Internet Measurements

[Then he explained that what can be observed is really determined by the theory.
He said, you cannot first know what can be observed,
but you must first know a theory, or produce a theory,
and then you can define what can be observed....]

Heisenberg’s recollection of a meeting with Einstein

Internet measurements-Why?

- Why measure?
  - What’s the need? Do we need to measure?
  - Can we just google it?
  - What is the motivation? Who gains?

Why?

- Characterize behavior
  - Performance
  - Understand growth
  - Properties
- Design issues
  - Identify problems, suggest improvements
  - Understand protocol issues
  - Understand behavior
- Manage the network
  - Diagnose and correct
  - Traffic patterns
  - Nielsen rating for network performance

Internet measurements- What?

- What to measure?
- – need to determine X (X -a property of the internet)
- What property?
  - Performance
    - Latency, throughput, link B/W, jitter,…..
  - Structure
    - Topology, relationships, characteristics
  - Behavior
    - Stability, failures, pathological, misconfigurations
  - …

Internet measurements-How?

- How to measure?
- – need techniques and tools to measure, gather data, analyze data, present data

Network Measurement

- Measurement techniques
- Passive techniques
  - Collect traces and analyze or build special tools
  - Monitor application traffic, route announcements and analyze
  - Number of point to collect, single vantage point, several vantage points
  - Across ASes, or within AS (Gao 01, labovitz 97, labovitz 10)
- Active Techniques
  - Measure by probing (ping, traceroute, pathchar or build special tools(e.g., Bolot93, paxson 97, rocketfuel 02)
  - Inject probes/packets or remove packets and measure impact
  - Inject faults at various points BGP beacon

Issues in measurement

- To collect traces need mechanism to capture packets and dump information
- Design packet filters, should be able to take in a spec and collect trace
- Data collection and analysis are independent
- Share traces for different experiments
- Manipulation and storage cost for traces
- How to analyze packet traces?
  - Time-sequence plots
  - Plot sequence number as a function of time
  - Repeat for different time scales

Internet measurements- Challenge?

- Routing/announcement and data interaction
  - Control plane vs data plane
- Causality: cause and effect analysis
- Coverage and completeness
- Internet is too big, not everything is public
  - Commercial interests conflict
  - Access to data difficult (ASes hide stuff)
- Huge amounts of data

Network measurement Papers

- Jean Bolot SIGCOMM 93
- Vern Paxson SIGCOMM 96
  - 2006 test of time award paper
- Craig Labovitz SIGCOMM 97
  - 2008 test of time award paper
- Craig Labovitz SIGCOMM 2010 Internet Inter-domain Traffic
  - Current trends in Internet traffic, topology, content
- Faloutsos et al., Power-law relationships of internet topology SIGCOMM 99
- Additional reading: Rocket fuel SIGCOMM 02

Active Measurement

- Common tools:
  - ping
  - traceroute
  - scriptroute
  - Pathchar/pathneck/… BW probing tools
  - Fault injection: BGP beacons

End-to-end delay and loss behavior in the Internet

- Simple experiments, interesting way of arriving at the results
  - Presentation makes all the difference
- Simple analysis and model fitting
- Use end-to-end measurements to derive bottleneck link behavior
- Use active measurement tool
- Send UDP packets at regular intervals and measure RTT. Repeat experiments for different intervals
- Data analysis using phase plots

Time-series plot

- For every sample, plot the rtt_n
- Not much to gain (except some have zero or packet loss)
- Average, max rtt_n, min rtt_n
- Instead plot rtt_n, rtt_{n+1} (x,y)
- If constant, point.
- D, μ
- RTT= Prop delay + Queuing delay + Transfer delay
- Where does the time go? ➔D + w_{n+1} + Service time
- Probe size is P
- Both see same load, one sees load, interarrival time has a significant impact on probe compression.

Delay analysis

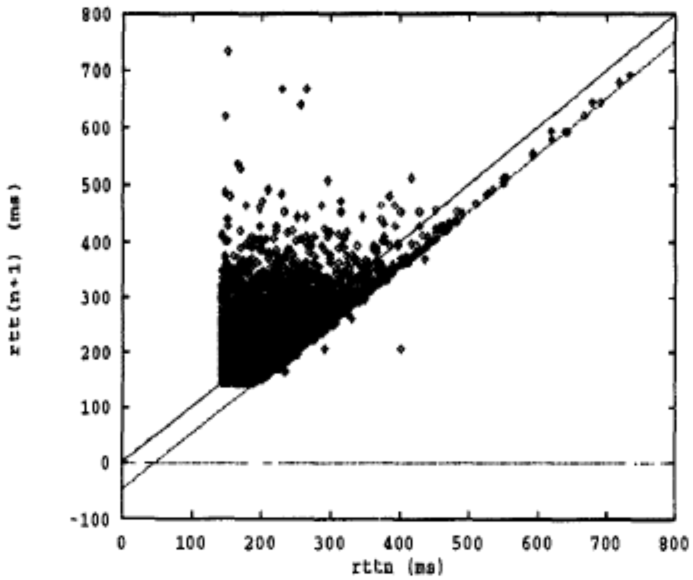
- s_n = time of nth probe, r_n = time at which echo of the nth probe is received, $rtt_n = r_n - s_n$ and $\delta = S_{n+1} - S_n$
- For various values of δ determine the distribution of rtt_n
- Need to figure the impact of load on the network
- Assume that there is very light load

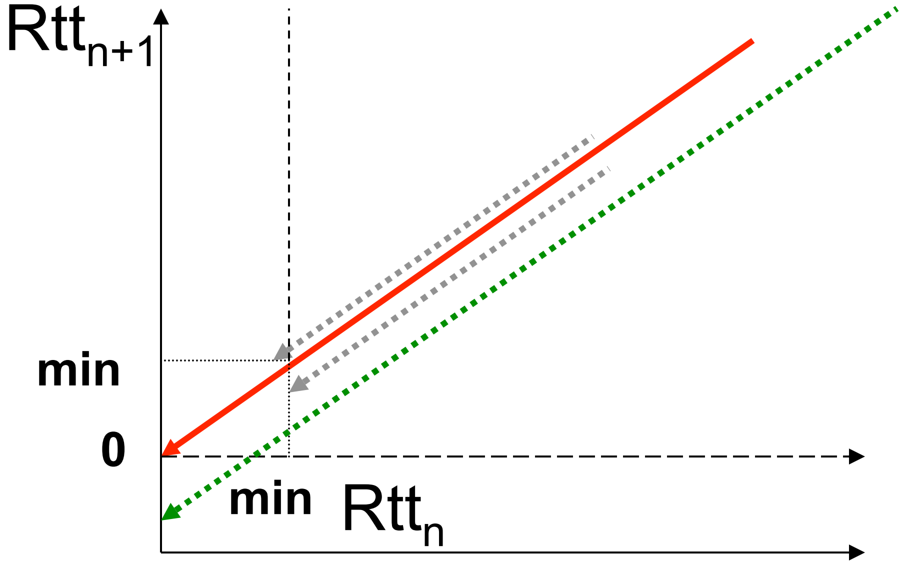
$$w_{n+1} \approx \mathrm{w}_{\mathrm{n}} \text{ or } \mathrm{w}_{\mathrm{n+1}} = \mathrm{w}_{\mathrm{n}} + \varepsilon_{\mathrm{n}}$$

