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

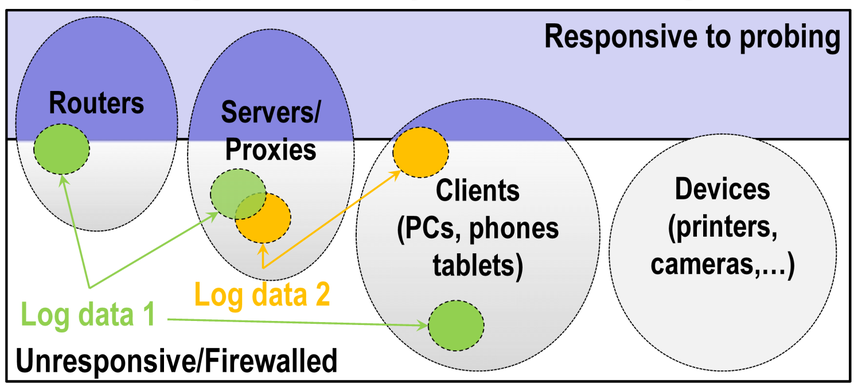
Sebastian Zander, Lachlan L. H. Andrew and Grenville Armitage
Centre for Advanced Internet Architectures (CAIA)
Swinburne University of Technology
{szander,landrew,garmitage}@swin.edu.au

MOTIVATION

- How much "IPv4 reserves"?
  - Predict IPv6 deployment time frame
  - Predict potentially emerging IPv4 address market
- Measure progressive exhaustion of IPv4 space after it is allocated completely

CHALLENGES

- Can probe all IPv4s, but many **don't** reply [1]
- No sharing of **unanonymised** IPv4 logs



APPROACH

- Combine probing ("ping") with several traffic/server logs (multiple IPv4 sources)
- Estimate **unseen** IPv4 addresses with Capture-Recapture (CR) based on
  - Size of sources (assume **not** private)
  - Size of overlap of all source combinations
- Compute overlap without revealing IPv4s (secure set intersection size)

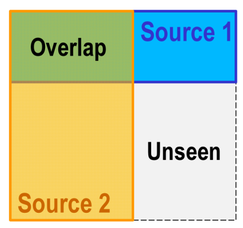
CAPTURE-RECAPTURE

- Predict total IPv4s N from multiple incomplete sources S_i

Two-source Lincoln-Petersen [2]

- Two independent sources
- Same sample probability for all IPv4s (no heterogeneity)

$$\widehat{N} = \frac{|S_1||S_2|}{|S_1 \cap S_2|}$$



Log-linear models [3]

- Multiple dependent sources
- Models source dependence due to heterogeneity

Source 1	Source 2	Source 3	Count
0	0	1	Z_{001}
0	1	1	Z_{011}
…	…	…	…
1	1	1	Z_{111}

Example: three sources counts

Step 1: Fit log-linear model based on counts

$$\log\left(E\left(Z_{ijk}\right)\right) = u + u_1 I(1,0,0) + u_2 I(0,1,0) + u_3 I(0,0,1) + u_{12} I(1,1,0) + u_{13} I(1,0,1) + u_{23} I(0,1,1) + u_{123} I(1,1,1)$$

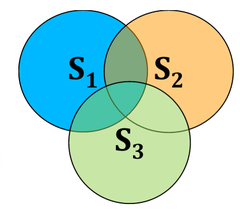
Step 2: Project model to estimate unseen/total

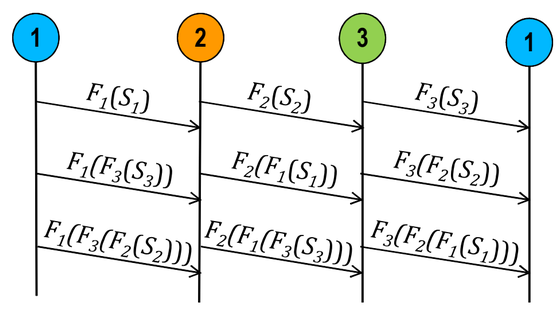
$$\widehat{N} = \sum Z_{ijk} + \hat{Z}_{000}$$

- Stratification by allocation RIR, country, age, prefix size, industry
- CR estimates IPv4s unseen by chance (focus on **routed** space)

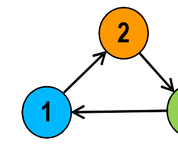
SECURE SET INTERSECTION SIZE

- Securely compute intersection sizes using permutation and **commutative encryption [4]** function F_i (with private key k_i)
  - Pohlig-Hellman encryption: $C = \left(P^{k_1}\right)^{k_2} \bmod n = \left(P^{k_2}\right)^{k_1} \bmod n$
- Compute IPv4s only in S_i, e.g. $Z_{001} = |S_3| - Z_{011} - Z_{101} - Z_{111}$
- Securely compute counts Z_{ijk} for all intersections
  - Collaborators permute and encrypt each (encrypted) data source starting with their own



