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

Sebastian Zander[1], Lachlan Andrew[2], Grenville Armitage[1]

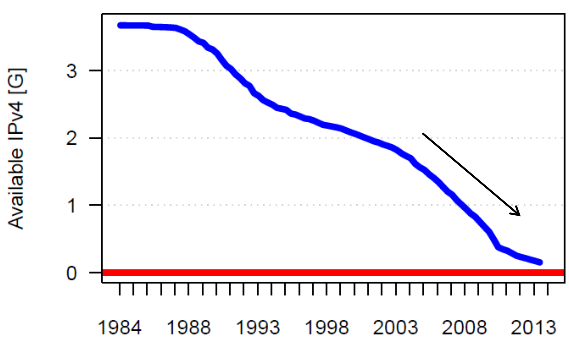
[1]Centre for Advanced Internet Architectures (CAIA)
Swinburne University of Technology

[2]Faculty 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

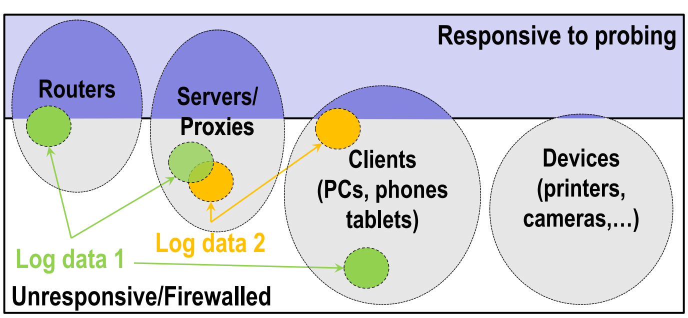
RIPE /20: US$15 per IP

APNIC /20: US$13 per IP

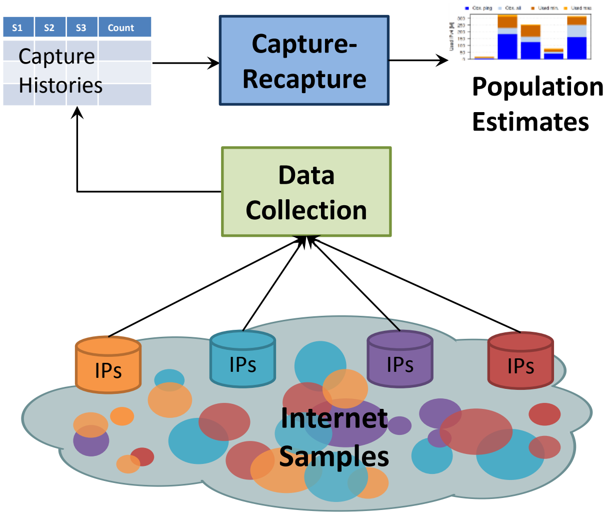
Sales data from 2014
Source: http://ipv4marketgroup.com

Main Challenges

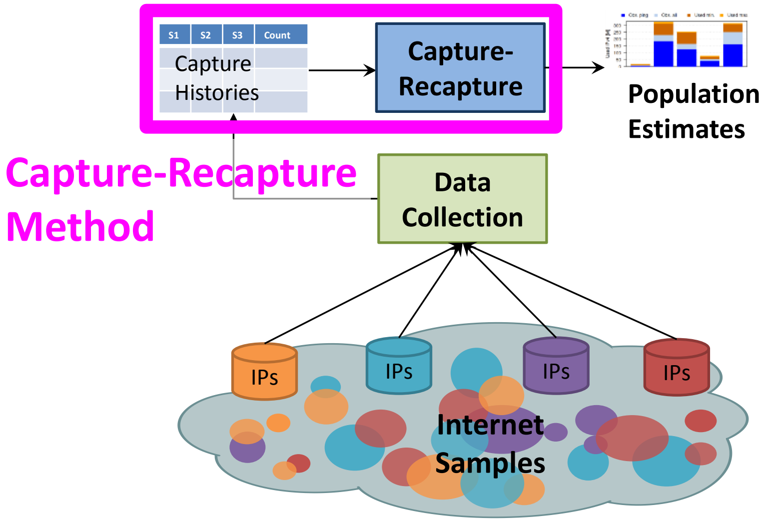
- Previous research focused mainly on active probing, but many hosts do not respond to active probing
- Passive measurements capture only parts
- **Combine many sources and estimate unseen (ghosts)**



