Internet measurement: what have we learned?

kc claffy
kc@caida.org
7 dec 07

scope of field

- **workload**
- **topology**
- **routing**
- **performance**
- **security**
- **geolocation**

also:
standards,
software,
storage,
statistics.
and recently,
lawyers.

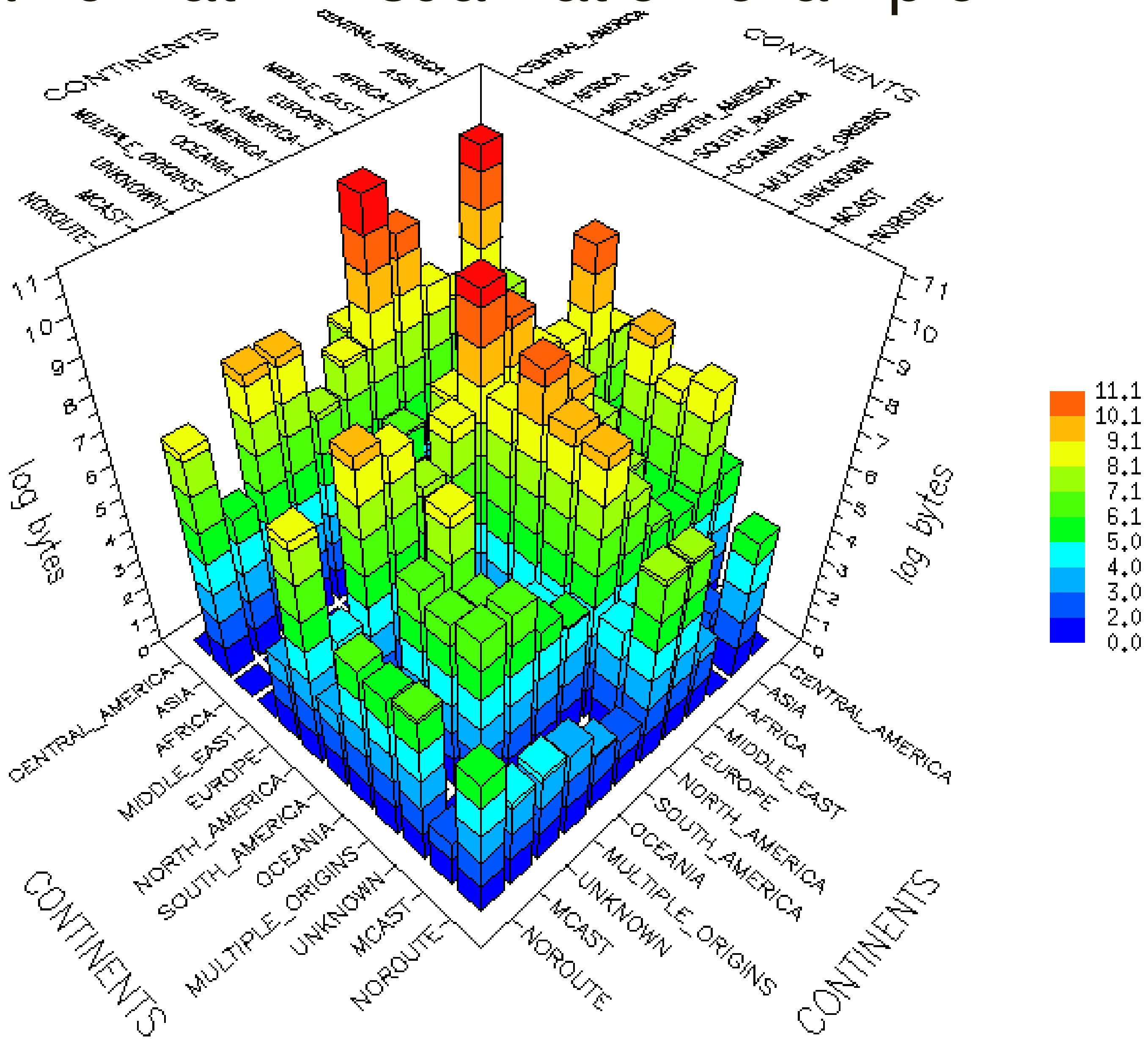
workload characterization & modeling

- traffic matrix inference (on small scale..we think)
- cross-section of core (failure, but lesson)
- intelligent sampling
- anonymization methods

