

CS 268: Lecture 9 (Routing Behavior in the Internet)

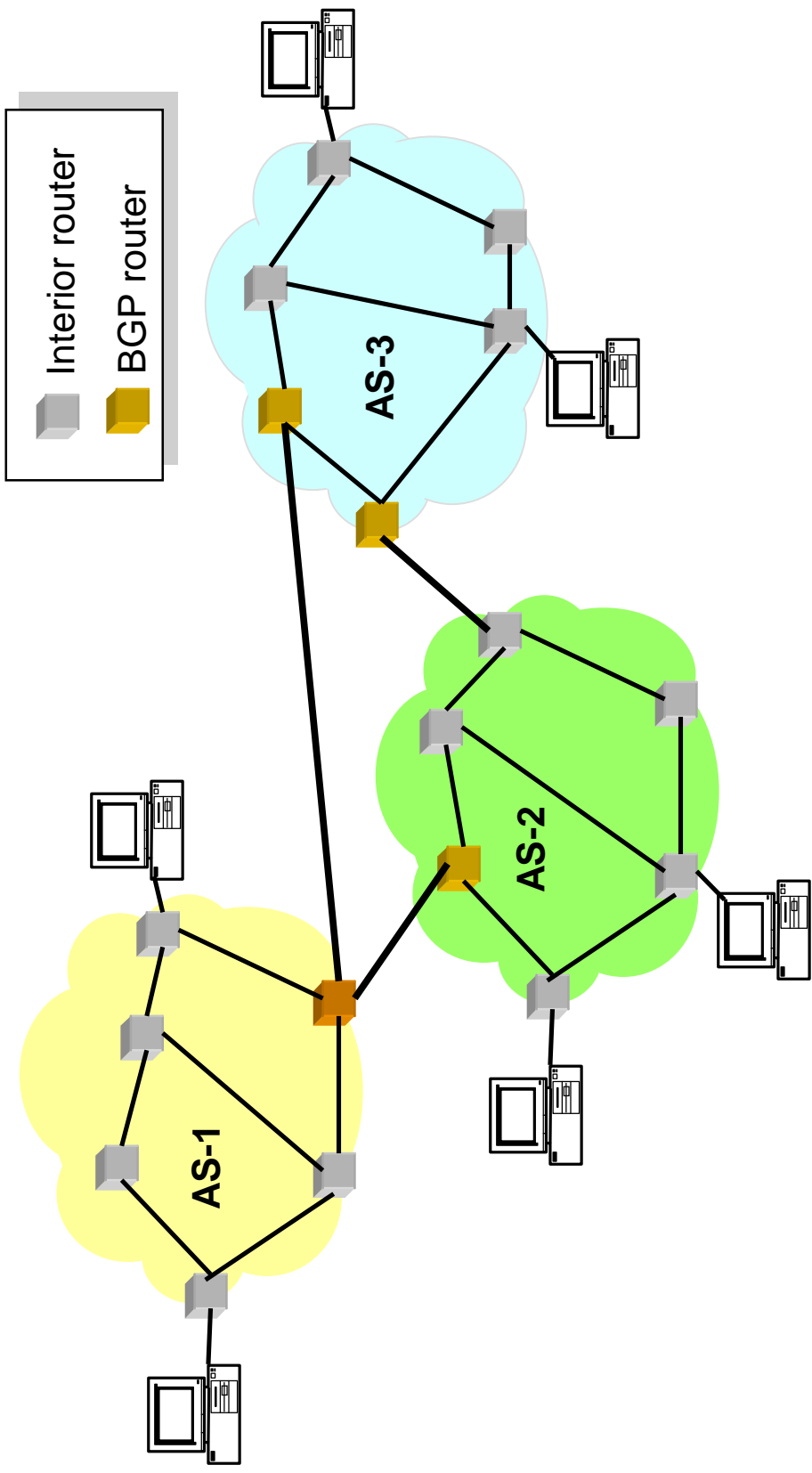
Ion Stoica

February 25, 2002

Internet Routing

- Internet organized as a **two** level hierarchy
- First level – autonomous systems (AS's)
 - AS – region of network under a single administrative domain
- AS's run an intra-domain routing protocols
 - Distance Vector, e.g., RIP
 - Link State, e.g., OSPF
- Between AS's runs inter-domain routing protocols, e.g., Border Gateway Routing (BGP)
 - De facto standard today, BGP-4

