Internet measurement: state of DeUnion

the so-called science of poll-taking is not
a science at all but a mere necromancy.
people are unpredictable by nature,
& though you can take a nation's pulse,
you can't be sure that the nation
hasn't just run up a flight of stairs.
--e.b.white, New Yorker, Nov 1948.

2 apr 01
kc claffy, UCSD/SDSC/CAIDA
kc@caida.org
www.caida.org

abysmal but unsurprising

- *little capacity to predict, depict, or even measure traffic behavior on current and advanced networks*
- *few tools to engineer/operate networks or identify traffic anomalies in real time*
- *doesn't stop researchers from building junk*
- *doesn't stop random users from doing random junk (no dearth of activity)*
- *increasing risk to infrastructure*

Internet's resistance to modeling/measurement

evolution-based (good!) reasons

- protocols, technologies, applications
  - independently developed and deployed
  - by no means synergistic
  - by all accounts rapid
  - `punctuated' but no equilibrium
  - "have done fine without modeling so far"
  - (let's wait till modeling cheaper than bandwidth)

but simulation/analysis validation (& lately other stuff) needs data

- right granularities hard to come by
- measurement technology just not there
- argument for it also not there
- "helps everyone", but who pays?
- losing battle?

Internet's resistance to measurement

many would benefit

- vendors, users, researchers, ISPs

ISPs would bear cost

- multiple media: atm, pos, dwdm, mpls
- logistics/management
- privacy implications
- analysis/research obsolete after (before) done

.....how to justify measurement??

one answer:

tools affecting ISPs financial bottom line

payoffs

- *insights for vendors re next generation hw/sw requirement*
- *calibration for users, e.g., monitoring service level agreements*
- *diagnostic and planning tools for ISPs*
- *windows into the infrastructure for researchers*

measurement tools lack

- well-defined traffic metrics
  - e.g supporting SLAs or billing
- uniformly applied methodologies
  - varied topologies, equipment, ISP practices
- clear definition of measurement hypotheses or goals
- measurement scalability
- ability to explain phenomena
  - topology changes, routing loops, black holes
- relevance to actual ISP problems or mechanisms for fixing
- communication of useful results

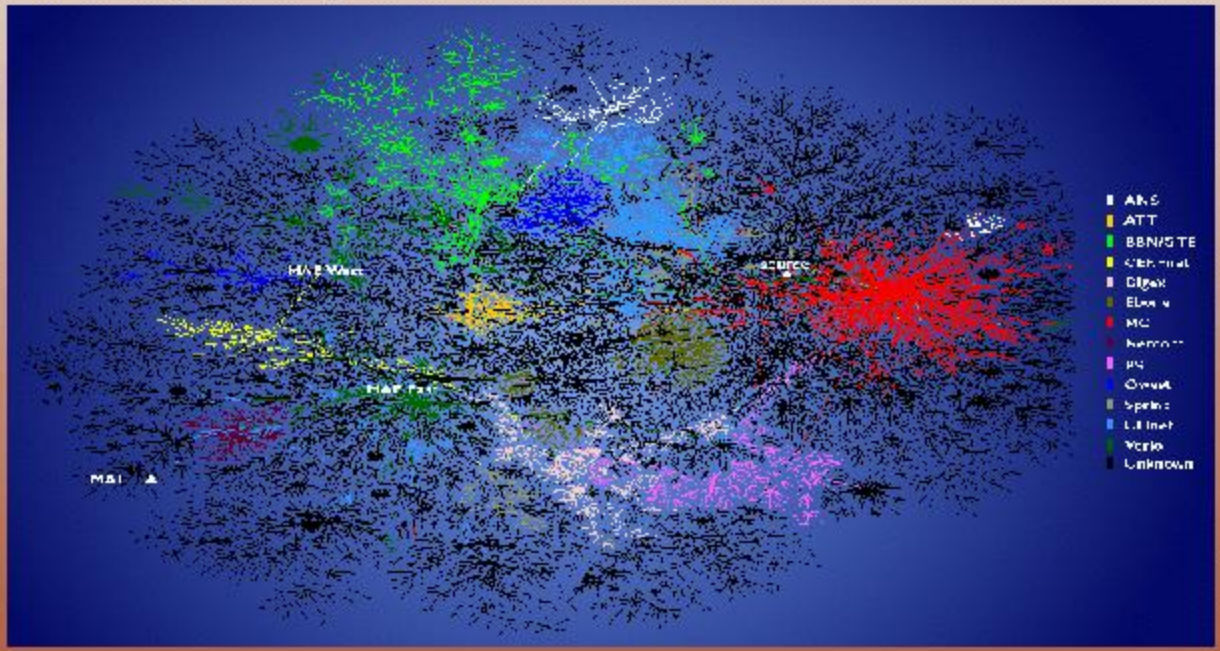
four areas of measurement

- topology (mapping)
- workload characterization (passive)
- performance evaluation (active, passive)
- routing (dynamics)

will show examples, priorities, obstacles

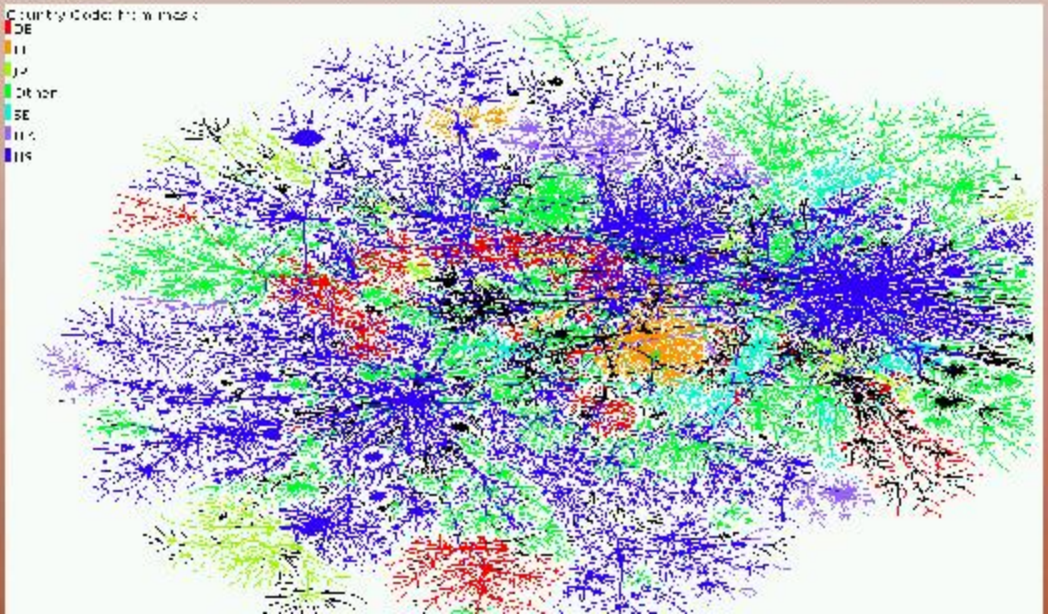
topology: skitter

- macroscopic, infrastructure-wide
- dynamically discover/depict topology (& b/w)
- correlate path perf. w events, e.g. BGP
- identify critical pieces of infrastructure



skitter: infrastructure-wide measurements

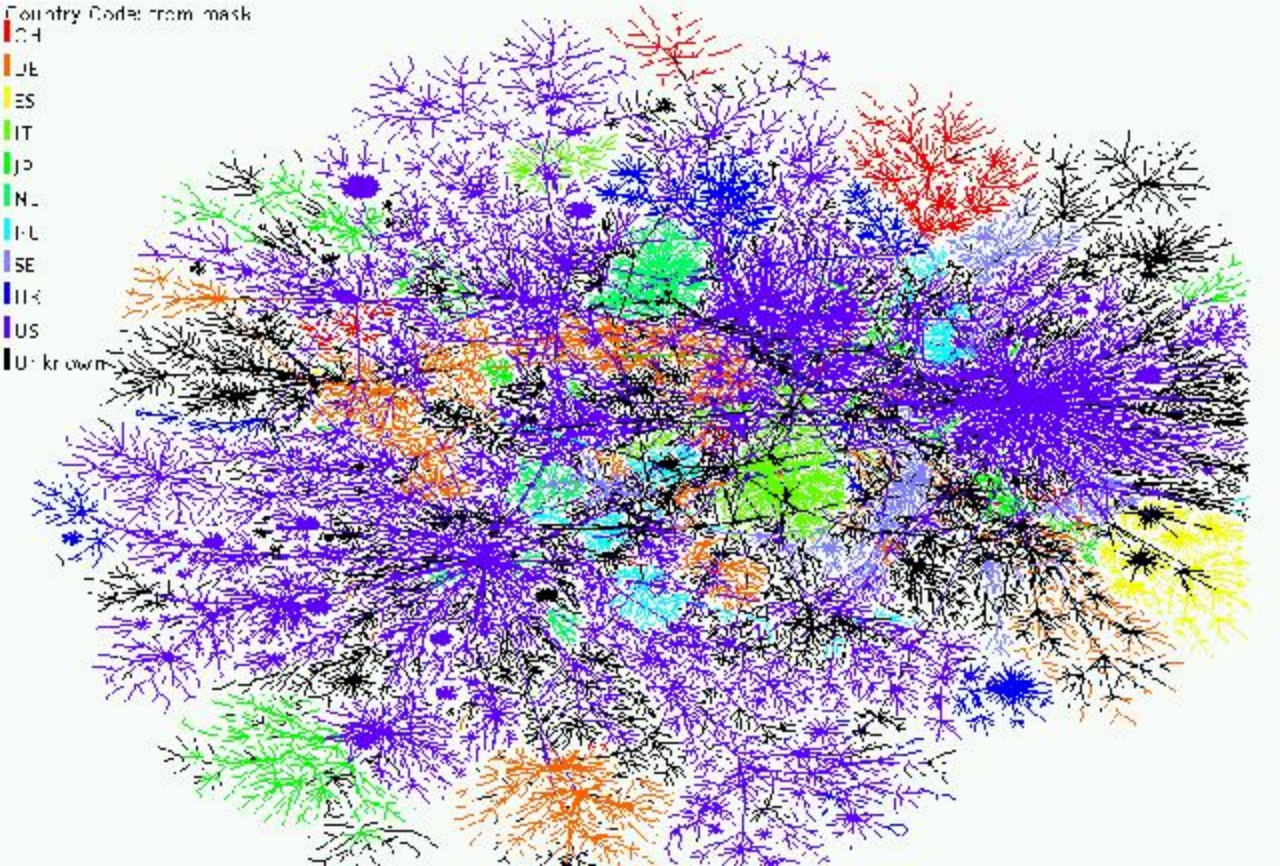
- *17 monitors (inc. 1 root name server)*
- *multiple dst lists (29k servers, 36k dns)*
- *architecture:*
  - *parallel ICMP probes*
  - *52-byte packets*
  - *kernel time stamping*
  - *ssh / Kerberos*



topology data: skitter

- O(20) sources around world
- 400K destinations (diff by box)
- almost 50% of prefixes covered (still building lists)
- ICMP echo request reply
- lightweight, coarse temporal granularity
- used for variety of macroscopic studies
- most comprehensive set in world (low bar...)
- available to researchers
- www.caida.org/tools/measurement/skitter/