- A phase plot should give us points on or around the diagonal

Phase plot

- $rtt_{n+1}=rtt_n \pm \varepsilon$



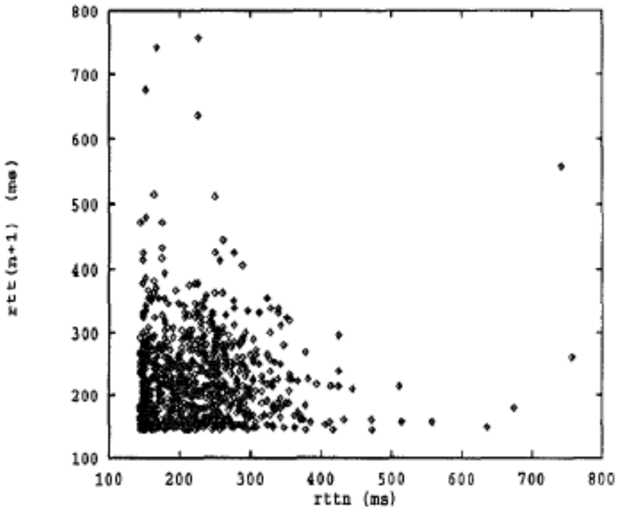


Delay analysis

- Assume a bottleneck link
- rtt_n = prop delay + queuing delay
- rtt_n = D + wait time + service time
  - $rtt_n = D + w_n + P/\mu$
- Heavy cross traffic
  - $w_{n+1} = w_n + B/\mu$ or $rtt_{n+1} - rtt_n = B/\mu - \delta$ or approx B/μ
- During this time k probes arrive
- For these probes the waiting times are P/μ
  - $rtt_{n+k} - rtt_{n-(k-1)} = P/\mu - \delta$

Delay analysis

- $\delta > P/\mu$, RHS is negative
  - $rtt_{n+1} - rtt_n < 0$, probe compression (traffic light analogy)
  - The line $y=x+P/\mu - \delta$ with a large number of points shows this effect
  - Find the x intercept, we know P and δ hence calculate μ
  - Figure 2, δ = 50 msec
  - Bottleneck link = 128 Kbps
- $\delta >> P/\mu$ RHS is also negative
  - But two probes never get behind each othe
  - Say δ = 500msec , then $\delta - P/\mu$ = 498 mse
  - Find how many points on the line $y = x - 4$
  - Figure 4, δ = 500 msec
- $\delta < P/\mu$ probe saturates the queue



Bottleneck link capacity

- Repeated experiments for different values of δ over a different path (umd, pittsburgh)
- For δ = 8 msec, y intercept = -7.8 msec, P/μ = 0.2
  - Given P = 32 bytes, we get m = 1.28 Mbps (Figure 5)
- For δ = 50 msec, phase plot consists of points scattered around y=x line. No probe compression
  - Figure 6

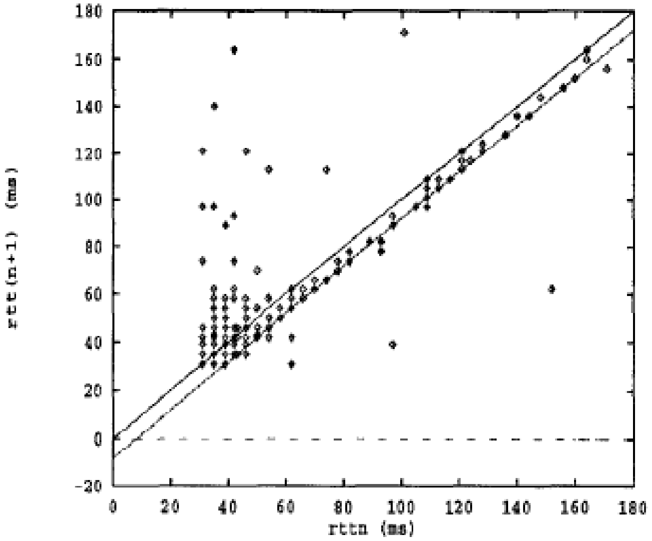
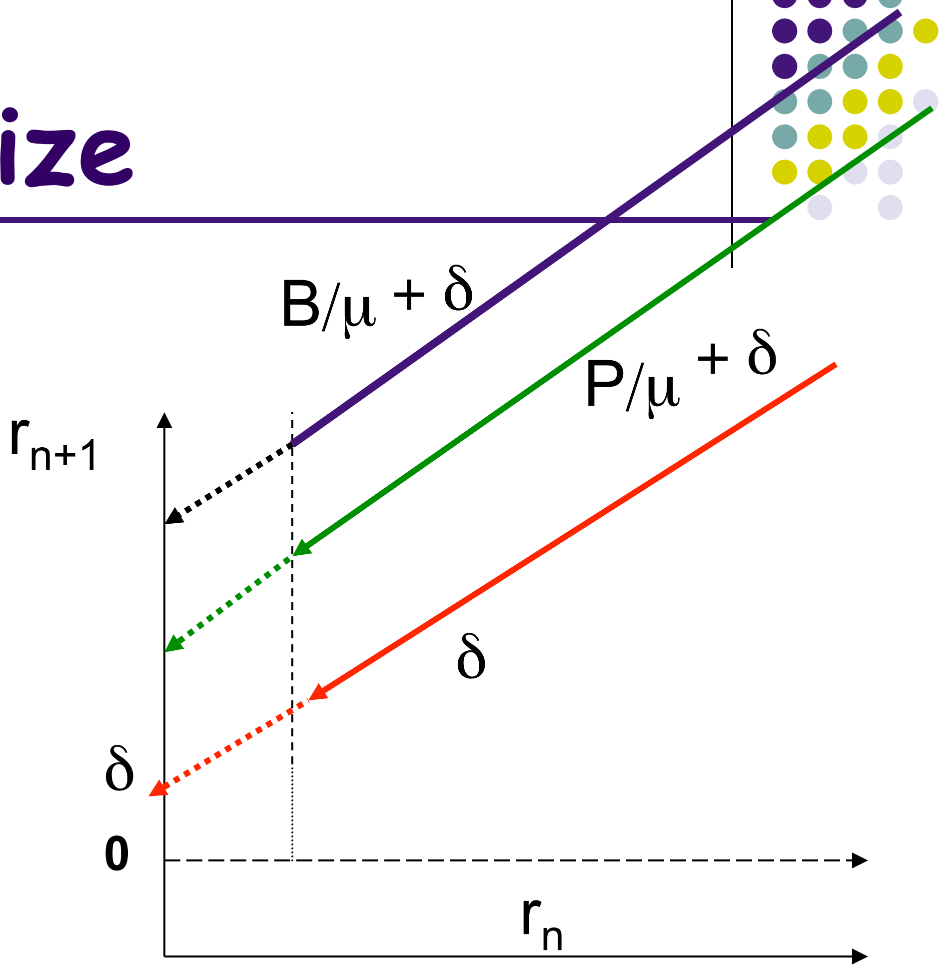


Figure 5

Packet size of cross traffic

- Assume cross traffic between two probes δ units apart
- w_{n+1} = service time for b_n and w_n + wait time for w_n
- $= (B + P)/\mu - \delta + w_n$
- $B = \mu(w_{n+1} - w_n + \delta) + P$
- $w_{n+1} - w_n + \delta = (B+P)/\mu$
- Interarrival time between successive probe returns is $(r_{n+1}-r_n)$ is $w_{n+1} - w_n + \delta$
  which is the same as $rtt_{n+1} - rtt_n + \delta$
- Plot CDF of this function

