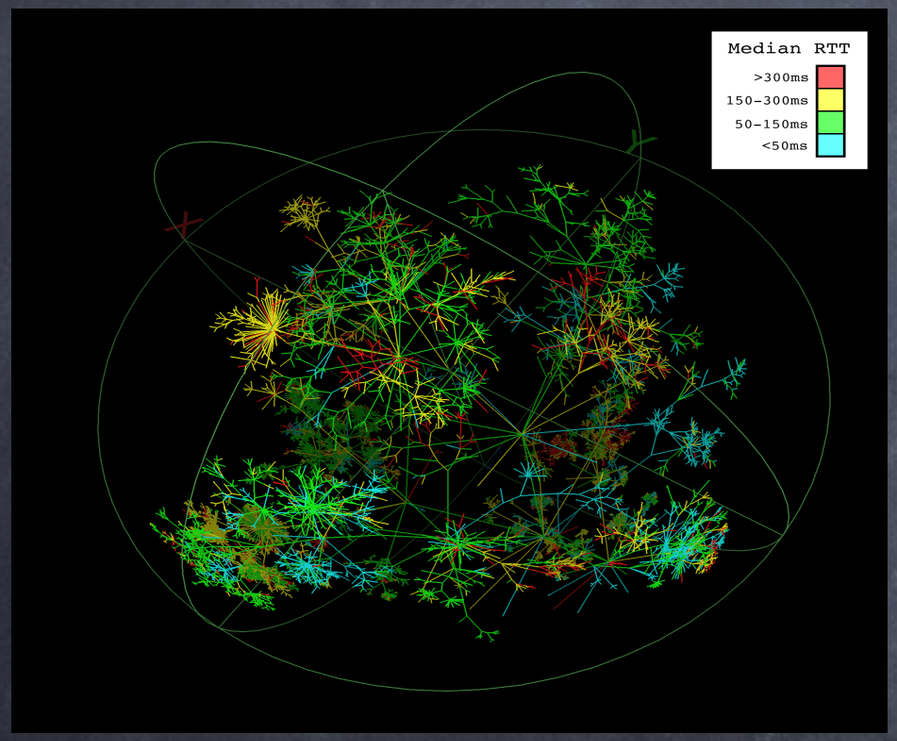
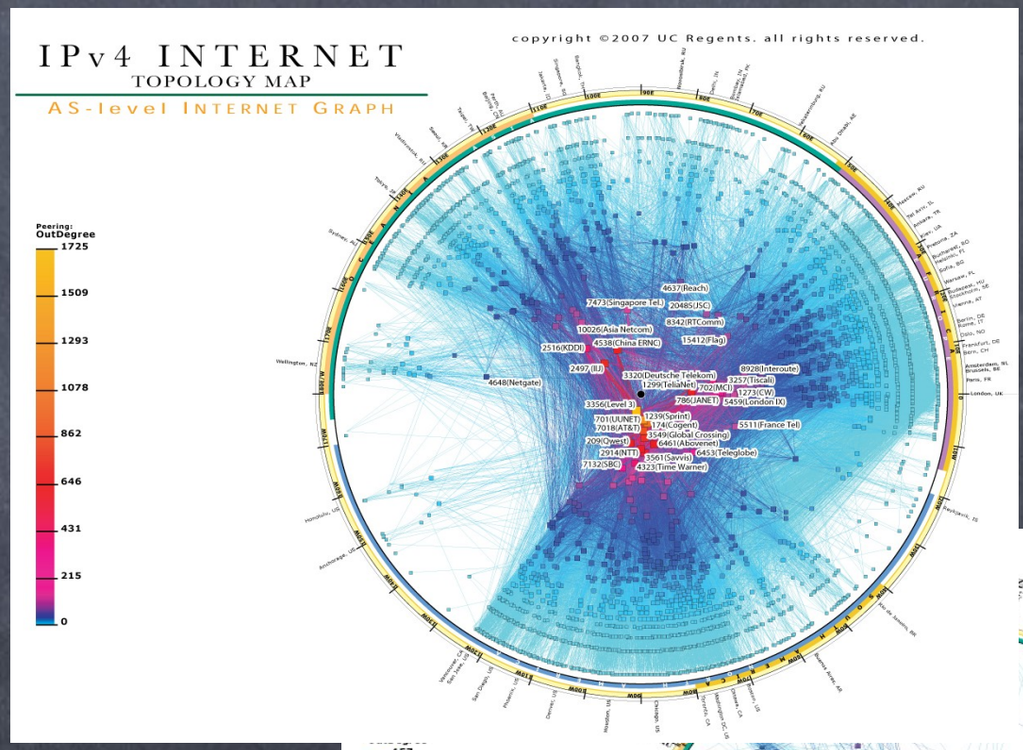
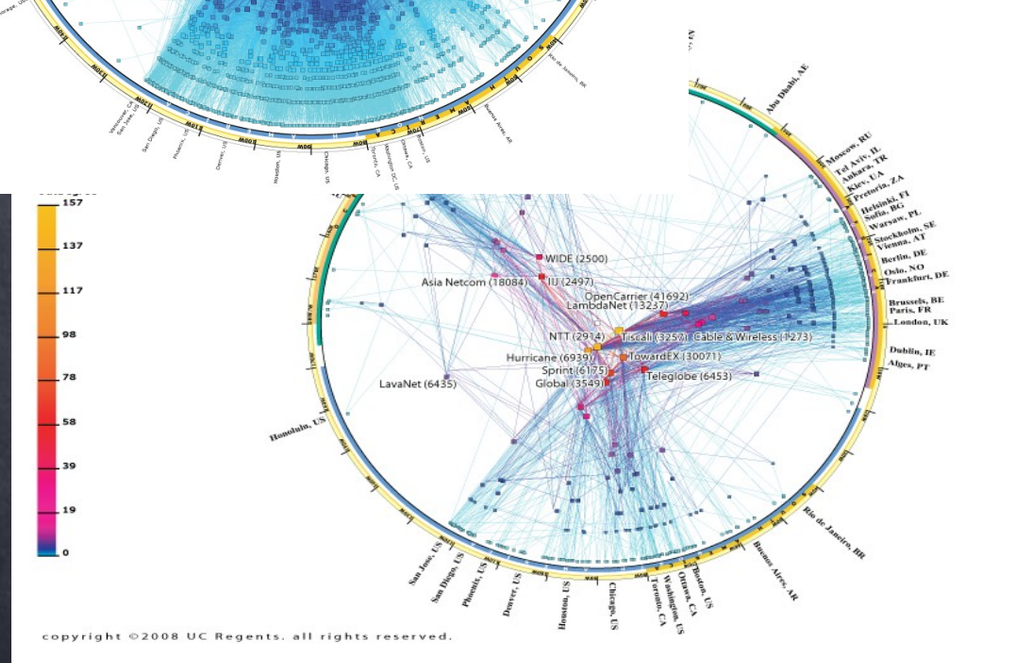
Internet science: why Wall Street and Main Street should care (a survey of caida activities)