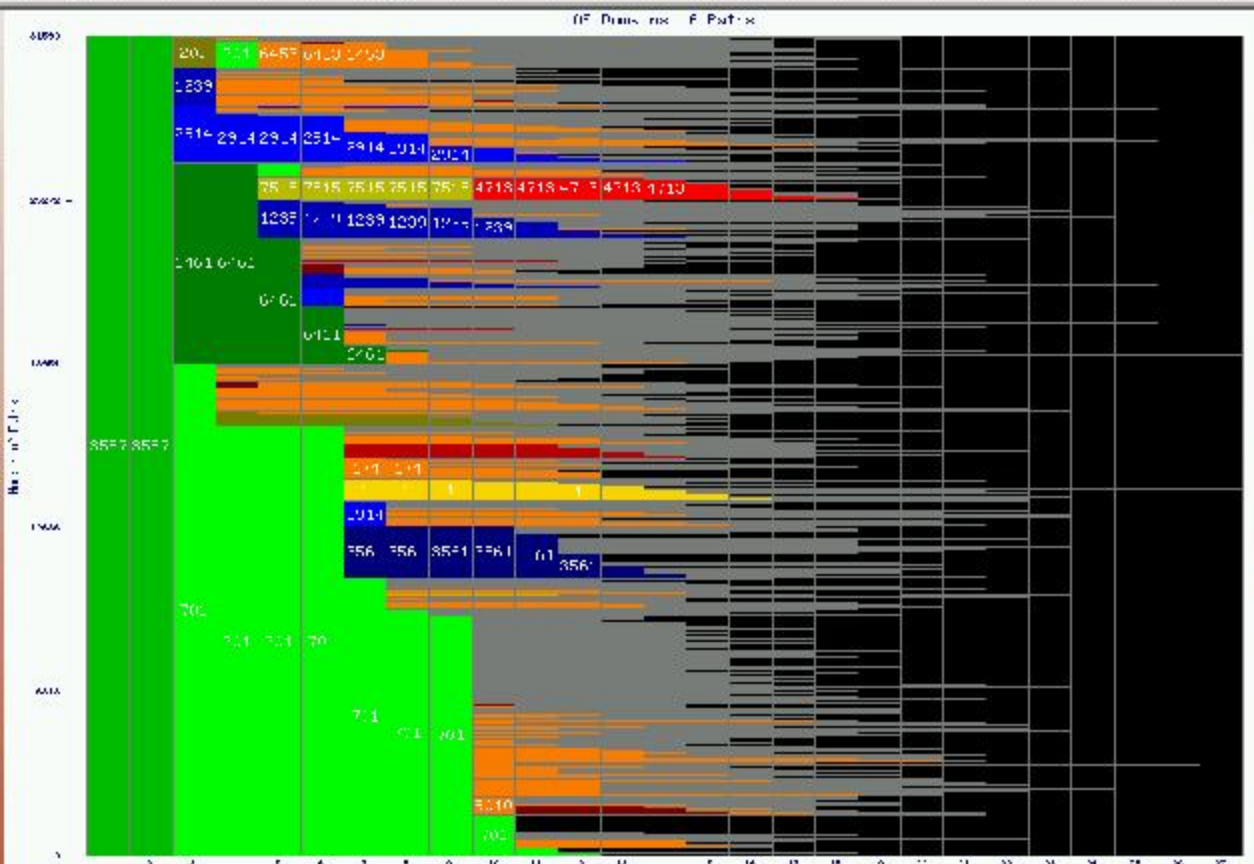
skitter: colored by more countries



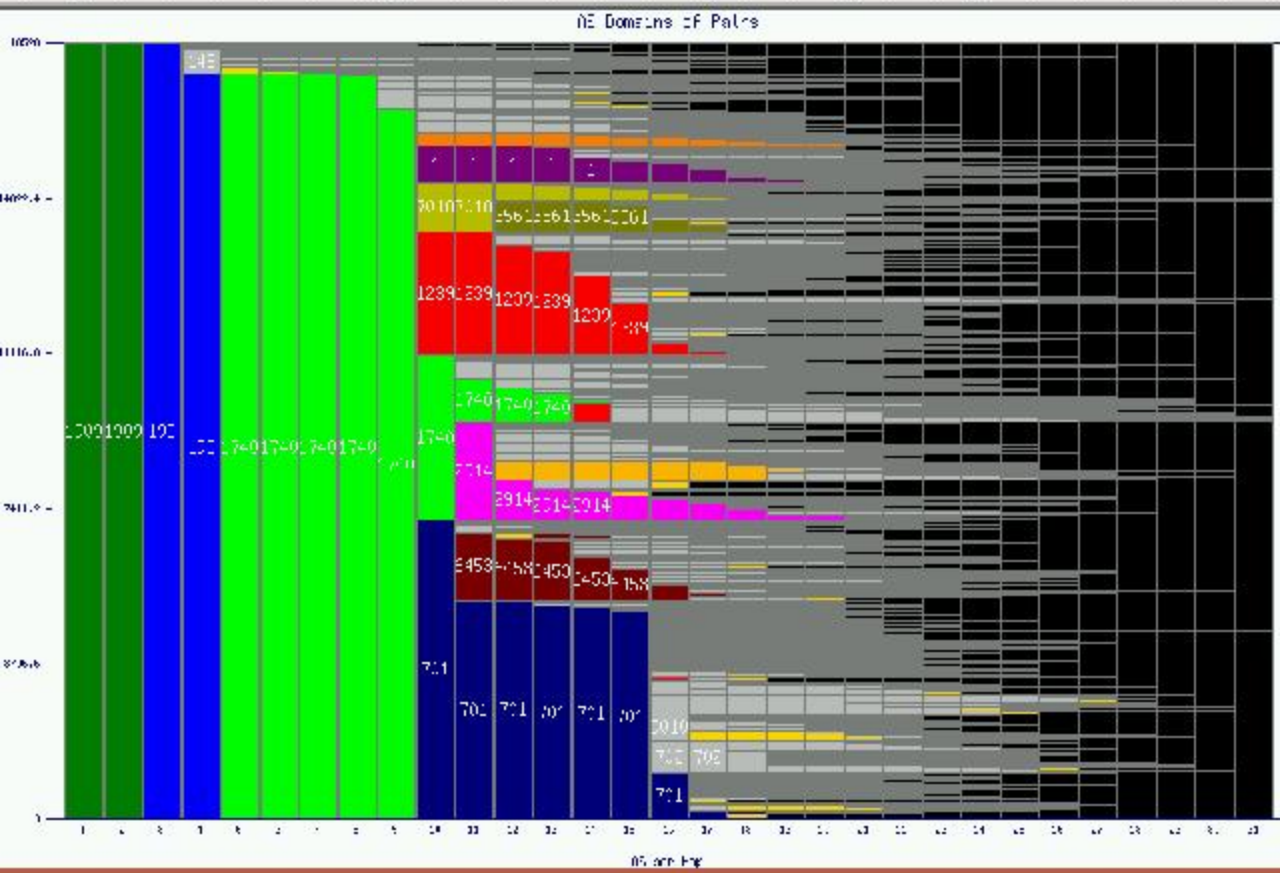
Internet topology graphs

- graph: set of nodes and edges/links
- directed edges
- infrastructural dispersion (AS, country)
- in- and outdegrees
- one- and two-way connectivity
- connected components
- shortest and longest paths
- combinatorial core
- giant component
- may not capture all properties
- correlate with routing (BGP) data later