Our Approach



Overview



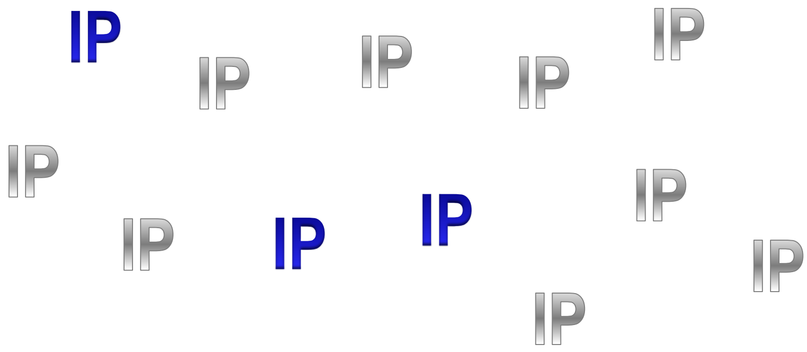
Capture-Recapture (CR) Method

- 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

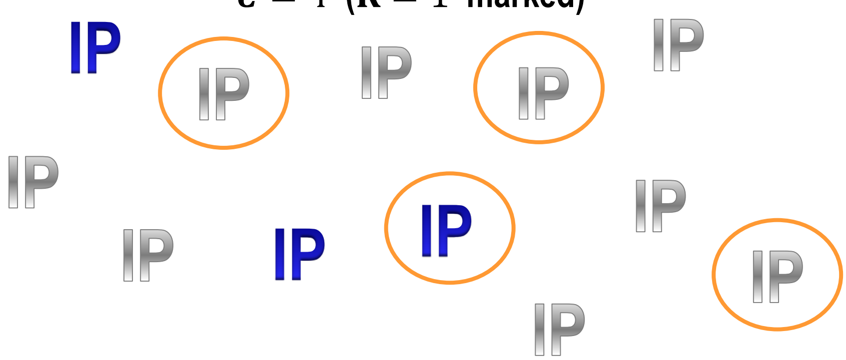


We know how **many** IPs marked ($M = 3$)

If we knew marked **fraction** $\frac{M}{N}$ could find population

Second Sample: Sample C IPs

$C = 4$ ($\mathbf{R} = 1$ **marked**)



Est. marked fraction: $\frac{R}{C}$ **Population**: $N = \frac{MC}{R} = \frac{3 \cdot 4}{1} = \mathbf{12}$

Limitations of Lincoln-Petersen

- **Homogenous population**: same sample probability for all individuals

  Clients Servers Cameras Phones Printers

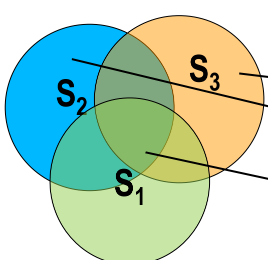
- **Independent sources**: Inclusion in one source does not affect inclusion in other source ( )

- Heterogeneity causes "apparent source dependence" (both effects confounded)

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

Log-linear Models (LLMs)

- Illustrate LLMs with 3 data sources



S_1	S_2	S_3	Count
0	0	1	Z_{001}
0	1	0	Z_{010}
…	…	…	…
1	1	1	Z_{111}

Capture Histories

$Z_{000} = ?$

Log-linear Models (LLMs)

- System of $2^3 - 1 = 7$ equations

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

- 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}: $\hat{\boldsymbol{Z}}_{\mathbf{000}} = \exp(\boldsymbol{u})$