none generally implemented by vendors

intellectual achievements

traffic matrix visualization example



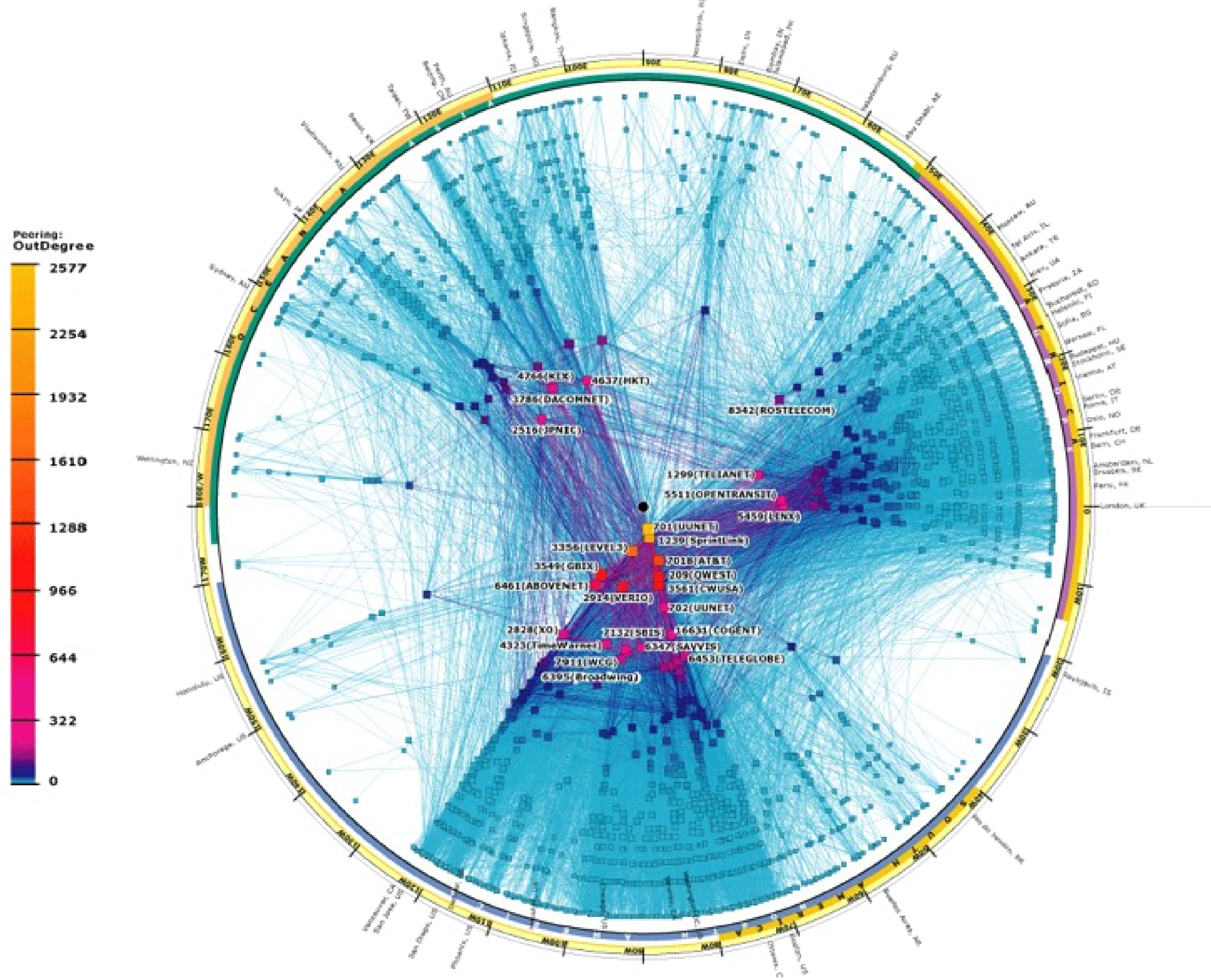
workload characterization & modeling

- flow menagerie (traffic engineering challenge)
- relentless growth in p2p, spam, worms, viruses (faster than traffic)
- critical infrastructure (dns roots) sees much (up to 95% of traffic) pollution
- people use connectivity once there (.jp study)

