cariden

the economics of network control

Peering Planning Cooperation without Revealing Confidential Information

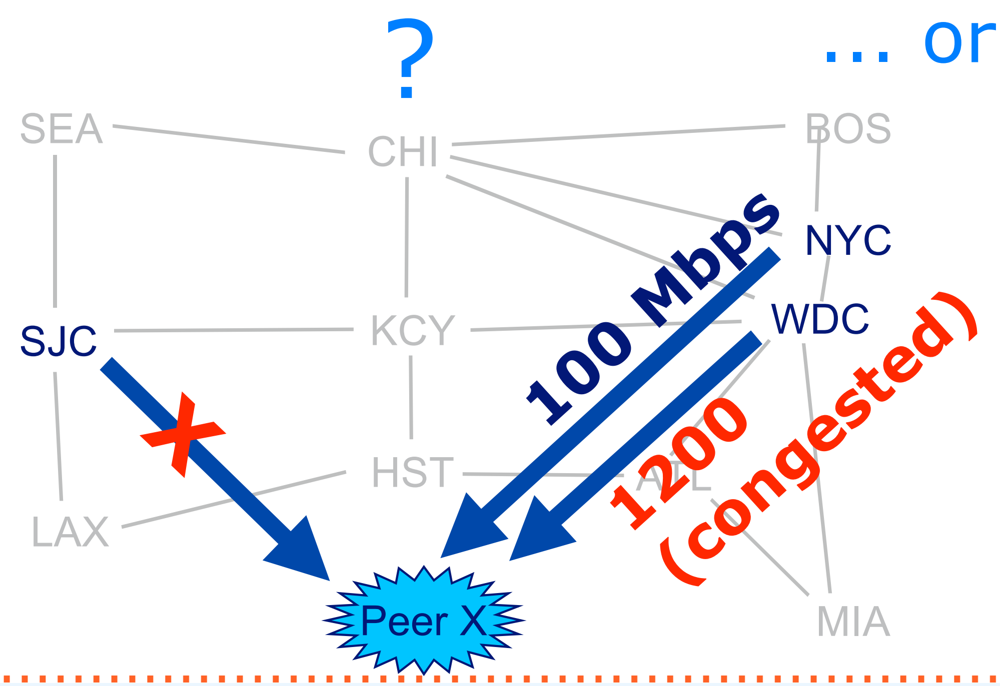
Thomas Telkamp
telkamp at cariden.com

RIPE 52
Istanbul, Turkey

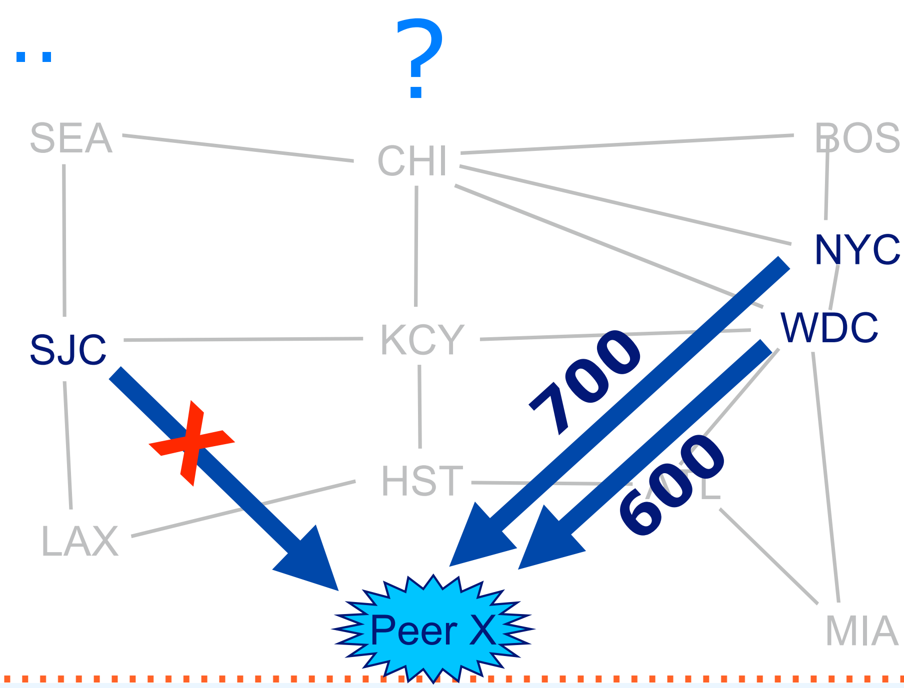
www.cariden.com

The Issue

- Multi-Homed Neighbor, 2 or more links > 50%
- Example
  - 1000Mbps connections to Peer X in 3 locations
  - SJC-to-Peer = 600Mbps, NYC = 100, WDC = 600