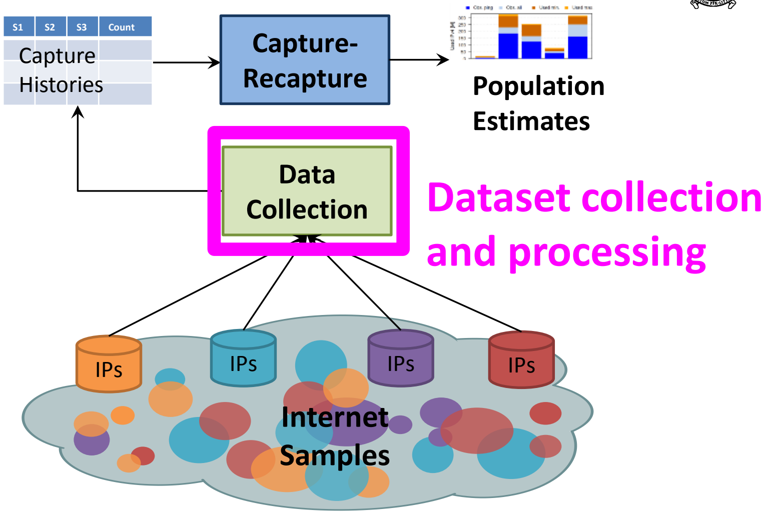
Log-linear Models (LLMs)

- System of $2^3 - 1 = 7$ equations

$$\log\left(E(Z_{ijk})\right) = u + u_1\mathbf{1}_{i=1} + u_2\mathbf{1}_{j=1} + u_3\mathbf{1}_{k=1} + \boldsymbol{u}_{\mathbf{12}}\mathbf{1}_{i=1 \wedge j=1} + \boldsymbol{u}_{\mathbf{13}}\mathbf{1}_{i=1 \wedge k=1} + \boldsymbol{u}_{\mathbf{23}}\mathbf{1}_{j=1 \wedge k=1} + \boldsymbol{u}_{\mathbf{123}}\mathbf{1}_{i=1 \wedge j=1 \wedge k=1}$$

- 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}: $\hat{\boldsymbol{Z}}_{\mathbf{000}} = \exp(\boldsymbol{u})$

Overview



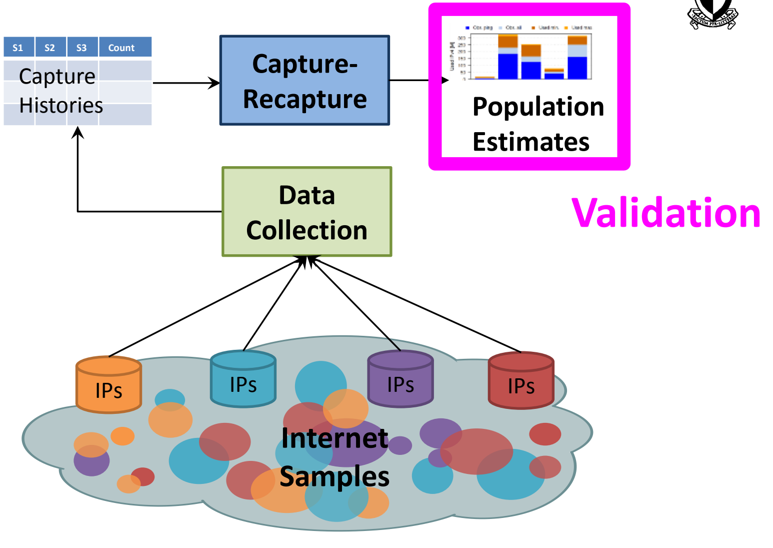
Collected IPv4s (Jan 2011 – Jun 2014)