topology structure and dynamics

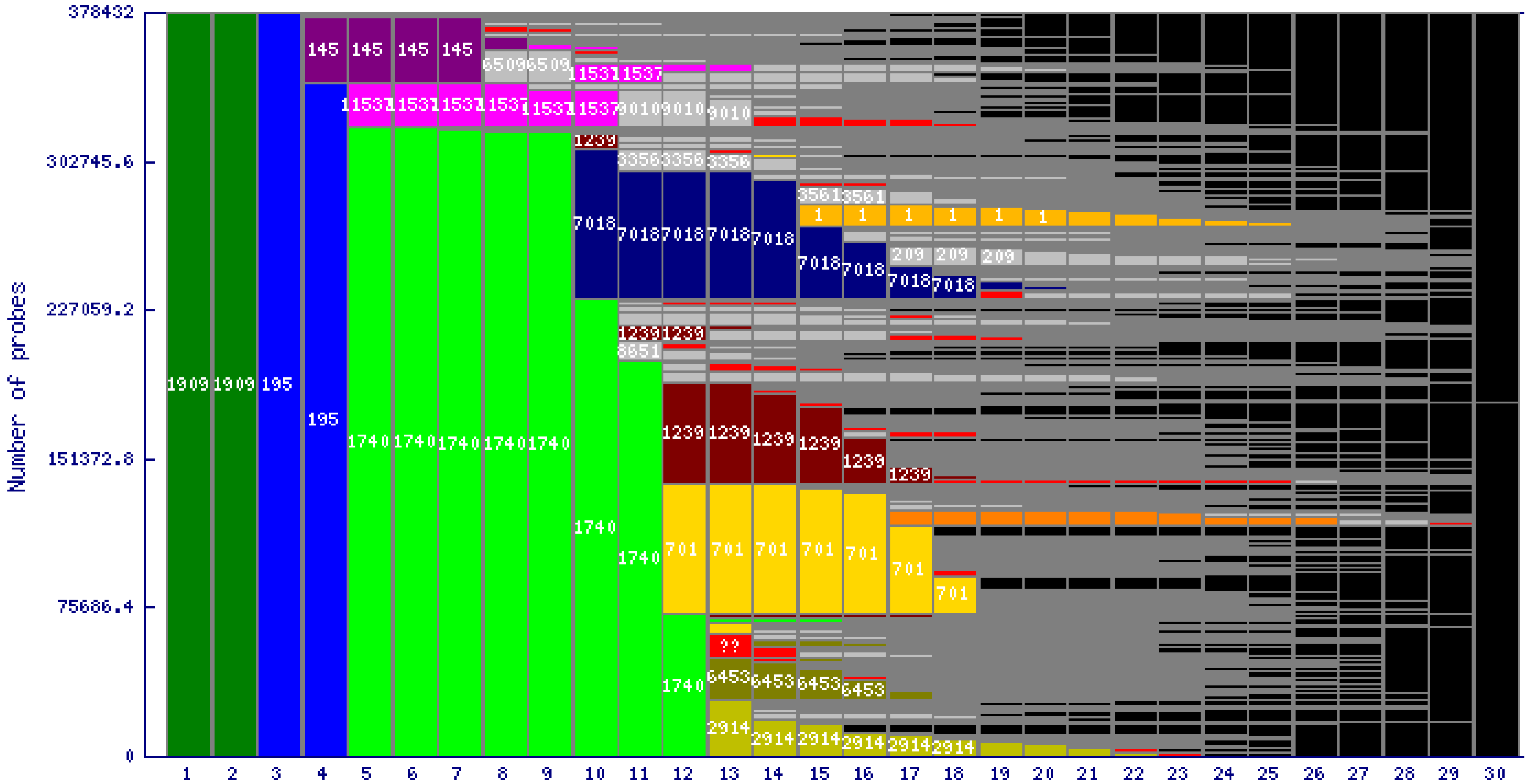
- not just random (see google) -- degree variability higher than expected.
- power laws abound?
- small distance distributions implies current (& proposed) routing architectures inherently poor fit