  - SJC-to-Peer link fails
- Are we in trouble?



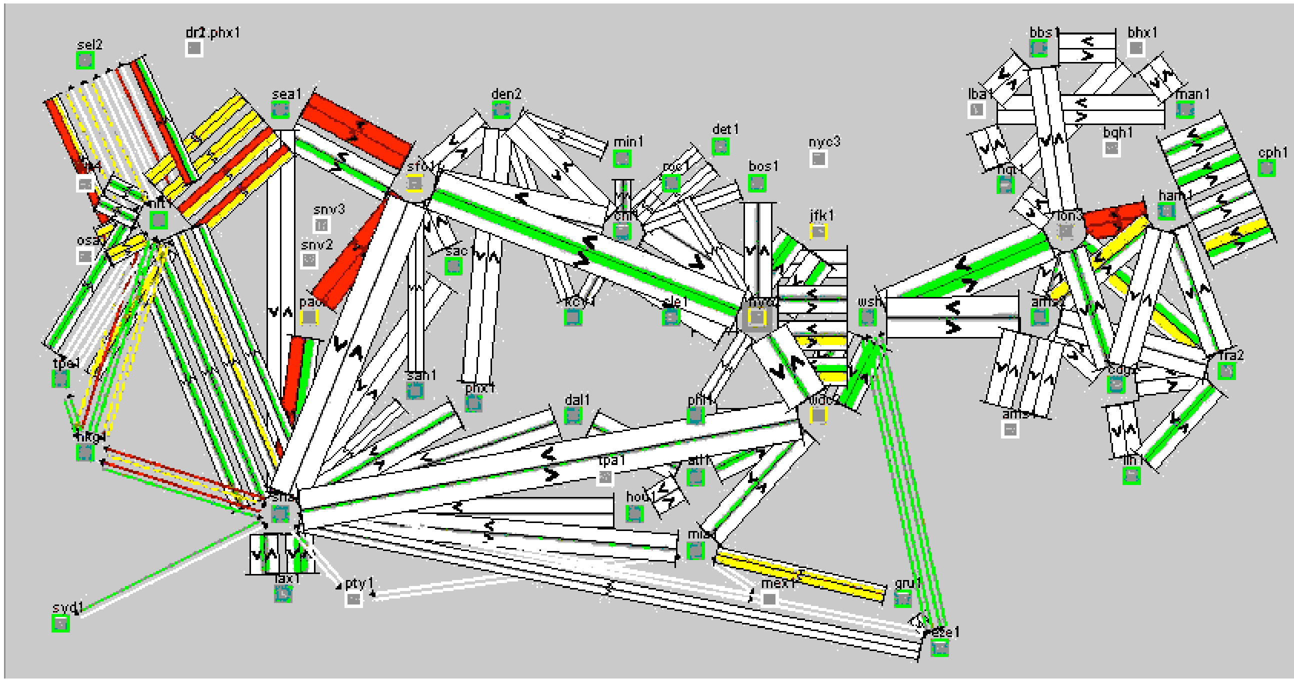
... or ...



Capacity Planning Utopia

- Uniform capacity links
- Diverse connections
(unlikely double failures at Layer 3)
- Upgrade at 50%
(planning objective is to be resilient to single failures)

Capacity Planning Reality



- Range of capacities
- Multiple Layer 3 failures
- Upgrade impediments (money, cable plant, ...)

