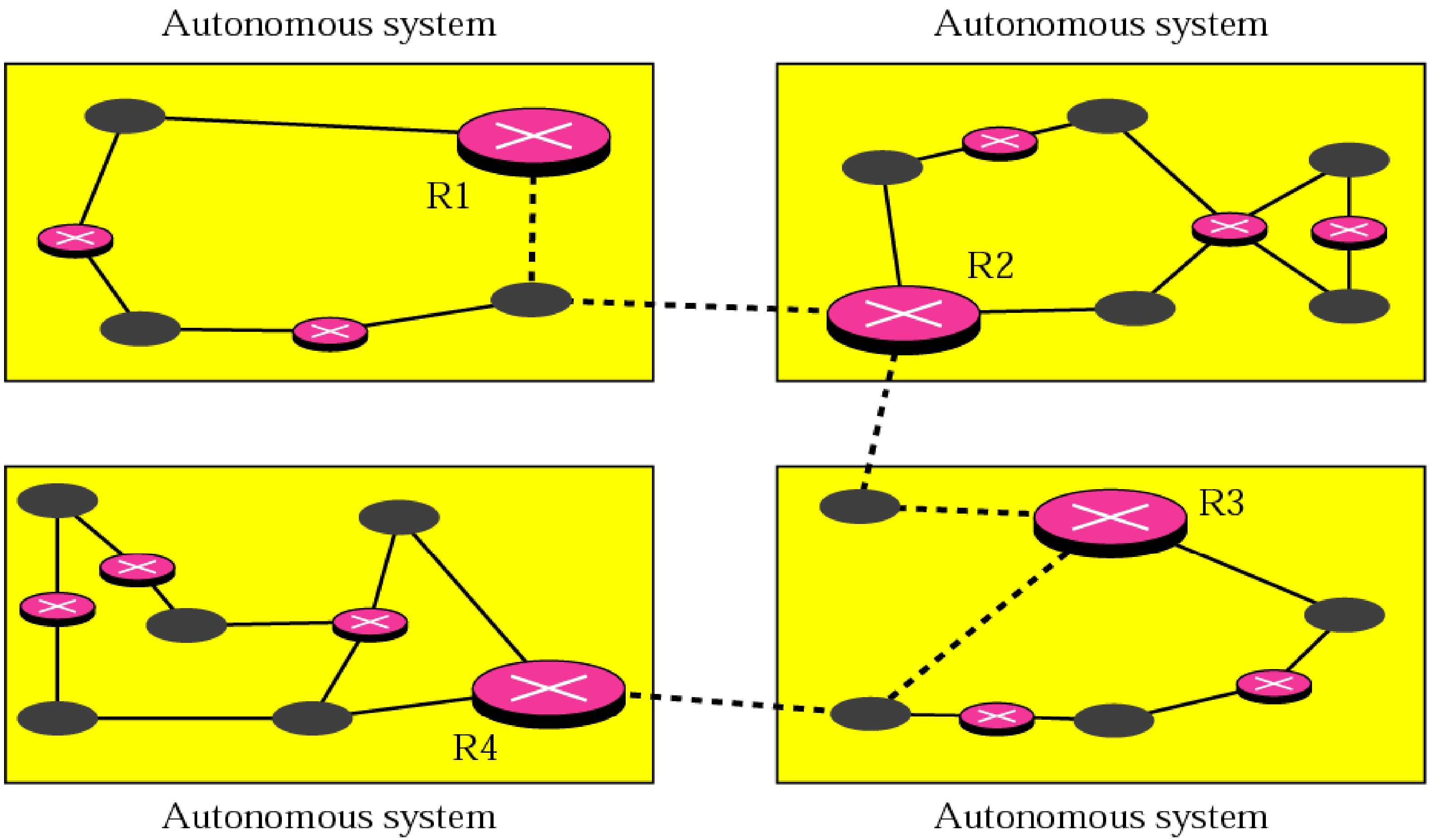
CS 5229

Routing

Dr. Chan Mun Choon
School of Computing
National University of Singapore

Autonomous systems



Vern Paxson, "End-to-End Routing Behavior in the Internet," Transaction of Networking, 1997.

Overview

- Measurements done at 1994, 1995
- Do not have access to routers, only end-to-end measurements
  - How to extract information from fairly “sparse” measurement data?
- Internet has changed substantially, but it is still good to know what was discovered then
- Even more relevant, what were the questions asked?

Measurements

- Run traceroute from different hosts
- Two data sets
  - D1: mean measurement intervals of 1-2 days (27 sites, Nov - Dec 1994)
  - D2: 60% with mean intervals of 2 hours, and 40% with mean intervals of 2.75 days (33 sites, Nov - Dec 1995)
- Measurement intervals are exponentially distributed, why?

Enough Data?

- In 1994 – 1995, there are
  - 6.6 million hosts
  - 1000 actives Ass
- Is the data collected sufficient?
- How the author argue:
  - With N sites/hosts, there are N^2 pairs
  - Traverse 8% of additional ASs
  - If AS is weighted by likelihood that it appears on the path, coverage is 50%,
- Due to difficulties in data collection, disconnection may be underestimated

Two Main Questions

1. What kinds of routing problems can be observed (routing pathologies)
2. Routing path characteristics

Loops

- Three kinds of loops:
  - Forwarding loop
  - Information loop
  - Traceroute loop
- Only traceroute loop can be directly observed
  - If the same router sequence appears 3 or more times, it is considered a forwarding loop

Loops

- Loops can persist for
  - Under 3 hours
  - > 0.5 days
- Some loops have observed to last for 14-17hrs, 16-32 hours
- Loops may come in geographical clusters