intellectual achievements



AS topology structure

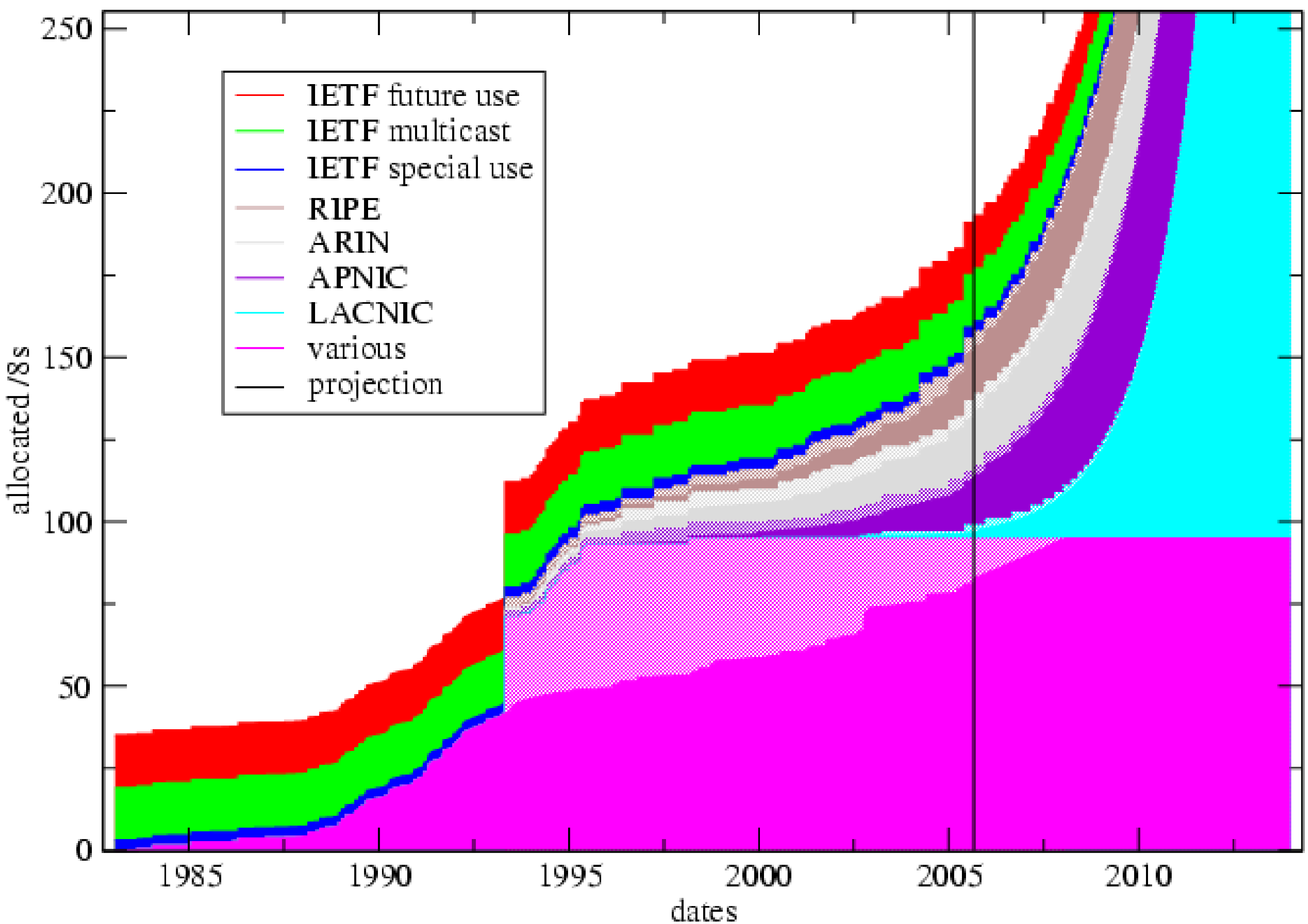
topology structure and dynamics



AS dispersion from single source/many dests

IPv4 allocated /8s (first)

RIR whois dumps and IANA table of top-level /8 allocations



routing

- BGP has inherently non-deterministic features (MEDs)
- oscillations observed, but if we follow simple rules, we can achieve stability. but no way to enforce simple rules.
- discovery: observed evolving topology diverging from current (and proposed) routing system.