IGP Different from BGP

- Data is more accessible
- Failure behavior is predictable
- Established process for within AS planning
  - Gather Data
    - Topology (OSPF, IS-IS, ...)
    - Traffic matrix [1]
    - Estimate growth
  - Simulate for failures
  - Perform traffic engineering (optional)[2]
  - Upgrade as necessary
- Commercial and free tools

The Trouble with BGP

- Data is larger and harder to access
- BGP decision process complicated
- Planning practices not well established
- Failure behavior often depends on someone else's network!
  - e.g., incoming traffic from a peer

subject of this talk

BGP Path Decision Algorithm[1]

1. Reachable next hop
2. Highest Weight
3. Highest Local Preference
4. Locally originated routes
5. Shortest AS-path length
6. IGP > EGP > Incomplete
7. Lowest MED
8. EBGP > IBGP
9. Lowest IGP cost to next hop
10. Shortest route reflection cluster list
11. Lowest BGP router ID
12. Lowest peer remote address

Respect MEDs

Shortest Exit Routing

[1] Junos algorithm shown here. Cisco IOS uses a slightly different algorithm.

Common Routing Policies

- Shortest Exit
  - Often used for sending to peers
  - Get packet out of network as soon as possible
  - Local Prefs used to determine which neighbor, IGP costs used to determine which exit
- Respect MEDs
  - Often used for customers who buy transit
  - Deliver packets closest to destination
  - Neighbor forwards IGP costs as MEDs (multi-exit discriminators)

Blind Spots

- Cannot predict behavior when routing depends on other network (see 3 cases below).

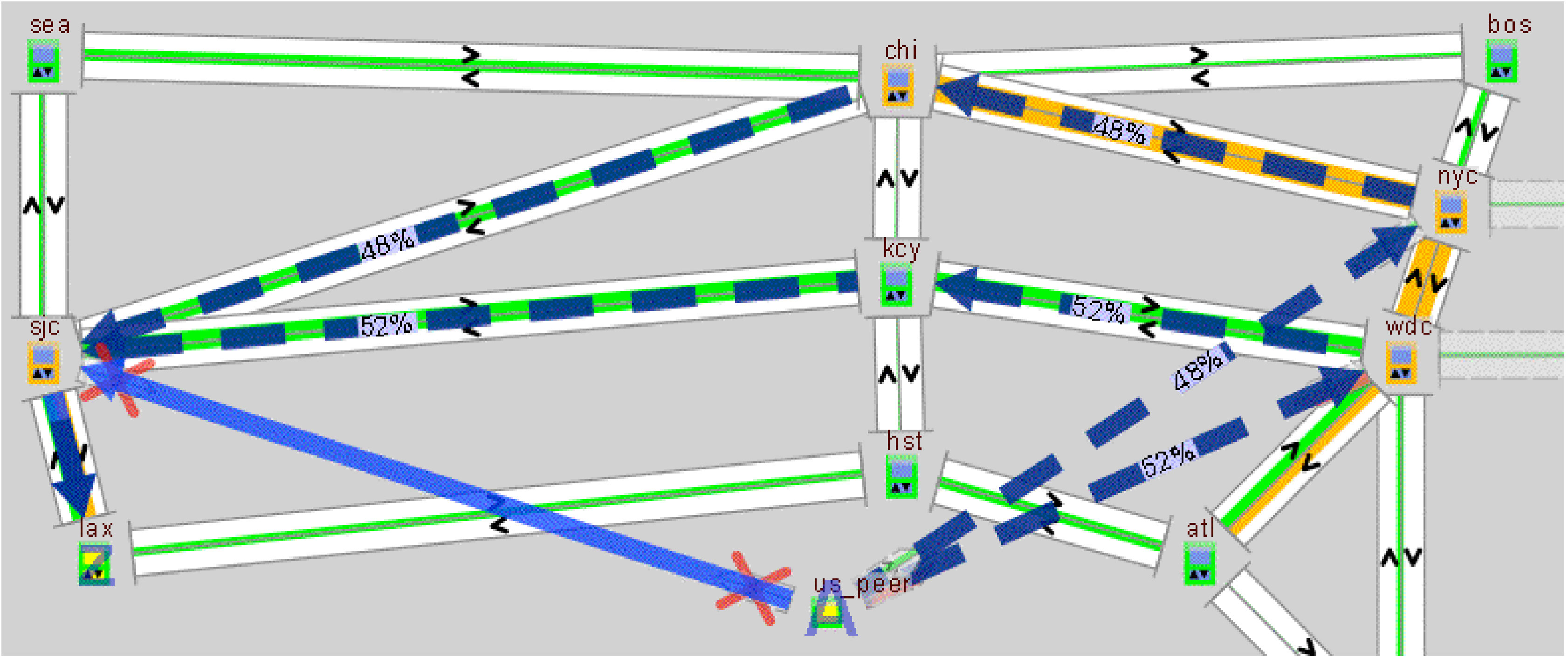
Relationship to Remote AS	Routing To Remote AS	Routing From Remote AS
Peer	Shortest Exit in known network	Shortest Exit in unknown network
Customer	Respect MEDs from unknown	Shortest Exit in unknown network
Transit Provider	Shortest Exit in known network	Respect our MEDs