Route Recovery

- Route recovery occurs when traceroute is being performed
- Recovery time is bimodal:
  - Less than 1 sec
  - Minutes
- Guess:
  - Short recovery is due to new routes available
  - Longer recovery is due to route repair

Fluttering/Oscillation

- Why is it bad?
- Cause: load balancing

Reachability

- Receive “host unreachable” message
- In D1, 0.21%
- In D2, 0.5%

Hop Count

- traceroute stops probing when the number of hops exceed 30
- In D1, mean hop count is 15.6
  - 30 is sufficient for all measurements
- In D2, mean hop count is 16.2
  - 6 measurements fail
- Hop count is not necessary related to geographic distance
  - 3 hops for a 1,500km route
  - 11 hops for a 3km route

Outage Duration

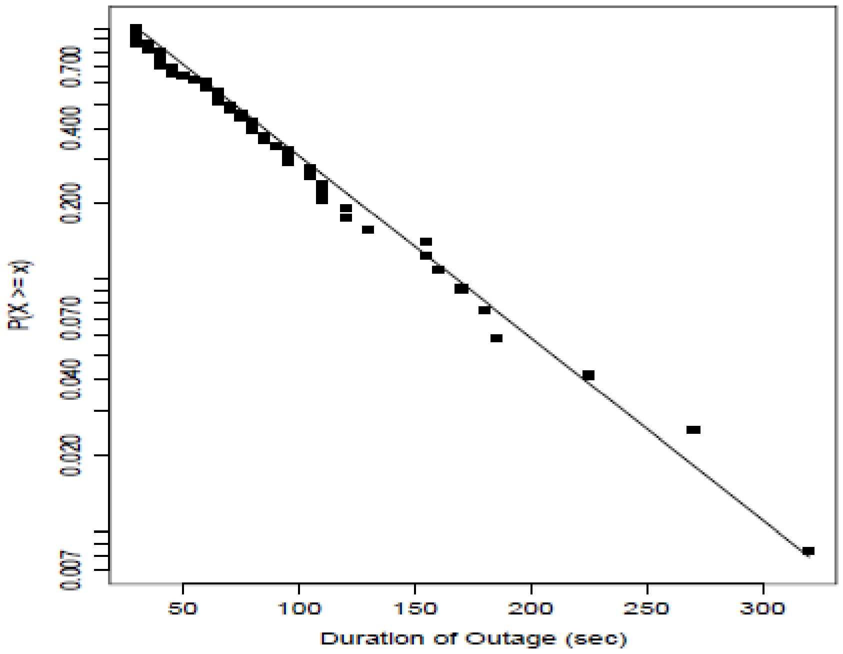


Figure 4: Distribution of long $\mathcal{D}_1$ outages

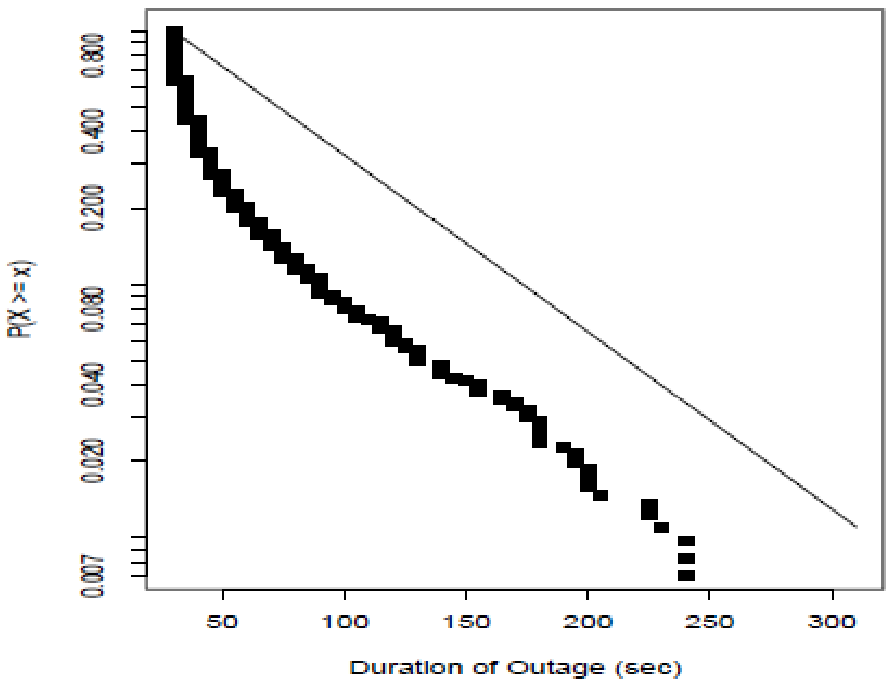


Figure 5: Distribution of long $\mathcal{D}_2$ outages

Time of Day

- Most common outage duration is 30s
  - Occurs most often during high traffic period (3pm to 4pm)
  - Less common during low traffic period (1am – 2am)
- Outage of a longer duration, probably due to node failure:
  - Most common during 3pm to 4pm
  - Second common during 6am to 7am (why?)
  - Least common during 9am to 10am

Routing Stability

- Two definitions
  - Prevalence
  - Persistent
- Example

Reducing the data

- Some level of aggregation may be helpful
- Three levels:
  - Host
  - City
  - AS
- Why is this useful?