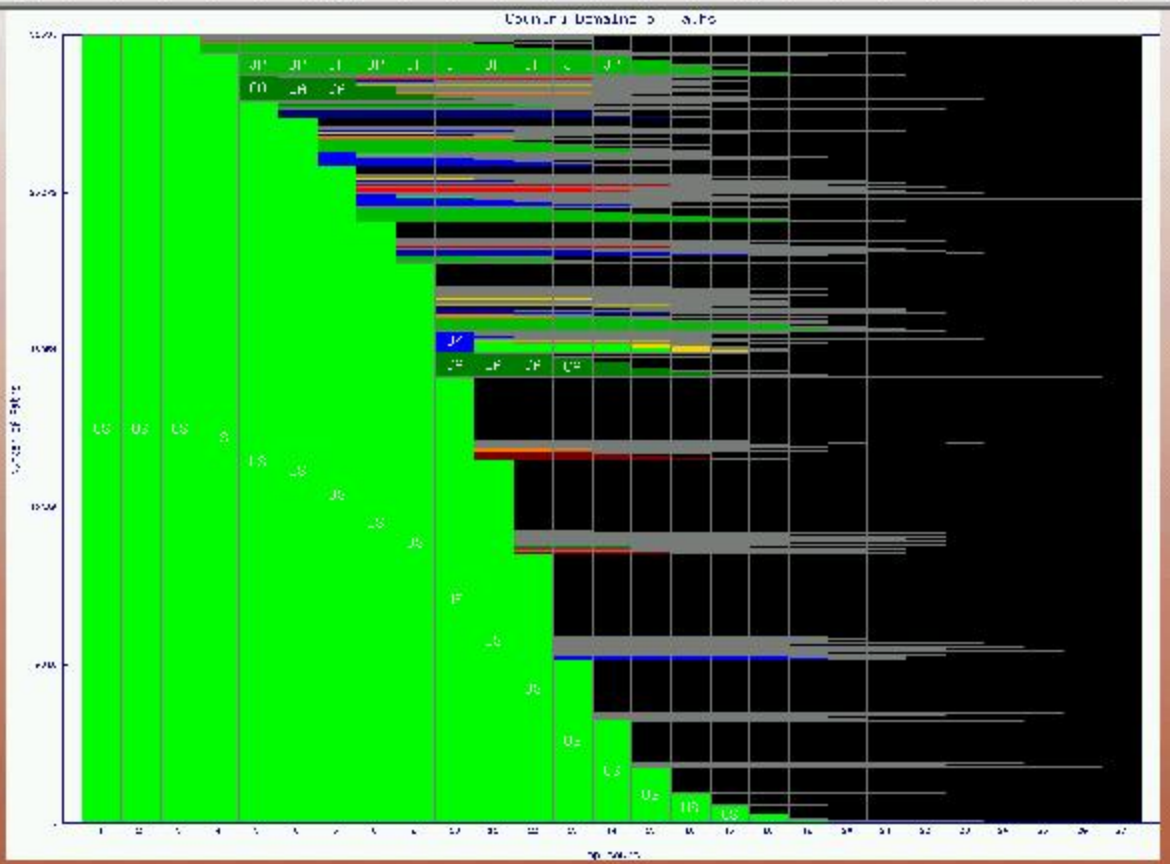
dispersion among ASes across paths



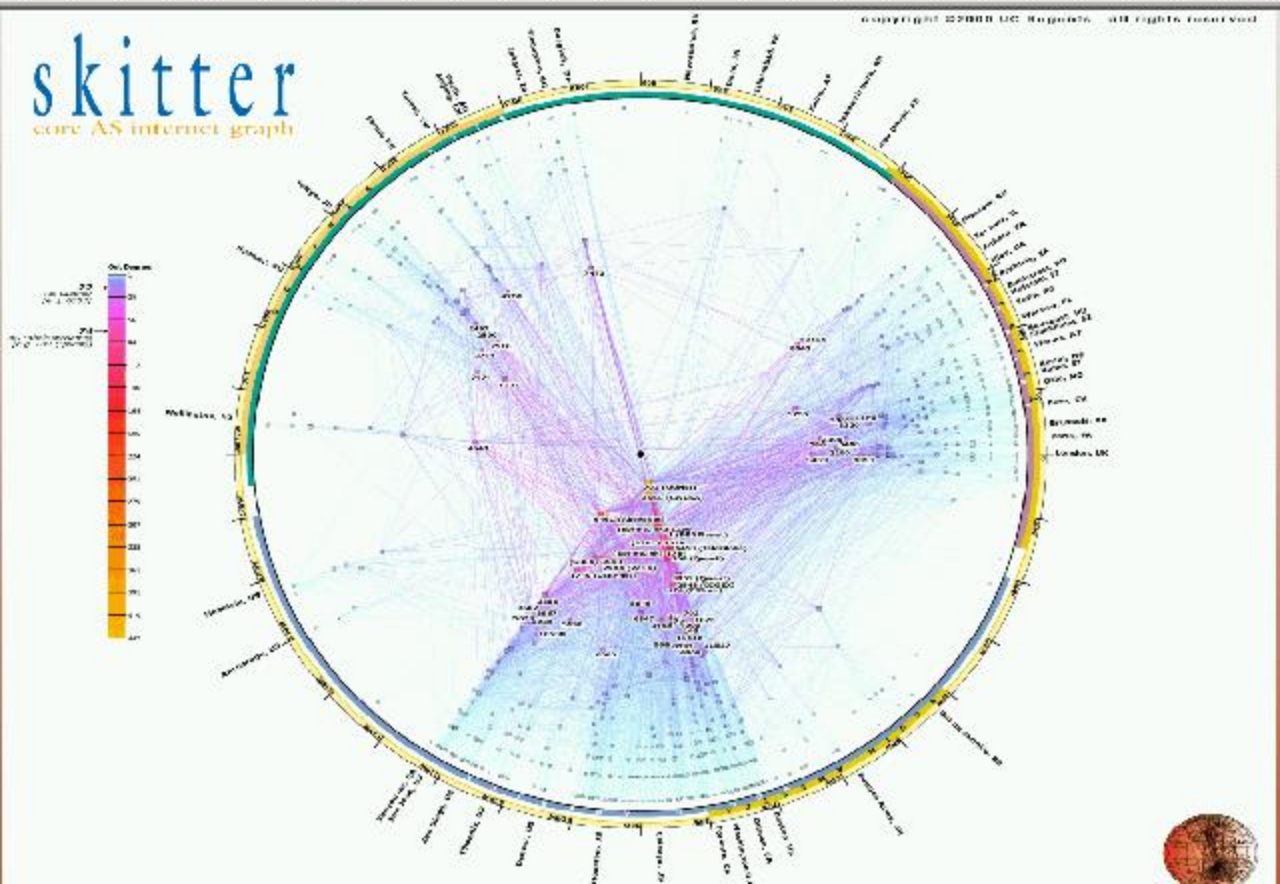
dispersion among ASes across paths (sdsc)



dispersion among countries across paths

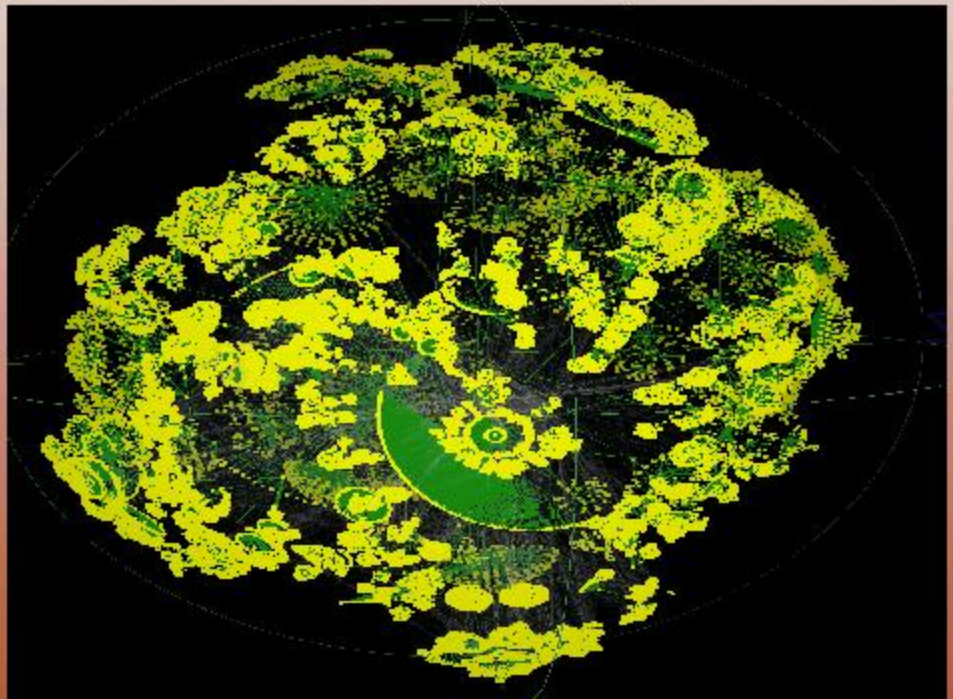


AS core peering richness visualization



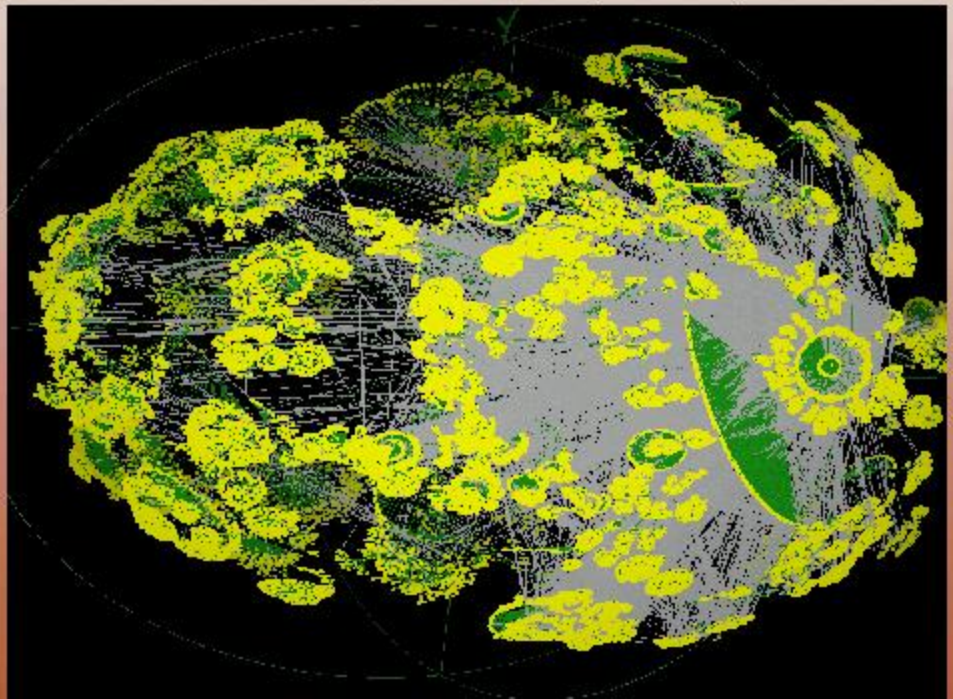
hyperbolic viewer (java 3D, 100,000s nodes)

- *from london skitter monitor in*
- *535,102 nodes, 535,101 tree links*
- *66,577 nontree links (transparent)*



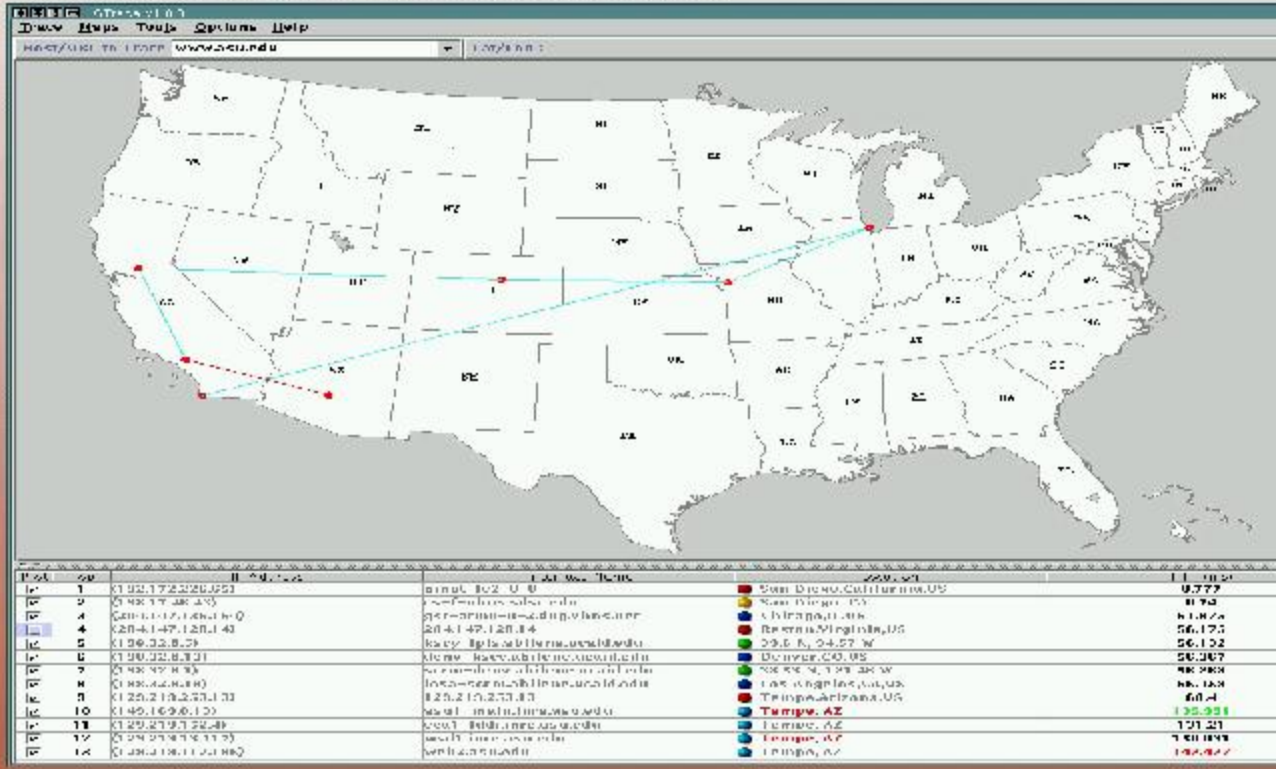
hyperbolic viewer (java 3D, 100,000s nodes)

- *from london skitter monitor in*
- *535,102 nodes, 535,101 tree links*
- *66,577 nontree links (less transparent)*



Routing: geographic traceroute

www.caida.org/tool/visualization//gtrace/



topology: research priorities

- visualization
  - latency
  - key routers/networks
  - AS granularity
  - geographic
  - integration w mgt tools

- obstacles:

mapping IP addresses to

  - router
  - geography
  - AS
  - service provider
  - country
  - anything...