recognized need for new routing architecture

(and yet no concerted effort)

performance

- ECN, RED, CBQ: developed, not deployed
- bandwidth estimation: failed at per-link, can do limited per-path, not deployed
- systems integration complexity hinders validation (instead we have keynote, internetweather, akamai, corporate SLAs

daunting place to do science

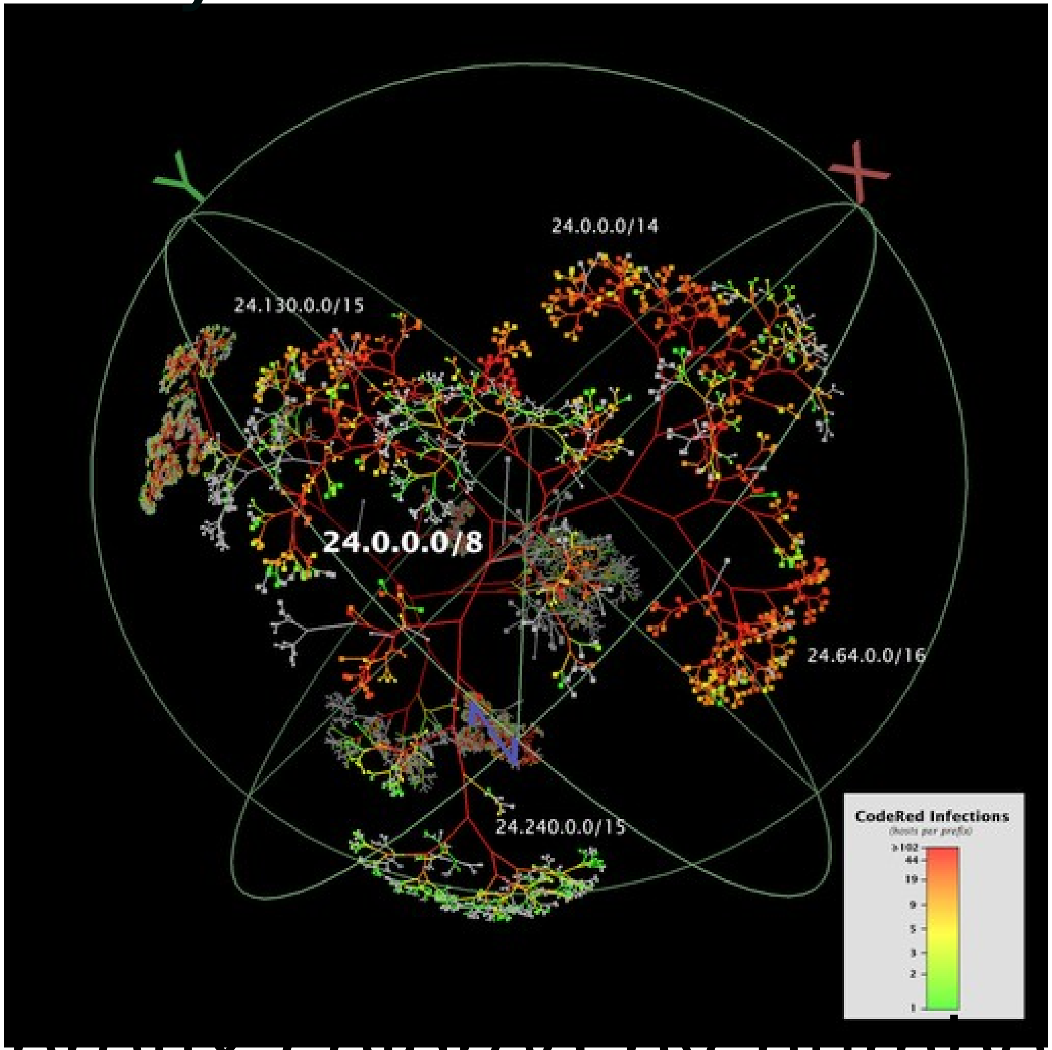
(don’t know congestion locations, lengths, or causes)