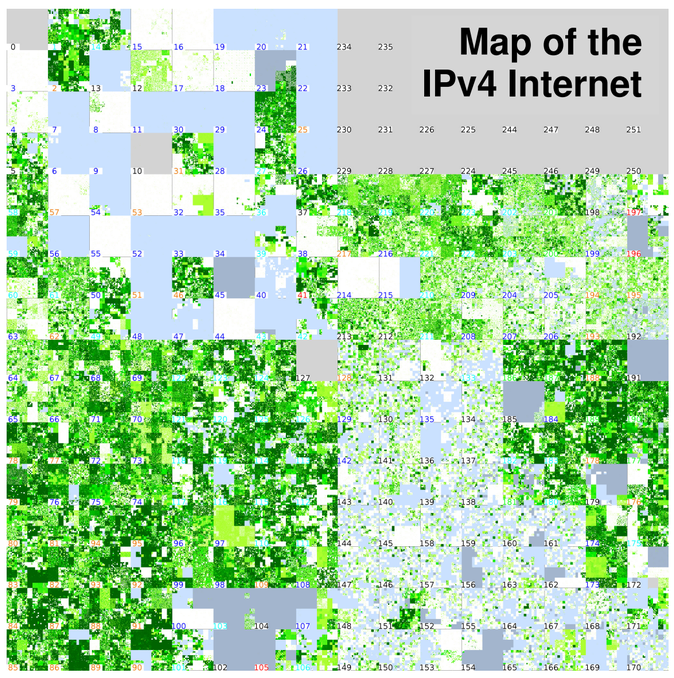


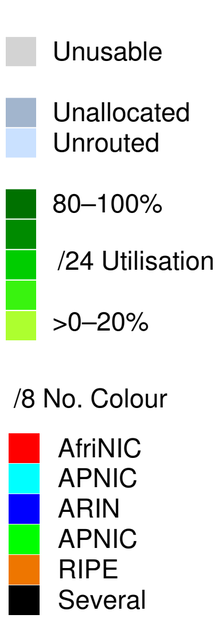
Ring topology

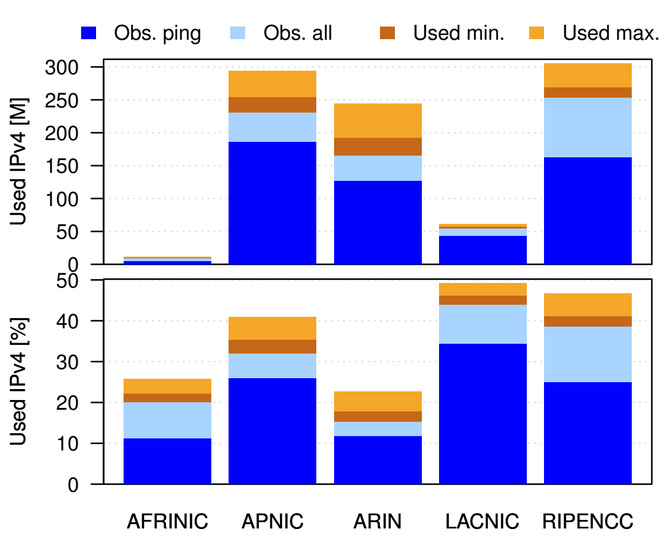


- Collaborators distribute *n*-party-encrypted datasets and intersect, e.g. $|S_1 \cap S_2| = |F_3(F_2(F_1(S_1))) \cap F_1(F_2(F_3(S_2)))|$
- Increase scalability with deterministic sampling of IPv4s

PRELIMINARY RESULTS







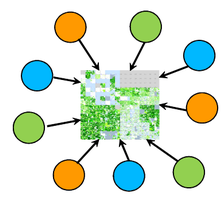
IPv4 address utilisation per Regional Internet Registry (RIR)

- **524** million addresses responsive (ICMP echo, TCP SYN port 80)
- **714** million alive addresses (all sources)
- **790–920** million used addresses estimated with CR (Lincoln-Petersen)
- **30–35%** of routed IPv4 space used

Data from Jan 2011 to Mar 2013. Collaborators: **APNIC**, **Caltech**, **Valve Corp.**

GETTING INVOLVED

- We seek more collaborators to share IPv4 address data
- Contribute unanonymised or anonymised data
- Contact: Sebastian Zander **szander@swin.edu.au**



http://www.caia.swin.edu.au/sting

REFERENCES

1. J. Heidemann, Y. Pradkin, R. Govindan, C. Papadopoulos, G. Bartlett, J. Bannister, "Census and Survey of the Visible Internet", ACM Conference on Internet Measurement (IMC), 2008.
2. F. C. Lincoln, "Calculating Waterfowl Abundance on the Basis of Banding Returns", U.S. Dept . Agric. Circ., vol. 118, pp. 1–4, 1930.
3. A. Chao, "An Overview of Closed Capture-Recapture Models", J. Agric. Biol. Envir. S., vol. 6, no. 2, pp. 158–175, 2001.
4. J. Vaidya, C. Clifton, "Secure Set Intersection Cardinality with Application to Association Rule Mining", J. Comput. Secur., vol. 13, pp. 593–622, July 2005.