Estimating the Used IPv4 Address Space with Secure Multi-party Capture-recapture

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

I. Introduction

Many people have data sources of used IPv4 addresses, e.g. server logs or network measurements. However, the challenge when estimating IPv4 address usage is to combine the data sources of multiple collaborators in a secure and efficient way. The number of observed addresses in one source is often not sensitive information, but most people do not want to share datasets of unanonymised IPv4 addresses. We propose using a secure and reasonably efficient protocol that combines the datasets while keeping the addresses of collaborators private. We are also looking for more collaborators willing to share their data under our scheme.

As of December 2012 over 90% of the available IPv4 address space has been allocated and the Regional Internet Registrars (RIRs) will run out of addresses in 2013–2014 [1]. While most of the address space has been *allocated*, it is unclear how many allocated addresses are actually *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 usage.

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] and the more recent related Heidemann *et al*. [3] and Cai *et al*. [4]. 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 to obtain a better estimate of the used IPv4 space by 1) combining several different data sources of observed used IPv4 addresses and 2) using the capture-recapture (CR) method [5]–[7] to estimate the total population of used addresses, including the *unobserved* used addresses. A diverse set of data sources is needed to get good "coverage" of the IPv4 space and produce a good CR estimate. For CR we need to know the number of observed used IPv4 addresses for each source and for all combinations of set intersections of all sources.

Section II describes how the used IPv4 address space can be estimated using CR. Section III describes how multiple parties can compute the set intersection cardinalities (needed for CR) without revealing any observed IPv4 addresses to each other.

II. Capture-Recapture Method

Approaches for estimating population sizes based on limited samples, either use problem-specific techniques that we cannot apply or use a technique called *capture-recapture* (CR).

To illustrate CR, consider the simple two-sample Lincoln-Petersen (L-P) method [5]. First, some individuals are sampled from the population, tagged, and then released into the population again (capture). Later another sample of the population is taken, and the marked and unmarked individuals are counted (recapture). Let M be the number of individuals captured and marked in the first sample, R be the recaptured marked individuals, and C be the total number of individuals sampled during recapture. Then the estimated population N is [5]:

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

The method can be applied to any situation where there are incomplete lists of individuals (*sources*). Then M, C and R are the number of individuals sampled by source one, source two and both sources respectively. In our context, the sources are different measurements of used IPv4 addresses. The two-sample method relies on two important assumptions:

1) For any source each individual has the same chance of being sampled by the source (*homogenous population*). This implies that any individual must have a positive sample probability in any source; absence is due to small chance, not due to impossibility.
2) 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*).

In our case the assumptions are likely violated, because of heterogeneity between used IPv4 addresses. Heterogeneity means different types of addresses have different capture probabilities, e.g. for the two groups of "home users" and "other users", the capture probabilities are likely different across sources. Even if sources are independent, they may appear dependent if there is heterogeneity (apparent source dependence). Also, if some addresses are systematically missed by any sources, they cannot be included in the target population [7].

But there are more sophisticated CR models that *can* cope with heterogeneity and/or source dependence. For example log-linear models or the sample coverage technique have been used successfully in epidemiology [6], [7], where researchers face similar challenges of very different types of sources with

Table I
EXAMPLE THREE-SOURCE CAPTURE HISTORY TABLE

Source 1	Source 2	Source 3	Frequencies
0	**0**	**0**	Z_{000}
0	**0**	**1**	Z_{001}
0	**1**	**0**	Z_{010}
0	**1**	**1**	Z_{011}
...	...	...	...
1	**1**	**1**	Z_{111}

various heterogeneities and dependencies. These models require more than two sources and knowledge of the aggregated capture histories of individuals. Let t be the number of sources and $Z_{s_1 s_2 \ldots s_t}$ be the number of IPv4 addresses with capture history $s_1 s_2 \ldots s_t$, where $s_j = 0$ means an address was not sampled by source j and $s_j = 1$ means an address was sampled by source j. For $t = 3$ there are seven capture frequencies $Z_{001}, Z_{010}, \ldots, Z_{111}$, and e.g. Z_{001} is the number of addresses in source 3 but not in sources 1 or 2 (see Table I).