route changes faster than can measure

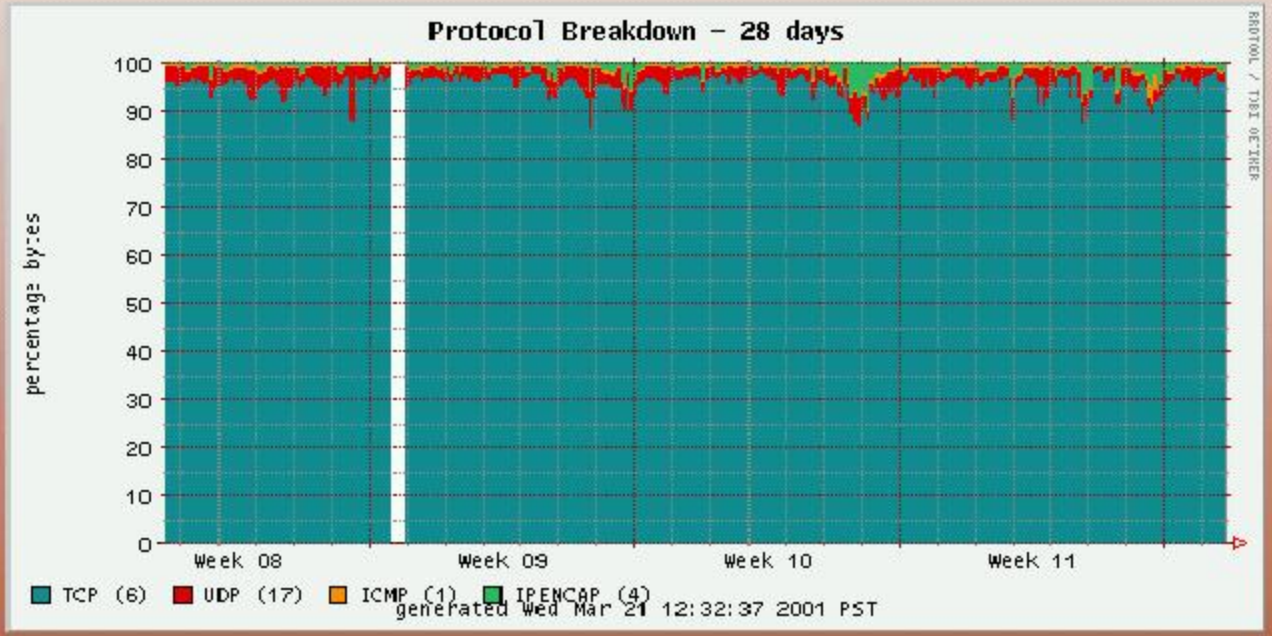
workload characterization

- workload profiling (s/w & h/w design, architecture optimizing, capacity planning
- security
- performance analysis
  - delay, loss, jitter?
- QOS assurance across ISPs
- accounting/billing

- tools: netramet, netflow, cflowd, coral
  - some suck less? ...evolution requires use

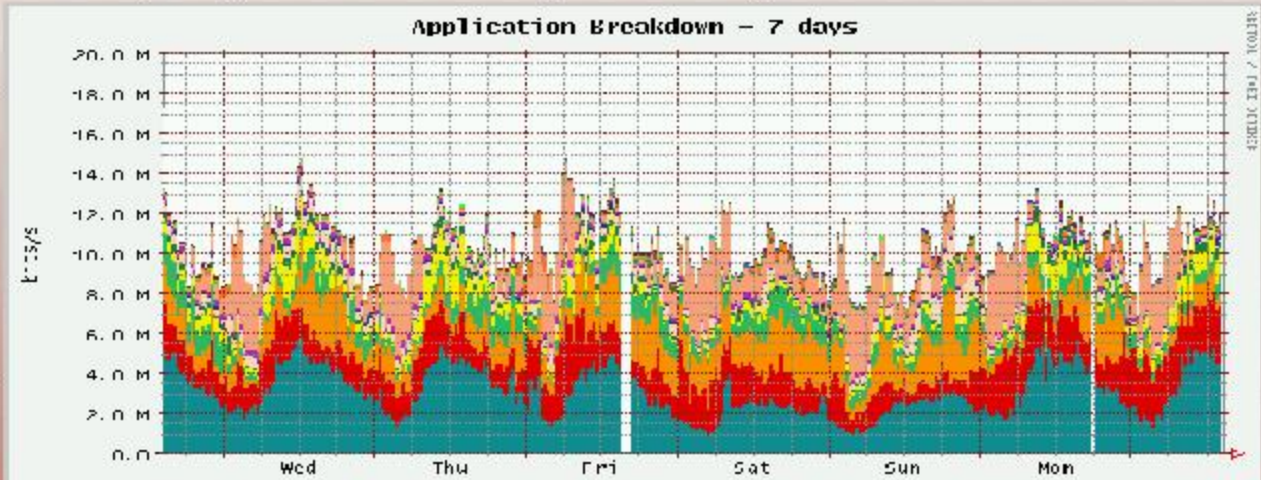
workload char.: protocol

23 may 01, ucsd-cerfnet (coralreef)



workload char.: applications

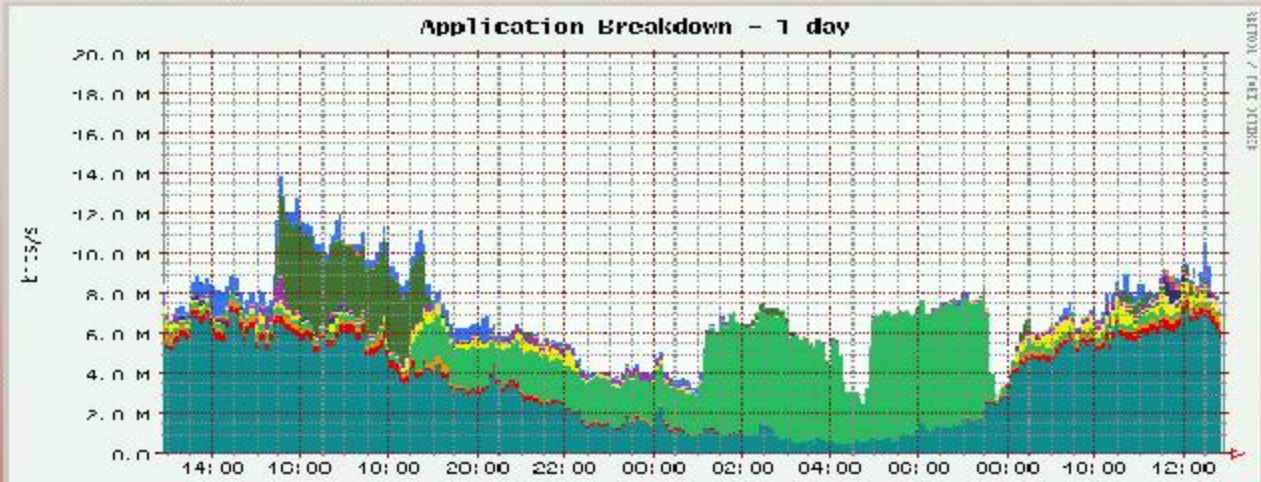
23 may 01, ucsd-cerfnet (coralreef)



	Min	Avg	Max
World Wide Web(80) (WWW)	756.43 k	3118.74 k	7206.46 k
Squid web cache (SQUID)	169.03 k	1741.37 k	5719.30 k
FTP Data Stream (FTP_DATA)	6.57 k	1511.34 k	5120.38 k
Napster MP3 (NAPSTER_DATA)	2.50 k	702.12 k	2000.07 k
Shoutcast MP3 (SHOUTCAST)	58.49 k	615.19 k	2106.40 k
SMTP (mail forwarding) (SMTP)	13.79 k	203.35 k	1746.26 k
Secure Shell (SSH)	6.66 k	121.72 k	2981.29 k
Secure Web (HTTPS)	0.81 k	74.80 k	854.41 k
GNUtella file sharing (GNUTELLA)	1.14 k	100.02 k	1010.00 k
USENET NNTP (NNTP)	35.29 k	1451.42 k	4984.95 k
Half Life game (HALFLIFE)	0.00 k	66.25 k	432.28 k
America Online (AOL)	6.58 k	130.56 k	1902.43 k
Domain Name Service (DNS)	25.02 k	64.66 k	275.66 k
ICMPECHOREQUEST (ICMPECHOREQUEST)	0.41 k	46.00 k	107.01 k

workload char.: applications

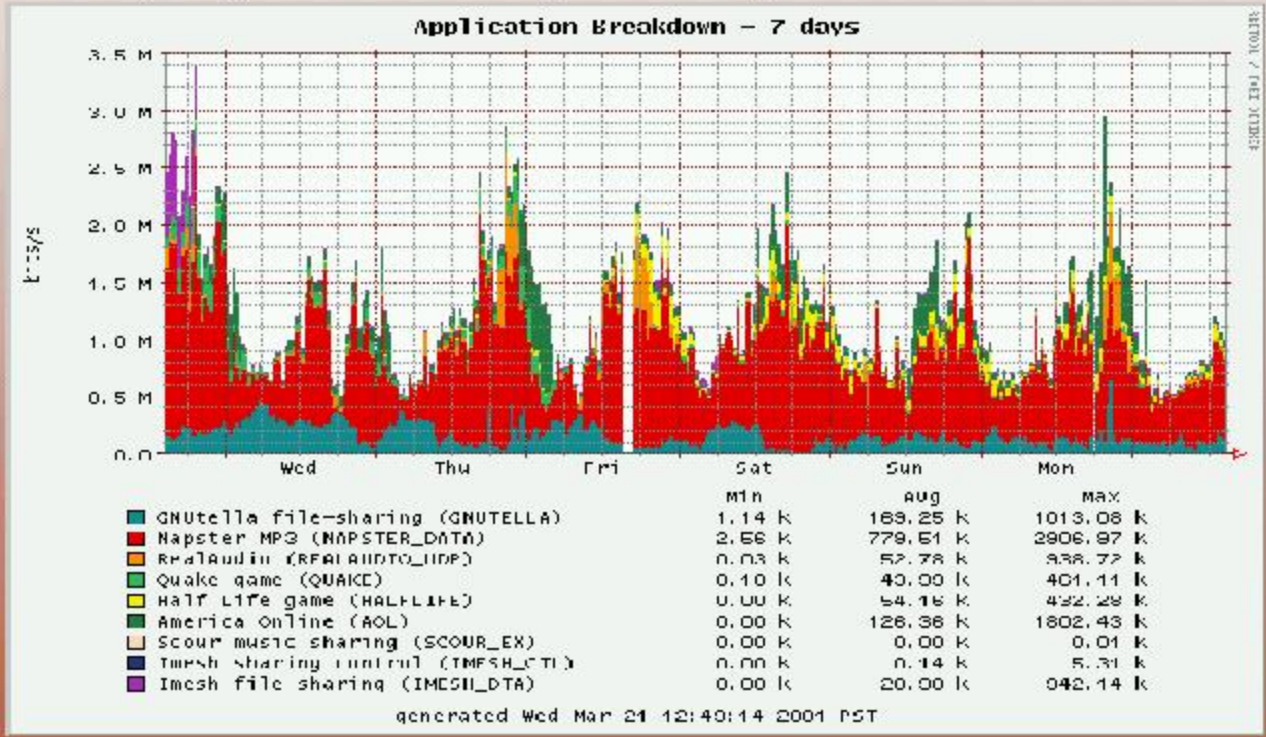
23 may 01, sdnap (coralreef), usenet peak



	Min	Avg	Max
World Wide Web(80) (WWW)	407.88 k	3493.03 k	7345.38 k
MS_MEDIA (MS_MEDIA)	0.03 k	193.21 k	471.00 k
RTSP Media Streaming (1) (RTSP)	5.52 k	130.92 k	461.35 k
USENET NNTP (NNTP)	100.05 k	2000.20 k	6005.20 k
Napster MP3 (NAPSTER_DATA)	0.04 k	238.70 k	1033.24 k
Hotline servers (HOTLINE)	0.02 k	43.81 k	191.61 k
Secure Web (HTTPS)	26.83 k	80.29 k	169.33 k
Tmesh file sharing (TMESH_DTA)	0.02 k	26.34 k	758.32 k
SMTP (mail forwarding) (SMTP)	4.04 k	122.11 k	1177.05 k
Domain Name service (DNS)	15.09 k	29.79 k	97.46 k
Secure Shell (SSH)	0.17 k	544.40 k	4421.29 k
GNUtella file-sharing (GNUTELLA)	0.00 k	12.39 k	269.50 k
SQUID (SQUID)	4.34 k	16.88 k	137.26 k
FTP Data stream (FTP_DATA)	0.01 k	302.07 k	1471.10 k

workload char.: emerging applications

23 may 01, ucsd-cerfnet (coralreef)



workload characterization: priorities

- coral/ocXmons (OC3,12,48, gigE)
- persistent, real-time, full-frame collection
- dynamic packet filtering triggered by attack precursors
- security policy
  - compliance auditing (passive)
  - enforcement (active)

obstacles
- hardware expensive
- privacy issues
- IPsec

workload char: working w/vendors

cflowd

- *www.caida.org/Tools/Cflowd*
- *primarily for capacity planning and trend analysis*
- *Cisco's netflow export*

