# Examining the Internet Address Space through Census and Survey

John Heidemann
USC/Information Sciences Institute

joint work with Xue Cai, Yuri Pradkin, Ramesh Govindan, Christos Papadopoulos, Genevieve Bartlett, Joseph Bannister

IPAM Workshop on Multi-Resolution Analysis of the Internet
Los Angeles, California, USA
3 November 2008 and 14 November 2008



Examining the Internet Address Space through Census and Survey

# (the multi-scale analysis remix)

John Heidemann
USC/Information Sciences Institute

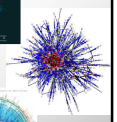
joint work with Xue Cai, Yuri Pradkin, Ramesh Govindan, Christos Papadopoulos, Genevieve Bartlett, Joseph Bannister

IPAM Workshop on Multi-Resolution Analysis of the Internet
Los Angeles, California, USA
3 November 2008 and 14 November 2008



## Our Goal: Study All of *T*oday's Internet

map **all edge hosts**



*and* scale to the **size of today's Internet**

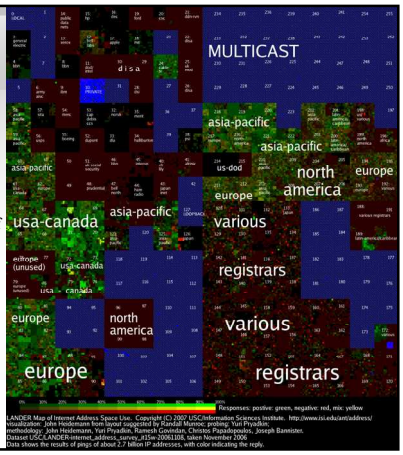
our approach:

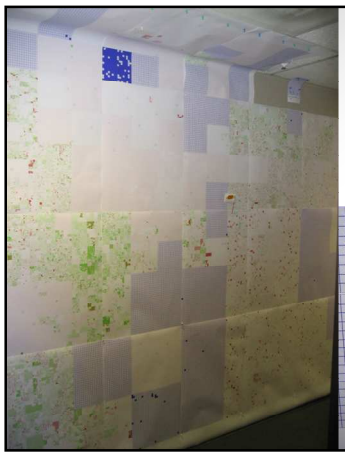
ping *all* addresses once (census)
*some* addresses many times (survey)
quantify sources of error



## The Internet

- average each /16
  - each pixel: 65k addresses
  - represents all $2^{32}$ addrs
- brightness: responsivenesses
- green/red-ness: degree of positive vs. negative replies
- blue: areas not probed
- layout: Hilbert Curve

## The *W*hole Internet

- here, 1 pixel is 1 address
- 9x9' at 600dpi
- green: positive, red: negative; white: no resp.

## Caveats

- *not* a perfect statement of truth
  - misses NAT'ed hosts
  - misses non-ICMP-responsive hosts (those behind firewalls)
  - some pings are lost (we estimate <
- *but* the *best current view* of the I and a *new methodology* to be refined

"Your data is useless, everybody blocks pings" –common first reaction

"ghetto science" – slashdot "discussion"

*We don't think so, and we'll show data to support our claim.*

## Outline

- introduction
- methodology
- **validation**
  - **variance**
  - **overcounting**
  - **undercounting**
- applications
- multi-resolution analysis



## Sources of Error

- variance
  - measurement location: *doesn't matter; normal error*
  - sampling error:
    - *can predict from theory*
    - *function of probe frequency*
    - *surveys within 0.4% (with 95% confidence)*
  - births/deaths during survey: *estimate in paper*
  - probe type (ICMP vs. TCP): *coming up*
- overcounting
  - routers and multi-homed hosts: *estimated at <6% in paper*
- undercounting
  - probe loss: *random due to probe order; use 1-repair process to recover single losses in survey*
  - firewalled hosts: *coming up*



## Comparing USC and a Random Sample

USC Survey (82k hosts)

| category: | | any | active |
|---|---|---|---|
| addresses probed | 81,664 | | |
| non-responding | 54,078 | | |
| responding any | 27,586 | 100% | |
| ICMP or TCP | 19,866 | 72% | 100% |
| ICMP | 17,054 | 62% | 86% |
| TCP | 14,794 | 54% | 74% |
| Passive | 25,706 | 93% | |
| ICMP only | 656 | | |
| TCP only | 1,081 | | |
| Passive only | 7,720 | | |

1M Random Addresses

| category: | | active |
|---|---|---|
| addresses probed | 1,000,000 | |
| non-responding | 945,703 | |
| responding either | 54,297 | 100% |
| ICMP | 40,033 | 74% |
| TCP | 34,182 | 62% |
| both ICMP and TCP | 19,918 | |
| ICMP only | 20,115 | |
| TCP only | 14,264 | |

Internet best estimate omits passive component
*but Internet results are consistent with USC*
(Internet is somewhat less responsive: 74% vs. 86%)

Q: additional validation data and approaches?