Example



Intra-domain Routing Protocols

- Based on unreliable datagram delivery
- Distance vector
 - Routing Information Protocol (RIP), based on Bellman-Ford
 - Each neighbor periodically exchange reachability information to its neighbors
 - Minimal communication overhead, but it takes long to converge, i.e., in proportion to the maximum path length
- Link state
 - Open Shortest Path First Protocol (OSPF), based on Dijkstra
 - Each network periodically **floods** immediate reachability information to other routers
 - Fast convergence, but high communication and computation overhead

Inter-domain Routing

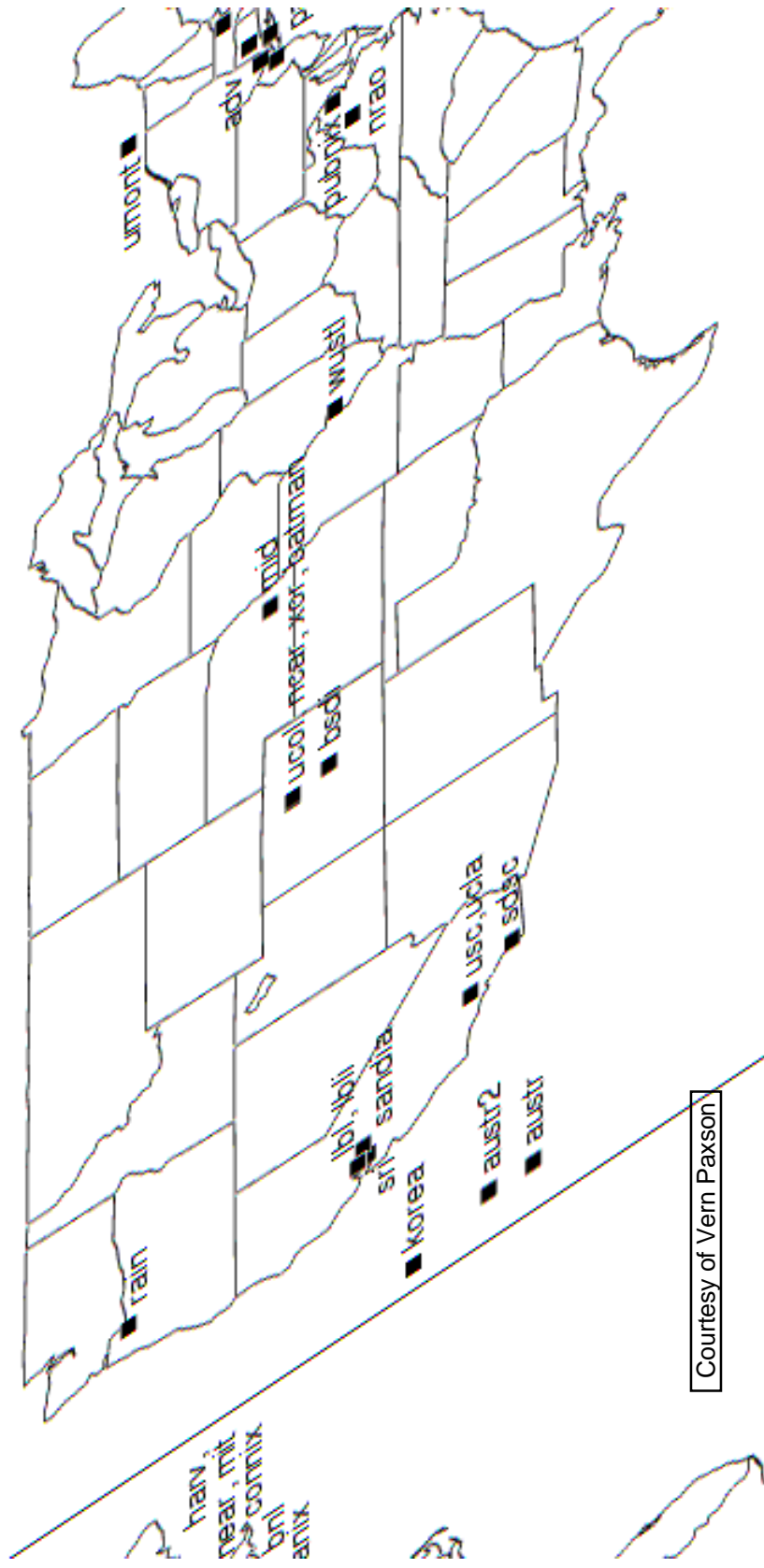
- Use TCP
- Border Gateway Protocol (BGP), based on Bellman-Ford path vector
- AS's exchange reachability information through their BGP routers, **only** when routes change
- BGP routing information – a sequence of AS's indicating the path traversed by a route; next hop
- General operations of a BGP router:
 - Learns multiple paths
 - Picks best path according to its AS **policies**
 - Install best pick in IP forwarding tables