Prevalence

- Since measurements is based on sampling, prevalence can be estimated directly

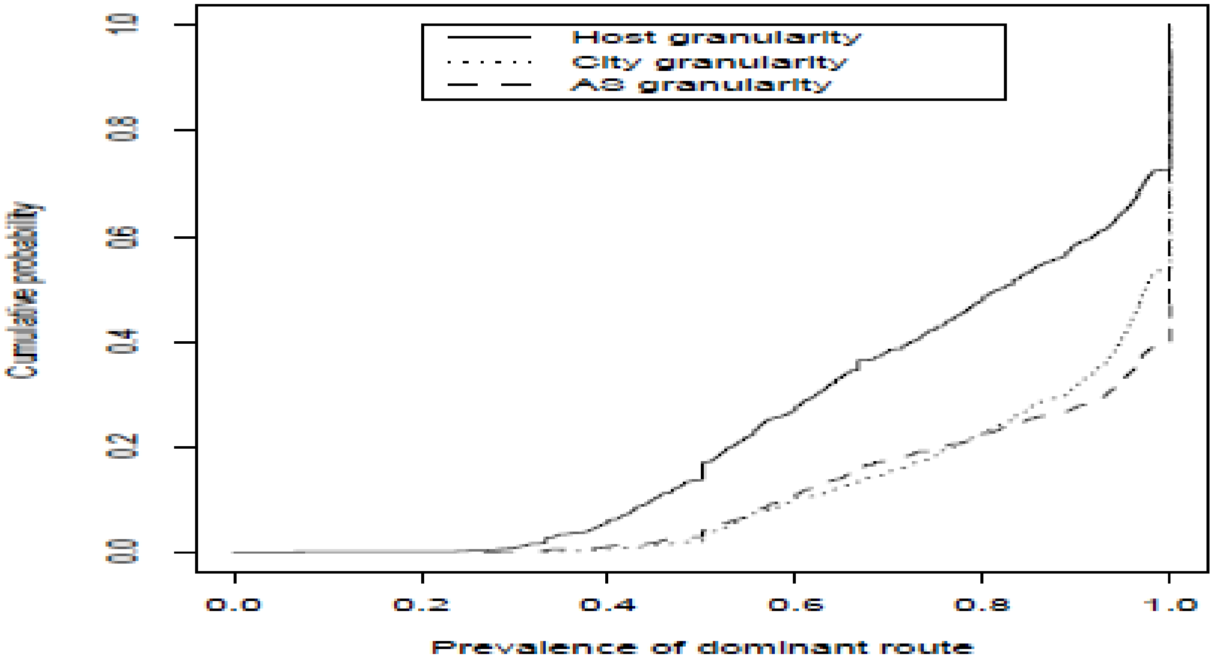


Figure 6: Fraction of observations finding the dominant route, for all virtual paths, at all granularities

Persistence

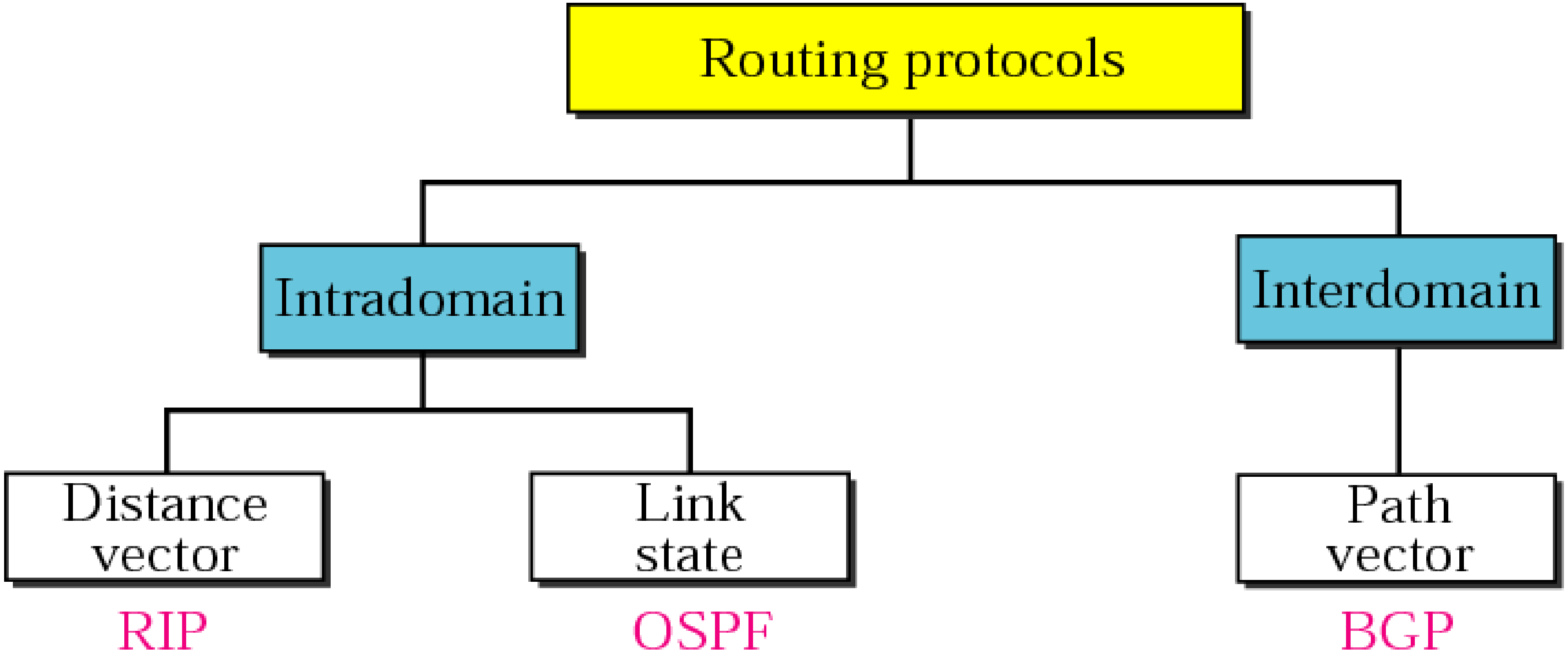
- Difficult to estimate using measurement obtained (why?)