- *AS-to-AS matrices*
- *net-to-net matrices*
- *port and protocol tables*
- *forward IP path*

measurement specifications to vendors

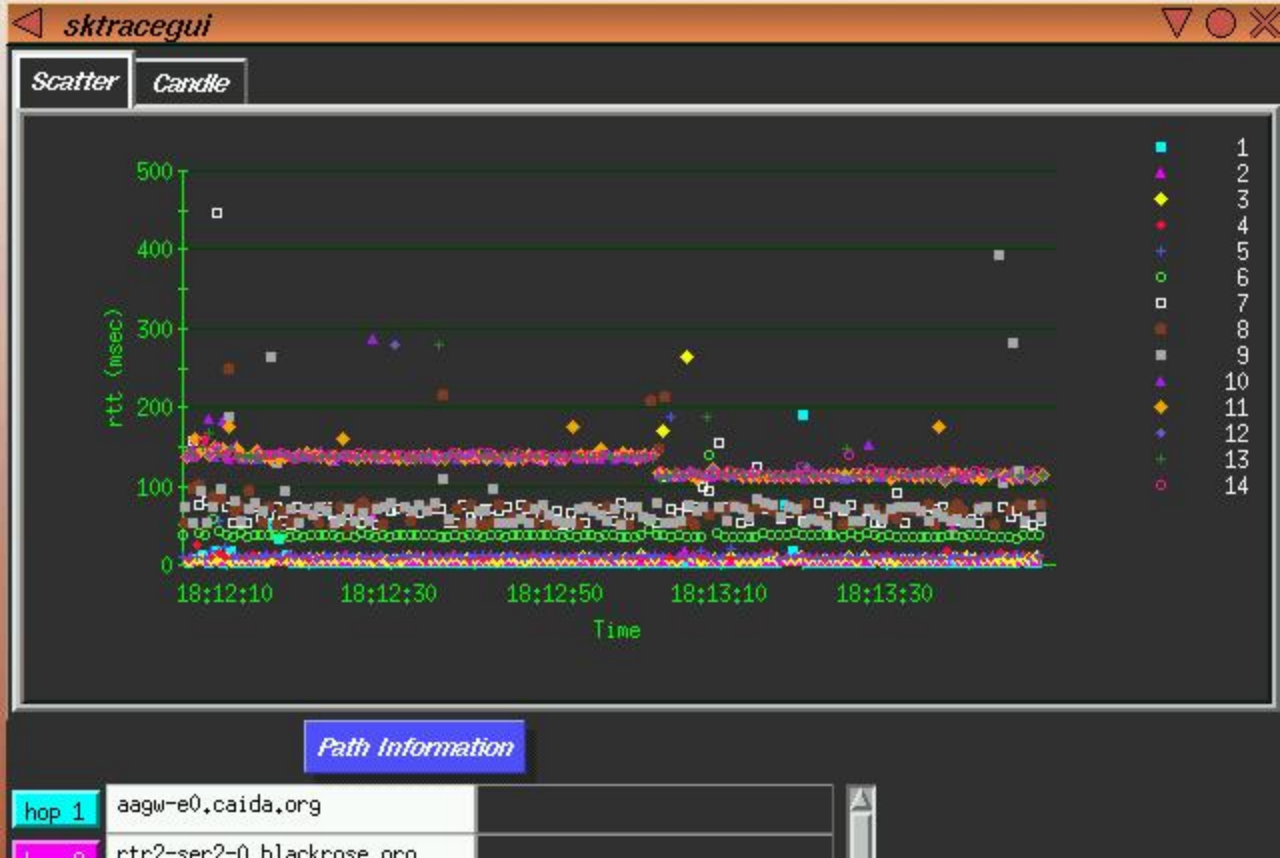
performance evaluation (active)

- *network engineers to diagnose problems*
- *ISPs & users to verify SLAs*
- *traffic engineering*
- *middleware, adaptive applications*
- *designers of real-time apps to predict software HCI*
- *Internet weather reports*

need attention to passive/hybrid approaches

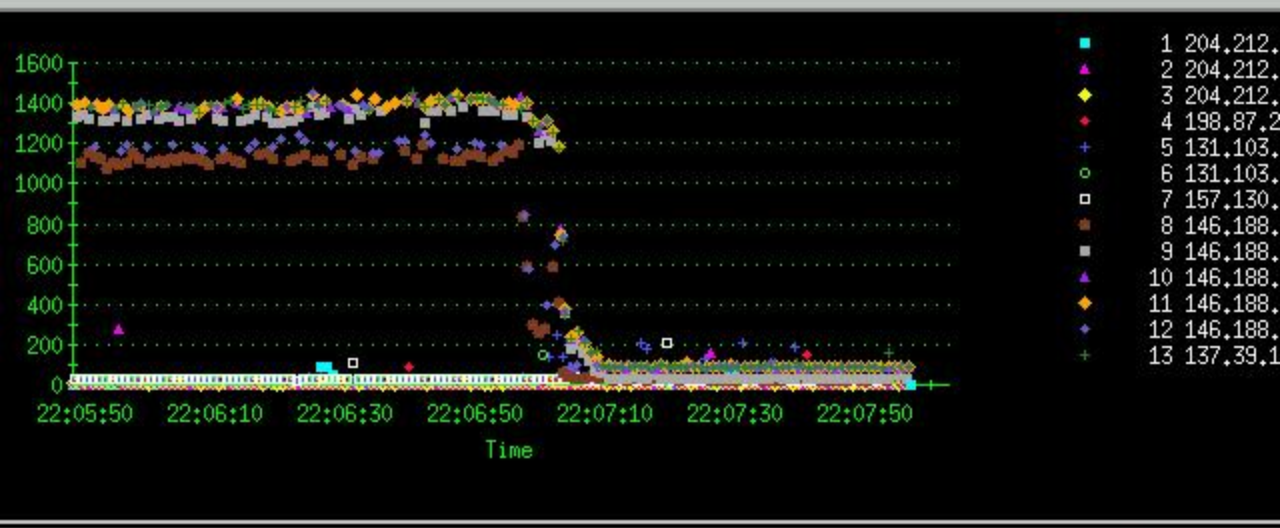
perf. eval: skping (www.freebsd.org)

perf.eval: routing (path change)



perf.eval: sktrace (www.cnet.com)

racegui



perf.eval: bandwidth estimation

- holy grail
- link or path
- available or capacity

would facilitate (allow)

- adaptive applications
- inter-domain traffic engineering / NMS
- flexible bandwidth brokering
- quality of service

would be turning point for infrastructure.
and almost completely unable to do now

perf.eval: bandwidth estimation

metrics

- capacity of a path
  - maximum IP-layer throughput path can provide to a flow given no competing traffic load (cross-traffic)
- available bandwidth of path
  - maximum IP-layer flow throughput to flow, given current cross traffic

- link with minimum transmission rate determines capacity (narrow link)
- link with minimum unused capacity limits the available bandwidth (tight link)

perf.eval: bandwidth estimation

basic model

- H hops
- Ci capacity transmission rate of link i
- C0 transmission rate of the source
- Ui utilization of link i

capacity of path

- C = min (Ci)'s

available bandwidth of path

- A = min (Ci * (1-Ui)) (per time interval)

perf.eval: bandwidth estimation

tools

- active or passive
  - no significant work on passive tecniques
- end-to-end or hop-by-hop (latter harder)
  - e-e: paxson, carter, dovrolis
  - h-h: jacobson, mah, downey
- SNMP gathered vs active probled
- network-intrusive or network-friendly
  - Iperf, ttcp vs pathchar, pchar

perf.eval: bandwidth estimation

two main methodologies

- *Variable Packet Size (VPS) probing*
  - *hop by hop*
  - *send many IP (UDP) pkts of varying size*
  - *use min RTT per size to filter out noise in b/w estimate*
  - *implemented in pathchar, pchar, clink*
  - *not great for high b/w or busy links*
    - *need lot of probes*
- *Packet Train Dispersion (PTD) probing*
  - *end-to-end metrics*
  - *based on packet pair*
  - *implemented in bprobe, cprobe, others*
  - *suggested change to TCP to base slow-start on disperson of first 3-4 ACKs*
  - *assumption: dispersion of long packet trains is inversely proportional to the available bandwidth*

.....caida finds this assumption wrong...

perf.eval: bandwidth estimation

caida (dovrolis) capacity estim. methodology (2000)

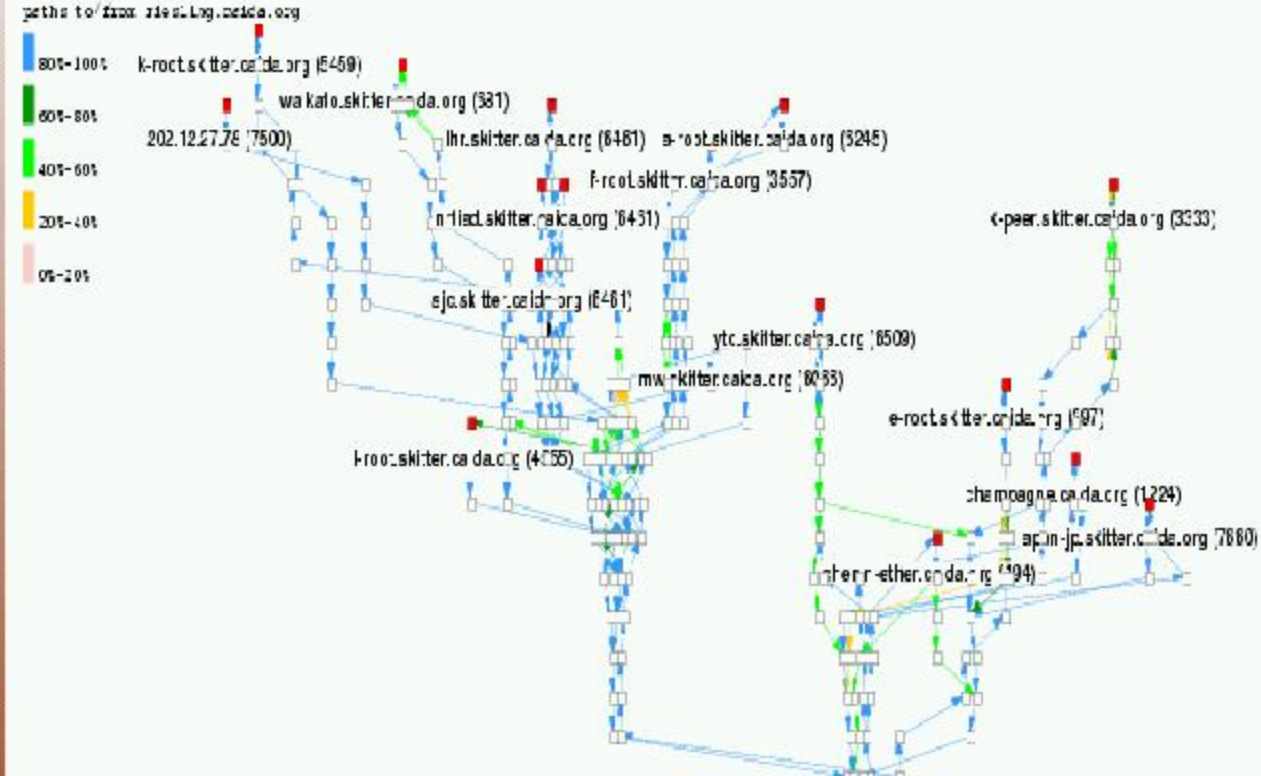
- effects of cross-traffic on measurement
- packet pair technique by itself doesn't work if cross-traffic ignored
  - distribution of b/w measurements multi-modal
  - local modes often stronger than capacity mode
- increasing train length helps
  - but estimates convert to under-estimate
  - implemented in pathrate (caida)
  - accurate for capacities 10-40 Mbps at 1Mbps resolution
  - higher capacity paths require larger estimate resolution

bandwidth estimation: priorities

- improve accuracy and robustness
  - layer-2
  - multipath forwarding
  - parallel links
  - qos packet scheduling
  - slow path for ICMP
- higher bandwidth measurements
- calibration against real paths
- visualization of output
- integration into apps, middleware, routing, traffic engineering/shaping