End-to-End Routing Behavior in the Internet [Paxson '95]

- Idea: use end-to-end measurements to determine
 - Route pathologies
 - Route stability
 - Route symmetry

Methodology

- Run Network Probes Daemon (NPD) on a large number of Internet 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 $A \rightarrow B$ and then $B \rightarrow A$

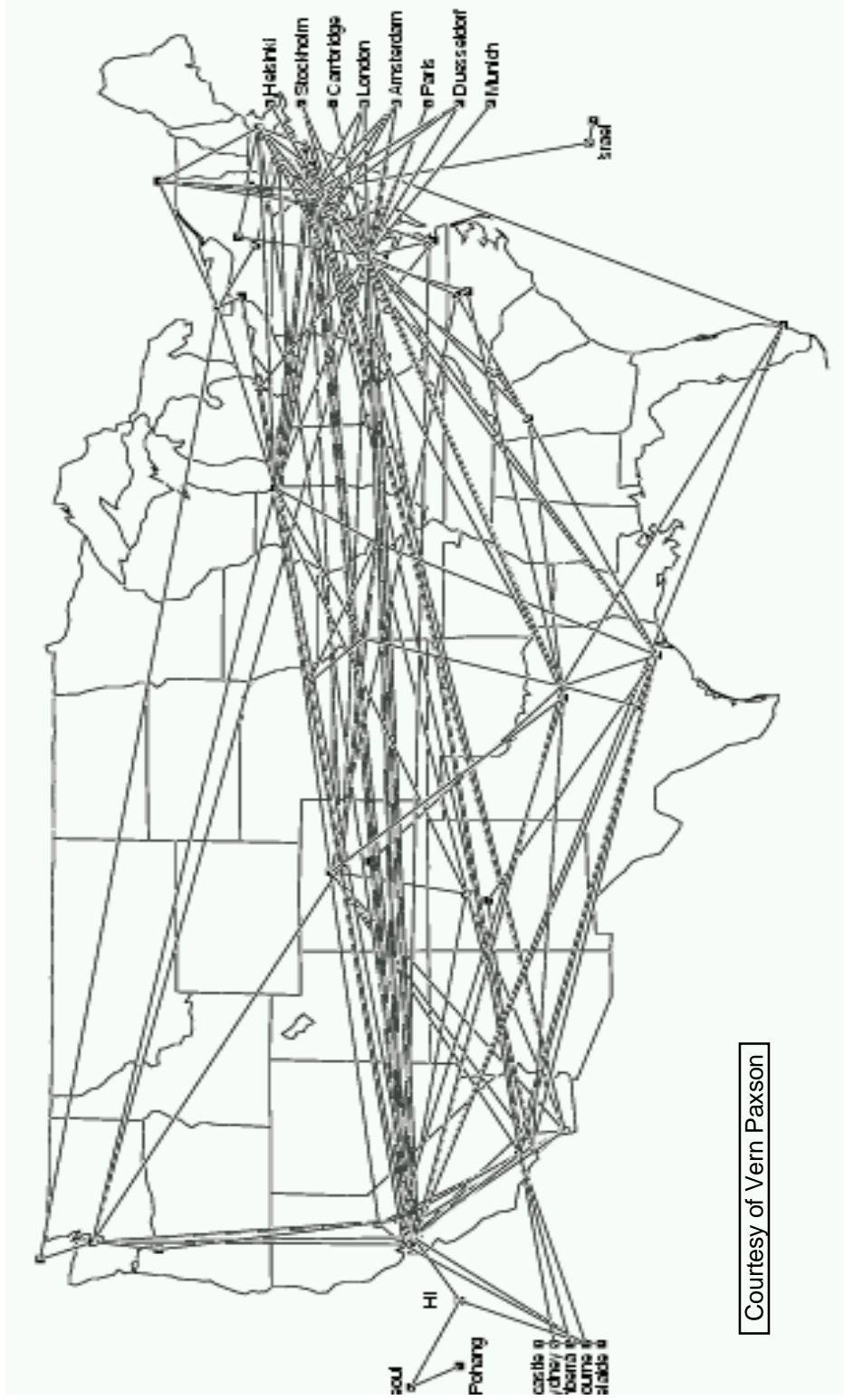
Traceroute Example

```
sky.cs.berkeley.edu → whistler.cmcl.cs.cmu.edu
```

```
eroute to whistler.cmcl.cs.cmu.edu (128.2.181.87), 30 hops max, 38 byte packets
snr45 (128.32.45.1) 0.570 ms 0.434 ms 0.415 ms
jig10-cnr1.EECS.Berkeley.EDU (169.229.3.65) 0.506 ms 0.513 ms 0.434 ms
jigE5-0-0.inr-210-cory.Berkeley.EDU (169.229.1.45) 0.726 ms 0.570 ms 0.553 ms
ast1-0-0.inr-001-eva.Berkeley.EDU (128.32.0.1) 1.357 ms 0.998 ms 1.020 ms