security

- detection & mitigation of specific (similar) threats
- worm propagation models, intrusion detection tools, startups
- discovery: patching model a failure
- discovery: monoculture a failure
- discovery: can't quarantine networks fast enough
- discovery: correlated attacks (e.g., botnets) prevalent
- discovery: little ingress filtering; vulnerable DNS resolvers

hard to measure progress, scope of attacks & number of vulnerabilities guarantees thriving industry w or w/o science

security: visualization example



- prefix colored by number of infected hosts

notable achievements under circumstances

for U.S. inter-domain internet science, the crash happened in 1994 when the nsfnet retired...

. can't figure out where an IP address is
. can't measure topology effectively in either direction, at any layer
. can't track propagation of a routing update across the Internet.
. can't get router to give you all available routes, just best routes
. can't get precise one-way delay from two places on the Internet
. can't get an hour of packets from the core
. can't get accurate flow counts from the core
. can't get anything from the core [used to have anonymized traces]
. can't get topology of core
. can't get accurate bandwidth or capacity info
    not even along a path, much less per link
. can't trust whois registry data
. no general tool for `what's causing my problem now?'
. privacy/legal issues deter research (was hard w enlight'd monarchy)

science abysmal, discouraging to remaining academics

NAS report on ‘network science’

1) networks are everywhere and thus important

2) we don't yet have any predictive power over complex networks

3) funding situation backwards: domain-specific (splintered) rather than fundamental

http://fermat.nap.edu/books/0309100267/

NAS report on ‘network science’

identifies as top three challenges:

1) characterization of dynamics and information flow in networked systems

2) modeling, analysis, & acquisition of data for extremely large networks

3) rigorous tools for the design and synthesis of robust, large-scale networks

http://fermat.nap.edu/books/0309100267/html

jarring observation from history of science

The modern field of elementary particle physics depended crucially on the establishment of a huge volume of data gathered mainly in the period 1945-65. Only then was it possible for the synthesis of the Standard Model to take place, 1967-74.

-- Peter Galison, Professor of History of Science and Physics, Harvard

(unfortunately, we're not doing research,
we're building critical infrastructure.
and it's riddled with structural problems.)

DatCat

Internet Measurement Data Catalog http://datcat.org/

- To facilitate searching for and sharing of data

  Index as much as possible, including datasets not publicly available. DatCat doesn't store any network data itself

- To enhance documentation of datasets via public annotations

  Easy for anyone (not just dataset creator) to annotate

- To advance network science by promoting reproducibility

  Paper X ran their detection algorithm on dataset X and had a false positive rate of 0.2. Using our algorithm on dataset Y, we get a false positive rate of 0.1. Therefore our algorithm is better. …

Persistent handles to allow for consistent citing and comparison:

http://imdc.datcat.org/collection/1-003M-5=AOL+500k+User+ Session+ Collection

broader impact

- what has happened to the Internet since the NSF transitioned it to the private sector “(commercialization and privatization”)?
- what false assumptions do we carry?
- for remaining problems, what prevents progress?
- how can we move forward?