Failover Matrices

- Solution to peering planning blind spots
- Procedure
  - Gather data
    - Topology, Traffic, Routing Configurations
  - Simulate knowable effects
    - Generate Failover Matrices
  - Share Failover Matrices for unknowables
    - e.g., peer gives failover matrix for traffic it delivers, we provide peer failover matrix for traffic we deliver
- Both sides benefit from cooperating
- AS-Internal information is kept confidential

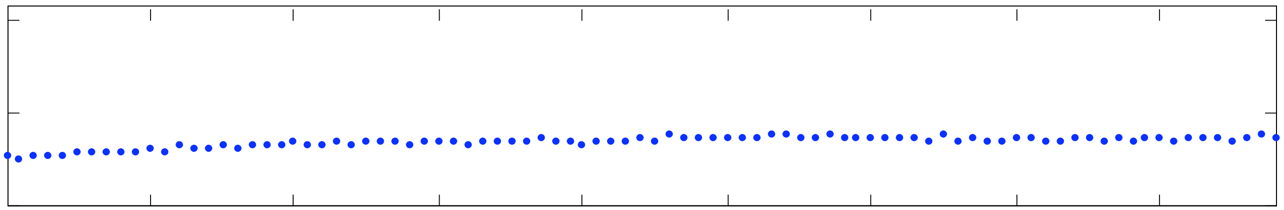
Failover Matrix Example

Node:Interface	Traffic: no failure	%Traffic: fail_SJC	%Traffic: fail_nyc	%Traffic: fail_wdc
ar1.sjc:Gig3/2	600	-	10% (610)	1% (606)
ar1.nyc:ge-2/1	100	48% (388)	-	95% (670)
ar2.wdc:ge-2/2	600	52% (912)	70% (670)	-

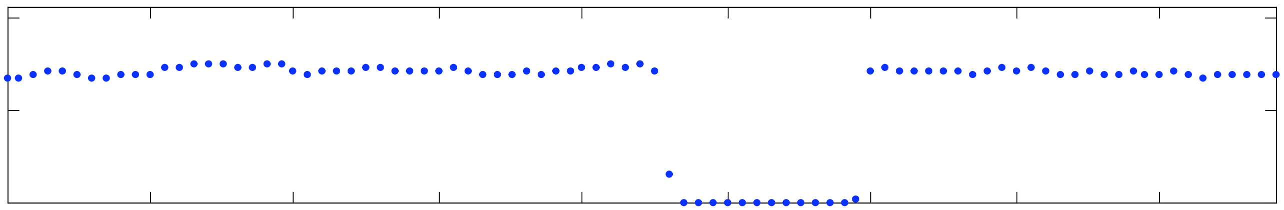


Note: 388Mbps=100Mbps+(0.48*600Mbps), 912=600+(0.52*600), ...