Finding packet size



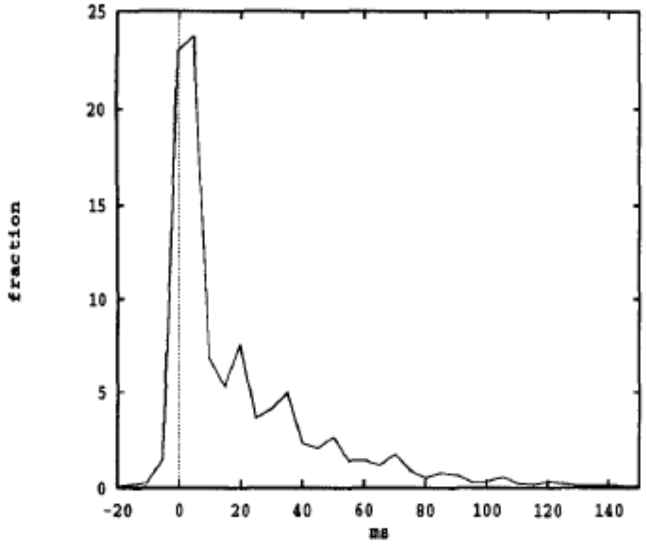


Figure 8: Distribution of $w_{n+1} - w_n + \delta$ for $\delta = 20$ ms

Three cases

- Probes that were subject to probe compression and were serviced according to probe packet size
  - Difference in Wait times should be around P/μ
  - First peak
- Probes that were not subject to probe compression
  - Difference in Wait times should be δ
  - Second peak
- Probes that were behind one internet packet
  - Difference in Wait times should $(B + P)/\mu$
  - Third peak
- Use the value of third peak in Equation 6. Knowing bottleneck link, probe size, we can calculate internet packet size

Packet loss

- rtt_n was considered to be zero for a lost packet
- Look at the number of packets that were lost for various values of δ
  - Smaller the value of δ larger the loss
- Find probability of correlated losses (burst size)
  - Further, conditional loss probability > loss probability
  - Because loss depends on queue length
- Burst loss size is about 1 for most cases
- Interesting observation
- Implications for FEC and Go-back-N

Discussion

- Simple experiments, clever analysis leads to interesting results
- Phase plots, very creative
- Assumptions: single bottle neck
  - Needed for simple analysis
- How about internet of today
  - Asymmetric links
  - Traffic patterns may not pan out as in Jean Bolot's paper
- What about packet size distributions?

Active Measurement (2)

- Using traceroute
- Every packet has TTL
- At each hop, TTL is decremented
- When TTL becomes 0, packet is dropped and an ICMP error message is sent back to the source
- Use this to determine routers along the path from source to destination
- www.traceroute.org

Traceroute

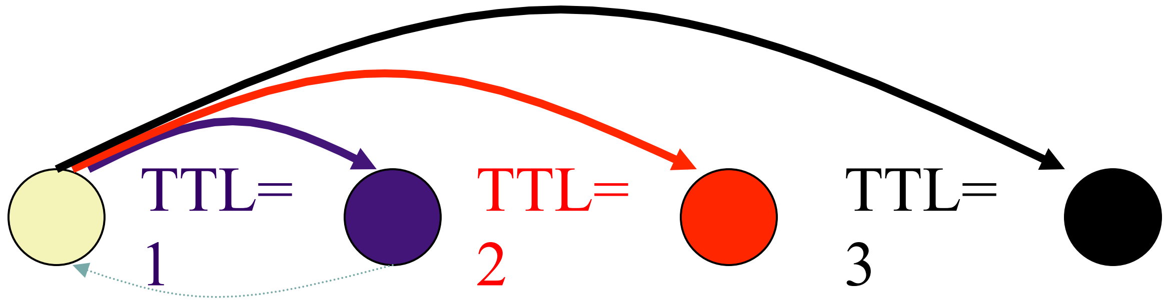
```
tracing path from www.net.princeton.edu to 128.6.25.18 ...
traceroute to 128.6.25.18 (128.6.25.18), 30 hops max, 40 byte packets
1 gigagate1 (128.112.128.114) 0.530 ms 0.300 ms 0.343 ms
 2 vgate1 (128.112.12.22) 0.360 ms 0.328 ms 0.348 ms
 3 local1.princeton.magpi.net (216.27.98.113) 23.504 ms 2.484 ms 2.490 ms
 4 remote2.njedge.magpi.net (216.27.98.82) 4.624 ms 4.217 ms 4.180 ms
 5 130.156.251.206 (130.156.251.206) 28.162 ms 4.967 ms 4.706 ms
 6 198.151.130.133 (198.151.130.133) 5.180 ms 22.034 ms 5.067 ms
 7 sr02-hill012-svcs.runet.rutgers.net (128.6.1.22) 5.391 ms 5.343 ms 5.316 ms
 8 * * *
9 * * *
10 * * *
11 * * *
12 rags.rutgers.edu (128.6.25.18) 6.534 ms 5.813 ms 5.650 ms
```

Traceroute

```
Warning: multiple IP addresses found for www.iitb.ac.in, using 203.78.217.179
tracing path from www.net.princeton.edu to 203.78.217.179 ...
traceroute to 203.78.217.179 (203.78.217.179), 30 hops max, 40 byte packets
 1 gigagate1 (128.112.128.114) 0.826 ms 0.329 ms 0.343 ms
2 vgate1 (128.112.12.22) 0.373 ms 0.587 ms 0.349 ms
3 gi-6-0-226.hse1.phlapa02.paetec.net (209.92.27.33) 2.188 ms 2.616 ms 3.321 ms
 4 ge-1-1-0-311.core02.phlapa02.paetec.net (169.130.105.19) 8.179 ms 2.550 ms 2.501 ms
 5 * * *
6 * * *
7 as0.core02.asbnva01.paetec.net (169.130.81.233) 5.919 ms 6.093 ms 6.013 ms
8 157.130.49.209 (157.130.49.209) 5.946 ms 6.119 ms 5.699 ms
9 0.so-1-0-3.XL4.IAD8.ALTER.NET (152.63.41.26) 6.560 ms 6.263 ms 6.234 ms
 10 0.so-3-0-0.IL4.NYC9.ALTER.NET (152.63.23.177) 14.967 ms 15.097 ms 14.756 ms
11 0.so-0-0-1.IR2.LND17.ALTER.NET (210.80.51.141) 87.187 ms 86.995 ms 85.641 ms
12 so-0-0-2.XT2.BOM1.ALTER.NET (210.80.37.198) 209.186 ms 210.845 ms
13 so-3-0-0.GW3.BOM1.ALTER.NET (210.80.37.114) 209.359 ms 209.795 ms 211.322 ms
14 iitb.ac.in-gw.customer.alter.net (203.78.210.150) 212.256 ms 211.991 ms 210.394 ms
15 * * * 16 * * * 17 * * * 18 * * * 19 * * *
```

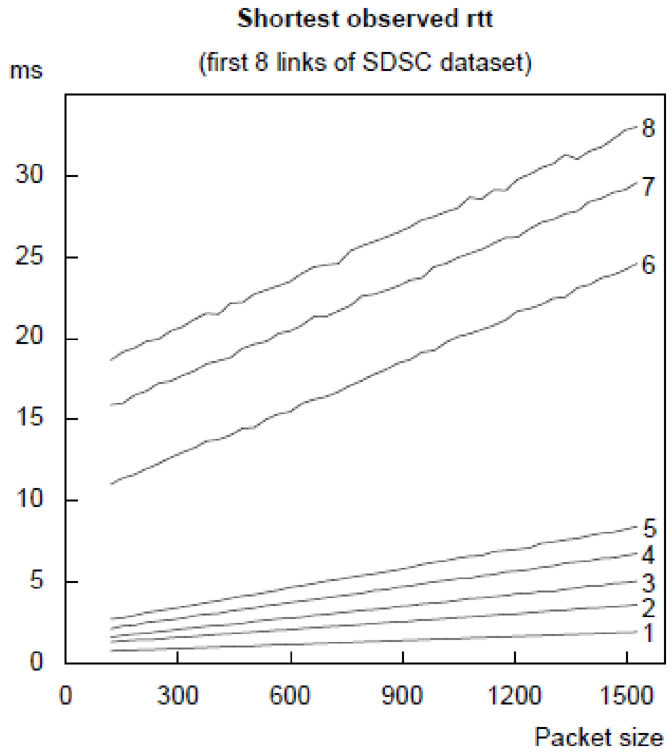
Known issues with Traceroute

- Not all routers accept ICMP messages
  - Filtered or given low priority
- Can’t identify one-way failures (* * *)
  - Forward link or reverse link
- IP address of “time exceeded” packet may be the outgoing interface of the return packet



