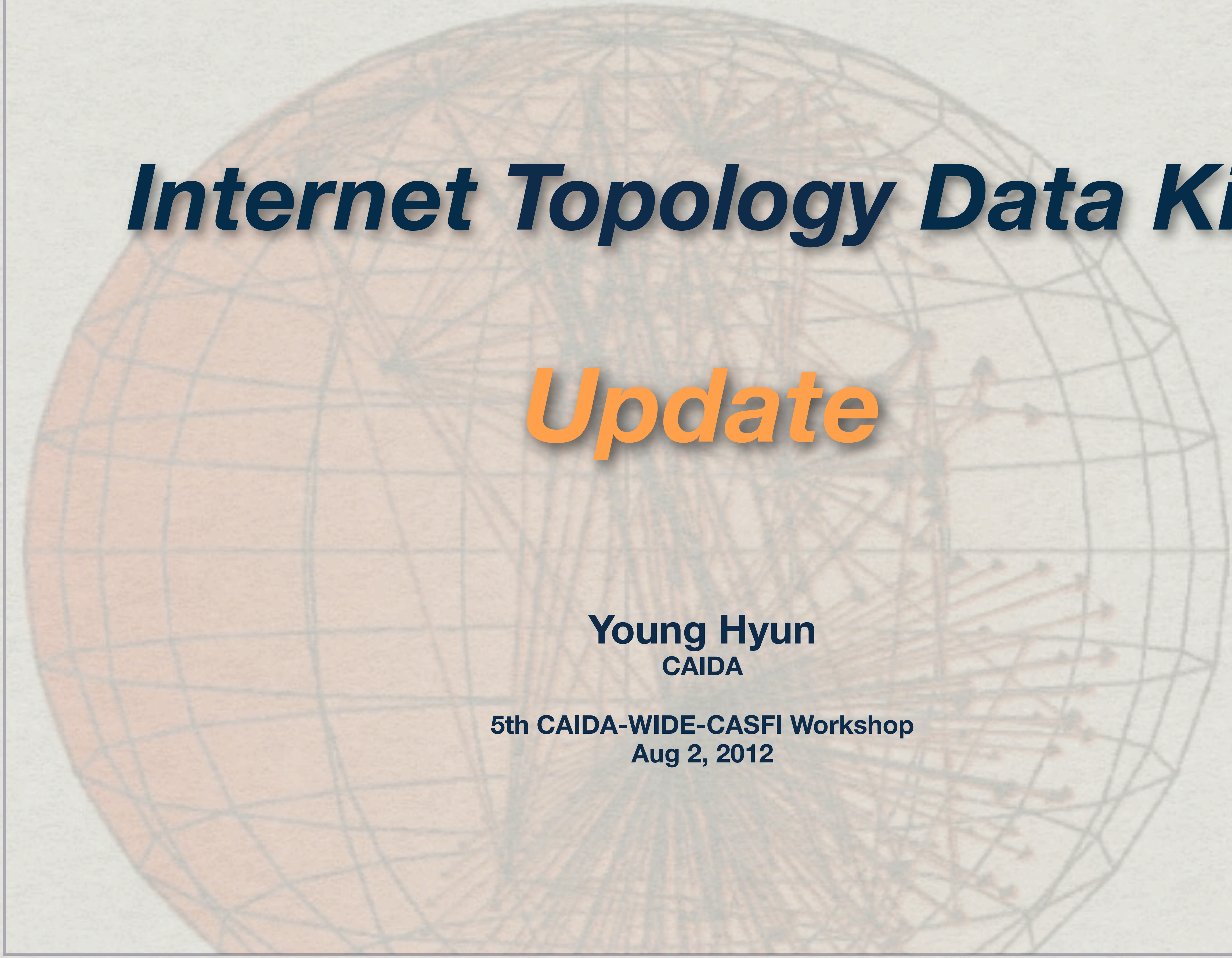


Internet Topology Data Kit

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

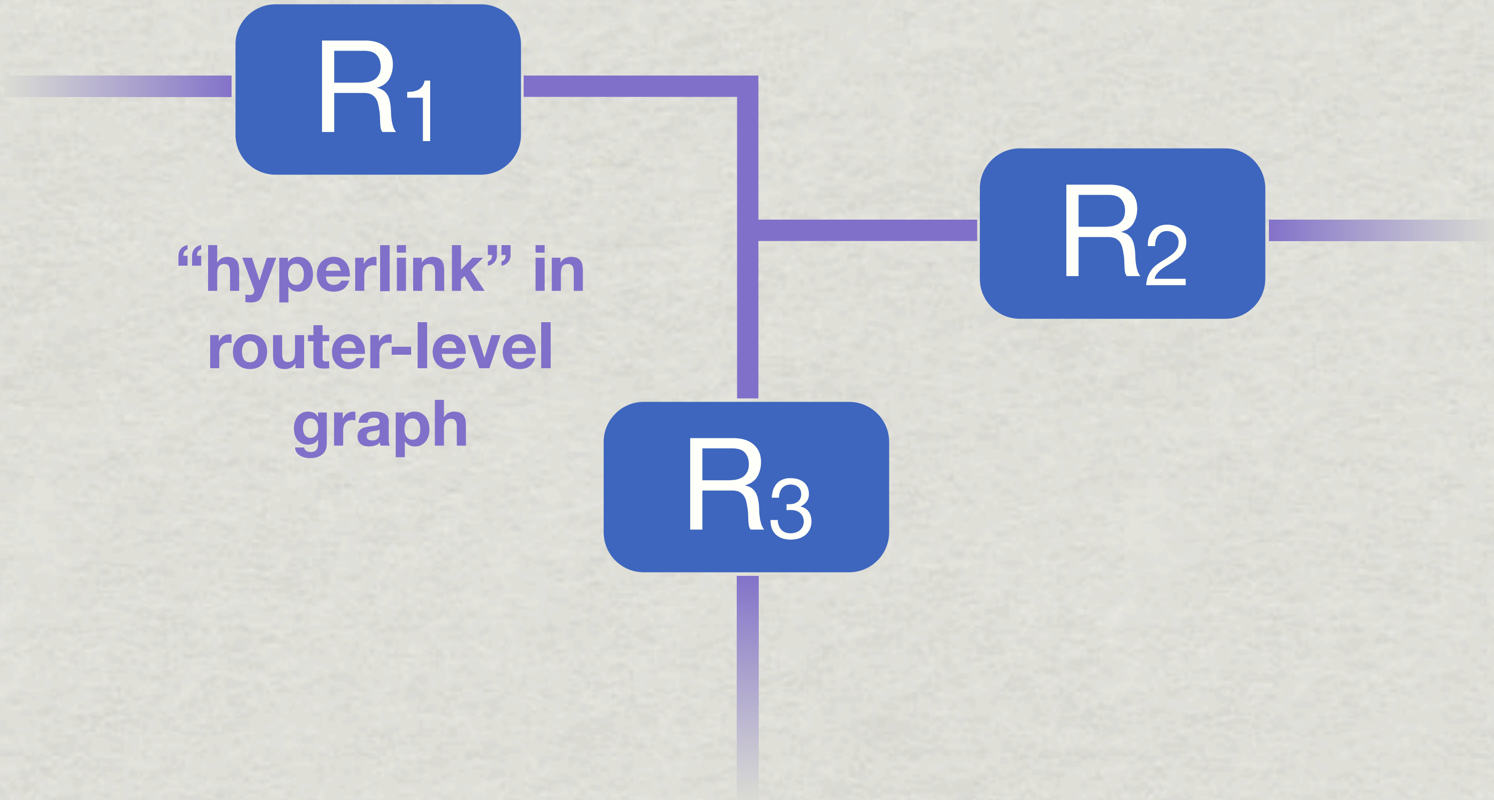
- router-level topology graphs
- router-to-AS assignments
- geographic locations of routers
- DNS names of observed IP addresses

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

Contents: Topology

- 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)