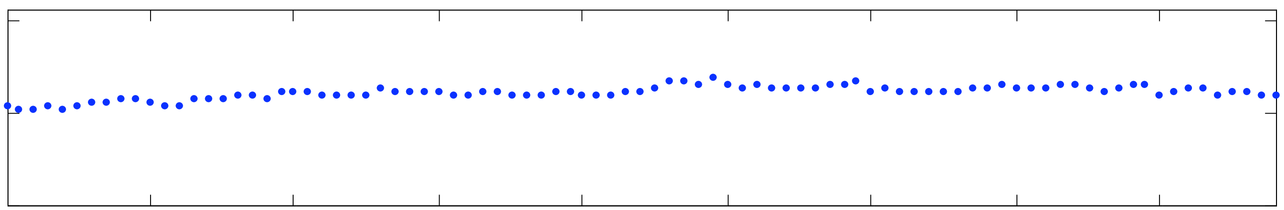
Failover Example (from real network)



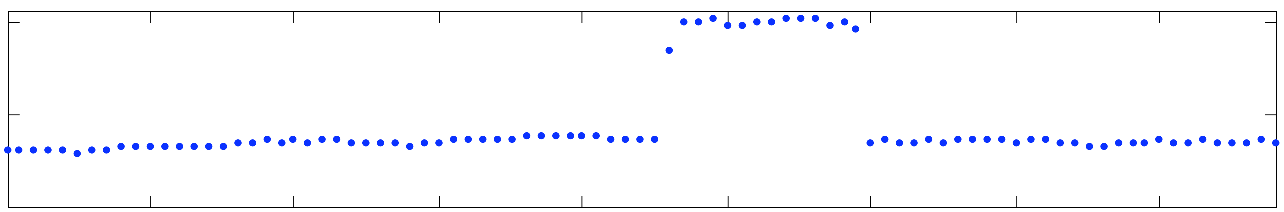
Peer Circuit 1: Traffic levels at five minute intervals



Peer Circuit 2: Traffic levels at five minute intervals



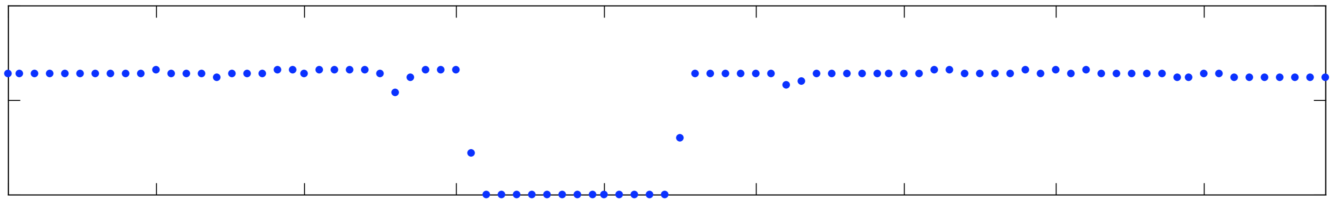
Peer Circuit 3: Traffic levels at five minute intervals



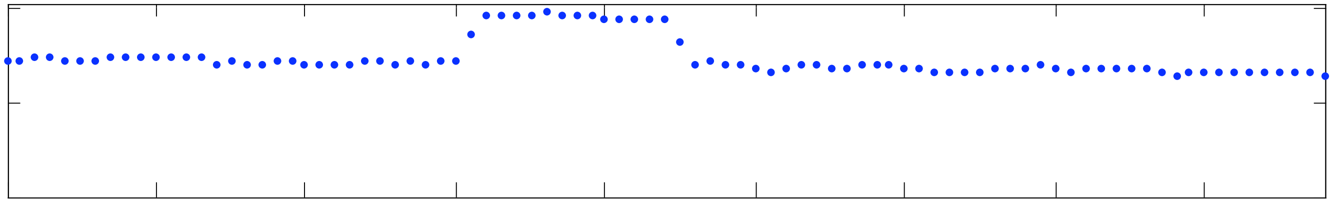
Peer Circuit 4: Traffic levels at five minute intervals

- Circuit 2 fails.
  Traffic shifts to circuit 4.