cos0-0.inr-000-eva.Berkeley.EDU (128.32.0.65) 1.459 ms 2.371 ms 1.600 ms
cos3-0.c2-berk-gsr.Berkeley.EDU (128.32.0.90) 3.103 ms 1.406 ms 1.575 ms
SUNV--BERK.POS.calren2.net (198.32.249.14) 3.005 ms 3.085 ms 2.407 ms
abilene--QSV.POS.calren2.net (198.32.249.62) 6.112 ms 6.834 ms 6.218 ms
dnvr-scrm.abilene.ucaid.edu (198.32.8.2) 34.213 ms 27.145 ms 27.368 ms
kscy-dnvr.abilene.ucaid.edu (198.32.8.14) 38.403 ms 38.121 ms 38.514 ms
ipls-kscy.abilene.ucaid.edu (198.32.8.6) 47.855 ms 47.558 ms 47.649 ms
clev-ipls.abilene.ucaid.edu (198.32.8.26) 54.037 ms 53.849 ms 53.492 ms
abilene.psc.net (192.88.115.122) 57.109 ms 56.706 ms 57.343 ms
cmu.psc.net (198.32.224.36) 58.794 ms 58.237 ms 58.491 ms
CS-VLAN255.GW.CMU.NET (128.2.255.209) 58.072 ms 58.496 ms 57.747 ms
WHISTLER.CMCL.CS.CMU.EDU (128.2.181.87) 57.715 ms 57.932 ms 57.557 ms
```