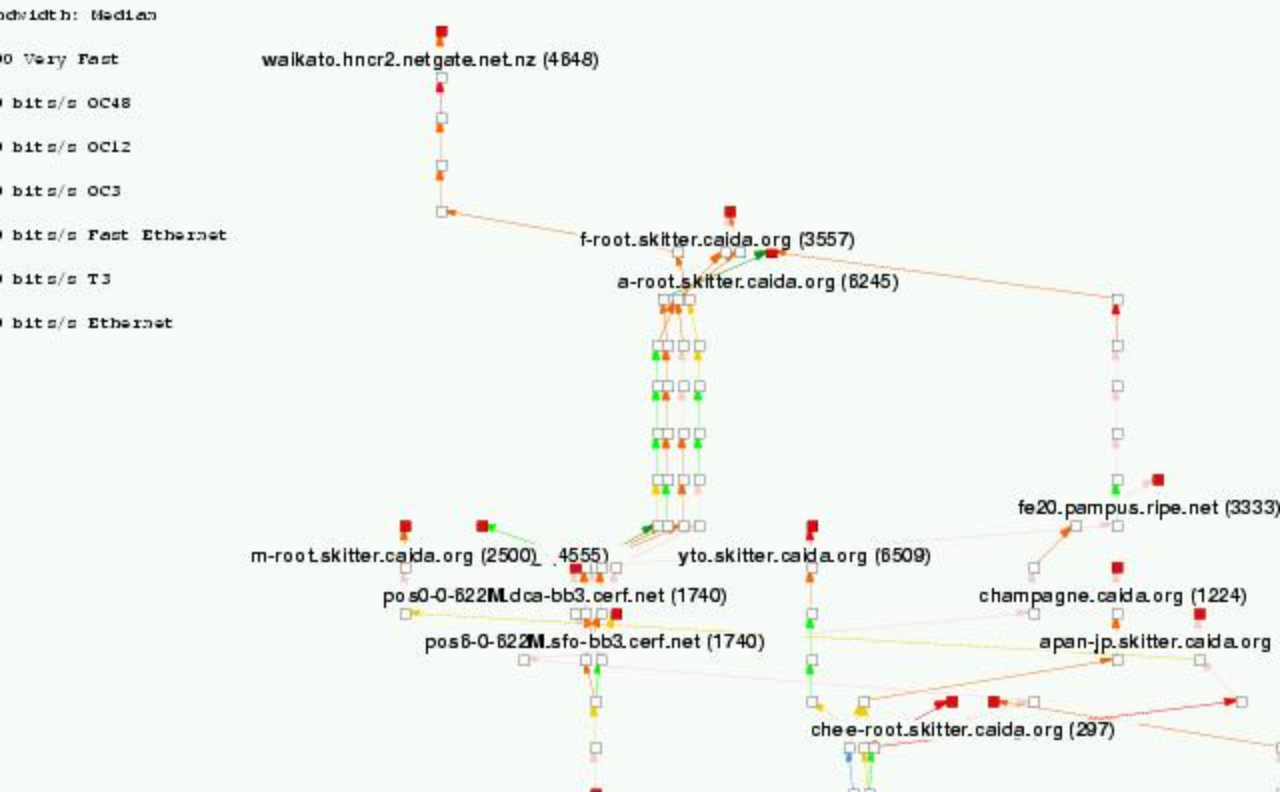
bandwidth estimation: topology dynamics (72 hrs)

dozens of forward paths per day (reverse unknown)



bandwidth estimation: topology dynamics (72 hrs)

only some of real capacity values available



caida's b/w estimation project

- improve VPS/PTD methods, extend tools
- distributed measurement infrastrcture
  - on commodity Internet
- data processing back end
- correlation with delay, workload data
- hybrid with passive techniques
- characterize non-stationarity of cross-traffic
  - (whacky variance)
- effect of routing dynamics and asymmetry
- visualization of output Internet

performance eval.: priorities

- definitions & metrics
- bandwidth assessment techniques
- correlation
  - across sources
  - with workload, routing data
- large scale deployment
- user interface to measurements

obstacles

- core infrastructural access
- mathematicians needed
- statisticians needed

routing dynamics

- *15-year-old technology*
- *admittedly seems like pretty good stuff*
- *sausage/laws....*
- *incredibly inefficient, incantation-driven*

Interdomain routing (BGP) data: sources

- BGP tables from David Meyer's Oregon Route Views
- http://moat.nlanr.net/Routing/rawdata
- MAE-East (Washington DC), MAE-West (Palo Alto), London, Amsterdam, Tokyo, Frankfurt, Ankara, Chicago, Johannesburg

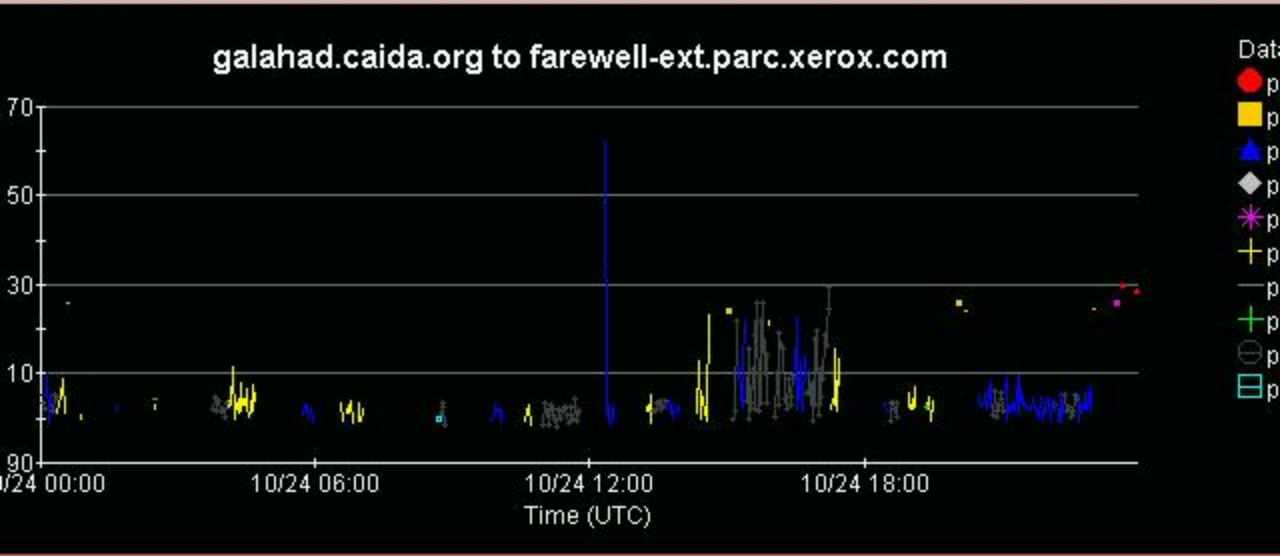
- Looking glasses (BGP-enabled traceroute servers)

Analysis:

- www.telstra.net/ops/bgp-as-paths.html
- mirror.caida.org/~broido/bgp/bgp.html
- CAIDA's "Arctic views" (longitude/degree)

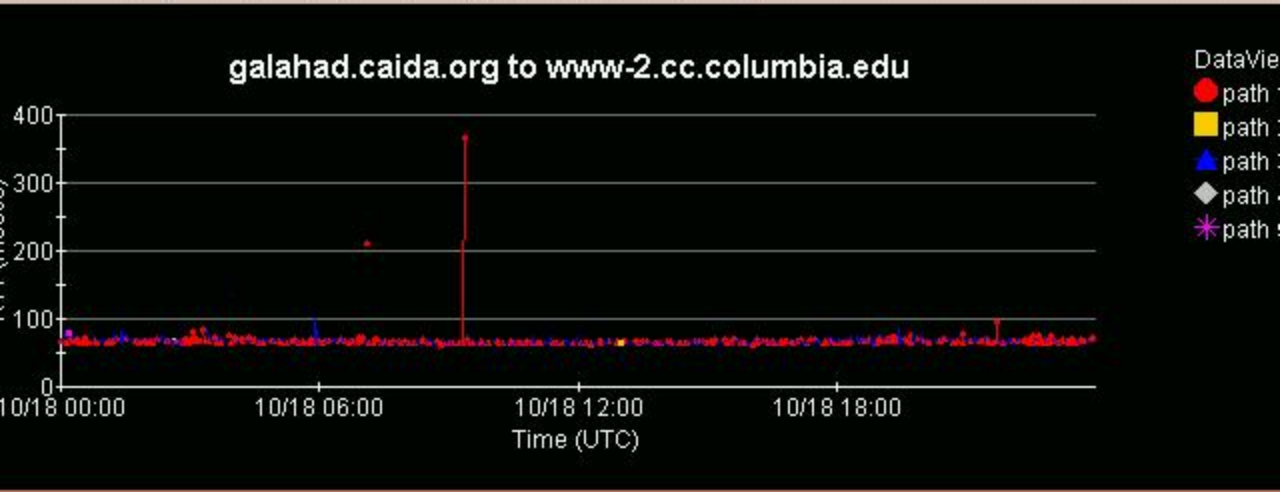
routing: microscopic example (instability)

- end-to-end RTT data changes color if path changes
- 10 unique paths over 24 hour period
- lots of jitter in data
  - unlikely to be intentional
- heavy tails predominate



routing: microscopic example (load balancing)

