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

Sebastian Zander¹, Lachlan Andrew², Grenville Armitage¹

¹Centre for Advanced Internet Architectures (CAIA)
Swinburne University of Technology

²Faculty of IT, Monash University

Slides as presented at IMC 2014

IPv4 Address Space Exhausted

- More than 96% of IPv4 space allocated
- RIRs, except AfriNIC, down to less than /8 prefix
- Rationing is prolonging life of remaining pools
But Allocated is not Actively Used

- How many unused IPv4 “reserves”?  
- Why care about actively used?  
  - Track progressive IPv4 exhaustion  
  - Predict size and costs of IPv4 market  
  - Assist planning for IPv6 transition

RIPE /15: US$10 per IP  
RIPE /20: US$15 per IP  
APNIC /20: US$13 per IP  

Sales data from 2014  
Source: http://ipv4marketgroup.com

Main Challenges

- Previous research focused mainly on active probing, but many hosts do not respond to active probing  
- Passive measurements capture only parts  
- Combine many sources and estimate unseen (ghosts)
Our Approach

Capture-Recapture Population Estimates

Data Collection

Internet Samples

Capture Histories

Capture Histories

Overview

Capture-Recapture Method

Population Estimates

Data Collection

Internet Samples
Capture-Recapture (CR) Method

- Multiple samples over time or multiple data sources

General assumptions

- Individuals can be matched between sources → YES
- Non-zero chance of sampling any individual
  - 25% of IPv4 space not publicly routed → excluded
  - Hidden specialized devices (e.g. printers) → downward bias

Simplest method: two-sample Lincoln-Petersen (L-P)

- Too restrictive assumptions but good for illustration of idea

Method we use: Log-linear models (LLMs)

Lincoln-Petersen Method Illustrated

Unknown population of $N$ IPs
First Sample: Sample and Mark M IPs

We know how many IPs marked \( M = 3 \)
If we knew marked fraction \( \frac{M}{N} \) could find population

Second Sample: Sample C IPs

\[ C = 4 \quad (R = 1 \text{ marked}) \]

\[
\text{Est. marked fraction: } \frac{R}{C} \quad \text{Population: } N = \frac{MC}{R} = \frac{3 \cdot 4}{1} = 12
\]
Limitations of Lincoln-Petersen

- **Homogenous population**: same sample probability for all individuals
- **Independent sources**: Inclusion in one source does not affect inclusion in other source
- **Heterogeneity causes “apparent source dependence”**: (both effects confounded)

Log-linear models use more than two sources to compensate (apparent) source dependence

Log-linear Models (LLMs)

- Illustrate LLMs with 3 data sources

Capture Histories

\[ Z_{000} = ? \]
Log-linear Models (LLMs)

- System of $2^3 - 1 = 7$ equations
  \[
  \log \left( E(Z_{ijk}) \right) = u + u_1 1_{i=1} + u_2 1_{j=1} + u_3 1_{k=1} + u_{12} 1_{i=1^j=1} + u_{13} 1_{i=1^k=1} + u_{23} 1_{j=1^k=1} + u_{123} 1_{i=1^j=1^k=1}
  \]

- Parameters $u$ model dependencies
- Maximum-likelihood estimation of $u$
- Model selection process selects $u$ to use
  - Select least complex model with “adequate” fit
- Estimate $Z_{000}$: $\hat{Z}_{000} = \exp(u)$
### Overview

![Diagram showing Capture-Recapture and Population Estimates](image)