kc claffy

CAIDA
DHS - ITTC
SRI
16 Oct 2008

recipe for disaster
(aka “you are here”)

- We now critically depend on the Internet for our professional, personal, and political lives.
- But what do we know about it? e.g, what keeps the system stable or drives it to instability?
- Researchers and policymakers currently analyze a trillion dollar industry in the dark.
- Few data points available suggest a dire picture.
- Agencies charged with infrastructure protection have little situational awareness regarding global dynamics and operational threats.

How did we get here?

- Telephone system: 140+ years of history, including regulated data collection requirements (and profits). and a precisely defined system.

- Data networks: 40 years old, ad hoc/hack, tossed to private sector before mature, with no govt support for research or metrics (or profit), ill-defined system.

- Current academic projects either lack sustainability (iplane) or ability to dedicate resources (PlanetLab)

- War: the best motivation so far for investing in situational awareness of critical infrastructure

CAIDA: background & history

Since 1997: narrowing the gap between Internet operations and science in face of global privatization

Largely US taxpayer funded (nsf, dhs), plus sponsors

Seek, analyze, communicate salient features of best available data on the Internet

Use this data to prepare for the future

Recent expansion of research agenda into policy and economics

CAIDA activities

- data sharing & curation for reproducible research (datcat, predict, ditl, commons)
- passive measurement: software support, h/w deployed
- dns traffic and vulnerability analyses
- active measurement, curation, analyses, modeling, simulation
- forward-looking: routing architecture for 1B nodes
- policy guidance: “top10 list”, IPv6 surveys, blog