16 operational internet problems

- security
- authentication
- spam
- scalable configuration management
- robust scalability of routing system
- compromise of e2e principle
- dumb network
- measurement
- patch management
- “normal accidents”
- growth trends in traffic and user expectations
- time management and prioritization of tasks
- stewardship vs governance
- intellectual property and digital rights
- interdomain qos/emergency services
- inter-provider vendor/business coordination

persistently unsolved problems for 10+ years

why we’re not making progress

- if providers are broke, they can’t invest in long-term health of infrastructure.
- so add to list of problems: **sustainability**
- top unsolved problems in internet operations and engineering are rooted in **economics, ownership, and trust (EOT).**

does not mean there aren’t useful technical problems to study. but there will be no technical solutions to these problems that don’t solve the EOT issues.

historical context

1966: Larry Roberts, “Towards a Cooperative Network of Time-Shared Computers” (first ARPANET plan)

(*we are still using the same stuff*)

1969: ARPANET commissioned by DoD for research

1977: Kleinrock’s paper “Hierarchical Routing for large networks; performance evaluation and optimization”

(we are still using the same stuff)

1980: ARPANET grinds to complete halt due to (statusmsg) virus

1986: NSFNET backbone, 56Kbps. NSF-funded regionals.
IETF, IRTF. MX records (NAT for mail)

1991: CIX, NSFNET upgrades to T3, allows .com. web. PGP.

1995: under pressure from USG, NSF transitions backbone to competitive market. no consideration of economics or security. kc proposes caida.org

2005: *Economist* cover: “*How the Internet killed the phone business*” (Sept)

what have we done?

we replaced a critical infrastructure with something not designed to be critical infrastructure

historical context explains it but does not address incongruities

and this decade, free markets go up against free speech

what have we learned?

most important thing we've learn so far: society has decided IP is like water.

"our best success was not computing, but hooking people together" --david clark, 1992 ietfplenary

strong implications for an industry structuring itself to sell wine. but that's what the data shows.

when you want to move water, you care about 4 things: safe, scalable, sustainable, stewardship.

the 4 S's

- safety: is the data toxic upon arrival?
- scalable: can we route/name/address earth's needs?
- sustainable: is it economically viable?
- stewardship: will the provisioning and legal frameworks we choose leave our children -- and democracies -- better or worse off?

none are purely technical, but all require technical understanding to get right.

and they're all connected.

how have we done?

- how safe is the Internet?
  - data doesn’t look good
- how scalable is the Internet?
  - data doesn’t look good
- how sustainable is the Internet?
  - data doesn’t look good
- how did we do on stewardship?
  - data doesn’t look good

there is good news

- we made something so great, everyone wants it.
- in fact many of us want it more than once! (um..)
- the current industry is a historical artifact of technical and (science & regulatory) policy ‘innovations’ in the 60s, 70s, 80s, 90s, and 00s
- people are starting to study interplay, but they’re undercapitalized
- in the meantime, it became global critical infrastructure. oops.

cataloguing lessons

- although the Internet has over-achieved on plenty, it has underachieved on: security, scalability, sustainability, and stewardship. substantial oversights.
- our ability to measure is surprisingly abysmal, although policy history explains
- cooperative, data-sharing approaches key to moving forward

we have learned more from our failures than from our successes...

measurement accuracy is the only fail-safe means of distinguishing what is true from what one imagines, and even of defining what true means.

..this simple idea captures the essence of the physicist's mind and explains why they are always so obsessed with mathematics and numbers: through precision, one exposes falsehood.

a subtle but inevitable consequence of this attitude is that truth and measurement technology are inextricably linked.

-- robert b laughlin, a different universe,

caida recent activities

- data sharing for reproducible research (datcat, PREDICT, Day in the Life of the Internet (2008 data available)
- hardware and software upgrades
- dns traffic and vulnerability analyses (/research/dns/)
- topology measurement, curation, analyses (as-relations, as-rank), modeling (dk-series), simulation
- next generation Internet routing architectures
- security: network telescope, cceid
- community and muni network support: commons
- policy guidance e.g., ipv4 consumption, blog.caida.org