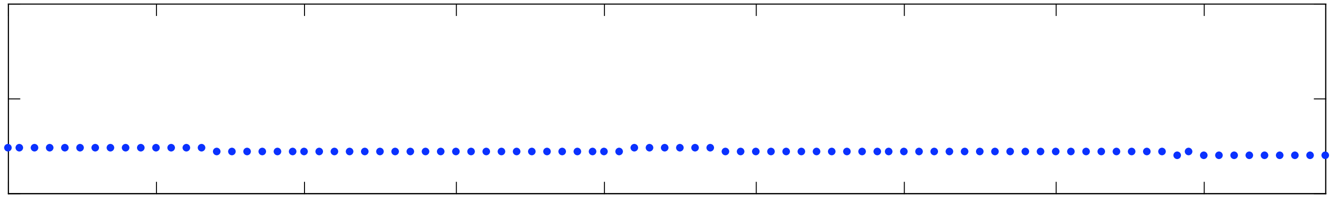
Failover Example (from real network)



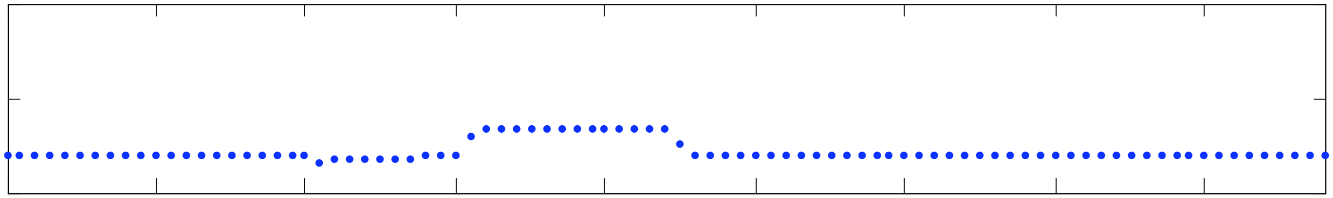
Peer Circuit 1: Traffic levels at five minute intervals



Peer Circuit 2: Traffic levels at five minute intervals



Peer Circuit 3: Traffic levels at five minute intervals



Peer Circuit 4: Traffic levels at five minute intervals

- Circuit 1 fails. Some traffic shifts to 2 & 4
- Some “leaks” to other AS’s

Questions

- How do I calculate a failover matrix?
- How do I use a failover matrix from a peer?
- What if my peer does not cooperate?
- What if a substantial amount of traffic “leaks” to another AS?

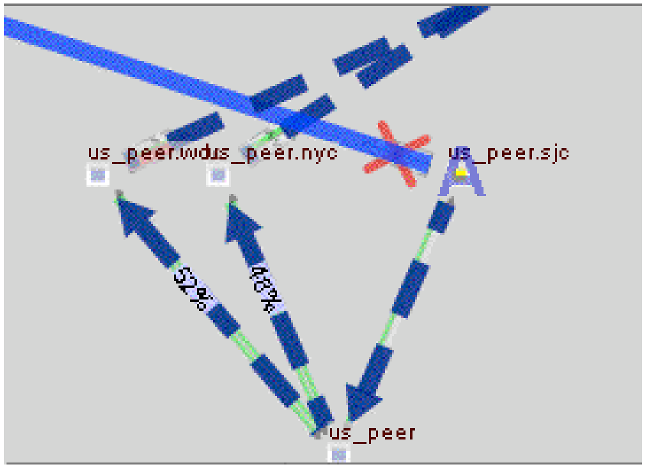
Calculating Failover Matrices

- Accurate and Detailed[3,4]
  - Per prefix routing and traffic statistics
  - Full BGP simulation
- Simple and Good Enough
  - Traffic matrix based on ingress-egress pairs
    - e.g., Peer1.LAX-AR1.CHI (measure and/or estimate) instead of 192.12.3.0/24-208.43.0.0/16
  - Limited simulation model
    - Shortest Path, Respect MEDs
    - “Our” AS plus immediate neighbors

