 and /16). The reason for this is unclear.

If the used but unobserved /8 to /24 subnets estimated by the model of Section 6.1 were divided into /24s, there would be 0.47 million /24s. This is consistent with the estimate of 0.4–0.5 million by the independent LLM model of Section 5.1, providing evidence for the validity of both models.

6.2.1 *Router FIB limitations*

One of the reasons that the distribution of prefix sizes is important is that each routed prefix requires an entry in a router's Forwarding Information Base (FIB), and there is a *prima facie* risk that allocating all unused prefixes could overflow the FIBs. Above we estimated that including the unrouted space there are 0.75 million prefixes that are /24 or larger. Currently, there are more than 0.5 million routed prefixes already (but a substantial fraction is unused). In 2007 Juniper [26] stated that its M120 and MX960 had FIB capacities of about 2 million IPv4 routes, and that IPv4 FIBs with approximately 10 million entries are feasible within a few years if demand exists. In addition, FIB compression techniques can reduce size of FIBs [26]. This suggests that it will be feasible to use and route all less than 1.25 million available prefixes. Even if unused prefixes are subdivided further, it appears feasible to route them all, although it may require upgrades of existing routers.

6.2.2 *Estimated value of potential IPv4 market*

The value of the unused IPv4 address market depends on the number of unused addresses and the average price per address. The price per address depends on the supply and demand of IPv4 addresses and on the size of the address space sold. The demand for IPv4 addresses depends on the availability of IPv6 as well as other factors, such as the region (e.g. higher prices in regions where RIRs have run out of IPv4 address space). The price per IP is generally higher for smaller blocks due to transactions overheads (see [27]) and possible discounting for larger blocks, since naturally a smaller pool of demand exists for larger blocks.

RIRs report the addresses that have been transferred through their paid transfer systems, but the RIR records lack information about prices [28]. In 2011 Microsoft paid $7.5 million for approximately 666,000 IPv4 addresses (something between a /12 and /13 prefix) equivalent to US$11.25 per address. Other transactions from sellers in bankruptcy have brought prices that are broadly similar [28]. In 2013, several /15 and /16 were sold for prices between US$8.50 and US$10.50 per IP, and for several /20 the price was between US$14 and US$17 per IP [27].

At a price of US$10 per IP address, the 4.3 million routed unused /24 subnets have a value of around US$11 billion. However, it is likely that only a small fraction of those will be sold, even if the price rises substantially, and so the eventual market value is likely to be much smaller.

6.2.3 *Estimated years of supply*

Even if all 4.3 million unused but routed /24 subnets could be prised away from their current owners, they would be exhausted in 2022 under the current growth trend of 0.5 million /24 subnets per year. Unused routed IPv4 addresses would be exhausted in 2024 given the current growth of 160 million addresses per year.[8]

If the overall utilization of routed /24 subnets remains below, say, 75%, the current growth rate suggests three years

[8] Suggesting that we are at least 2/3 of the way from the standardization of IPv6 in 1996 to its required adoption.

of remaining supply. Of course some parts of the world may be exhausted before this time. An open question is the large amount of unrouted IPv4 space, much of which has not been routed for years. Unused parts of the unrouted space might provide a short-lived increase in IPv4 supply.

Over the next 2 to 3 years, we expect IPv4 exhaustion to be increasingly felt, resulting in a brief growth in the IPv4 address market. Most organizations holding unused addresses do so for operational reasons – to allow expansion or flexibility, or in one case as a /8 darknet – but some may be holding them to sell if the market price rises sufficiently. The numbers in this paper may guide how long they can expect to be held for, assuming that the market will collapse once IPv6 is widely adopted. However this is complicated by the fact that the very act of selling a large block of IPv4 addresses will delay the implementation of IPv6, and hence prolong the IPv4 market.

7. RELATED WORK

The related measure of *routed* address space has been estimated based on prefixes advertised by BGP [29, 30]. However, estimation of the number of *actively used* addresses began with Pryadkin *et al*. [2], who used ICMP echo and TCP SYN probing to probe the allocated Internet. They discovered 62 million used IPv4 addresses in 2003 and 2004. Pryadkin *et al*. also showed that only a small number of allocated prefixes appeared to be heavily used, while a large part of the IPv4 space appeared unused or underutilized.