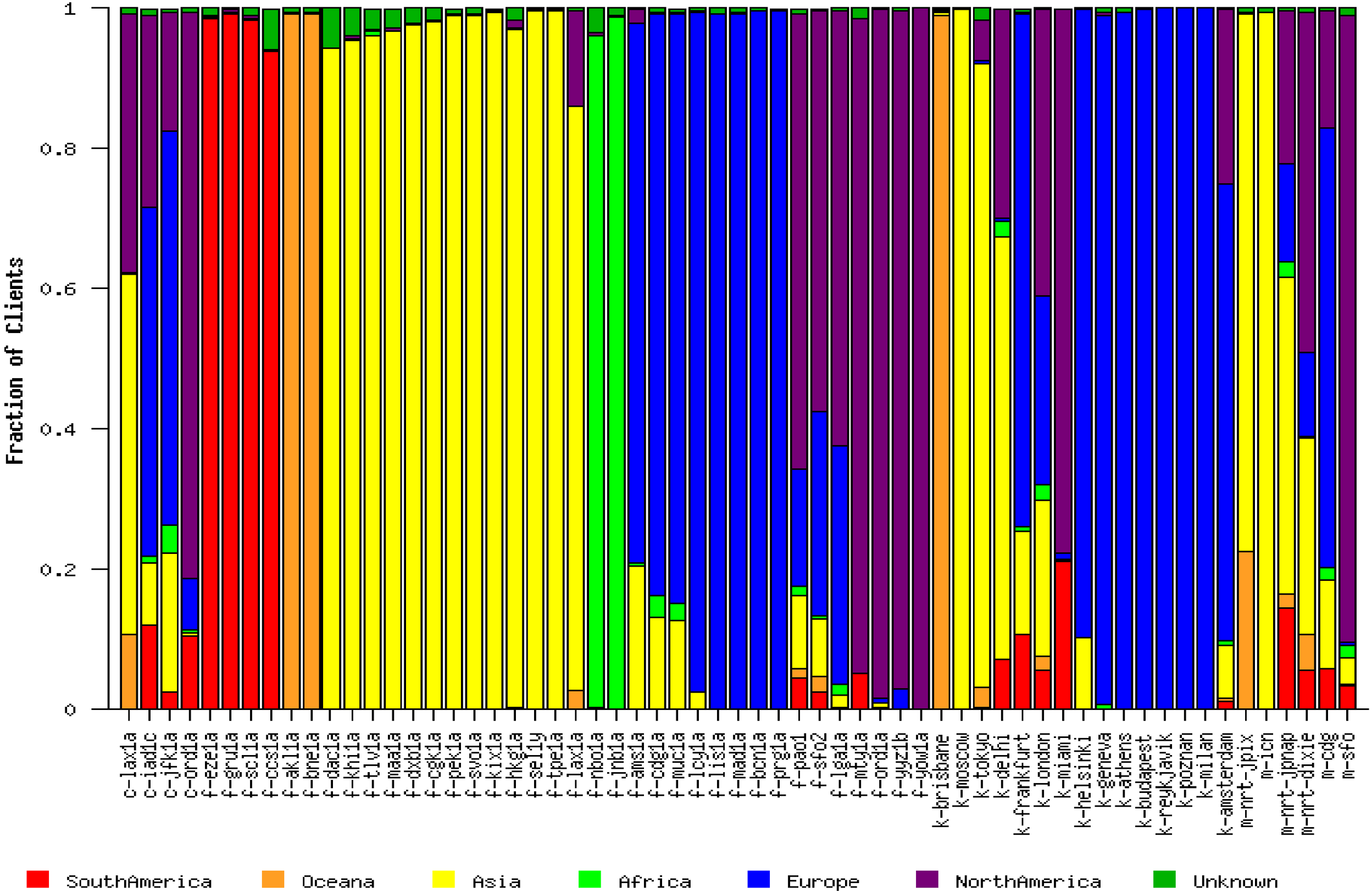
Internet measurement data catalog

- First catalog to support indexing and user annotations of Internet measurement data sets.
- DatCat: (http://imdc.datcat.org)
  - facilitates searching for and sharing data among researchers,
  - enhances documentation of datasets via a public annotation system, and
  - advances network science by promoting reproducible research and persistent references.