**Dataset collection and processing**

---

### Collected IPv4s (Jan 2011 – Jun 2014)

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Description</th>
<th>Unique IPs 2013 [M]</th>
</tr>
</thead>
<tbody>
<tr>
<td>IPING</td>
<td>ICMP Internet census</td>
<td>411</td>
</tr>
<tr>
<td>CALT</td>
<td>Caltech NetFlow data</td>
<td>356</td>
</tr>
<tr>
<td>GAME</td>
<td>Steam game server logs</td>
<td>confidential</td>
</tr>
<tr>
<td>SWIN</td>
<td>Swinburne NetFlow data</td>
<td>113</td>
</tr>
<tr>
<td>WEB</td>
<td>Web clients tested for IPv6</td>
<td>109</td>
</tr>
<tr>
<td>TPING</td>
<td>TCP port 80 Internet census</td>
<td>93</td>
</tr>
<tr>
<td>MLAB</td>
<td>Clients measurement lab</td>
<td>22</td>
</tr>
<tr>
<td>SPAM</td>
<td>Spam database</td>
<td>18</td>
</tr>
<tr>
<td>WIKI</td>
<td>Wikipedia page edit history</td>
<td>7</td>
</tr>
</tbody>
</table>
Dataset Pre-Processing

- Internet census (IPING, TPING)
  - Include only probed IPs that responded with ICMP echo replies or SYN/ACKs
  - Include router IPs that sent ICMP errors
- Passive datasets
  - Filter out private, multicast, unrouted addresses
  - Filter out spoofed unused addresses (NetFlow datasets)
  - Use 12-month time windows starting every 3 months
  - Analyse unique IP addresses, unique /24 subnets used

Overview

<table>
<thead>
<tr>
<th>Data Collection</th>
<th>Capture-Recapture</th>
<th>Population Estimates</th>
</tr>
</thead>
<tbody>
<tr>
<td>Capture Histories</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Internet Samples</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Validation</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
How Well Does Approach Work?

- Don't know number of used IPs (ground truth)
- Estimate IPs only in one source using all other sources

![Graph showing observed ping and LLM estimate range for different data sources.](image)

How Well Does Approach Work?

- For /24 subnets we have really high overlap
- Still LLM improves marginally

![Graph showing observed ping and LLM estimate range for different data sources.](image)
Comparison Ground Truth Samples

- Compare actual “peak usage” of few networks with LLM estimates

![Graph showing Comparison Ground Truth Samples]

Overview

- Capture-Recapture
- Data Collection
- Population Estimates
- Results

Internet Samples

IPs
Results

- Used IPv4 addresses and /24 networks
  - Overall
  - Depending on regions / RIRs
    - Depending on allocation age
    - Depending on allocation size
    - Depending on allocation country
  - Remaining unused prefixes distribution

Growth IP Addresses

<table>
<thead>
<tr>
<th></th>
<th>Estimated</th>
<th>Routed</th>
<th>Observed</th>
</tr>
</thead>
<tbody>
<tr>
<td>IPv4 addresses</td>
<td>1.2G</td>
<td>2.7G</td>
<td>0.7G</td>
</tr>
<tr>
<td>Growth rate</td>
<td>+170M/year</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
**Growth /24 Subnets**

- **Routed**
  - 10.6M
  - Observed
  - 5.9M
  - Estimated
  - +0.45M/year

**Absolute Growth IPs Regions/RIRs**

- Asia: 450M
- Europe: 300M
- North America: 290M
- South America: 220M
- Africa: 40M
Relative Growth IPs Regions/RIRs

How Long Until Exhaustion?

- Based on current estimated remaining unused+unallocated space and growth trends
  1. South America (LACNIC)  Exhausted soon (1-2 years)
  2. Asia (APNIC)
  3. Africa (AfriNIC)
  4. Europe (RIPE)
  5. North America (ARIN)  Enough supply (2 decades)

- BUT
  - Given IPv6, how much unused IPv4 space will be used?
  - What will growth trends be in future?
  - Future transfers of address blocks between RIRs?
Future Work

- More ground truth validation
- Estimate IPv6 space usage
- More data sources
  - Looking for collaborators
  - Developed secure scheme ensuring data anonymity

Summary

- Log-linear capture-recapture approach shows promising results for estimating used IPv4 space
- Estimated used IPs well over observed or pingable IPs, but observed /24 subnets close to estimated /24 subnets
- 1.2G IPv4 addresses used (45% publicly routed space)
- 6.2M /24 subnets used (60% publicly routed space)
- Significant unused IPv4 space (especially legacy blocks)
  - IPv4 address market, if regulators permit
  - Slower transition to IPv6

Supported by Australian Research Council grants LP110100240 (with APNIC Pty Ltd) and FT0991594.