Extend to multiple links

- Extend analysis and methodology of Bolot’s paper
- Send probes of varying sizes, across several links
- Use traceroute over N hops; repeat expt for various packet size



pathchar

- Traceroute includes Cumulative performance
- By subtracting successive link char you can get
- Subtract Y intercepts (RTT, RTT/2 gives one way delay)
- Subtract successive slopes (P/μ) you get slope which is 1/ link bandwidth
- Note, X-axis is packet size (y=P/μ)

Reference: Allen Downey Using pathchar to estimate internet link characteristics- SIGCOMM 1999

Active measurements

- Multiple vantage points
- Send probes from multiple destinations
- Cover different paths from various S, D pairs

End-to-end routing behavior [Paxson98]

- Use end-to-end measurements to answer questions about
  - Route failures and pathologies
  - Route Stability
  - Route symmetry
- Tool used
  - Traceroute between 37 internet sites
  - Network probe daemon (NPD)
  - Periodicity controlled trace routes to various destinations
  - Figure 1 in the paper Gives the sites

Methodology

- Each NPD site periodically measure the route to another NPD site, by using traceroute
- Two sets of experiments
- D_1 – measure each virtual path between two NPD's with a mean interval of 1-2 days, Nov-Dec 1994
- D_2 – measure each virtual path using a bimodal distribution inter-measurement interval, Nov-Dec 1995
  - 60% with mean of 2 hours
  - 40% with mean of 2.75 days
- Measurements in D_2 were paired
  - Measure A→B and then B→A

Shortcomings

- Just a small subset of Internet paths
- Just two points at a time
- Difficult to infer why something happened based on just end-to-end measurements
- NPD site down
  - 5%-8% of time couldn't connect to NPD's → Introduces bias toward underestimation

Routing Pathologies

- Persistent routing loops
- Temporary routing loops
- Erroneous routing
- Connectivity altered mid-stream
- Temporary outages (> 30 sec)

Routing Loops & Erroneous Routing

- Persistent routing loops (10 in D_1 and 50 in D_2)
  - Several hours long (e.g., > 10 hours)
  - Largest: 5 routers
  - All loops intra-domain
- Transient routing loops (2 in D_1 and 24 in D_2)
  - Several seconds
  - Usually occur after outages
- Erroneous routing (one in D_1)
  - A route UK→USA goes through Israel

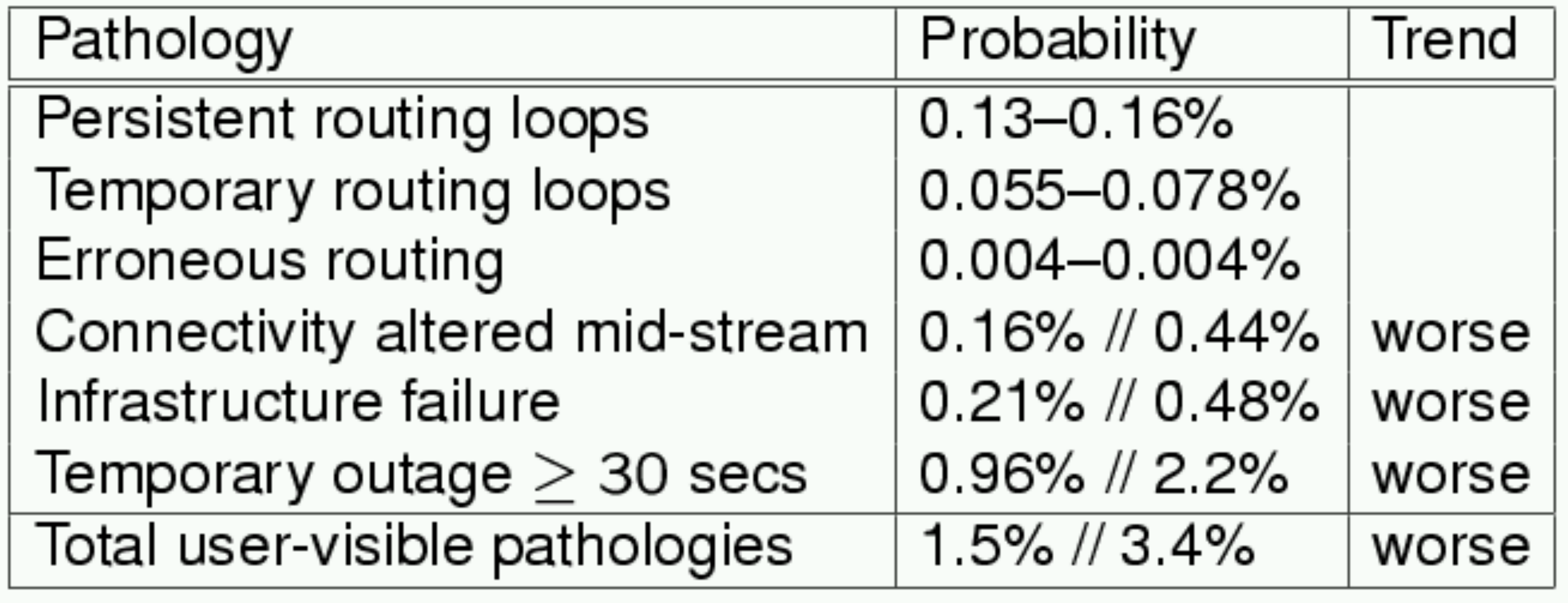
Problems with Fluttering

- One packet via DC, the other via Anaheim CA
- Path properties difficult to predict
  - This confuses RTT estimation in TCP, may trigger false retransmission timeouts
- Packet reordering
  - TCP receiver generates DUPACK's, may trigger spurious fast retransmits
- These problems are bad only for large scale flutter; for localized flutter is usually ok

Distribution of Long Outages (> 30 sec)