Methodology

- Links traversed during D_1 and D_2



Courtesy of Vern Paxson

Methodology

- Exponential sampling
 - Unbiased sampling – measures instantaneous signal with equal probability
 - PASTA principle – Poisson Arrivals See Time Averages
- Is data representative?
 - Argue that sampled AS's are on half of the Internet routes
- Confidence intervals for probability that an event occurs

Limitations

- Just a small subset of Internet paths
- Just two points at a time
- Difficult to say **why** is something happened, only with end-to-end measurements
- 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 mead-stream
- Temporary outages (> 30 sec)

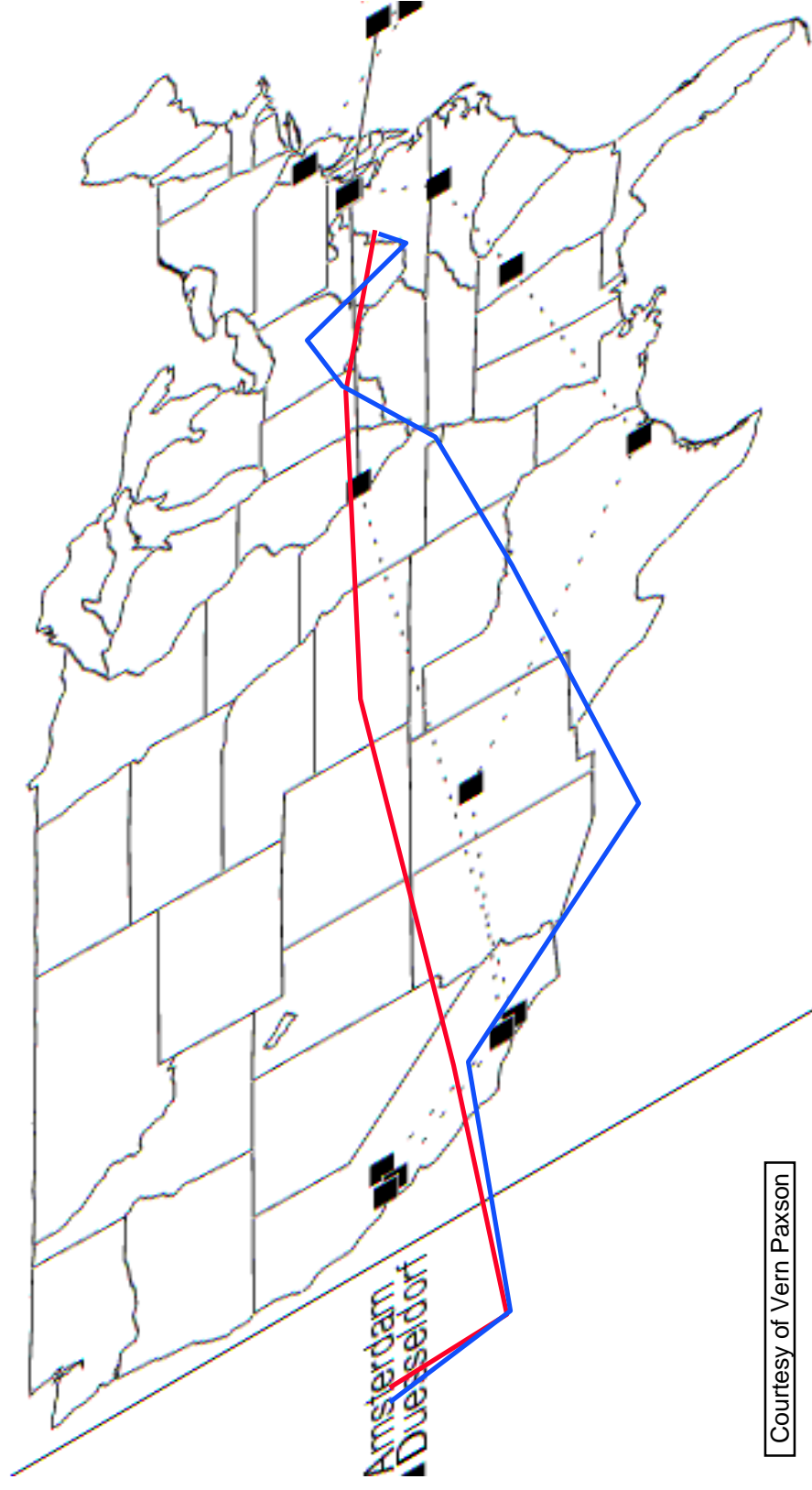
Routing Loops & Erroneous Routing

- 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

Route Changes

- Connectivity change in mid-stream (10 in D_1 and 155 in D_2)
 - Route changes during measurements
 - Recovering bimodal: (1) 100's msec to seconds; (2) order of minutes
- Route fluttering
 - Rapid route oscillation

