AMITÉ: Annotation and Mapping of Internet Topology at the Edges
BAA 07-09, TTA 5: Internet Tomography/Topography

Project Retrospective

John Heidemann and Ramesh Govindan
joint work with Xue Cai, Xun Fan, Zi Hu, Yuri Pradkin;
part with Bala Krishnamurthy and Walter Willinger (AT&T)
USC/ISI and USC/CS Dept.

9 October 2012

Copyright © 2012 by John Heidemann
Release terms: CC-BY-NC 3.0 unported
Yet We Must Understand the Internet

many network security questions
• how robust is…
  – the Internet routing topology?
  – cloud computing?
  – models of topology, traffic, …?
• trends and correlations in…
  – {network location, country, provider, users…}
  – to {compromised hosts, spam generation, botnet C&C, traffic generation, service use…}
• new technology deployment…
  – firewalls? content filters?
  – IPv6 and new protocols? new applications?

how did 2010 Japanese earthquake affect Internet connectivity?

• anecdotes are useful starters…
  – but data essential to
ex: Nov ‘08: take out a bad ISP (McColo)
how much did spam change?
how disproportionate was McColo’s spam?
compare McColo to daily spam ebb&flow?
answers must combine spam + population

how quickly is IPv6 deployment going? or DNSsec?
Our Insight

• we can directly measure the whole Internet
  – fast computers and networks
  – “only” 4 billion IPv4 addresses or 100 anycast nodes
• bring new techniques
• and validation and calibration
  – quantify uncertainty

privacy-sensitive collection of real network data

new measurement and analysis

understanding and knowledge
AMITÉ’s Approach to provide an **Internet map** that is

- **updated**
  - continuous probing and regular map updates

- **edge-conscious**
  - end hosts and services
    (not just routers and links)

- **annotated**
  - latencies, services, owners
    (not just connectivity)

to improve network understanding and security

**one map: pings to 3 billion addresses over 2 months**
Other data: allocation, usage, high/low bitrate…
Expected AMITÉ Benefits

- **new measurement tools**
  - run on the Internet or your network
  ⇒ new raw data

- **annotated maps**
  - knowledge in that raw data
  ⇒ informs simulations and studies

- **studies quantify accuracy**
  - answers with error bars
  ⇒ know bounds and limits

- **visualization and understanding**
  ⇒ browse, query, or reuse this new knowledge
AMITÉ Development Model

- from **basic research to operation**
  - prove new results via peer review
  - ongoing collection
  - results at range of maturities
- **provide data** to others
- **standardize new approaches**
  - where appropriate
- **today**
  - data is browsable on the web
  - datasets available for use (gratis under DHS PREDICT)
  - tools in use (data collection 24x7); available for others
  - mature tools open sourced
- **ongoing**
  - long-term data collection
  - as a service: we run our tools on your network
  - as a product: spinning out the tools
AMITÉ Challenge and Competitors

• what collection scales to the Internet?
  – many targets
  – many security policies (and paranoid and naïve administrators)
  – need for new infrastructure

• how does it change?
  – constant evolution
  – what are the invariants?

• how can we assess accuracy?
  – finding ground truth

• we complement other topology studies:
  – other researchers (CAIDA, iPlane, DIMES) look at core, 
    *us: edge and services*
  – companies (MaxMind, Neustar, etc.) focus on products, 
    *us: new approaches and free (gratis) data*
Specific AMITÉ Results

- artifacts
  - web-based Internet map
  - datasets
- new techniques
  - topology discovery
    - hitlist generation
    - AS-to-org mapping
  - understanding the edge
    - block-allocation policies
    - low-bitrate edge detection
    - anycast enumeration
    - mass geolocation
Specific AMITÉ Results

- artifacts
  - web-based Internet map
  - datasets
- new techniques
  - topology discovery
    - hitlist generation
    - AS-to-org mapping
  - understanding the edge
    - block-allocation policies
    - low-bitrate edge detection
    - anycast enumeration
    - mass geolocation
Integrated Visualization of Internet Edge

**need:** first responders, network admins, researchers need to understand Internet

**insight:** Google maps shows power of web browser

**approach:** use OpenStreetMaps, plus Hadoop-generated tiles

**benefits:**
- go-to tool to understand unknown IP address
- excellent PR for Internet mapping (poster, news coverage, understandable)

**responsiveness**

**latency**

**allocation**

**geolocation**
**Automatic Hitlist Generation**

**need:** all topology studies need *targets* *(a hitlist)* ideally addresses that are up

**insight:** our censuses tell the best targets

**approach:**
- study series of censuses *(data on all reachablity)*
- look at each /24’s history
- to find best representative for each /24 over whole Internet

**benefits:**
- data used by 5 groups
- analysis suggested improvements to Internet mapping

---

need: Autonomous Systems (ASes) are important to understand Internet topology.

insight: must consider that large ISPs often use many ASes.

approach: build an AS-to-organization map from whois.

benefits: better understanding of true ISP footprint and influence.

Cai, Heidemann, Krishnamurthy, Willinger: "An Organization-Level View of the Internet and its Implications (Extended)", ISI TR 2012-679 joint work with AT&T.

Our map: 49,262 ASes to 36,463 orgs. Only 11% of orgs have multiple ASes, but those orgs are important controlling 64% of addrs and 29% of ASes.

AS-to-org matters: compare “biggest AS” (today) vs. all ASes (us), without all ASes, often underestimate ISPs.

underestimates size of large organizations.
Anycast Enumeration

**need:** anycast is used in *most* root and TLD DNS requests. How big is the infrastructure? Are there anycast hijackers?

**insight:** we can map anycast servers by looking from many places

**approach:** probe anycast from 100k end users or 300k recursive DNS; define new data to assist auditing

**benefits:** better understanding of true ISP footprint and influence

*Fan, Heidemann, Govindan: “Characterizing Anycast in the Domain Name System”, ISI TR 2012-681*

---

**dns?** I’ll use closest (anycast)

**our approach is complete (recall > 0.8)**

but requires at least 20k vantage points
**need:** where (*physically*) is each IP address? Today’s geolocation is often approximated for whole IP blocks.

**insight:** we can scale up geolocation to study *every* IPv4 address

**approach:** use many vantage points, but pick the *right ones for each target*

**benefits:** public geolocation data for every IPv4 address

High-Impact Results

• deep-edge mapping
  – IP census/survey data* [14 external users]
  – block allocation* [6 external users of data]
  – low-bitrate block detection*
  – IP hitlists* [7 external users of data]
• wide-area service discovery
  – anycast discovery† [standardization underway; 2 external users of ideas]
  – as-to-org mapping† [4 external users of data; 1 external user of ideas]
• map views and data export
  – IPv4 address browser
  – lots of data to research [over 23 unique research groups using our data]
• wide-area geolocation
  – geolocation* in progress: 54 /8s = 24% allocated, 55% geolocatable
  – datasets now public
• impact via
  – ideas (talks plus *peer reviewed and †in-progress papers)
  – datasets
  – public code and services
Do Our Results Inform Your Work?

• browse our data on the web
  – http://www.isi.edu/ant/address/browse/

• use our ideas?  http://www.isi.edu/ant/pubs/

• data for your simulations and models?
  – data is free (gratis): http://www.isi.edu/ant/traces/
  – approval at http://predict.org

• collaboration or to extend our tools?

http://www.isi.edu/ant/