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

Dataset Pre-Processing

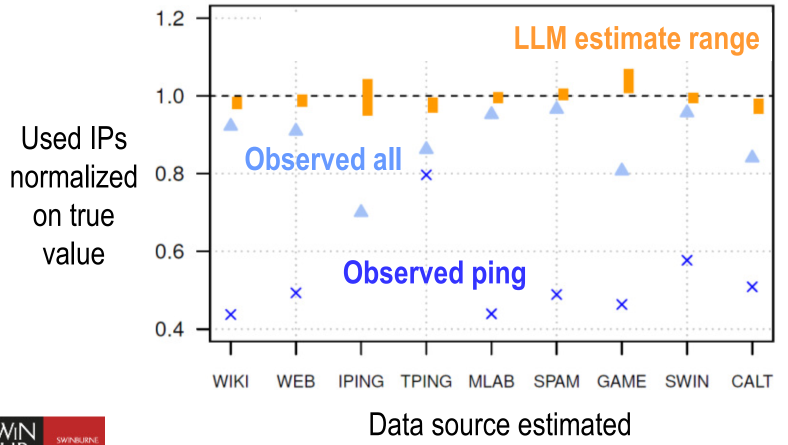
- Internet census (IPING, TPING)
  - Include only probed IPs that responded with ICMP echo replies or SYN/ACKs
  - Include router IPs that sent ICMP errors
- Passive datasets
  - Filter out private, multicast, unrouted addresses
  - Filter out spoofed unused addresses (NetFlow datasets)
- Use 12-month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

Overview



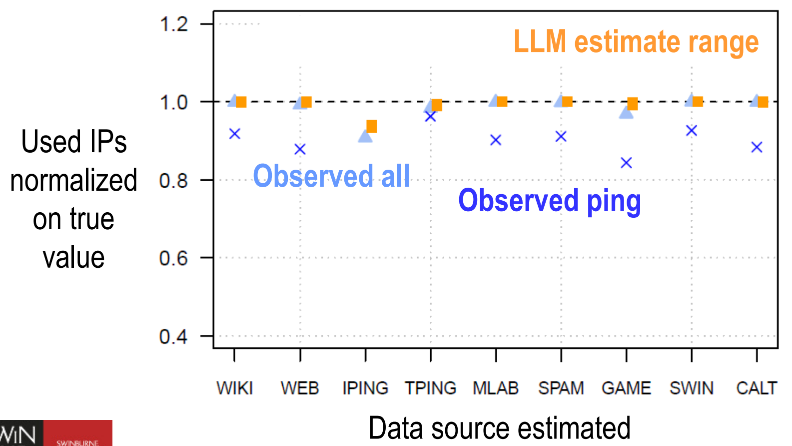
How Well Does Approach Work?

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



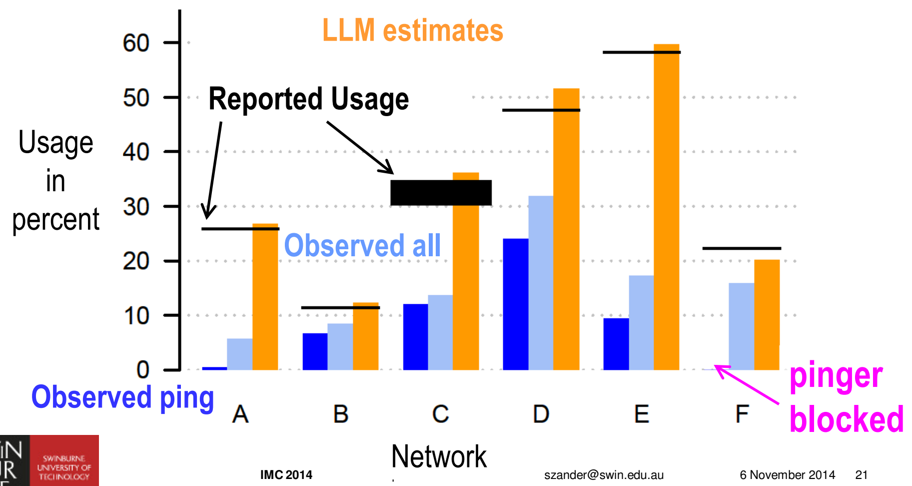
How Well Does Approach Work?