Example of Route Fluttering



Courtesy of Vern Paxson

Problems with Fluttering

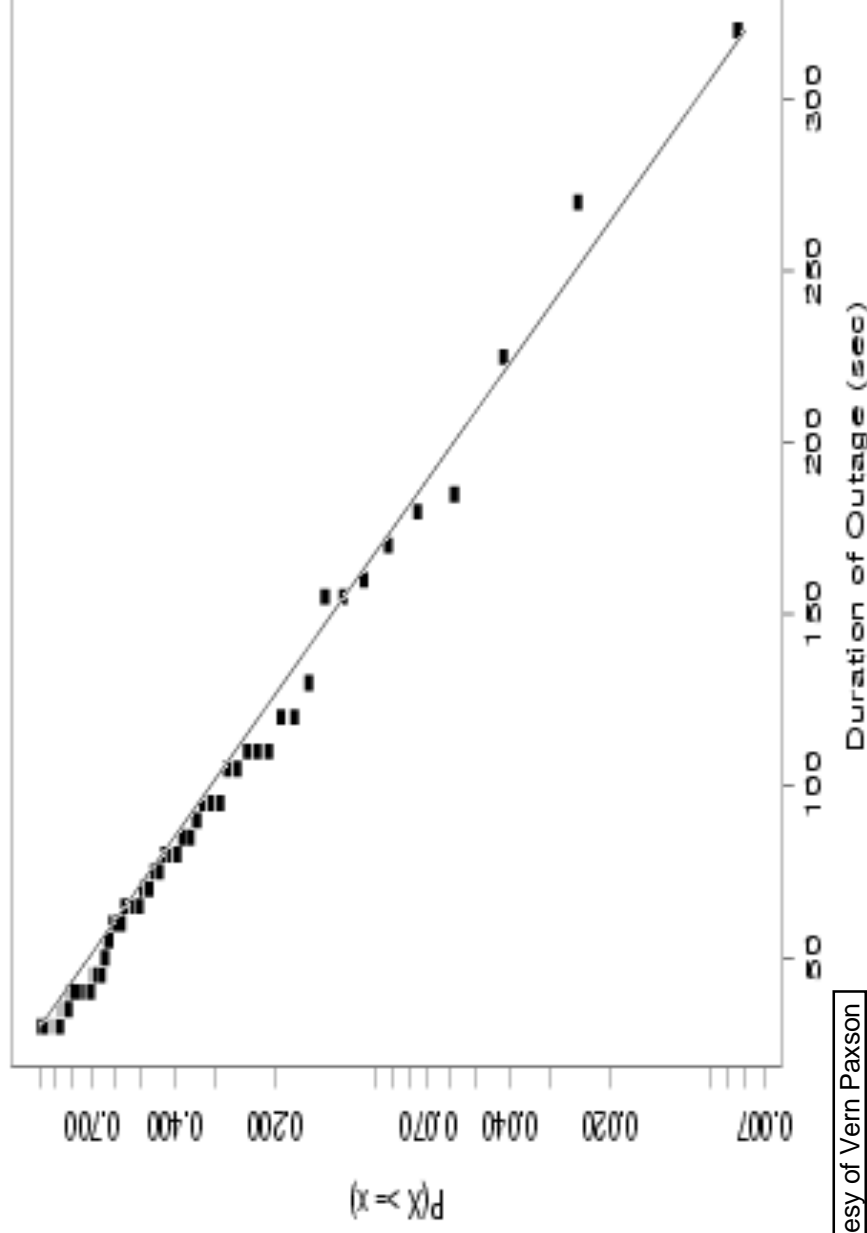
- Path properties difficult to predict
 - This confuses RTT estimation in TCP, may trigger false retransmission timeouts
- Packet reordering
 - TCP receiver generates DUPACK's, may trigger spurious fast retransmits
- These problems are bad only for large scale flutter; for localized flutter is usually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

Infrastructure Failures

- NPD's unreachable due to many hops (6 in D_2)
 - Unreachable → more than 30 hops
 - Path length not necessary correlated with distance
 - 1500 km end-to-end route of 3 hops
 - 3 km (MIT – Harvard) end-to-end route of 11 hops
- Temporary outages
 - Multiple probes lost. Most likely due to:
 - Heavy congestions lasting 10's of seconds
 - Temporary lost of connectivity

Distribution of Long Outages (> 30 sec)