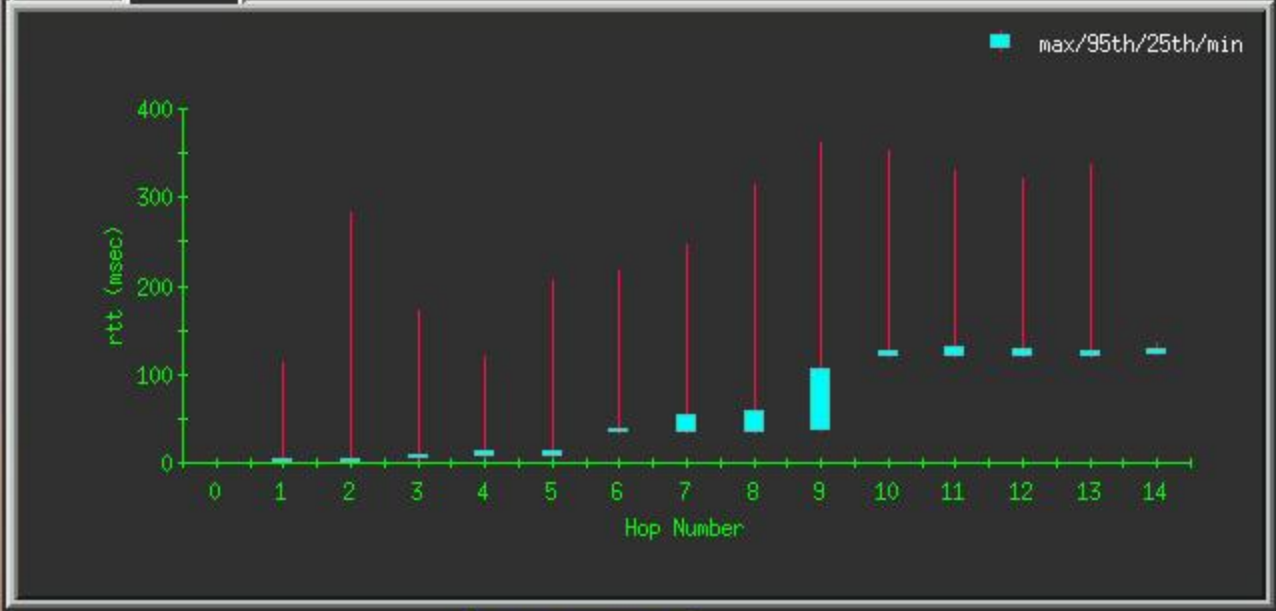
- *end-to-end RTT similar over predominantly two paths*
- *likely intentional load balancing*



routing: sktrace (parc.xerox.com)

sktracegui

Scatter | Candle



Path Information

1	aagw-e0.caida.org		
2	rtr2-ser2-0.blackrose.org		

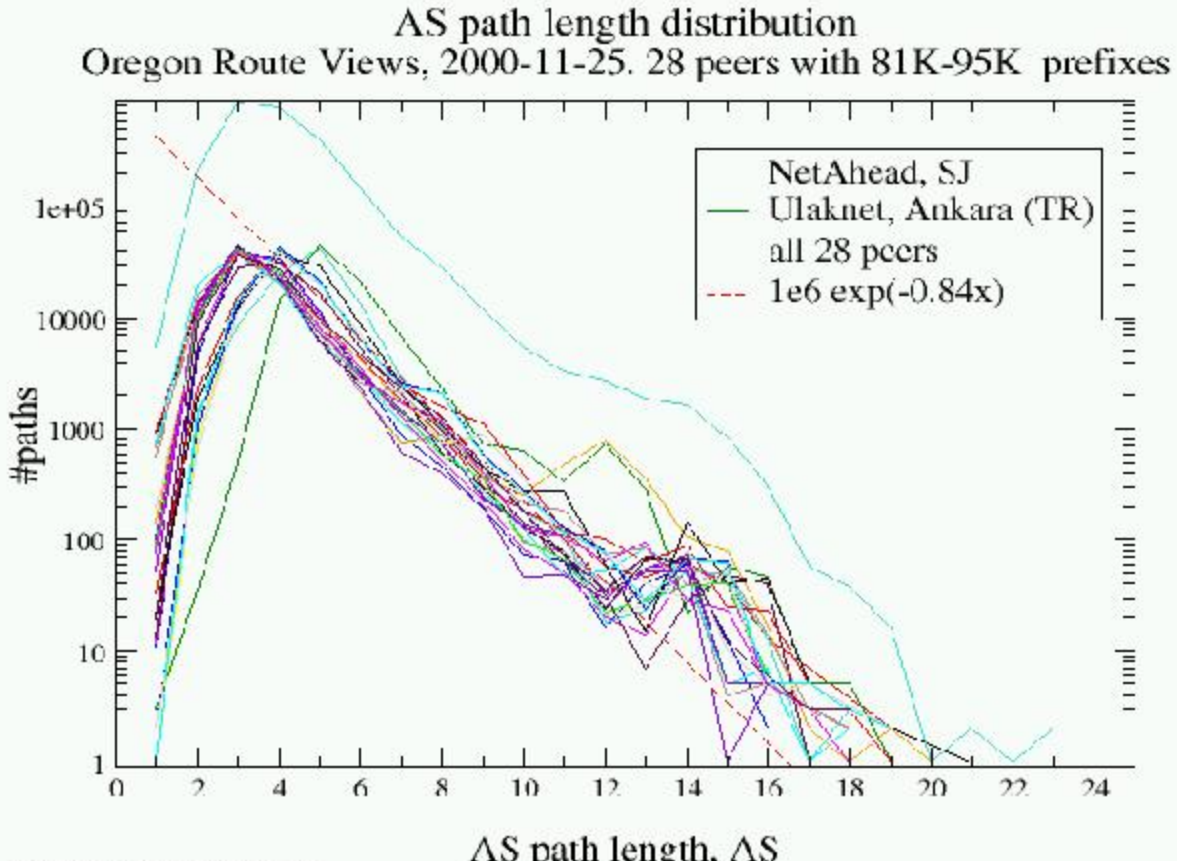
routing: macroscopic questions

BGP routing tables

- single measurement point with many feeds, e.g, routeviews (42)

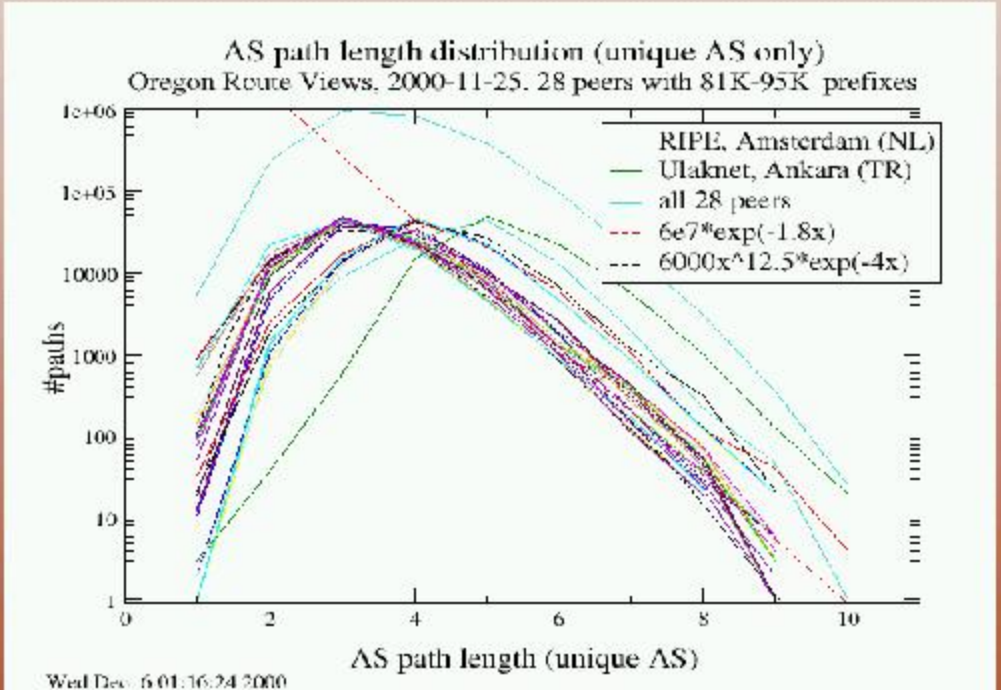
sampled data... warning

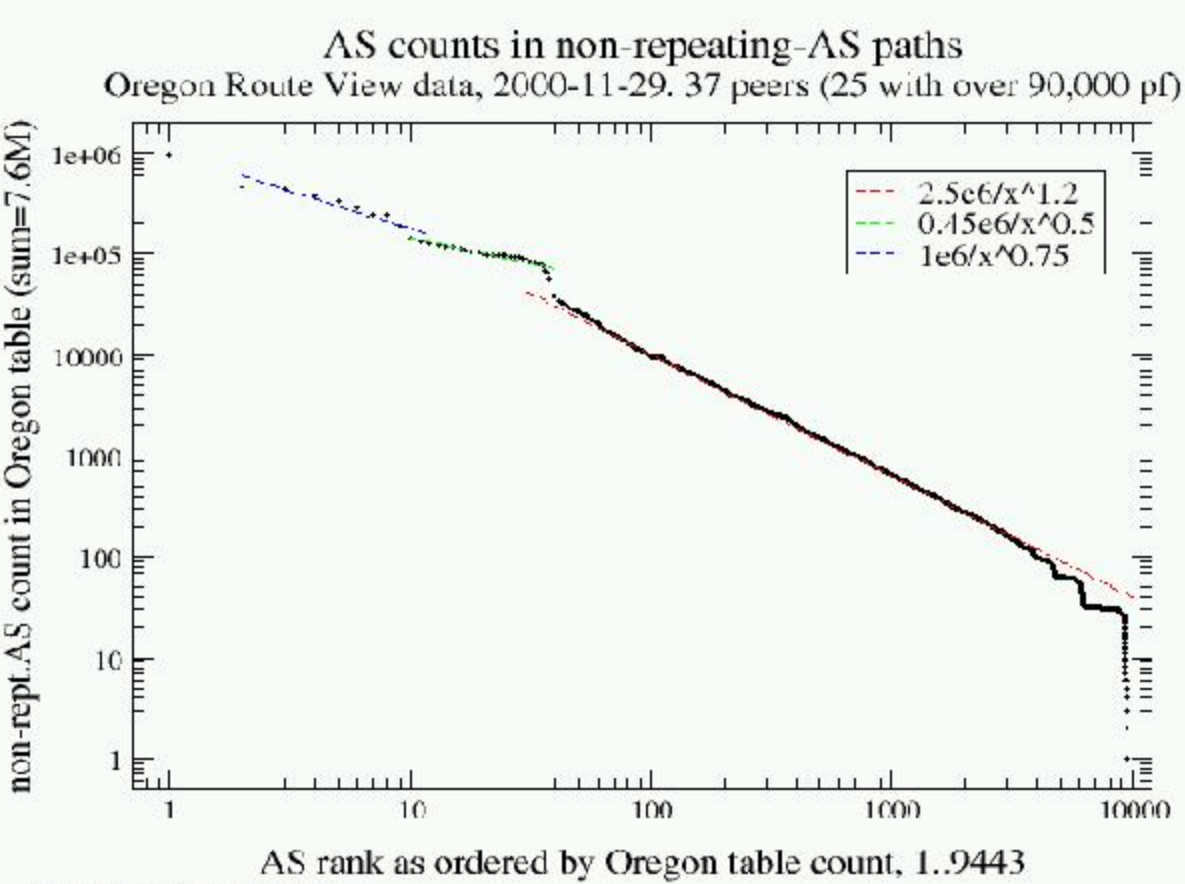
- not link state protocol... no synchronized view of entire topology & policy state
- every viewpoint contains a filtered view of network
- not seeing it doesn't mean it doesn't exist
- seeing it doesn't mean anyone else does


AS path length distribution
Oregon Route Views, 2000-11-25. 28 peers with 81K-95K prefixes
NetAhead, SJ
Ulaknet, Ankara (TR)
all 28 peers
1e6 exp(-0.84x)
#paths
1e+05
10000
1000
100
10
1
0
2
4
6
8
10
12
14
16
18
20
22
24
AS path length, AS
Wed Dec 6 01:15:37 2000