“Day in the life of the Internet”

- initially, NSF/DNS OARC project to measure as many root servers as possible over same interval
- DNS roots (c,f: all; k: 17; m 3; b.osrn, m.orsn)
- collaborators: ISC, NAMES, ORSN, Kaist, Postech, WIDE, Keio (all non-US)
- data from: passive hosts, topology probes
- genuine success blocked on policy changes

“DITL07: dns query sources by region”



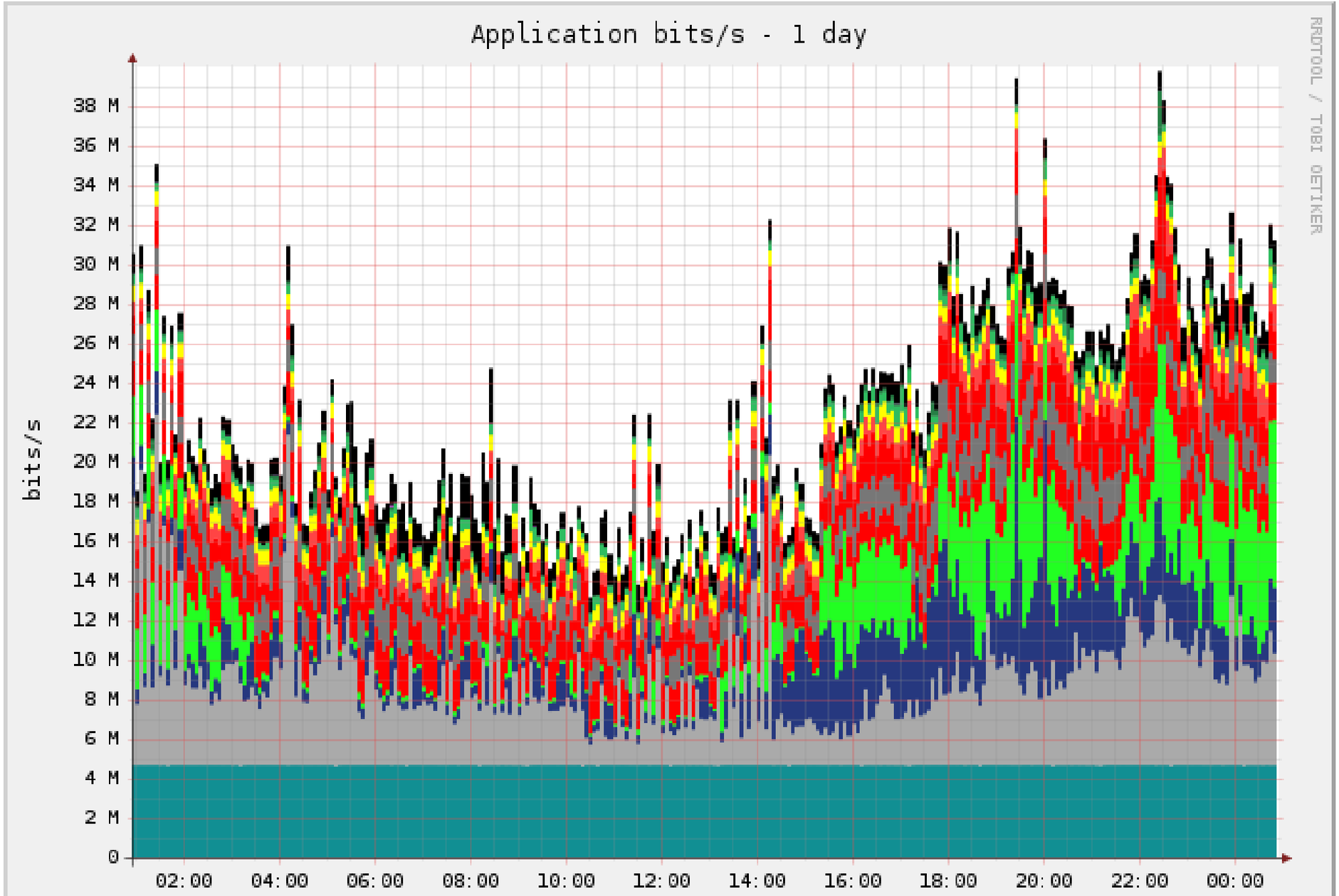
data sharing: muni/community wireless

- recognition that commercial data sharing poses persistently daunting if not insurmountable challenges
  - which means science is largely stalled
- COMMONS: cooperative measurement and modeling of open networked systems
  - "bandwidth for data trade" backbone bandwidth surplus in academic community, data dearth
  - valid on int'l, national, state or local level
- also essentially blocked on policy

Traffic measurement software: CoralReef

- collects and analyze data from passive Internet traffic monitors, in real time or from trace files.
- programming APIs for C, Perl; applications for capture, analysis, and web report generation.
- CAIDA developers maintain with help from Internet measurement community.