Time scale	%	Notes
seconds	N/A	"Flutter" for purposes of load balancing. Treated separately, as a pathology, and not included in the analysis of persistence.
minutes	N/A	"Tightly-coupled routers." We identified five instances, which we merged into single routers for the remainder of the analysis.
10's of minutes	9%	Frequent route changes inside the network. In some cases involved routing through different cities or AS's.
hours	4%	Usually intra-network changes.
6+ hours	19%	Also intra-network changes.
days	68%	Two regions. 50% of routes persist for under 7 days. The remaining 50% account for 90% of the total route lifetimes.

Table 3: Summary of persistence at different time scales

Brief Revision on Internet Routing

Routing protocol classification



Internet inter-AS routing: BGP

- BGP (Border Gateway Protocol): *the* de facto standard
- BGP provides each AS a means to:
  1. Obtain subnet reachability information from neighboring ASs.
  2. Propagate the reachability information to all routers internal to the AS.
  3. Determine "good" routes to subnets based on reachability information and policy.
- Allows a subnet to advertise its existence to rest of the Internet: *"I am here"*

Why different Intra/Inter-AS routing ?

Policy:

- Inter-AS: admin wants control over how its traffic routed, who routes through its net – policy based routing.
- Intra-AS: single admin, so no policy decisions needed

Scale:

- hierarchical routing saves table size, reduced update traffic

Performance:

- Intra-AS: can focus on performance
- Inter-AS: policy may dominate over performance

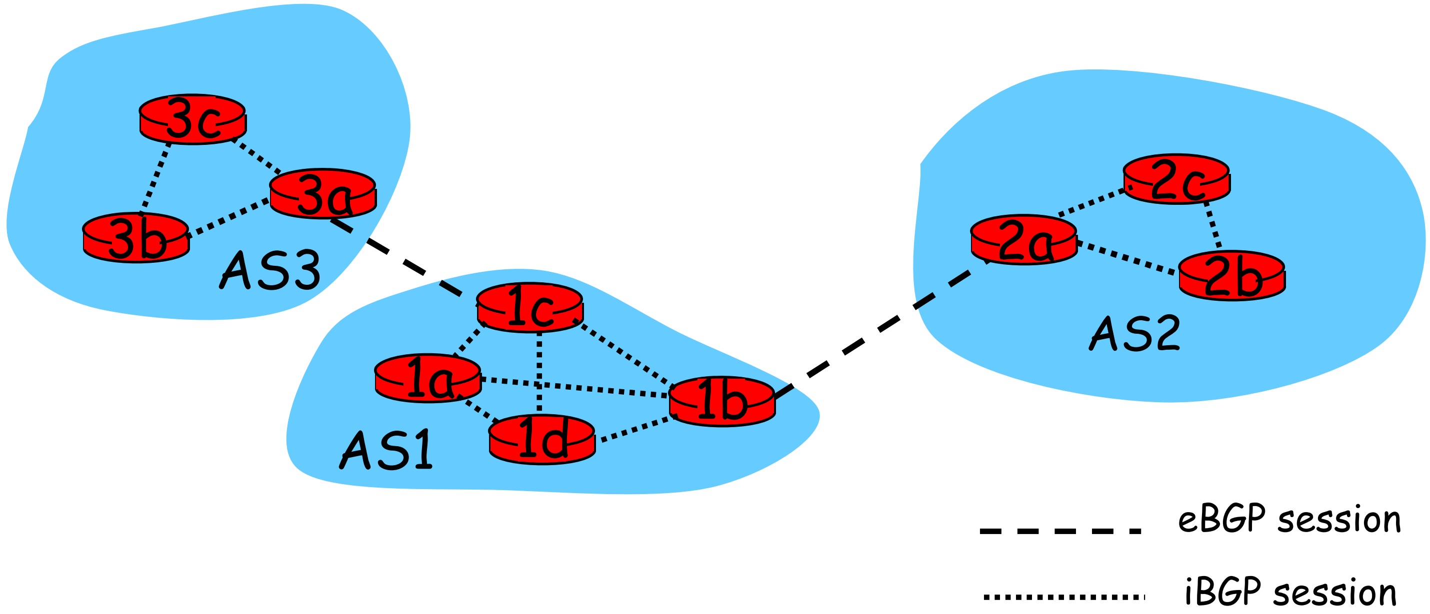
Internet inter-AS routing: BGP

- **Path Vector** protocol:
  - similar to Distance Vector protocol
  - each Border Gateway broadcast to neighbors (peers) *entire path* (i.e., sequence of AS's) to destination
  - BGP routes to networks (ASs), not individual hosts
  - E.g., Gateway X may send its path to dest. Z:

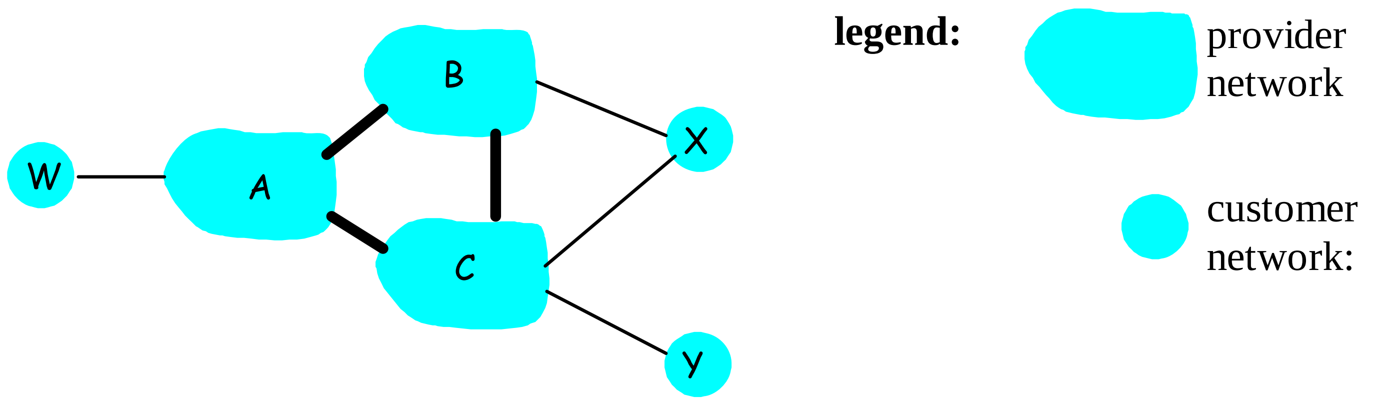
Path (X,Z) = X,Y1,Y2,Y3,...,Z

BGP basics

- Pairs of routers (BGP peers) exchange routing info over semi-permanent TCP conconnections: BGP sessions
- Note that BGP sessions do not correspond to physical links.
- When AS2 advertises a prefix to AS1, AS2 is *promising* it will forward any datagrams destined to that prefix towards the prefix.
  - AS2 can aggregate prefixes in its advertisement



BGP: controlling who routes to you



- A,B,C are provider networks
- X,W,Y are customer (of provider networks)
- X is dual-homed: attached to two networks
  - X does not want to route from B via X to C
  - .. so X will not advertise to B a route to C

Feng Wang, et.al, "A Measurement Study on the Impact of Routing Events on End-to-End Internet Path Performance", SIGCOMM 2006.

Overview

- If BGP route changes can be controlled, one could study in much more detail the effect of route changes on end-to-end delivery performance
- Requires access to ISP routers and BGP protocol, can only be done with help of ISP

About BGP

- To limit number of updates to be processed in a given time, using a rate-limiting timer called Minimum Route Advertisement Interval (MRAI) timer
- MRAI determines the minimum interval between route updates
- For eBGP (external, to other ASs), MRAI is 30s
- For iBGP (internal, within AS), MRAI is 5s
- “No valley” routing policy - no packet arriving from a provider may be forwarded to another provider
  - does not transit packet from one peer to another

Measurements

- Measure performances with UDP probes (send every 50ms when active), ping and traceroutes
- Use Beacon prefix to initiate routing events every two hours
  - withdrawal routes
  - Restore routes

Table 1: Classification of Beacon routing events

Beacon events	*BGP updates*	*Time schedule (GMT)*
Failover 1	Withdrawing route via $ISP1$	00:00, 04:00
Failover 2	Withdrawing route via $ISP2$	12:00, 16:00
Recovery 1	Restoring route via $ISP1$	02:00, 10:00
Recovery 2	Restoring route via $ISP2$	14:00, 22:00

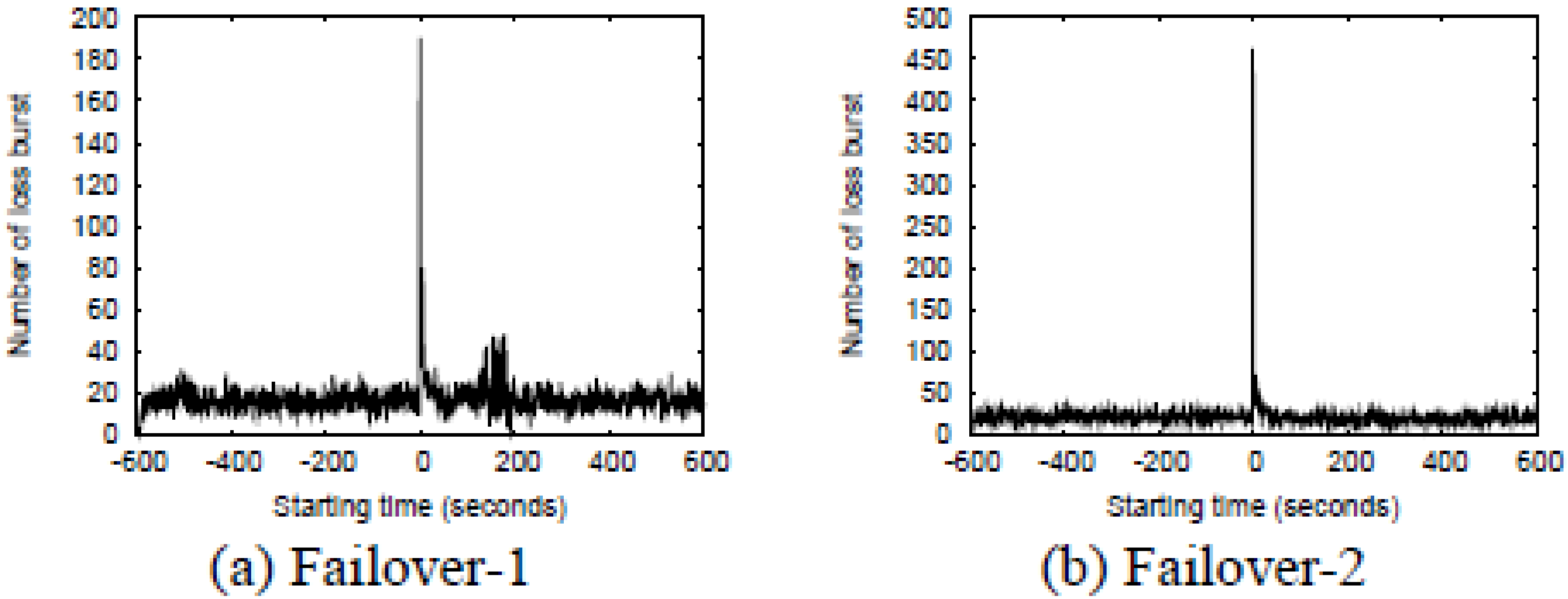


Figure 2: Number of loss bursts starting at each second.