- X > 6 = 0.92
- X > 30 approx 0.1, X >= 60 approx 0.01
- Geometric distribution

Pathology Summary

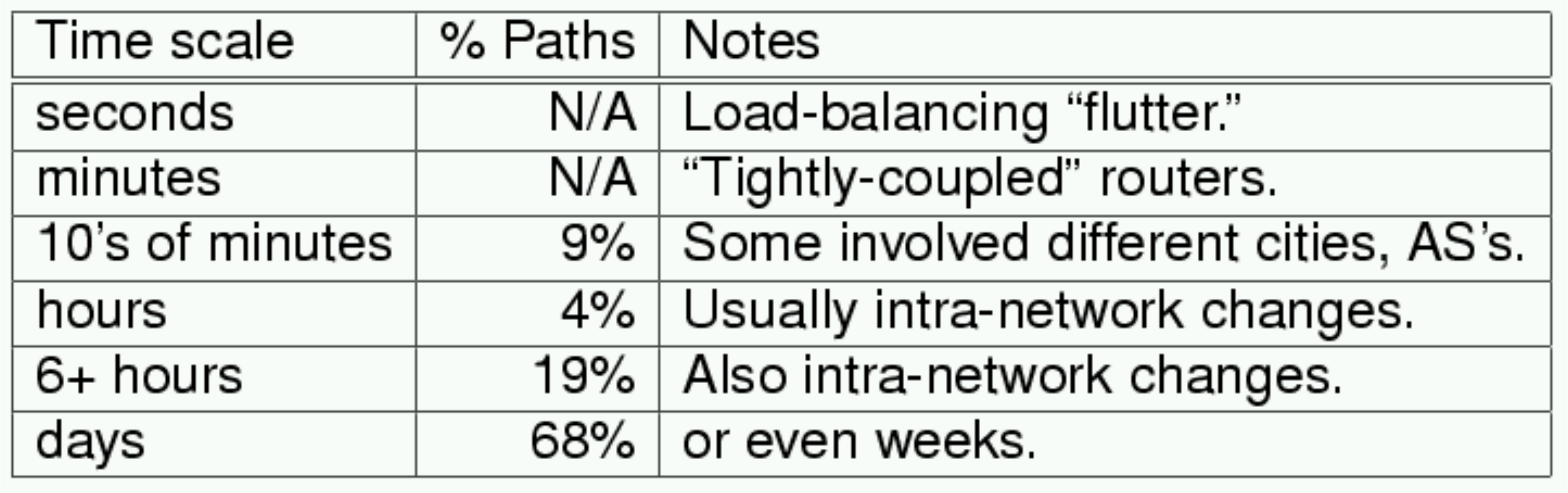


Pathology	Probability	Trend
Persistent routing loops	0.13–0.16%	
Temporary routing loops	0.055–0.078%	
Erroneous routing	0.004–0.004%	
Connectivity altered mid-stream	0.16% // 0.44%	worse
Infrastructure failure	0.21% // 0.48%	worse
Temporary outage $\geq$ 30 secs	0.96% // 2.2%	worse
Total user-visible pathologies	1.5% // 3.4%	worse

Routing Stability

- Prevalence: likelihood to observe a particular route
  - Steady state probability that a virtual path at an arbitrary point in time uses a particular route
  - What is the likelihood of observing a particular route?
  - Conclusion: In general Internet paths are strongly dominated by a single route
- Persistence: how long a route remains unchanged
  - Affects utility of storing state in routers
  - Sampling intervals should be able to capture events
  - Conclusion: routing changes occur over a wide range of time scales, i.e., from minutes to days

Route Persistence



Time scale	% Paths	Notes
seconds	N/A	Load-balancing “flutter.”
minutes	N/A	“Tightly-coupled” routers.
10’s of minutes	9%	Some involved different cities, AS’s.
hours	4%	Usually intra-network changes.
6+ hours	19%	Also intra-network changes.
days	68%	or even weeks.

Route Aymmetry

- 30% of the paths in D_1 and 50% in D_2 visited different cities
- 30% of the paths in D_2 visited different AS’s

- Problems:
  - Break assumption that one-way latency is RTT/ 2

Summary of Paxson’s Findings

- Pathologies doubled during 1995
- Asymmetries nearly doubled during 1995
- Paths heavily dominated by a single route
- Over 2/3 of Internet paths are reasonable stable (> days). The other 1/3 varies over many time scales

Passive measurement

- TCP dump
- Route table dumps
- Route views
- Single vantage point
  - Connect to one router

- Multiple vantage points

Internet Routing Instability [Labovitz et al '97, 99]

- Methodology
  - Collect routing messages from five public exchange points over nine months
- Problems caused by routing instability
  - Increased delays, packet loss and reordering, time for routes to converge (small-scale route changes)

Labovitz Study: BGP Background

- BGP is an incremental protocol
  - In theory, no update messages in steady state
- Two kinds of update messages
  - Announcement: advertising a new route
  - Withdrawal: withdrawing an old route
- Study saw an alarming number of updates
  - At the time, Internet had around 45,000 prefixes
  - Routers were exchanging 3-6 *million* updates/day
  - Sometimes as high as 30 million in a day
- Placing a very high load on the routers

Route Stability

- Routing instability: rapid fluctuation of network reachability information
- route flapping: when a route is withdrawn and re-announced repeatedly in a short period of time
  - happens via UPDATE messages
- because messages propagate to global Internet, route flapping behavior can cascade and deteriorate routing performance in many places
- Effects: increased packet loss, increased network latency, CPU overhead, loss of connectivity

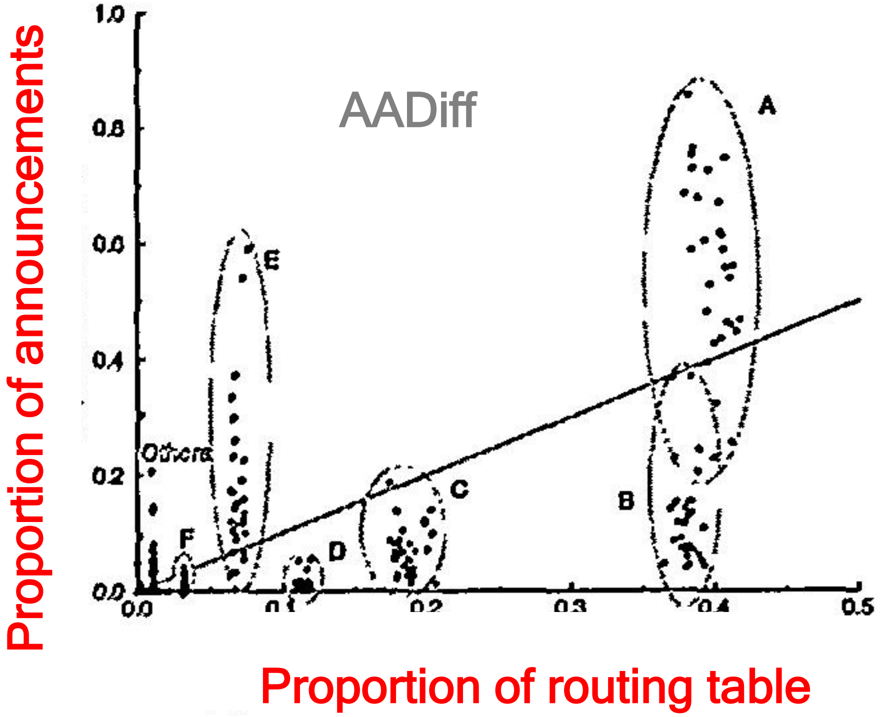
Types of Routing Updates

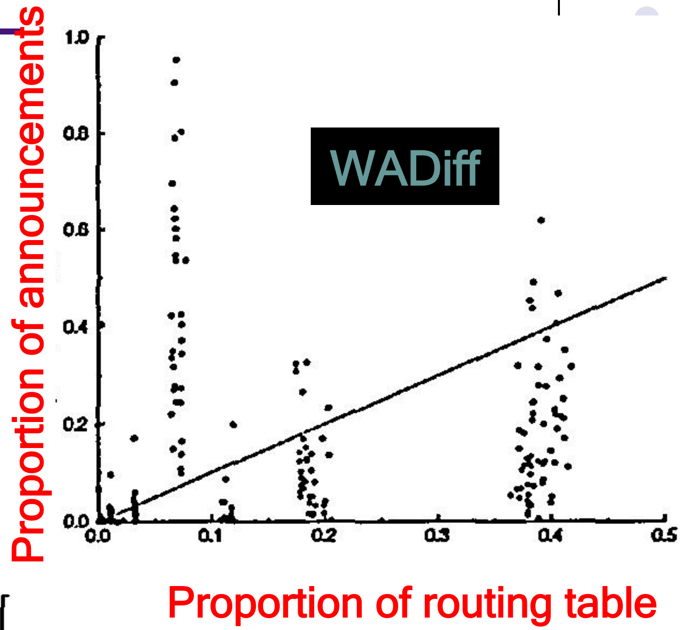
- Forwarding instability
  - reflects legitimate topology changes
  - e.g., changes in Prefix, NEXT_HOP and/or ASPATH
  - affects forwarding paths used
- Policy fluctuation
  - reflects changes in policy
  - e.g., changes in MED, LOCAL_PREF, etc.
  - may not necessarily affect forwarding paths used
- Pathological
  - redundant messages
  - reflect neither topology nor policy changes