- *http://www.caida.org/tools/measurement/coralreef/*



September 03 2007 - September 04 2007

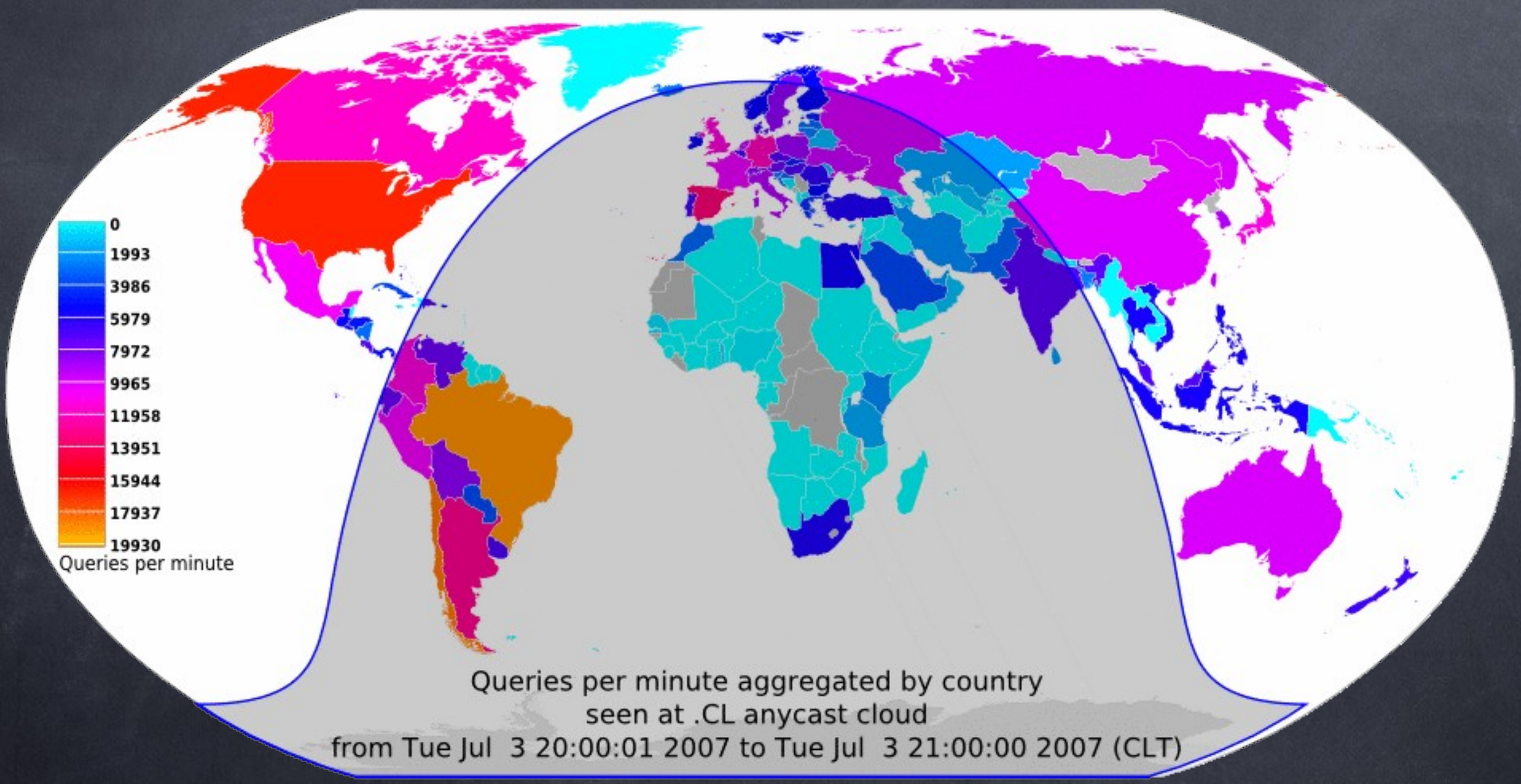
Application	Min	Avg	Max
RCMI_CIAR	4.63M	4.69M	4.70M
UNKNOWN_TCP	1.01M	4.44M	17.74M
IPSEC	344.61k	2.28M	13.08M
SSH	93.50k	2.24M	9.68M
UNIDATA_LDM	1.48M	1.94M	2.39M
UNKNOWN_UDP	916.48k	1.67M	3.05M
HTTP	277.94k	1.58M	5.14M
HTTPS	546.36k	889.34k	1.81M
YAHOO_MESSENGER	792.46k	796.90k	801.28k
SQUID	156.45k	507.31k	1.46M
SMTP	7.35k	135.22k	2.27M
RSYNC	72.05	1.41k	98.64k
other	440.67k	1.20M	3.57M