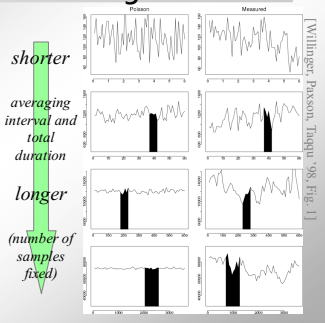
## Outline

- introduction
- methodology
- validation
- applications
- **multi-resolution analysis**
  - **what does it mean?**
  - **time vs. space**
  - **spatial analysis**



## Multi-Resolution Analysis in Networking

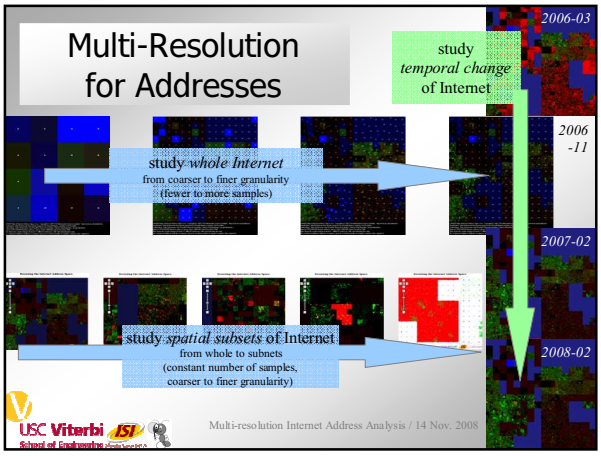
- network traffic:
  - vary *timescale* of study
- network router or AS topology
  - look for power laws and models
- what about for addresses?




[Willinger, Paxson, Taqqu '98, Fig. 1]

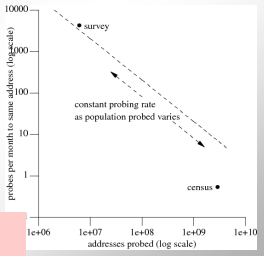


## Multi-Resolution for Addresses



## Trading Time and Space

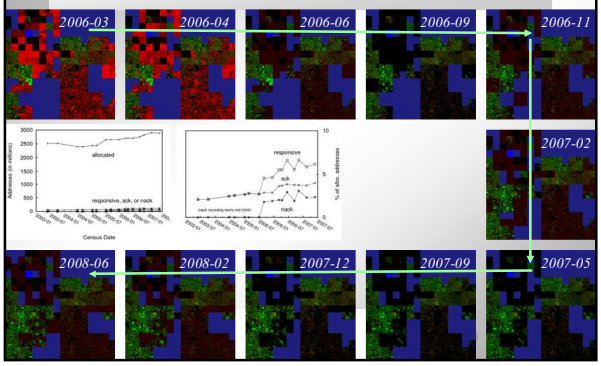
- given fixed probing resources
  - 4 machines
  - ~6k probes/s
- how to allocate?
  - trade off coverage vs. temporal precision




Q: what are the interesting points in this space?

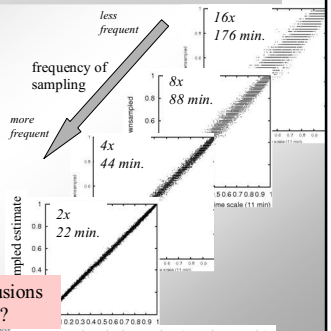


## Internet Evolution Over Time




## How Often to Probe?

- what is accuracy if we probe less frequently?
- method
  - study survey data
  - take most frequent as ground truth
  - downsample data (discard every other observation)
  - error := |decimate – true|
- less frequent sampling => more error (of course)




Q: any stronger conclusions about probe frequency?

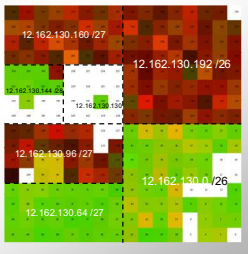


## Spatial Analysis

- *space* more interesting dimension
  - address block: *a.b.c.d/n:* all addresses with common *n*-bit prefix
  - 127/8: 127.0.0.0 through 127.255.255.255
- hypothesis:
  - if address *a.b.c.d* is used for purpose *X*
  - then *a.b.c.*(*d*+1) is likely *also* used for *X*
  - => **spatial correlation of address blocks**
- why?
  - Internet addresses are allocated in blocks
    - ICANN to regional registries to ISPs to you
  - *but* no guarantee they're used consistently

## Spatial Correlation: Preliminary Evidence

- survey /24 network
- plot those addresses
  - *green:* availability (A)
  - *red: volatility* (V): normalized number state changes (up-down)
  - layout: Hilbert curve
- many /24s show some blocks with consistent usage






## Exploiting Spatial Correlation

*if* spatial correlation is strong, then maybe…

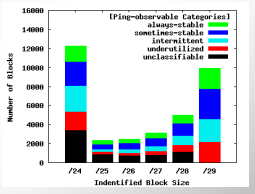
- exploit spatial correlation to get effect of more frequent probing?
- use block correlation to determine block size?
- relate correlated blocks to block usage?