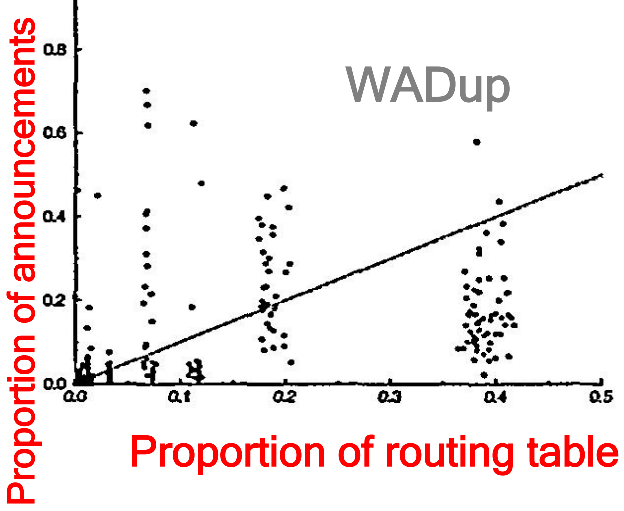
Event class

Name	Type	Result
WADiff	Explicit withdrawal followed by announcement. Replace route with different path	Forwarding instability
AADiff	implicit withdrawal. Same prefix with different path.	Forwarding instability
WADup	Explicit withdrawal followed by announcement. Replace route with same path.	Forwarding instability or pathological
AADup	Announced twice (implicit withdrawal). Replace prefix with same path.	Policy change or pathological
WWDup	Repeated duplicate withdrawals	Pathological

Instability Vs AS size







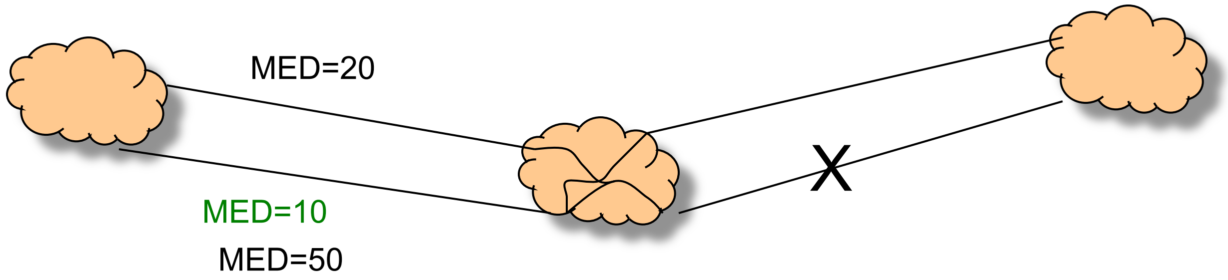
Instability Problem and Cause. Example 1.

- *Problem*: ½ to 6 million duplicate withdrawals
- *Cause*: stateless BGP implementation
  - time-space tradeoff: no state maintained on what advertised to peers
  - when receive any change, transmit withdrawal to all peers regardless of whether previously notified or not
  - O(P*U) updates sent, P is the number of peers
- lately, most vendors have BGP implementations with partial state. Transmit update if the route affects local and peer routers
- Result: number of WWDups reduced by an order of magnitude

Instability Problem and Cause. Example 2

- Problem: duplicate announcements (AADUP)
- Cause: min-advertisement timer & stateless BGP
  - min-adv timer: wait 30 seconds. Combine all received updates in last 30 seconds into single outbound update message (if possible).
  - within 30 seconds route can be withdrawn and re-announced so that there is no net change to original announcement but a dup announcement sent at end of timer expiry
- Solution: Have BGP keep some state about recently sent messages to peers. Avoid sending duplicate messages

Instability Problem and Cause. Example 3



AAdiffs due to MED value oscillation

Example: improper configuration between IGP
And BGP
policy: set MED using local metrics, such as shortest path

Sources of instability

- AS’s
  - No single AS dominates instability statistics
  - No correlation between the size of an AS and its share of updates generated.
  - Correlated to version of the router
- Prefixes
  - Instability is evenly distributed across prefix space

Conclusions

- BGP updates more than one order of magnitude larger than expected
- Routing information dominated by pathological updates (99%)
  - Implementation problems:
    - Routers do not maintain the history of the announcements sent to neighbors
    - When a router gets topological changes they just sent these announcements to all neighbors, irrespective of whether the router sent previous announcements about that route to a neighbor or not
  - Self-synchronization – BGP routers exchange information simultaneously → may lead to link periodic link/router failures
  - Unconstrained routing policies may lead to persistent route oscillations

Conclusions

- Things are getting better (stability)
  - router software and configuration management are maturing
  - increased emphasis on aggregation and route dampening are helping
- Things are getting worse (scalability)
  - multihoming is still growing (problems but useful)
    - Works against aggregation
  - internet topology growing less hierarchical
  - Too many configuration parameters

Questions on internet topology

- What does the internet connectivity graph look like?
  - Stub, tree, mesh
- What are the patterns
  - Invariants, growth (edges, core)
- Growth trends
  - Scale
- Realistic topologies for simulation
  - Is it random? Uniformly distributed links realistic?

Why do we need the topology?

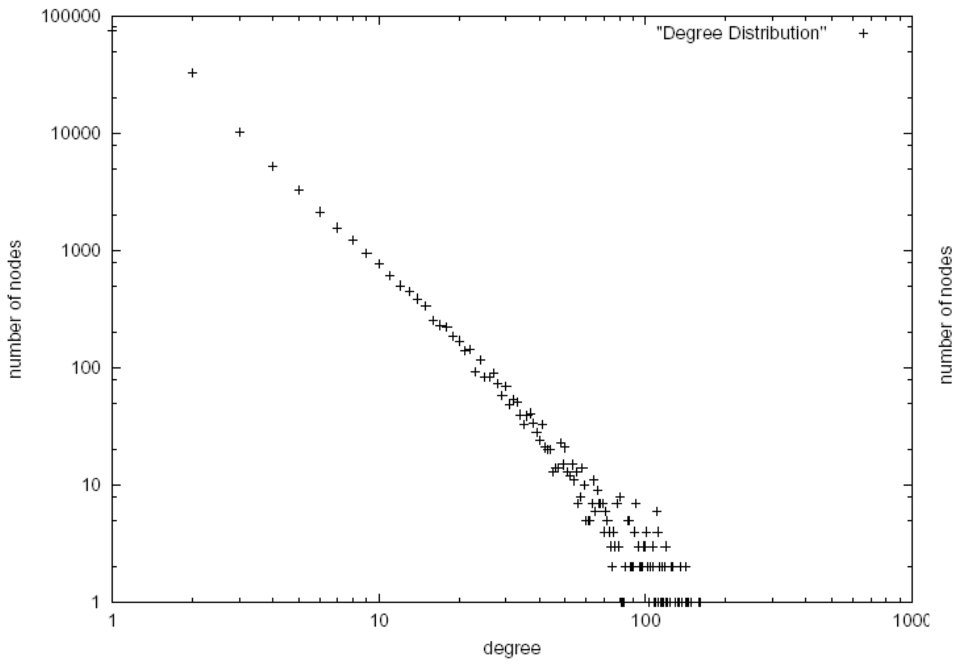
- Understand the macroscopic properties of the Internet physical structure
- Network management
- Topology-aware algorithms
- Simulation and topology generation tools