Path length measured in unique AS

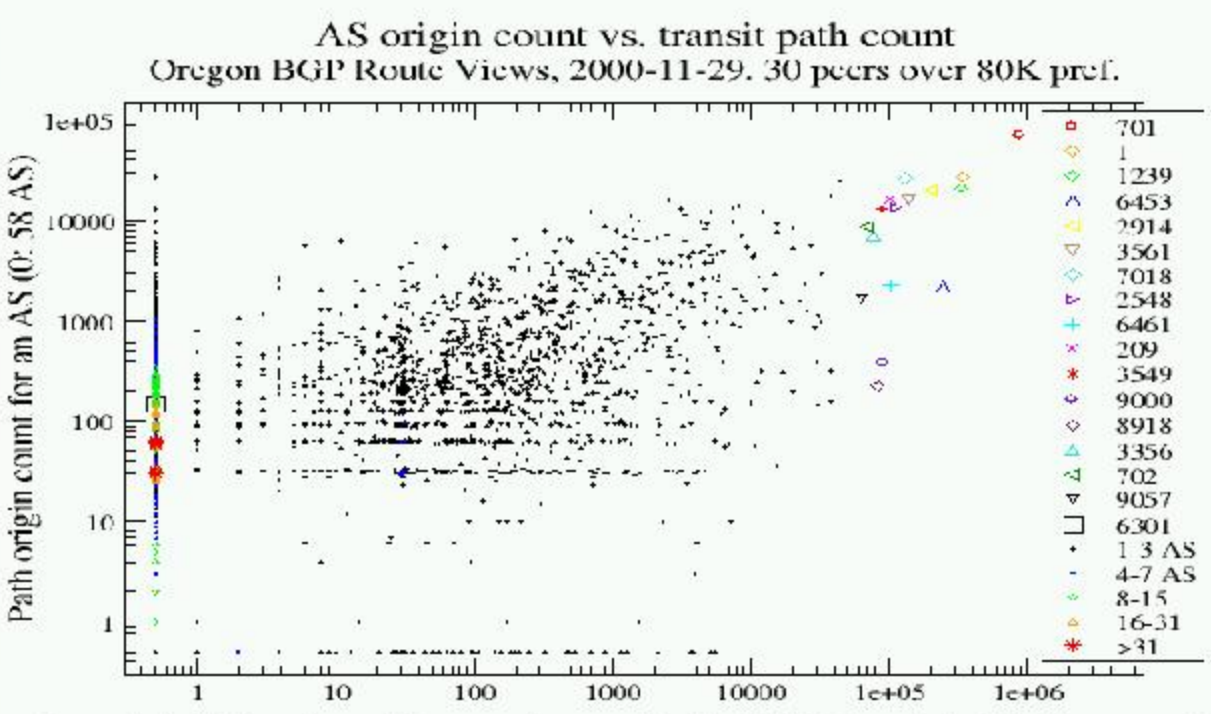
- much smaller tail: exp(-1.8x) vs. exp(-0.84x)
- much smoother
- close to Gamma distribution




AS counts in non-repeating-AS paths
Oregon Route View data, 2000-11-29. 37 peers (25 with over 90,000 pf)
non-rept.AS count in Oregon table (sum=7.6M)
1e+06
1e+05
10000
1000
100
10
1
2.5e6/x^1.2
0.15e6/x^0.5
1e6/x^0.75
1
10
100
1000
10000
AS rank as ordered by Oregon table count, 1..9443
Thu Jan 11 05:02:11 2001

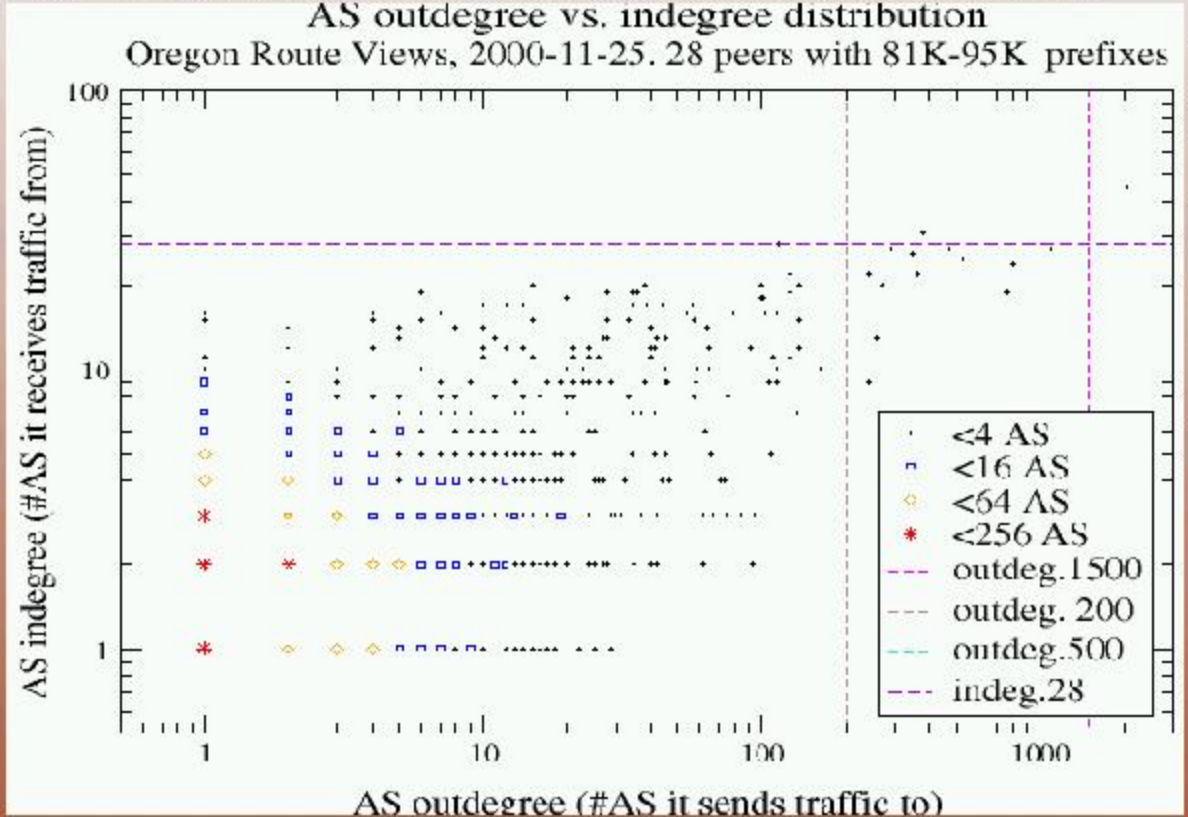
transit versus origin role of an AS

AS 701 is in 31% of all paths (origin for about 2.5% of prefixes)



routing: macroscopic questions

Are there some natural knees ('tiers') in AS outdegree distribution?



Uses of BGP routing data

- *correlation to topology data (congruent?)*
- *mapping topology data:*
  - *aggregating IP to network prefixes*
  - *aggregating prefixes to origin AS*
- *inferring contractual relations*
- *"bird's eye view" of the net - (AS polar graph)*
- *predicting AS path taken by a packet???*

important question: can you get essentially the same information from either dataset?

skitter versus BGP (topology vs routing)

indeed, even though often covering fewer ASes than a full BGP table, skitter data shows bidirectional and transit connectivity for a significantly more ASes than BGP data of the best available quality and sampling.

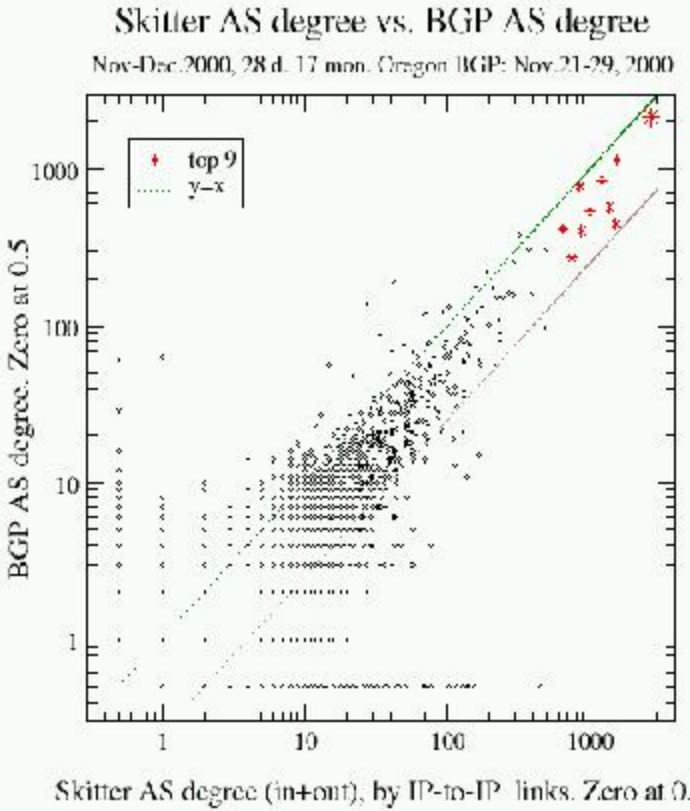
we did not expect this!

skitter versus BGP (topology vs routing)

metrics for comparison

- outdegree distribution
- combinatorial core
  - iteratively strip leaves
- giant connected component (almost 90% of skitter core)
  - largest bidirectionally connected component
  - 200X larger than any other component

skitter versus BGP (topology vs routing)



skitter vs BGP (topology vs routing)

- BGP routeviews: 37 peers, 9.4K ASes, ~100K prefixes
- skitter: 17 monitors, 400K dsts, 50K prefixes

to equalize, have to reduce skitter to AS graph:

- 0) pick one trace per prefix per day
  - ignore nodes next to non-responding hops
- 1) strip IP leaves
- 2) convert IP -> prefixes
- 3) strip prefix leaves
- 4) convert prefix -> AS

5) strip AS leaves

skitter vs BGP route-views

- basically a ton of decimation on a day of skitter data
- skitter AS graph: 6949 AS nodes, 16145 AS links (peering sessions)
- remove any potential advantage of skitter probing
  - frequency
  - finer (IP) granularity
- nonetheless, skitter captures much larger share of the Internet's bidirectional connectivity.
- skitter combinatorial core (full-transit portion) contains 988 AS nodes,

...as opposed to BGP's 299 nodes (3.3X more)

new units of connectivity analysis

proposed in course of our early analysis

- **BGP atoms**
  - equivalence class of IPv4 network address prefixes
- **cones**
  - nodes that wholly or in part depend on a given node (tip of cone) for connectivity
- **intra-prefix connected subcomponents**
  - deaggregate prefixes into truly connected subcomponents
- **dual AS graph**
  - inverted node graph

routing: research priorities

- effects of outages on surrouding ISPs
- effects of topology changes on Internet performance
- unintended consequences of new policies
  - MPLS
  - traffic engineering
- dynamic detection/response to congestion & topology changes
- identifying vulnerabilities created by dependencies on critical paths
- utilization of address space
- efficiency of routing table
- asymmetric, instabilities in routing across providers
- effects of unicast/multicast incongruity

routing: research obstacles

- canonical BGP (route table) data (not so much anymore)
- mapping IP address to anything
  (we've been here before)
- prudent security dictates making research difficult

now what?

- `seamless': no such thing
- measurement tools/infrastructure
  - well-considered
  - strategically deployed
  - collaboratively maintained
- more infrastructure-relevant research on resulting data
  - feedback into tool design
- correlation among data sources/types, simulation, visualization
- proactive participation
  - top-down (app developers scope constraints)
  - bottom-up (ISP cooperation)

www.caida.org/Presentations/

kc claffy
UCSD/SDSC/CAIDA
kc@caida.org
www.caida.org