"hyperlink" in router-level graph

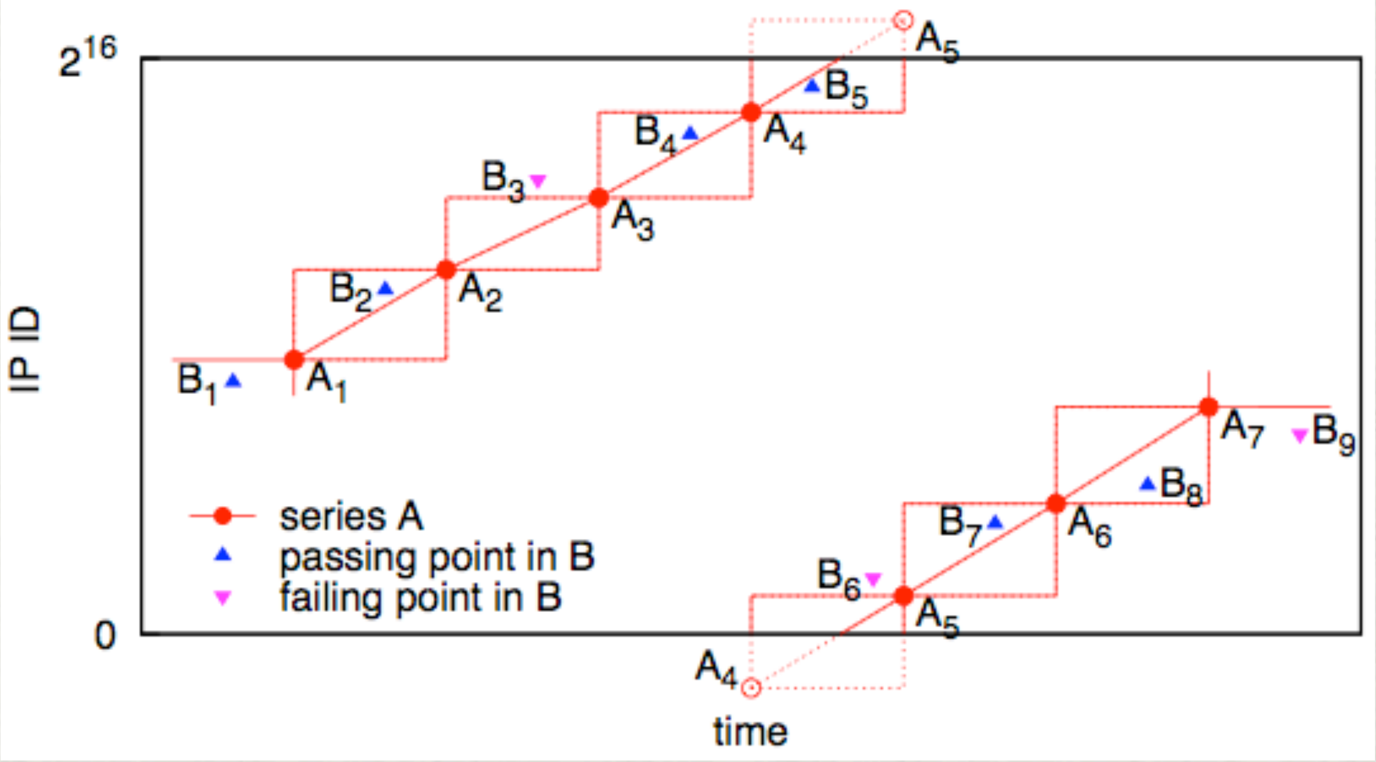
Contents: Topology

		2010-01	2010-04	2010-07	2011-04	2011-10
input topology traces		4 weeks	4 weeks	2 weeks	2 weeks	2 weeks
optimized for **accuracy**	nodes	3.33 M	4.41 M	3.34 M	3.38 M	3.25 M
	links	3.34 M	4.43 M	3.50 M	3.60 M	3.47 M
optimized for **completeness**	nodes	3.26 M	4.20 M	2.96 M	3.02 M	2.92 M
	links	3.30 M	4.32 M	3.38 M	3.48 M	3.36 M

- 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

MIDAR

- **M**onotonic **ID**-Based **A**lias **R**esolution (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

- K. Keys, Y. Hyun, M. Luckie, and k. claffy, “**Internet-Scale IPv4 Alias Resolution with MIDAR**”, to be published in IEEE/ACM Transactions on Networking, 2012.
  - http://www.caida.org/publications/papers/2012/alias_resolution_midar/
- 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