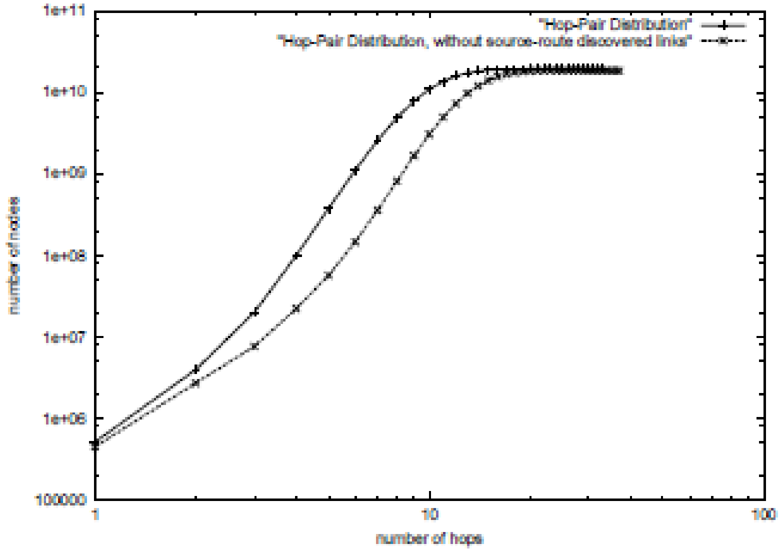
Active measurement

- Use Hop-limited probes from a single vantage point
- Informed random probing
  - Probe multiple IP addresses, adjacent prefixes
- Use Source Routing (specify intermediate nodes)
  - An option in the IP packet that forces routes to use the IP addresses specified in the packet
- Degree distribution
- Hop-pair distribution

Heuristics for internet map discovery,Infocom 2000

Degree, hop-pair distribution





(b) The hop-pair distribution on a log-log scale

Power-Laws and Internet

- What can be said about the structure of the internet?
- Average out-degree
- Average # of nodes within K hops
- Power law
- $y = k\,x^{\alpha}$ $\log(y) = \log(k) + \alpha \log(x)$
- On a log log scale the function is linear

On power law relationships of the internet topology-SIGCOMM2009

Power-laws

- Found in many different areas
- Economics (wealth distribution)
- File sizes, process (thread) execution times
- Sociology
- Probability distribution (pareto, weibull, log normal)
- Intuition: Large occurrences are very rare and small occurrences are very common

March 16: WSJ article

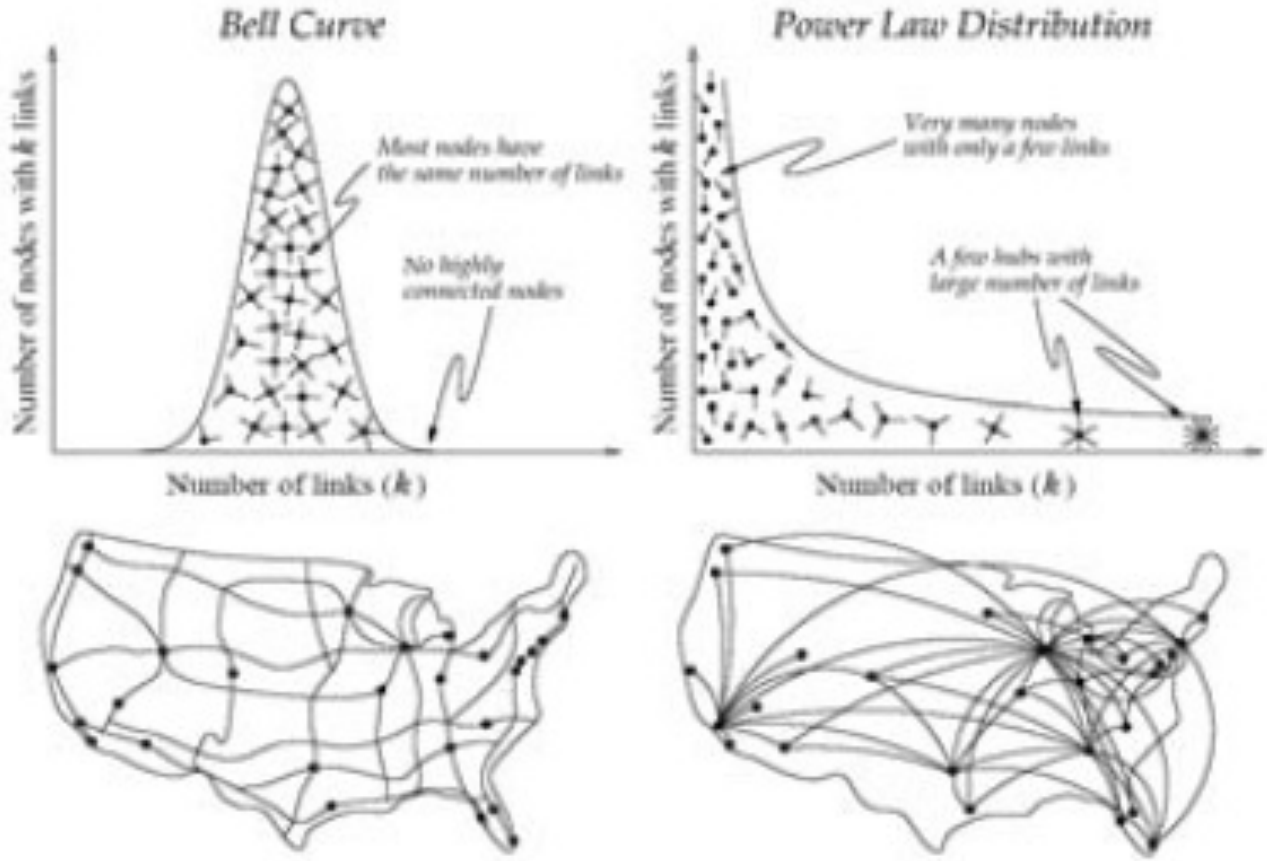
- Why the industry is evolving this way is rooted in balance sheets. Over the past two years, Apple Inc., Oracle Corp., Google Inc., Microsoft Corp. and six other large tech companies have generated $68.5 billion in new cash, compared with just $13.5 billion for the other 65 tech companies in the S&P 500 Index combined, according to a Wall Street Journal analysis of data provided by Capital IQ.

World’s richest people

No.	Name	Net worth (USD)	Age	Citize
1 ▲	Carlos Slim Helú and family	$53.5 billion ▲	70	Me
2 ▼	William Henry Gates III	$53.0 billion ▲	54	Un
3 ▼	Warren Buffett	$47.0 billion ▲	79	Un
4 ▲	Mukesh Ambani	$29.0 billion ▲	52	Inc
5 ▲	Lakshmi Mittal	$28.7 billion ▲	59	Inc
6 ▼	Lawrence Ellison	$28.0 billion ▲	65	Un
7 ▲	Bernard Arnault	$27.5 billion ▲	61	Fra
8 ▲	Eike Batista	$27.0 billion ▲	52	Br
9 ▲	Amancio Ortega	$25.0 billion ▲	74	Sp
10 ▼	Karl Albrecht	$23.5 billion ▲	90	Ge