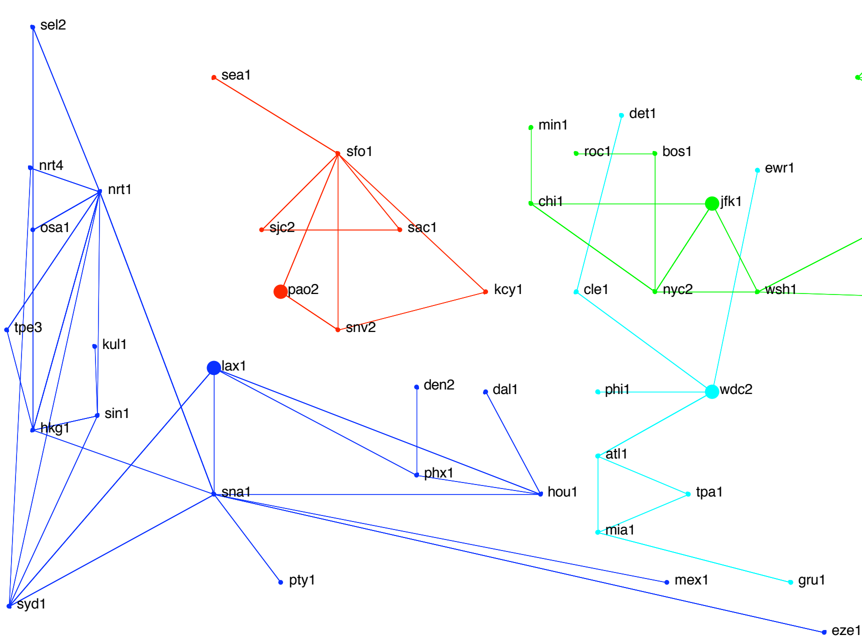
Using Failover Matrix from Peers

- Peer calculates failover matrix
- Peer exports failover matrix using IP addresses of peering links

- We import failover matrix
- We include in a representative model of peer network
- Use Failover Matrix in simulation