Heidemann *et al*. [3] infrequently probed all allocated IPv4 addresses (census) and frequently probed selected address samples (survey) with ICMP echo pinging to study usage, availability and up-time of addresses. Their last census in 2007 accounted for 112 million used addresses. Heidemann *et al*. compared ICMP probing with TCP port 80 probing and passive measurements based on small samples. They proposed a correction factor of 1.86, thus estimating the total number of used IPv4 addresses 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 typical address block sizes and their characteristics. They did not estimate the used IPv4 address space, but observed: "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 commodity routers to perform a port scan of the IPv4 Internet [5]. They detected 420 million addresses that responded to ICMP echo, which is broadly consistent with our two ping censuses that detected 360 million addresses between March and September 2012. They also detected another 36 million addresses that only responded to TCP SYN probes on several hundred ports. In our censuses 15–20 million IPs reacted to port 80 TCP SYNs but not to ICMP – probing hundreds of ports merely doubles this number.

In 2013 we initially proposed using a log-linear CR model to estimate the true population of used IPv4 addresses from multiple sources of IPv4 addresses [10]. Our preliminary workshop paper found that our log-linear CR estimate is significantly higher (one billion used IPv4 addresses in mid 2013) than the aggregate number of observed IPv4 addresses from multiple measurement sources.

Dainotti *et al.* [6] used darknet data from July to September 2012 to estimate the number of used /24 networks. They developed techniques to filter out spoofed IPv4 addresses from the darknet data. With the combined filtered darknet and ping census data [3], the number of observed /24 subnets was 4.8 million (47% of the routed space). This is broadly consistent with the 5.2 million /24 subnets we observed in the year to September 2012 (c.f. Figure 4). The difference is likely due to the larger number of sources and the larger time window we use.

8. CONCLUSIONS AND FUTURE WORK

Our key contribution is describing and demonstrating a new statistical *capture-recapture* technique for improved estimation of the true population of both observed and unobserved (yet still active) IPv4 addresses from diverse sources of active and passive measurement data. This technique refines our community's understanding of IPv4 address space exhaustion and consequent incentives for IPv6 adoption.

Data from nine sources over the past three years suggests 5.8 million used /24 subnets and 740 million used IPv4 addresses. Yet our CR technique indicates a significantly higher 1.1 billion IPv4 addresses in use across 6.3 million /24 subnets, with usage growing at around 0.5 million /24 subnets and 160 million IPv4 addresses per year. Europe and Asia have the highest utilisation, while Africa and South America show the fastest growth. At this rate all remaining /24 subnets will be used by 2022. If only 75% of routed /24 subnets could be used, supply will be exhausted in 2017.

Our ability to collect more IP data for validating or improving our estimates, or potentially detecting more hosts (e.g. private servers), is limited by common privacy restrictions. We plan to explore an enhanced method [31] for securely applying CR to multi-source measurement data without revealing which IPv4 addresses each source contain.

Acknowledgements

This research was supported by Australian Research Council grants LP110100240 (with APNIC Pty Ltd) and FT0991594. We thank Geoff Huston, George Michaelson, Valve Corporation, A. Reynolds, Swinburne ITS, Caltech IMSS, D. Buttigieg, C. Tassios, R. Bevier, B. Mattern, USC/ISI and J. Heidemann for providing data.

9. 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)*, pages 169–182, 2008.

[4] X. Cai, J. Heidemann. Understanding Block-level Address Usage in the Visible Internet. In *ACM SIGCOMM Conference*, pages 99–110, 2010.

[5] Internet Census 2012 – Port scanning /0 using insecure embedded devices, 2012. `http://internetcensus2012.bitbucket.org`.

[6] A. Dainotti, K. Benson, A. King, kc claffy, M. Kallitsis, E. Glatz, X. Dimitropoulos. Estimating Internet Address Space Usage Through Passive Measurements. *ACM Computer Communication Review (CCR)*, 44(1):42–49, Jan. 2014.