generated Mon Sep 3 23:45:40 2007 UTC

dns data: measurement & analysis

- analysis of root server traffic
- dns surveys (open resolvers, cache poisoning, server software)
- anycast measurement, modeling & simulation
- dns measurement software: dsc
- collaboration with NIC chile on measuring .cl
- http://www.caida.org/research/dns/

dns traffic: queries to .cl server



active measurement: ***archipelago*** (ark)

- CAIDA's new measurement infrastructure
- 'operating system' for measurement
- launch 12 Sept 2007
- 28 active probers
- 5 are IPv6-capable
  - collaborators can run vetted measurements on security-hardened platform through simple API
  - general public can perform restricted measurements
  - support for meta-data mgt, analysis, and infoviz

Connect with SA requirements

What makes ark an ‘operating system’ for (active) measurement?

- **Archipelago** provides a unique enabling infrastructure, featuring the Miranda tuple space, that allows researchers to quickly design, implement, and easily coordinate the execution of experiments across a widely distributed set of dedicated resources (monitors). Ark coordination facilities also enable ease of data transfer, indexing, and archival.

Enabling Internet science: example

- Need: What probing method discovers most topology?
  - Do per-flow load balancers implement different forwarding policies for TCP and UDP? Need experiment!
- Approach: Archipelago measurement platform
- Matthew Luckie, Young Hyun, and Bradley Huffaker, “Traceroute Probe Method and Forward IP Path Inference”, IMC 2008.
  - ICMP-based traceroute methods tend to successfully reach more destinations, as well as collect evidence of a greater number of AS links.
  - UDP-based methods infer the most IP links, despite reaching the fewest destinations.

What makes ark a breakthrough?

