Internet Topology Data Kit

Update

Young Hyun
CAIDA

5th CAIDA-WIDE-CASFI Workshop
Aug 2, 2012
Introduction: ITDK

goals:
- provide curated data for studying Internet topology
  - interface-, router-, and AS-level topology
- employ best available measurement and analysis techniques
- release 2-3 ITDKs per year
Introduction: ITDK

* motivation:
  * overwhelming amount of raw data
    • e.g., TB’s of raw traceroute data over a decade
  * researchers often interested in derived data
    • e.g., AS level, not interface level
  * valuable for multiple researchers to study same dataset
    • build upon each other’s work (explore different facets)
    • cross validation
History

- historical ITDK releases in 2002 and 2003
  - traceroute topology from skitter
- revived ITDK in 2010
  - same goals but significantly different contents
  - traceroute topology from Ark and other complementary data
- six releases:
  - 2010: 01, 04, 07 (Jan, Apr, July)
  - 2011: 04, 10
  - 2012: 07 (in progress)
Contents

* router-level topology graphs
* router-to-AS assignments
* geographic locations of routers
* DNS names of observed IP addresses
Contents: Topology

* router-level topology graphs
  * derived from IPv4 Routed /24 Topology Dataset
    * used two weeks of traceroutes to every routed /24
    * probed 9.5 million /24’s from 54 monitors in 29 countries (Oct 2011)
  * resolved interfaces into routers by combining multiple techniques
    * iffinder: implements Mercator technique
    * MIDAR: IP-ID based technique
    * kapar: extended APAR technique
graph components:

* node = router with list of interface addresses
* link = connection between routers

• may have >2 routers per link due to layer 2 and other causes (such as data collection/analysis artifacts)
two router-level topology graphs:

- **accuracy**: midar+iffinder: highest confidence alias resolution
- **completeness**: midar+iffinder+kapar: more alias coverage but also more false positives
  - kapar provides analytic alias resolution for targets unusable with measurement-based techniques
Monotonic ID-Based Alias Resolution (MIDAR)

Monotonic Bounds Test
- for two addresses to be aliases, their combined IP-ID time series must be monotonic

sliding-window probe scheduling for scalability

4 probing methods
- TCP, UDP, ICMP, “indirect” (traceroute-like TTL expired)

multiple sources


MIDAR v0.3.0 released Jul 11, 2012 (GPLv2)

- http://www.caida.org/tools/measurement/midar/
MIDAR Software

* three front-ends to MIDAR
  
  * **midar-cor**
    - testing a small (< 200) set of IP addresses
      - efficient testing of all possible pairs of single suspected alias set
    - corroboration stage only; single probe method; single host
    - can be used to test/verify aliases obtained by other means
  
  * **midar-full: local mode**
    - testing a medium-size (< 40,000) set of IP addresses
    - all MIDAR stages; multiple probe methods; *single* host

  * **midar-full: distributed mode**
    - testing an Internet-scale (2 million+) set of IP addresses
    - all MIDAR stages; multiple probe methods; *multiple* hosts
# MIDAR Results

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Input addresses</strong></td>
<td>1.12 M</td>
<td>1.50 M</td>
<td>1.90 M</td>
<td>2.32 M</td>
<td>2.19 M</td>
</tr>
<tr>
<td><strong>Monotonic addresses</strong></td>
<td>0.99 M</td>
<td>1.20 M</td>
<td>1.44 M</td>
<td>1.87 M</td>
<td>1.83 M</td>
</tr>
<tr>
<td><strong>Possible pairs</strong></td>
<td>486 G</td>
<td>724 G</td>
<td>1038 G</td>
<td>1754 G</td>
<td>1676 G</td>
</tr>
<tr>
<td><strong>Shared pairs after</strong></td>
<td>1.63 M</td>
<td>4.00 M</td>
<td>5.49 M</td>
<td>6.83 M</td>
<td>7.00 M</td>
</tr>
<tr>
<td><strong>Discovery stage</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Final Results</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>- Shared pairs</td>
<td>0.433 M</td>
<td>1.36 M</td>
<td>1.67 M</td>
<td>2.49 M</td>
<td>2.68 M</td>
</tr>
<tr>
<td>- Routers</td>
<td>69 k</td>
<td>108 k</td>
<td>121 k</td>
<td>125 k</td>
<td>118 k</td>
</tr>
<tr>
<td>- Addresses on routers</td>
<td>189 k</td>
<td>383 k</td>
<td>426 k</td>
<td>413 k</td>
<td>403 k</td>
</tr>
</tbody>
</table>

- continually improved MIDAR over time
- increasing input size
- improving accuracy and effectiveness
Contents: AS Assignments

* goal: determine which AS owns each router

Contents: Geolocation

* geographic location (at city granularity) of routers in the router-level graphs
  * MaxMind's free GeoLite City database
* procedure:
  * map each interface on a router to a location
  * if all interfaces map to same location, then use that location
  * otherwise, no assigned location for router
Contents: DNS Lookups

- use HostDB, CAIDA’s bulk DNS lookup service
- two datasets:
  - DNS lookups within days of observing an address in a traceroute path
  - DNS lookups during alias resolution runs
    - better matches alias resolution results
Future Work

- AS-level topology overlaid on router-level topology
- AS relationships
- IPv6 topology
Thanks!

For more information or to request data:

For questions: data-info@caida.org