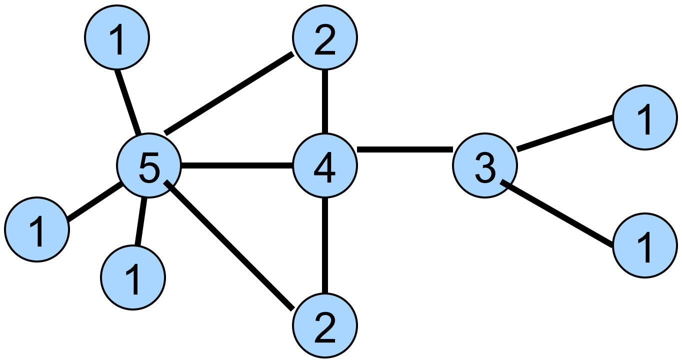
Random

Scale Free



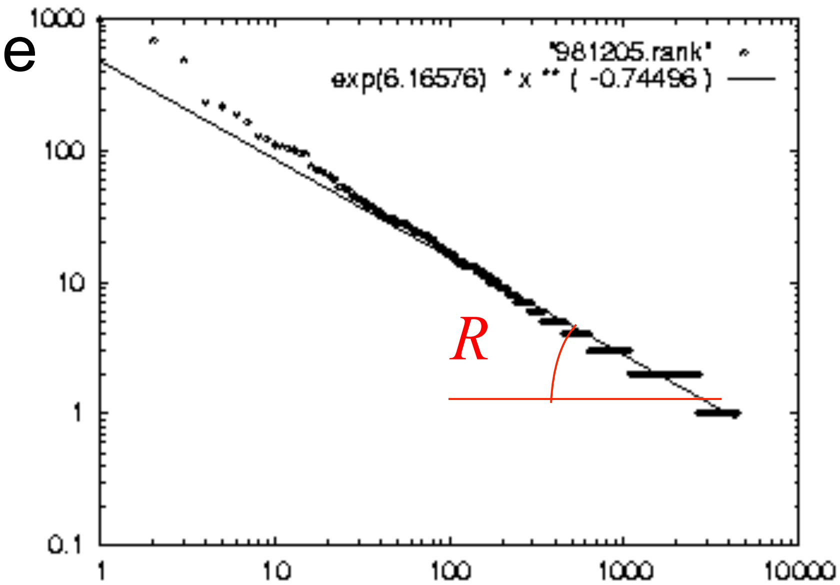
Power law 1: rank exponent

- Study out-degree of nodes
- Definition rv = The ***rank*, *rv*** ,of a node v is its index in the order of decreasing out-degree.



5	4	3	2	2	1	1	1	1
1	2	3	4	5	6	7	8	9

Power-law 1 : rank *R*



outdegree

Exponent = slope

R = -0.74

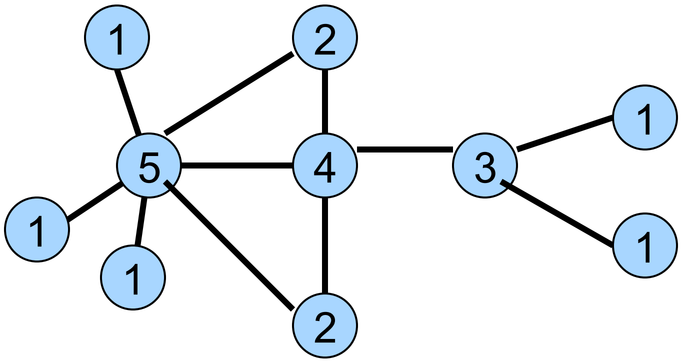
Dec’ 98

Rank: nodes in decreasing outdegree order

- The plot is a line in log-log scale
- Rank exponent R is the slope of the line deg vs rank

Power law 2: Outdegree O

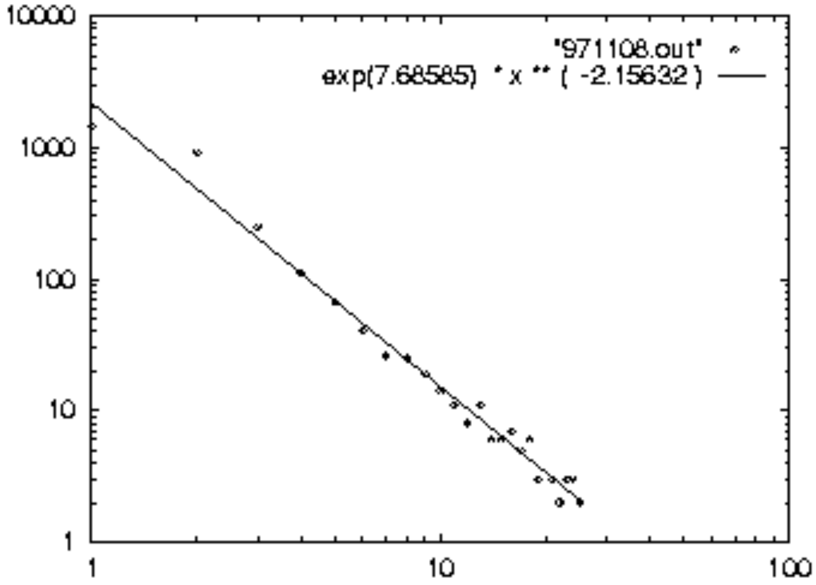
- Distribution of out-degree of the graph
- The ***frequency of an out-degree d, fd*** , is the number of nodes with out-degree d.



fd	5	2	1	1	1	0	0	0
Out-degree	1	2	3	4	5	6	7	8

Power-law 2: outdegree O

Frequency



Outdegree

Exponent = slope

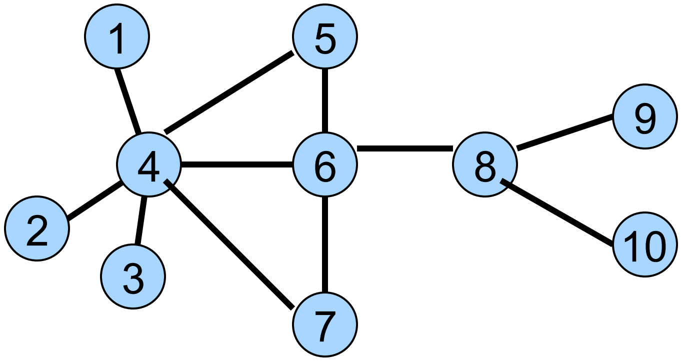
O = -2.15

Nov' 97

- The plot is linear in log-log scale
- Out degree exponent O is the slope of f vs O

Power law 3: hop-plot

- Distribution of number of nodes within h hops
- Include self pairs (i,j is pair so is j,i)



P(h)	10	10+22=32	32+38=70	70+18=88	88+12=100	0	0	0
hop	0	1	2	3	4	5	6	7

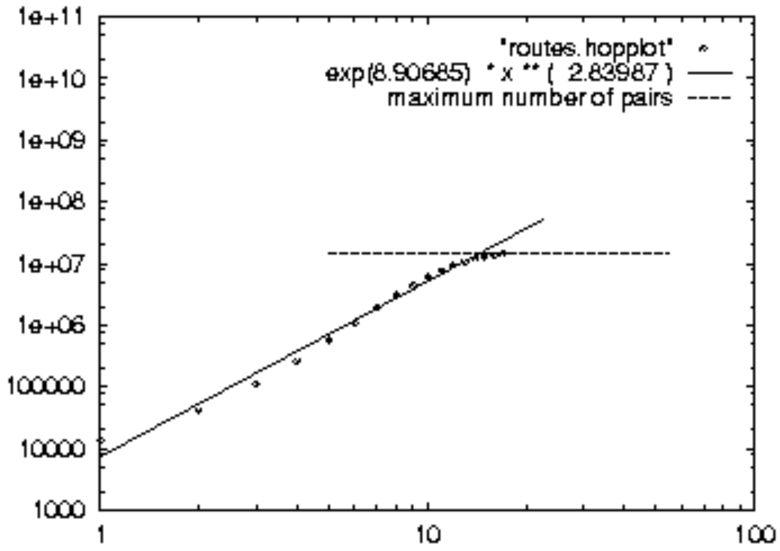
Power law 3- hop-lot

- Hop-Plot Exponent *H*: The total number of pairs of nodes *P(h)* within h hops, is proportional to the number of hops to the power of a constant, H, δ is the diameter of the graph

- $P(h) \propto h^{H}, \qquad h \ll \delta$

Power-law 3: hopplot *H*

No. of Pairs $H = 2.83$



Hops Router level ’95

- Pairs of reachable nodes as a function of hops

  Simpler: neighborhood size vs hops

Difficulties in topology discovery

- Shared media
- Backup links
- Router Identification and annotation
- Alias resolution
- Completeness Validation

Currently, none of them is completely solved!

Active perturbation

- Inject traffic, route announcements
- Inject faults
- Bring down machines/routers for predefined time intervals
- Study the affect
- See how the network behaves
- BGP beacon

What is a BGP Beacon?

- An unused, globally visible prefix with **known** Announce/Withdrawal schedule
- For long-term, public use for analyzing the behavior of the BGP