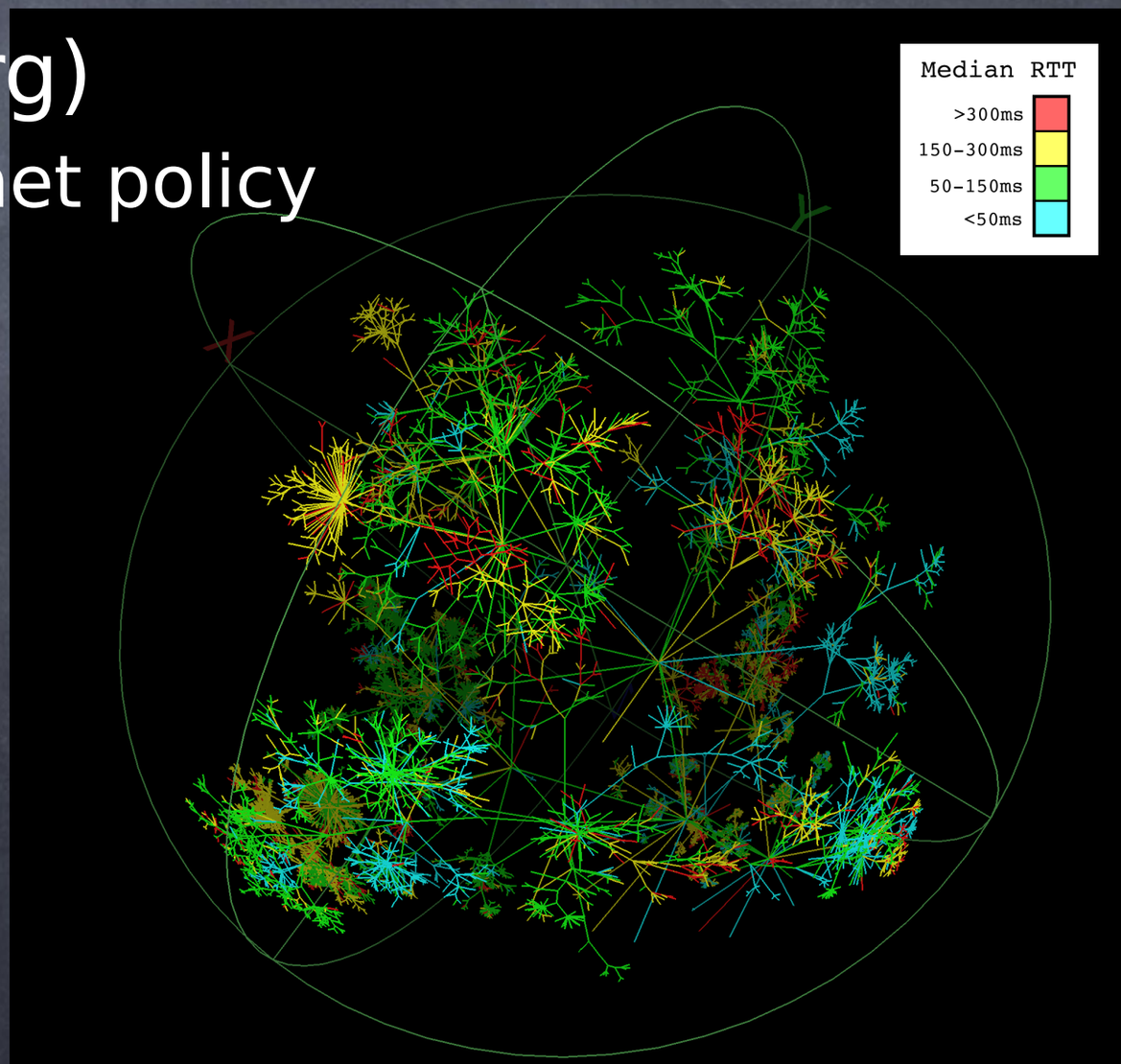
- Benefits:
  - Ease of experiment design, implementation, and coordination.
  - Dedicated resources (monitors).
  - No restrictive intellectual property.
  - Multiple levels of trust and access
- Competition:
  - Other platforms:
    - do not provide dedicated resources.
    - cannot guarantee the veracity of the collected data.
    - lack fine granularity access control
  - Other data collected on these platforms suffer the constraints of the underlying platform.

CAIDA's Internet mapping with ark

- Integrate 6 strategic measurement & analysis capabilities for dhs 'situational awareness' needs:
  - new architecture for continuous topology measurements,
  - IP alias resolution techniques,
  - dual router- and AS-level graphs,
  - AS taxonomy and relationships,
  - geolocation of IP resources, and
  - graph visualization.

ark: profound insights enabled

- incongruity between topology and routing system
  - topology evolving away from what routing system needs
  - radical implication for future of the Internet (IP)
- concentration of ISP ownership (as-rank.caida.org)
  - Inform communications, Internet policy
- inconsistencies between topology and routing data
  - still no guaranteed way to capture Internet topology
  - blocked on policy (really!)
  - but some methods are better than others, e.g., ICMP



ark platform: value summary

- improve critical national capabilities:
  - situational awareness for homeland security purposes
  - richly annotated maps of the Internet to support better understanding of infrastructure for national security
  - internet measurement, analysis and inference techniques
  - empirical basis for federal communications policy
- help address network science crisis
  - scalability in system management, monitor deployment, measurement efficiency, resource utilization
  - flexibility in measurement method, scheduling, data collection
  - let researchers spend less time on non-research