Example

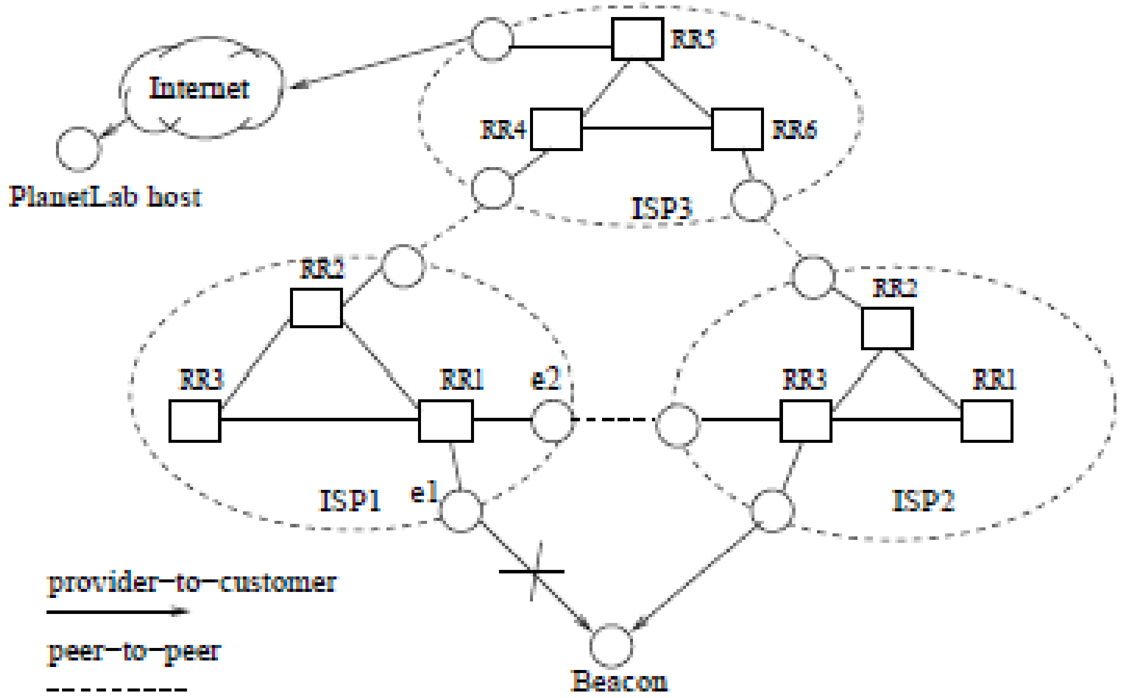


Figure 7: Topology of routers on the path from "planet02.csc.ncsu.edu" to the Beacon.

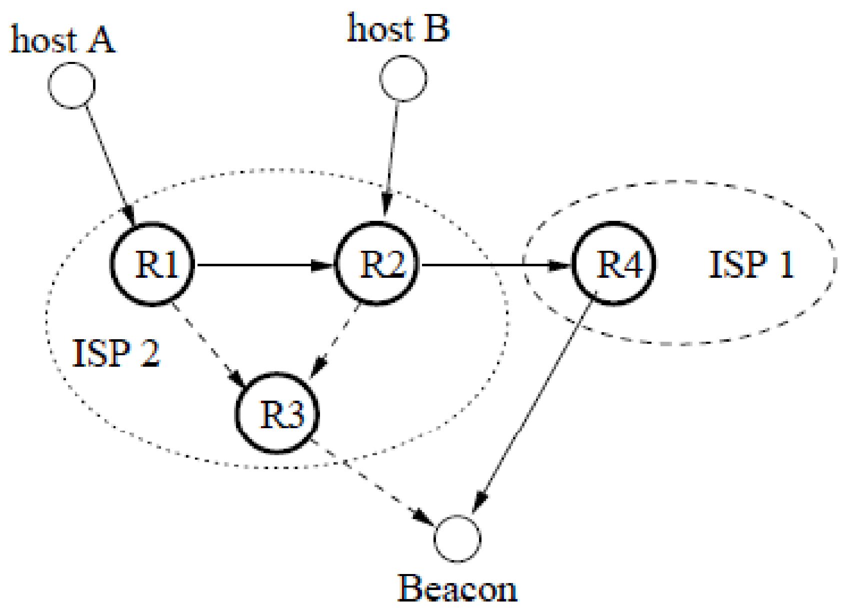
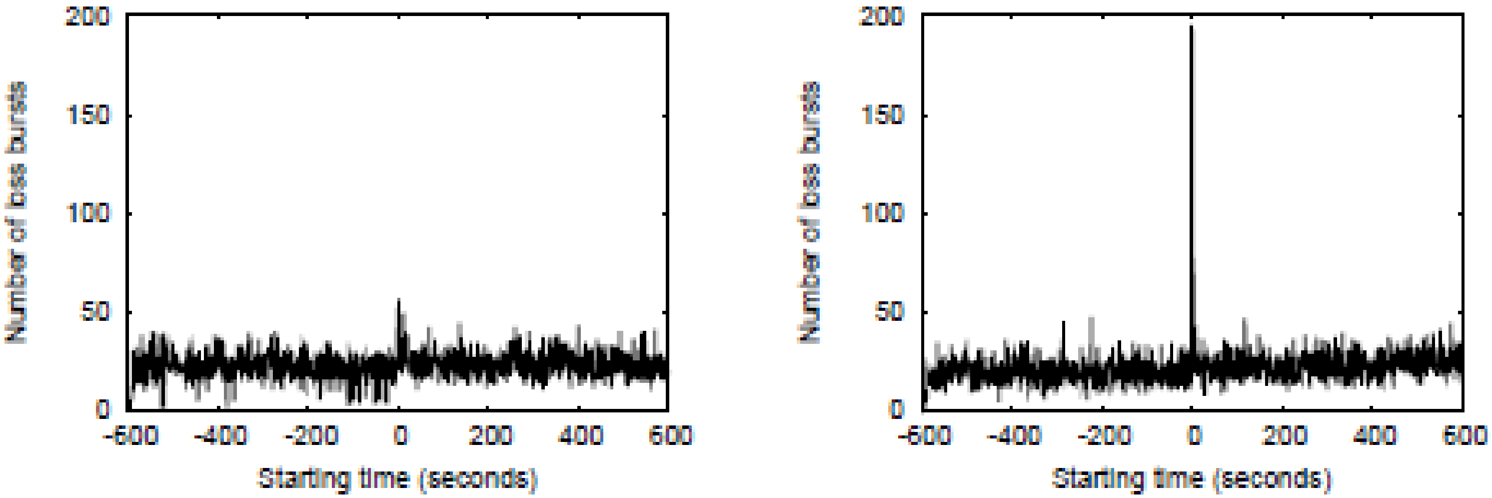