(learn about the hard-to-observe Internet)

## Sizes of Observed Blocks

- what sizes of blocks show correlated use?
- method: find "natural" block boundries
  - split blocks until variance in observed (A, V, M) falls

=> 59% of /24s are used consistently

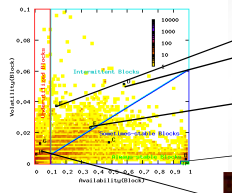



(data from Survey17w; consistent with Survey16w)



## Use of Observed Blocks

Density of /24s (Survey17w)



Maps of Sample /24s

dynamic

*green: A; red: M*

*compare to DNS-name implied usage*

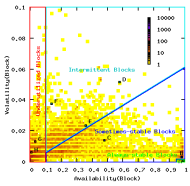
always stable
~> servers

underutilized
~> often dialup



## Ping-Inferred Use Categories

Density of /24s (Survey17w)



***Underutilized*:**
**A < 0.1**

***Intermittent*:**
**MedianUp < 6hours**

***Sometimes-stable*:**
**MedianUp >= 6hours**

***Unclassifiable*:**
**< 20% responses**

***Always-stable*:**
**A > 0.95, V < 0.0016 (Nu=1)**

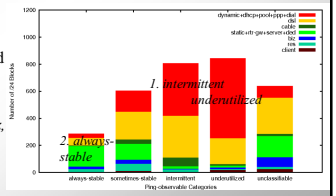
Q: better ways to extract information from data? or better data? Interesting sub-classes (servers, dynamic)?



## Preliminary Validation of Block Use

- examine hostnames as evidence of use
  - mail.example.com
  - dsl55.lax.example.net
  - dialup15.example.edu
- preliminary observations
  1. **intermittent** and **underutilized** has expected hostnames: *dynamic, dhcp, pool, ppp, dial, dsl*
  2. **always-stable** mostly expected: *static, rtr-gw, biz, server*
     - but some *dsl*

- hostnames: ISC Domain Survey (ISC_DS-2007JAN)
- pings: Survey17w






## Validating the Spatial Correlation Hypothesis

- we *assume* address blocks often have common use
- to test, network operations at USC provided ground truth
  - not perfect: often *they don't know* because of sub-delegation
  - evaluation is in progress



## USC Validation of Block Sizes

- block size id on 249 of USC's /24s
  - 6 not allocated by block
  - 147 are grouped correctly

**63%**
- 132 correct (true positives)
- *9 false positives—allocated in smaller increments, but adjacent blocks used in similar ways* *[not really incorrect observations]*

105 false negatives
- *18 not used*

**34%**
- 29 not responding (firewalled) [firewalls are known problem]
- 16 with few responses (firewalls?)
- **8 had sub-blocks of mixed usage**
- **34 had inconsistent usage** **[17% VIOLATE HYPOTHESIS]**

## USC Validation of Block Usage

- ping-infer usage on 138 USC /24s

**60%** – 83 correctly classified
  - 4 dynamic labeled intermittent
  - 4 dynamic labeled sometimes-stable
  - 3 VPN & PPP labeled intermittent or underutilized
  - 71 building-specific blocks labeled always-stable or sometimes-stable

**38%** – 52 labeled unclassifiable
  - 2 should have been identified (servers and routers)

**2%** – 3 incorrectly classified
  - all are dynamic blocks labeled always-stable

- rough classification of use pretty reliable, when use is by /24
  - but 38% /24s unknown, and worse for smaller blocks



## Spatial Analysis Conclusions

- studying the Internet is tough
  - limited ability to observe
  - (at least, without triggering intrusion detection!)
- careful analysis of survey data seems promising
- exploiting spatial correlation is key
  - seems to match real-world use (majority of time)
  - makes the most of limited data

Q: relationship of edge response with traditional router graph?

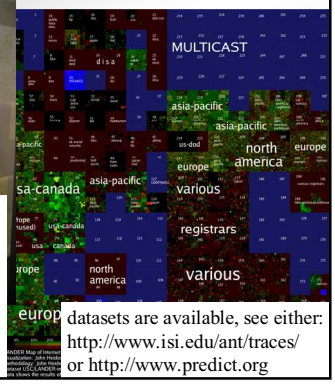


## Overall Conclusions

- we *can* study the edge of the Internet
  - early estimates of size of the Internet
  - ongoing work on use of edge hosts
- census and survey provide a basis to build on
  - reasonable method of active probing
  - new application of population sampling
- much still to do



## Where Next?




additional information:
http://www.isi.edu/ant/address/

"Census and Survey of the Visible Internet", Heidemann et al, IMC 2008

"Active Probing to Classify Internet Address Blocks (poster abstract)", Cai and Heidemann, SIGCOMM 2008

datasets are available, see either:
http://www.isi.edu/ant/traces/
or http://www.predict.org