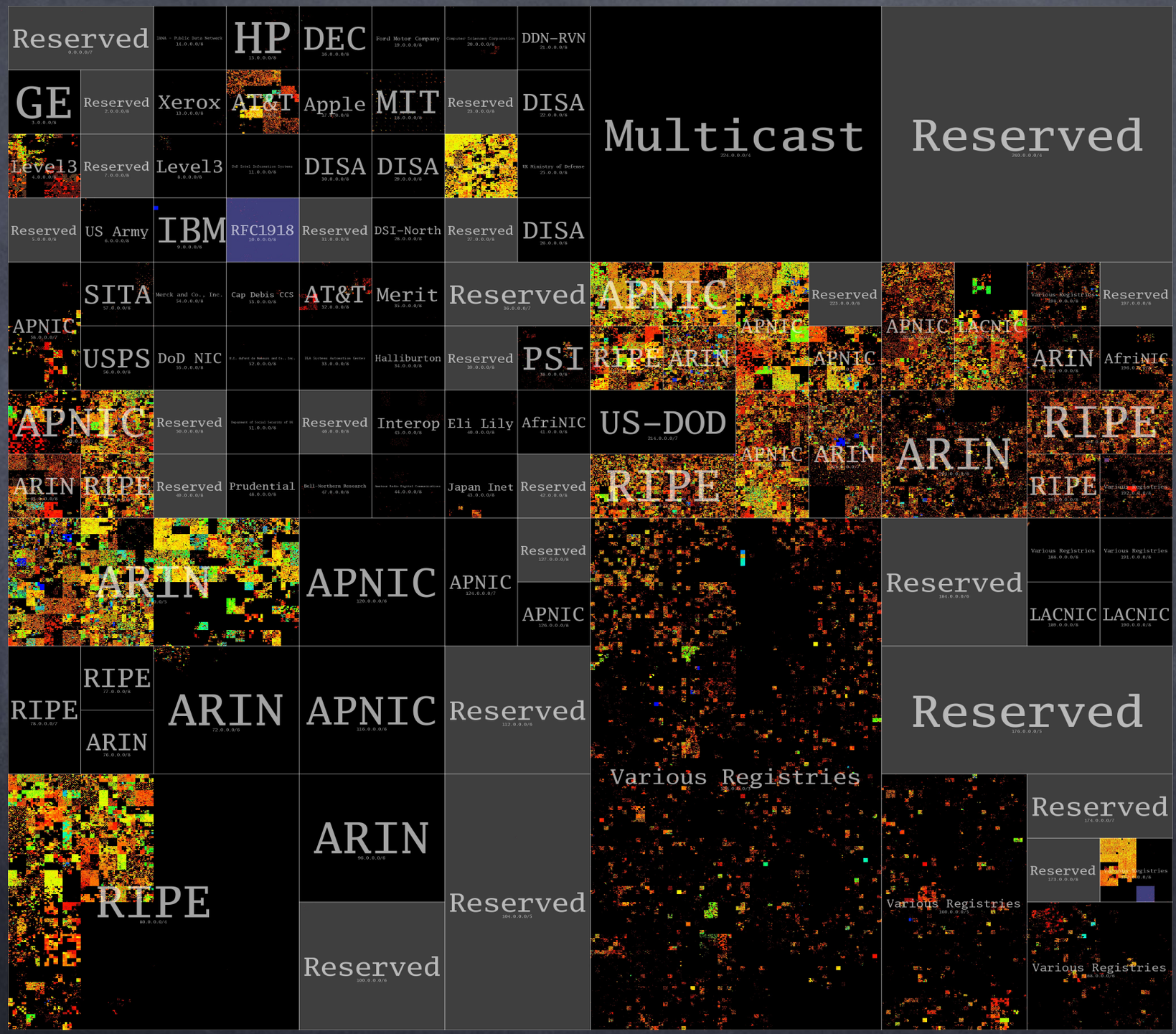
Theory: network topology calculus

- set of metrics that define all others, even future ones
- end few-decade quest of network and graph theorists
- provide framework to formalize correlations in graph structure (local and global properties)
- topology generator capable of constructing graphs that reproduce properties of target (observed) topologies with arbitrary degree of accuracy
- applies to variety of scientific research including, e.g., cs (systems and theory), physics, biology, social, economic, political sciences
- adding link and node semantics (annotations, e.g., bandwidth. validation blocked on policy)

Theory: ISP topology evolution

- first Internet growth model to combine following three properties:
  - realistic – based on a formalization of Internet economics,
  - analytically tractable, and
  - all its external parameters are measurable.
- prediction of value of exponent of the power-law node degree distribution sheds new light on an area that intersects with many different complex networks.
- other contributions of the model include:
  - an attempt to derive standard, i.e., linear, preferential attachment from economic realities of the Internet
  - predicts that topological awareness of Internet players leads to super-linear preference and consequently to deviations from power laws to condensed monopoly states.

Internet policy: address exhaustion



[ping data from isi.edu; poster by Duane Wessels@TMF]

CAIDA: summary of goals

Since 1997: narrowing the gap between Internet operations and science in face of global privatization

Seek, analyze, communicate salient features of best available data on the Internet

Forward-looking architectural research

Navigate data-sharing challenges, by lowering technology barriers

Support empirical needs of public sector