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

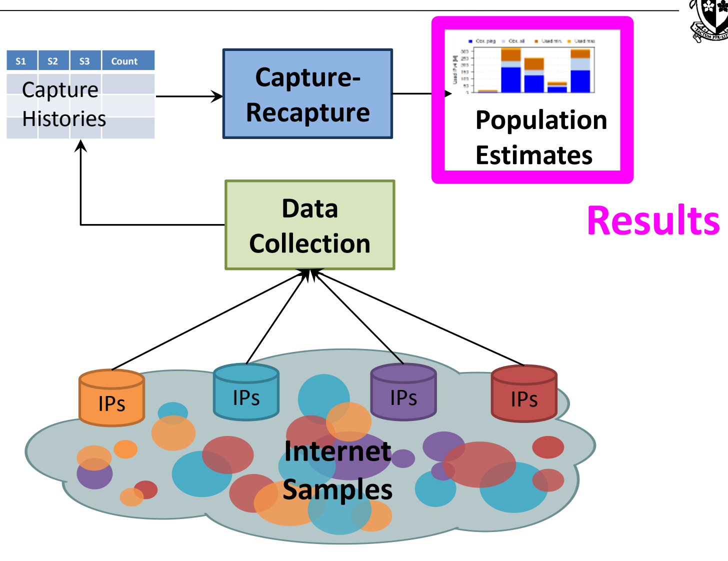


Comparison Ground Truth Samples

- Compare actual "peak usage" of few networks with LLM estimates



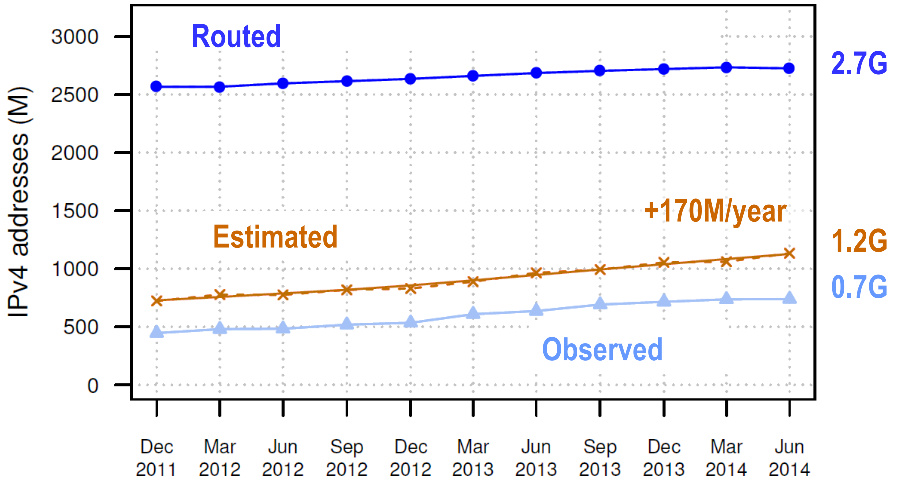
Overview



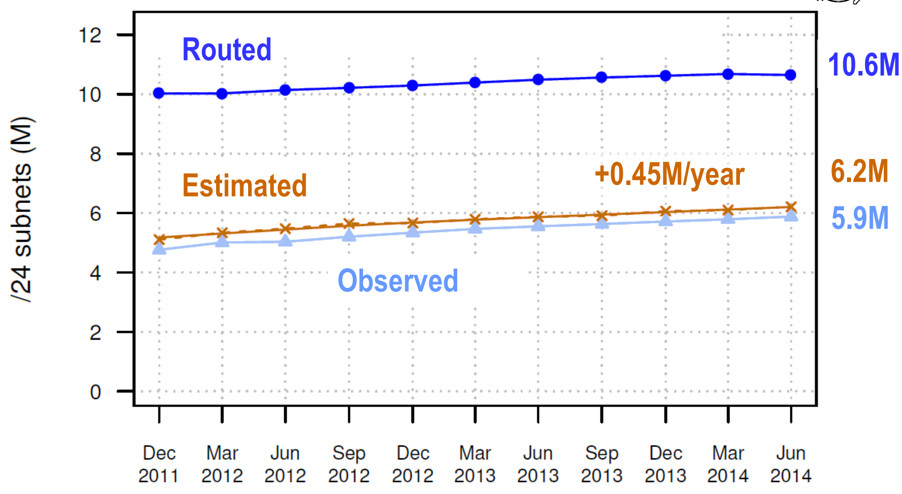
Results

- 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

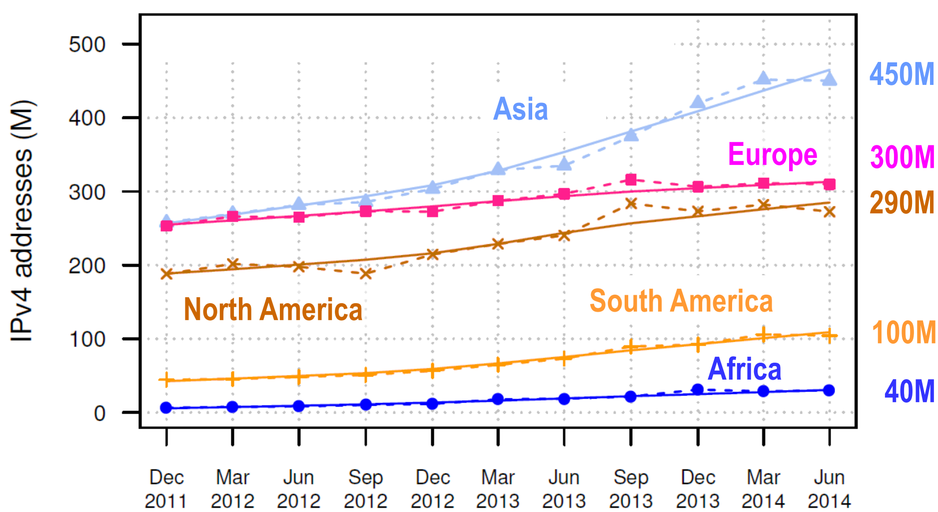
Growth IP Addresses