- Geometric distribution



Courtesy of Vern Paxson

istoica@cs.berkeley.edu

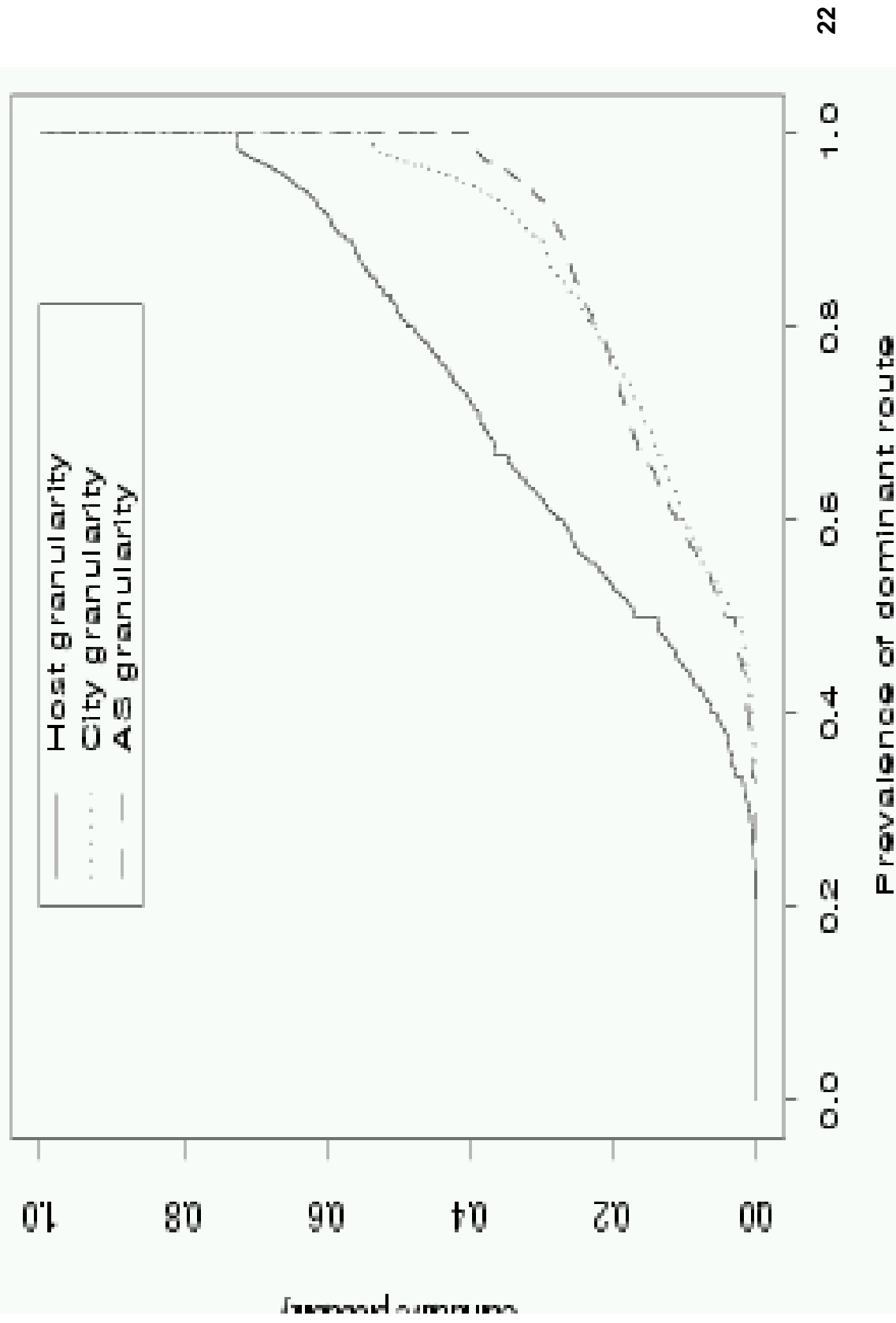
Pathology Summary

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

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
 - 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
 - Conclusion: routing changes occur over a wide range of time scales, i.e., from minutes to days