Figure 14: Topology for explaining packet loss burst during recovery.



(a) Recovery-1 (b) Recovery-2

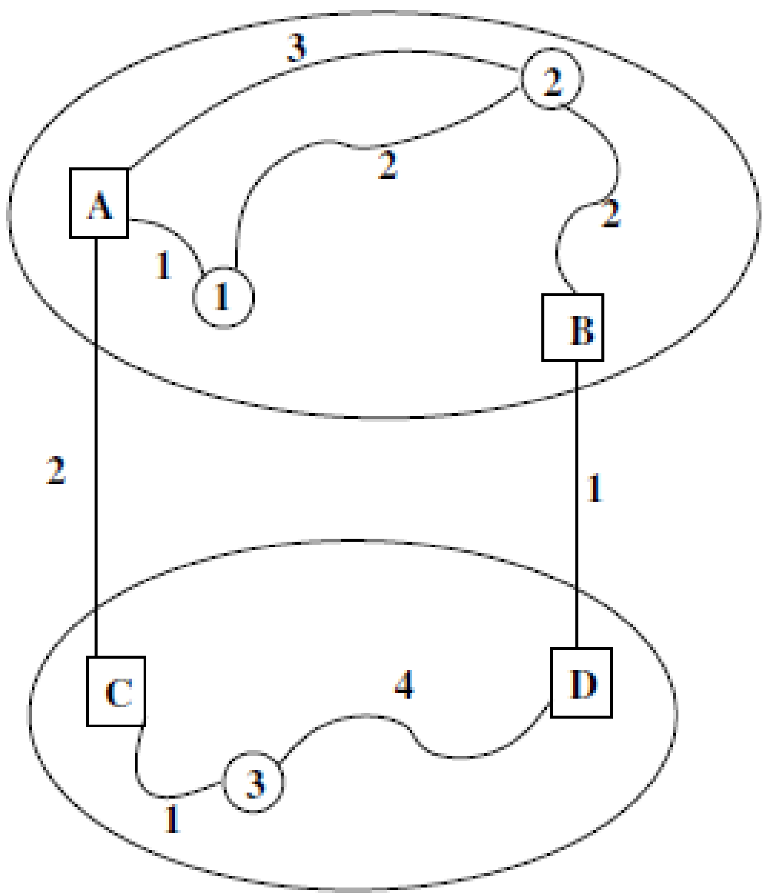
Figure 10: Number of loss bursts starting at each second during recovery events.

H Zheng, EK Lua, M Pias, TG Griffin, "Internet routing policies and round-trip-times," PAM 2005.

Triangle Inequality Violation (TIV)

- There are 3 hosts (A,B,C)
  - RTT between A and B is x
  - RTT between A and C is y
  - RTT between B and C is z
- There is a TIV in the RTT if $x > y + z$
- There is a problem for Internet Coordinate System based on RTT
- TIV can be explained based on routing policies

Example – Hot Potato Routing



$$13 = d(2,3) > d(2,1) + d(1,3) = 12$$

Example – BGP routing

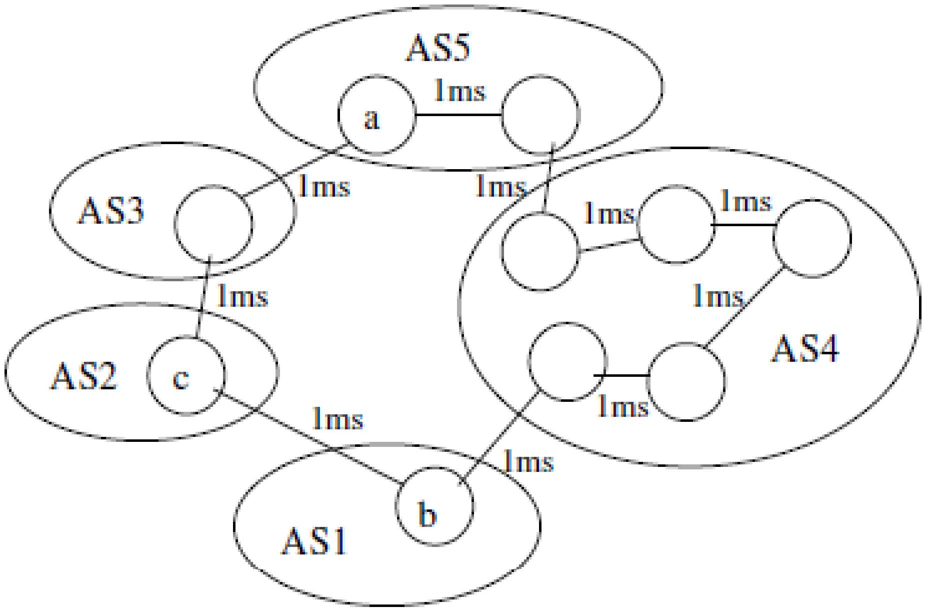


Fig. 3. When choosing the AS-level path between nodes a and b, BGP prefers AS 4 1 to AS 3 2 1, although the router-level path along AS4 is much longer.