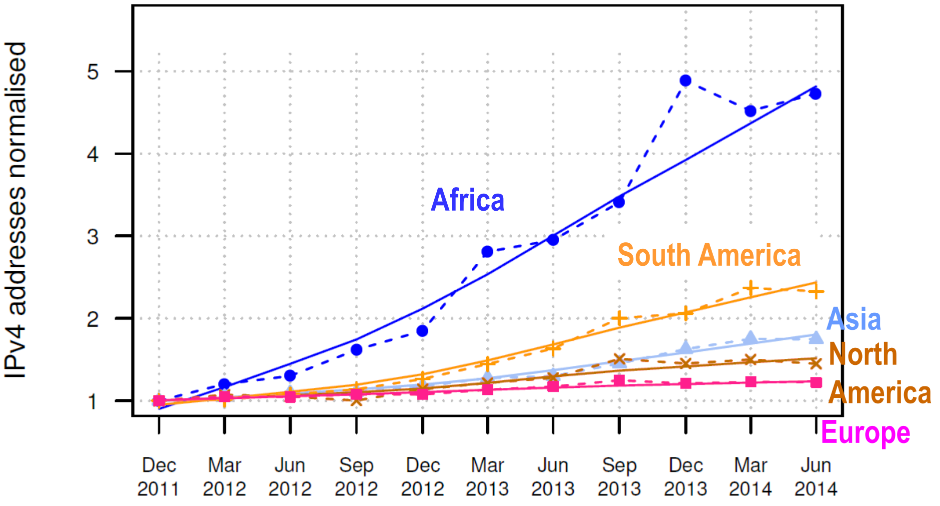
Growth /24 Subnets



Absolute Growth IPs Regions/RIRs



Relative Growth IPs Regions/RIRs



How Long Until Exhaustion?

- Based on current estimated remaining unused+unallocted space and growth trends

  1. South America (LACNIC)
  2. Asia (APNIC)
  3. Africa (AfriNIC)
  4. Europe (RIPE)
  5. North America (ARIN)

  Exhausted soon (1-2 years)

  ↕

  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.