USC/LANDER Passive and Active Data Collection

John Heidemann
joint work with Genevieve Bartlett, Yue Cai,
Maureen Dougherty, Ramesh Govindan, Christos Papadopoulos (co-PIs),
Liz Quinn, Yuri Pradhan
USC/ISI, USC/ITS, CSU
12 February 2009

LANDER in One Slide

- http://www.isi.edu/lan/lander
- passive data collection (ongoing)
  - snooping all packet headers with 2-levels of anonymization
  - Los Nets (in LA regional net)
  - FrontRange Gigapop (Colorado academic)
  - SurvPop (broadband commercial)
  - eventually plan to allow user-provided analysis code on our boxes
  - curating some datasets (DoS, etc.)
- active data collection (since 2003 and ongoing)
  - IP address census: ping the world (all allocated IPv4 addresses, every quarter or so)
  - IP address survey: ping some of the world, often 1% of v4, every 11 minutes, for 1 week
- support from DHS (infrastructure) and NSF (analysis)

The Internet

- average each /16
  - each pixel: 65k addresses
  - represents all 2^64 addresses
- brightness:
  - responsiveness
  - greyscale: net degree of positive vs. negative replies
  - blue: areas not probed
  - layout: Hilbert Curve

The Whole 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 <5%)
- but the best current view of the Internet,
  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 have data to support our claim.

Are Pings Useful at All?

<table>
<thead>
<tr>
<th>USC Survey (82k hosts)</th>
<th>1M Random Addresses</th>
</tr>
</thead>
<tbody>
<tr>
<td>category:</td>
<td>any</td>
</tr>
<tr>
<td>addresses picked</td>
<td>16,664</td>
</tr>
<tr>
<td>corresponding pairs</td>
<td>8,051</td>
</tr>
<tr>
<td>ICMP-only</td>
<td>456</td>
</tr>
<tr>
<td>TCP only</td>
<td>1,681</td>
</tr>
<tr>
<td>Passive only</td>
<td>7,410</td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td>YES! (given error estimates)</td>
<td></td>
</tr>
</tbody>
</table>
- USC says 24% low vs. solid baseline
- random Internet says ~25% low vs. weaker baseline
- ICMP strictly better than active TCP probing
What Good Are They?

- pretty pictures
- trying to get a handle on size of the Internet
- estimating how the net is used (work in progress with Xue Car; figure at left)
- building a hit list of live edge hosts (ex: for Ark and other topology probers)

- estimating how the net is used (work in progress)
  - red is dial-up and dynamic; green is servers

Additional LANDER Information

- http://www.isi.edu/ant/lander/
- part of PREDICT: http://www.predict.org
- all datasets are available

- active probing more info: see Heidemann et al.; “Census and Survey of the Visible Internet”, ACM IMC 2008; doi:10.1145/1452520.1452542