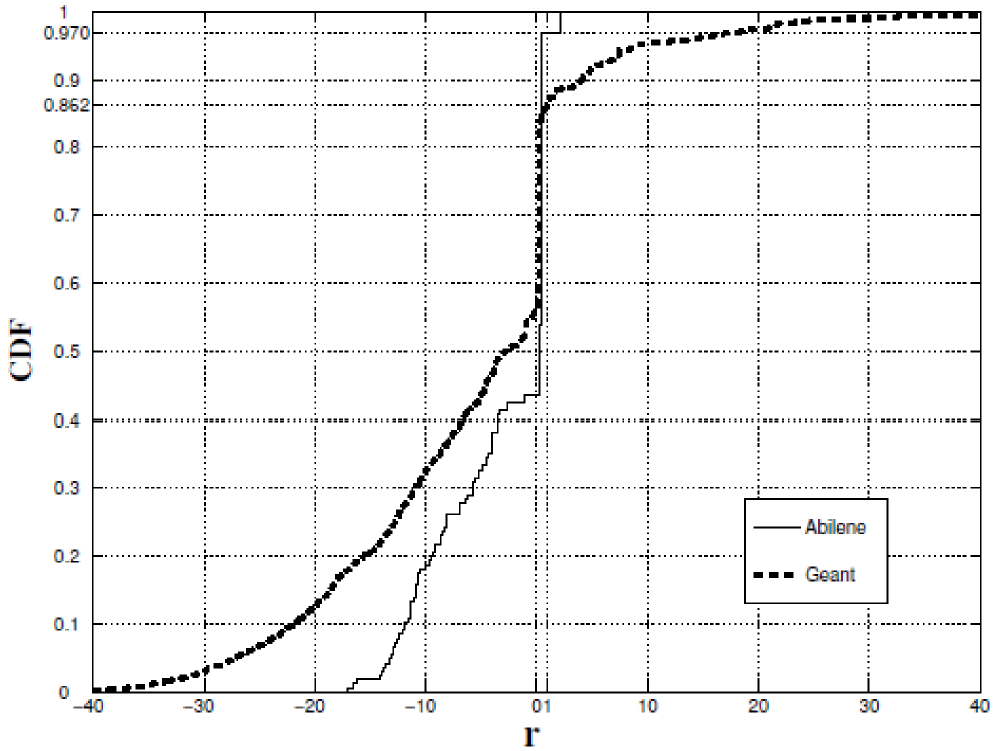


Fig. 9. CDF of the r metric for both Abilene and GEANT. TIVs are signified by $r > 1$, so it can be seen that GEANT exhibits a much higher percentage and magnitude of TIVs.

r = a/(b+c) * (1 + (a – (b+c)))
where a is the longest side of the triangle

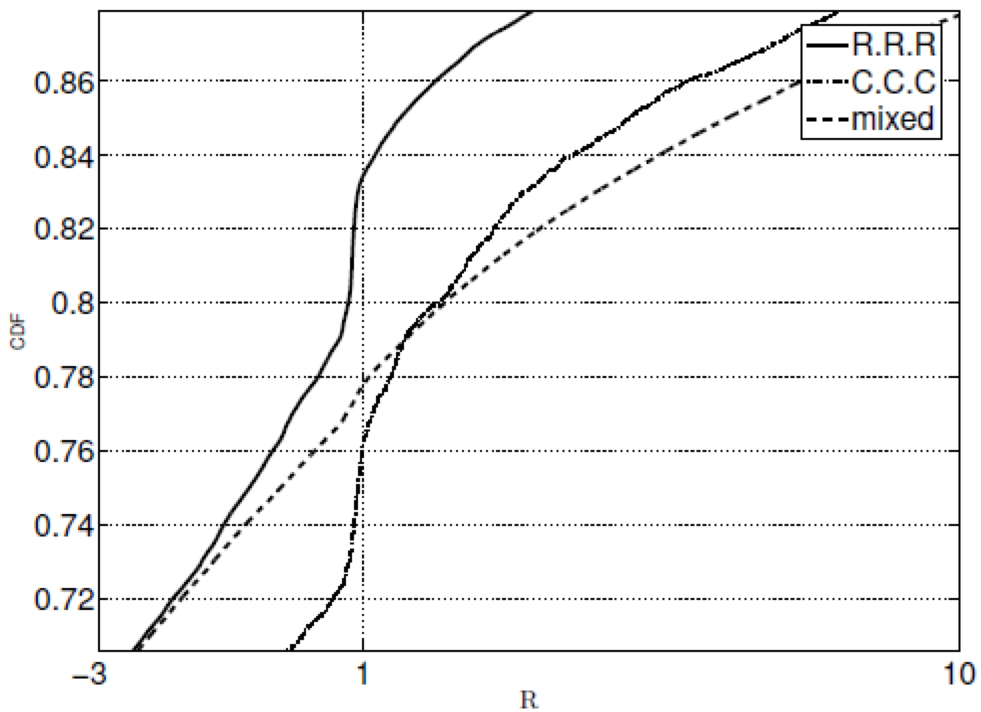


Fig. 14. The CDF distributions of r values for the three categories of triangles formed by Planet-Lab *nodes*, when zoomed in to areas around $r = 1$.