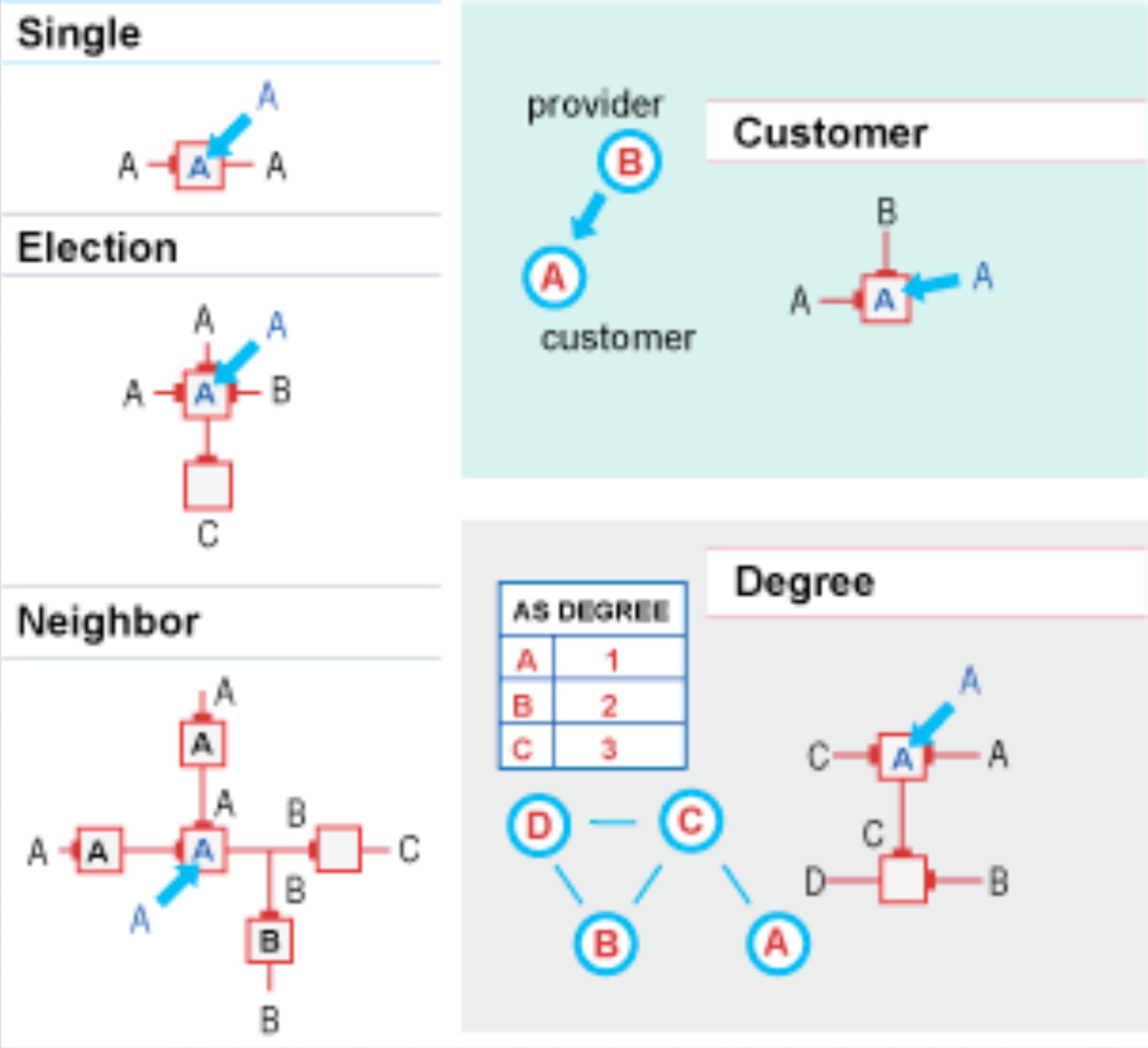
MIDAR Results

	2010-01	2010-04	2010-07	2011-04	2011-10
Input addresses Monotonic addresses Possible pairs	1.12 M 0.99 M 486 G	1.50 M 1.20 M 724 G	1.90 M 1.44 M 1038 G	2.32 M 1.87 M 1754 G	2.19 M 1.83 M 1676 G
Shared pairs after Discovery stage	1.63 M	4.00 M	5.49 M	6.83 M	7.00 M
Final Results • Shared pairs • Routers • Addresses on routers	 0.433 M 69 k 189 k	 1.36 M 108 k 383 k	 1.67 M 121 k 426 k	 2.49 M 125 k 413 k	 2.68 M 118 k 403 k

- continually improved MIDAR over time
    - increasing input size
    - improving accuracy and effectiveness

Contents: AS Assignments

- goal: determine which AS owns each router
- Huffaker, *et al*, “**Toward Topology Dualism: Improving the Accuracy of AS Annotations for Routers**,” in PAM 2010.



AS	DEGREE
A	1
B	2
C	3

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:

www.caida.org/data/active/internet-topology-data-kit

For questions: data-info@caida.org