[7] C. G. J. Petersen. The Yearly Immigration of Young Plaice into the Limfjord from the German Sea. *Rept. Danish Biol. Sta.*, 6:1–77, 1895.

[8] F. C. Lincoln. Calculating Waterfowl Abundance on the Basis of Banding Returns. *U.S. Dept . Agric. Circ.*, 118:1–4, 1930.

[9] A. Chao. An Overview of Closed Capture-Recapture Models. *Journal of Agricultural, Biological, and Environmental Statistics*, 6(2):158–175, 2001.

[10] S. Zander, L. L. H. Andrew, G. Armitage, and G. Huston. Estimating IPv4 Address Space Usage with Capture-Recapture. In *7th IEEE Workshop on Network Measurements (WNM)*, Oct. 2013.

[11] S. Zander, L. L. H. Andrew, G. Armitage, G. Huston, and G. Michaelson. Mitigating Sampling Error when Measuring Internet Client IPv6 Capabilities. In *ACM Internet Measurement Conference (IMC)*, Nov. 2012.

[12] DNS-based Blacklist of NiX Spam. `http://www.dnsbl.manitu.net/`.

[13] Measurement Lab. `http://www.measurementlab.net/`.

[14] University of Oregon Route Views Project. `http://www.routeviews.org/`.

[15] E. B. Hook, R. R. Regal. Capture-Recapture Methods in Epidemiology: Methods and Limitations. *Epidemiologic Reviews*, 17(2):243–264, 1995.

[16] M. Roughan, J. Tuke, O. Maennel. Bigfoot, Sasquatch, the Yeti and other missing links: what we don't know about the AS graph. In *8th ACM Internet Measurement Conference (IMC)*, pages 325–330, Oct. 2008.

[17] S. E. Fienberg. The Multiple Recapture Census for Closed Populations and Incomplete 2k Contingency Tables. *Biometrika*, 59(3):591–603, Dec. 1972.

[18] 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*, 20:3123–3157, Oct. 2001.

[19] S. Baillargeon, L.-P. Rivest. Rcapture: Loglinear Models for Capture-Recapture in R. *Journal of Statistical Software*, 19(5):1–31, Apr. 2007.

[20] E. Cooch, G. C. White. *Program MARK: A Gentle Introduction*. Cornell University, 2009.

[21] K. P. Burnham, D. R. Anderson. Multimodel Inference - Understanding AIC and BIC in Model Selection. *Sociological Methods & Research*, 33:261–304, 2004.

[22] Internet Addresses Census dataset, PREDICT ID: USC-LANDER/internet_address_census_it40c-20110406. Provided by the USC/LANDER project. `http://www.isi.edu/ant/lander`.

[23] ITU key 2006-2013 ICT data for the world, 2013. `http://www.itu.int/en/ITU-D/Statistics/Documents/statistics/2013/ITU_Key_2005-2013_ICT_data.xls`.

[24] Wikipedia. List of countries by number of households, Oct. 2013. `http://en.wikipedia.org/w/index.php?title=List_of_countries_by_number_of_households&oldid=576467223`.

[25] Wikipedia. Employment-to-population ratio, Mar. 2014. `http://en.wikipedia.org/w/index.php?title=Employment-to-population_ratio&oldid=598945003`.

[26] J. Scudder. Router Scaling Trends. Presentation at RIPE-54, May 2007. `http://meetings.ripe.net/ripe-54/presentations/Router_Scaling_Trends.pdf`.

[27] S. Brown. IPv4 Trading in Review, Jan. 2014. `http://ipv4marketgroup.com/blog/`.

[28] B. Edelman and M. Schwarz. Pricing and Efficiency in the Market for IP Addresses. Working Paper Number: 12-020, Nov. 2011. `http://hbswk.hbs.edu/item/6849.html`.

[29] 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)*, 35(1):71–80, 2005.

[30] 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)*, pages 501–508, Jul. 2007.

[31] S. Zander, L. L. H. Andrew, and G. Armitage. Estimating the used IPv4 address space with secure multi-party capture-recapture. In *INFOCOM (poster)*, Turin, Italy, 15-18 Apr 2013.