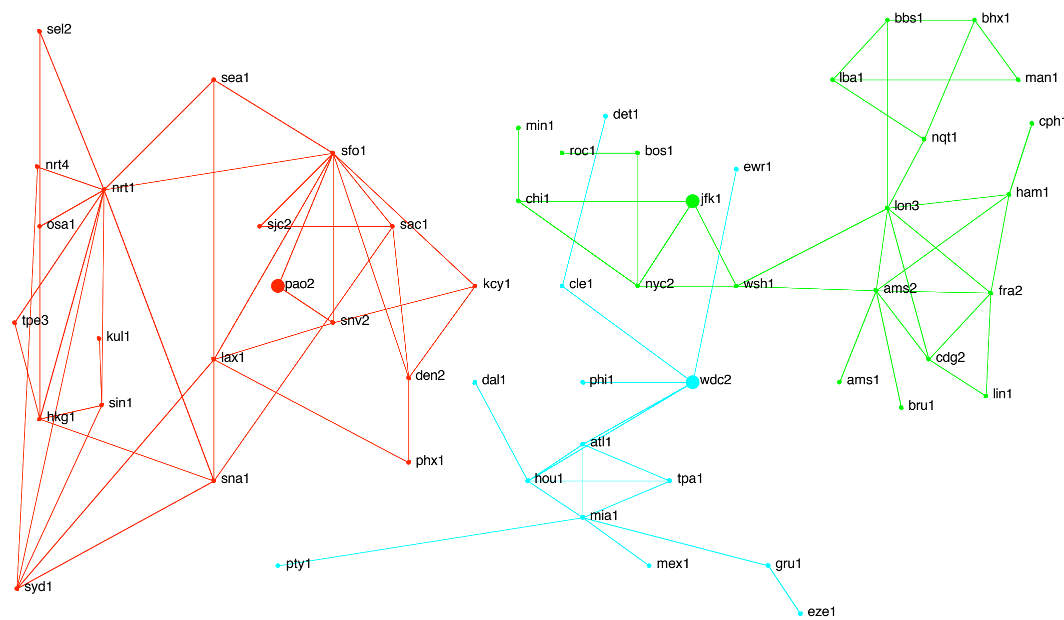


Estimate if Peer not Cooperate



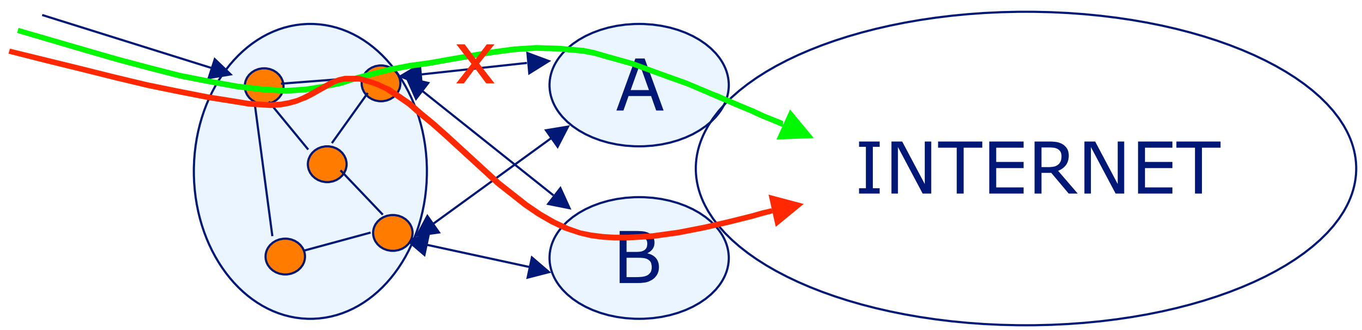
← Group own sources based on exit location (4 groups here)

→ Quantify shift (to 3 groups) after failure
Assume similar for other side



- Valid if topology and traffic distributions are similar

Leaks to Other AS's



- Simple option
  - Leaks between peers relatively small
    - Ignore
  - Shifts between transit providers can be large
    - Equal AS-path length to most destinations:
    - Assume complete shift (easy to model)
- Accurate option
  - Extend model to more than one AS away
  - Add columns in traffic matrix to designate extra traffic in case of other network failures

Work in Progress

- Evaluating goodness of models
  - Compare actual failures to models
- Evaluating goodness of failover estimates
  - Work with both sides of a peering arrangement, compare failover estimates to simulations
  - Compare estimated failover matrices to actual failures
- Streamlining sharing of information
- Looking for more participants
  Contact me to participate in the above

Summary

- Peering/transit links are some of the most expensive and difficult to provision links
- We can improve capacity planning on such links by modeling the network
- BGP modeling can be much more complex than IGP modeling
  - Some required information is not even available
- Failover Matrices provide a simple way to share information without giving away details
- Failover Matrices can be estimated using one’s own network details

Acknowledgments

- Jon Aufderheide (Global Crossing)
- Clarence Filsfils (Cisco)

References

[1] APRICOT 2005 tutorial: Best Practices for Determining the Traffic Matrix in IP Networks

[2] APRICOT 2004 tutorial: Traffic Engineering Beyond MPLS

[3] "Modeling the routing of an Autonomous System with C-BGP," B. Quoitin and S. Uhlig, IEEE Network, Vol 19(6), November 2005.

[4] "Network-wide BGP route prediction for traffic engineering," N. Feamster and J. Rexford, in Proc. Workshop on Scalability and Traffic Control in IP Networks, SPIE ITCOM Conference, August 2002.

Tutorials [1] and [2] are available at:

http://www.cariden.com/technologies/papers.html