If M is the total number of observed used IPv4 addresses, then the estimated population size is $\hat{N} = M + \hat{Z}_{000}$. The variable we want to estimate, the unobserved addresses, is the capture frequency Z_{000}. All other capture frequencies need to be computed based on the number of addresses n_i sampled by source i and the number of addresses in all combinations of intersections of the sources. For example, in Table I the capture frequency $Z_{001} = n_3 - Z_{011} - Z_{101} - Z_{111}$.

III. Secure Multi-party Set Intersection Cardinality

While people do not share unanonymised IPv4 addresses for privacy reasons, we assume in most cases they would share the number of addresses in their dataset n_i. Then a secure set intersection cardinality protocol can be used to compute all capture frequencies without revealing the actual IPv4 addresses.

We assume that all parties participating in the set intersection are potential honest-but-curious adversaries; they run the protocol correctly, but try to learn as much information as possible. We assume that man-in-the-middle attacks by 3rd parties are not an issue (properly configured SSL/TLS guards against these). Our protocol must work with two or more parties (but we expect no more than 10 parties), and it must not rely on trusted 3rd parties. The protocol must compute the cardinality of the intersecting sets, but not reveal any IPv4 addresses. The protocol should be reasonably efficient and either simple or a well proven technique to facilitate acceptance.

There are a number of secure set intersection algorithms that broadly fall into two categories: encryption-based and secret sharing [8]. We propose using a simple encryption-based protocol that is computationally-secure and works for two or more parties [9]. Let E be a secure, collision resistant and commutative encryption function, such as Pohlig-Hellman (PH) or commutative RSA. Each party has their own secret encryption key k_i (a decryption key is not needed). Then for some plaintext x we have $E_{k_1}(E_{k_2}(x)) = E_{k_2}(E_{k_1}(x))$. In the two-party case, each party permutes and encrypts their set of IPv4 addresses and sends the encrypted set to the other party. Then each party permutes and encrypts the received (encrypted) set and sends it back to the other party. Since the encryption is commutative and collision resistant, the cardinality of the set intersection is the number of encrypted values equal in both double-encrypted sets. Assuming the encryption is secure neither party learns any of the IPv4 addresses. The scheme can be extended to multiple parties by forming a ring topology.

Using commutative encryption is more efficient than other solutions [10]. Although with secure key and modulus sizes for PH/RSA the overhead is substantial, the scheme is practical, given that frequent computations are not required. We propose to improve the scheme's efficiency by down-sampling the original data sources with deterministic hash-based sampling. This would significantly reduce the data to be encrypted and exchanged at the cost of a manageable sample error.

One party could generate sets with mostly invalid (e.g. unrouted) IPv4 addresses to probe whether some address is in another party's set. This can be prevented by agreeing beforehand on a set of valid (e.g. routed) addresses ("reference set"), and requiring each party to provide a set of only valid addresses larger than a preset minimum. To test for validity, a large subset of the reference set is permuted and encrypted by all parties beforehand (as described above). If a party's dataset contains a too small proportion of addresses of the reference set it is deemed to be a probe. Not returning a fully-encrypted dataset to a prober prevents the attack.

Our approach computes the multi-source capture history for CR while ensuring the privacy of the observed IPv4 addresses.

Acknowledgements

This research was supported under Australian Research Council's Linkage Projects funding scheme (project LP110100240) in conjunction with APNIC Pty Ltd and by Australian Research Council grant FT0991594.

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

[6] E. B. Hook, R. R. Regal, "Capture-Recapture Methods in Epidemiology: Methods and Limitations," *Epidemiol. Rev.*, vol. 17, no. 2, pp. 243–264, 1995.

[7] A. Chao, "An Overview of Closed Capture-Recapture Models," *J. Agric. Biol. Envir. S.*, vol. 6, no. 2, pp. 158–175, 2001.

[8] T. B. Pedersen, Y. Saygm, E. Savas, "Secret sharing vs. Encryption-based Techniques For Privacy Preserving Data Mining." Joint UNECE/Eurostat work session on statistical data confidentiality, UK, Dec. 2007.

[9] J. Vaidya, C. Clifton, "Secure Set Intersection Cardinality with Application to Association Rule Mining," *J. Comput. Secur.*, vol. 13, pp. 593–622, July 2005.

[10] E. De Cristofaro, P. Gasti, G. Tsudik, "Fast and Private Computation of Cardinality of Set Intersection and Union," in *11th International Conference on Cryptology and Network Security (CANS)*, 2012.