Route Prevalence



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 Symmetry

- 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

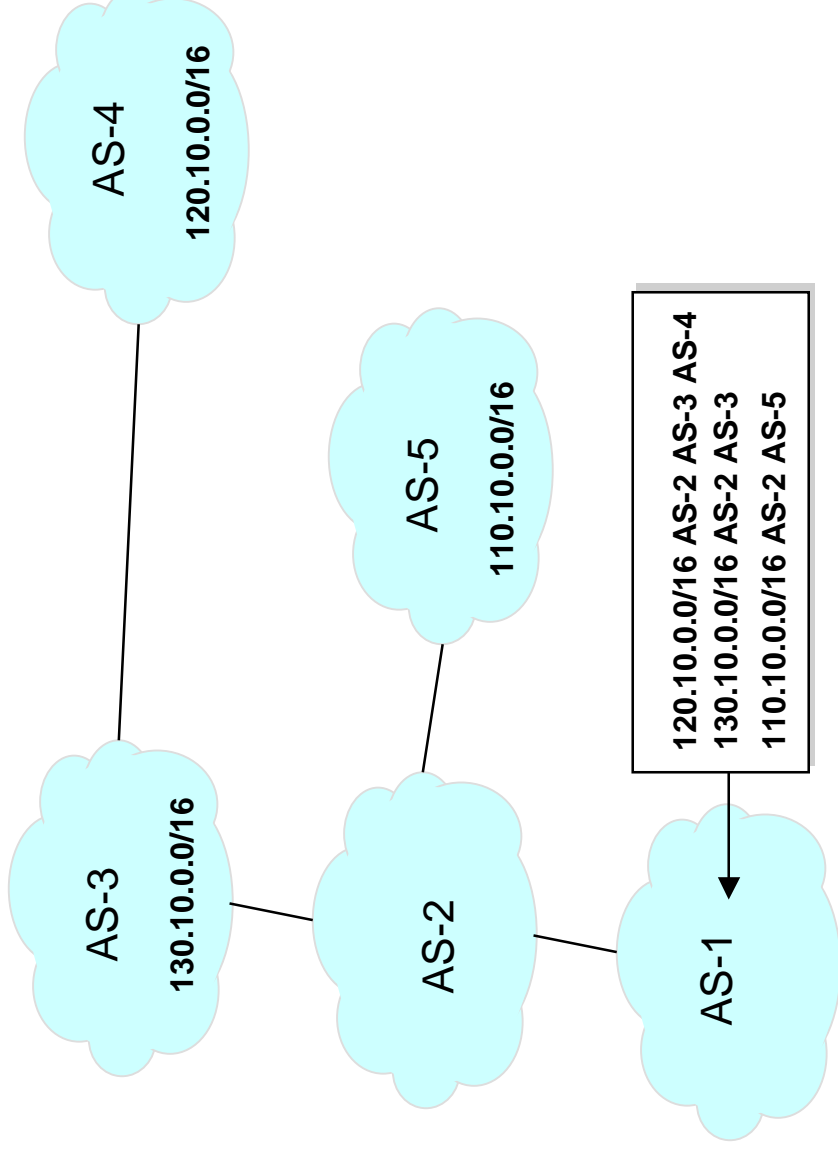
Internet Routing Instability

[Labovitz et al '96]

- 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)
- Relevant BGP information
 - AS-Path (see next slide)
 - Next hop: Next hop to reach a network
- Two routes are the same if they have the same AS-Path and Next hop

AS-Path

- Sequence of AS's a route traverses
- Used for loop detection and to apply policy



BGP Information Exchange

- Announcements: a router has either
 - Learned of a new route, or
 - Made a policy decision that it prefers a new route
- Withdrawals: a router concludes that a network is no longer reachable
 - Explicit: associated to the withdrawal message
 - Implicit: when a route is replaced as a result of an announcement message
- In steady state BGP updates should be the result **only** of infrequent policy changes
 - Update rate → measure of network stability

Types of Inter-domain Routing Updates

- Forwarding instability: may reflect topology changes
- Policy fluctuations (Routing instability): may reflect changes in routing policy information
- Pathological updates: redundant updates that are neither routing nor forwarding instability
- Instability: forwarding instability and policy fluctuation → change forwarding path

Routing Successive Events (Instability)

- WADiff: a route is explicitly withdrawn as it becomes unreachable, and is later replaced with an alternative route (forwarding instability)
- AADiff: a route is implicitly withdrawn and replaced by an alternative route as the original route becomes unavailable or a new preferred route becomes available (forwarding instability)
- WADup: a route is explicitly withdrawn, and reannounced later (forwarding instability or pathological behavior)

Routing Successive Events (Pathological Instability)

- AADup: A route is implicitly withdrawn and replaced with a duplicate of the original route (pathological behavior or policy fluctuation)
- WWWDup: The repeated transmission of BGP withdrawals for a prefix that is currently unreachable (pathological behavior)

Findings

- BGP updates more than one order of magnitude larger than expected
- Routing information dominated by pathological updates
 - Implementation problems:
 - Routers do not maintain the history of the announcements sent to neighbors
 - When a router gets topological changes they just send 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

Findings

- Instability and redundant updates exhibits strong correlation with load (30 seconds, 24 hours and seven days periods)
 - Overloaded routers fail to respond an their neighbors withdrawn them
- Instability usually exhibits high frequency
- Pathological updates exhibits both high and low frequencies
- No single AS dominates instability statistics
- No correlation between size of AS and its impact on instability statistics
- There is no small set of paths that dominate instability statistics

Conclusions

- Routing in the Internet exhibits many undesirable behaviors
 - Instability over a wide range of time scales
 - Asymmetric routes
 - Network outages
 - Problem seems to worsen
- Many problems are due to software bugs or inefficient router architectures

Lessons

- Even after decades of experience routing in the Internet is not a solved problem
- This attests the difficulty and complexity of building distributed algorithm in the Internet, i.e., in a heterogeneous environment with products from various vendors
- Simple protocols may increase the chance to be
 - Understood
 - Implemented right