BGP Wedgies, and how to avoid them

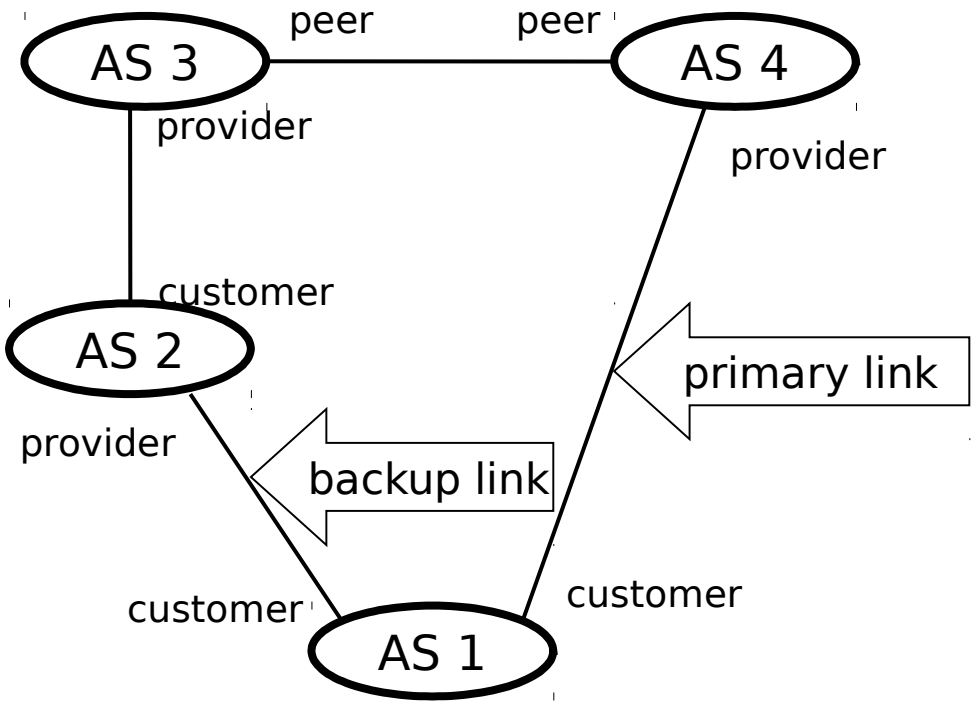
Timothy G. Griffin
tgg22@cam.ac.uk
Computer Lab
Cambridge UK

ENOG 15
June 4, 2018
MOSCOW

BGP Wedgies (Informational RFC 4264)

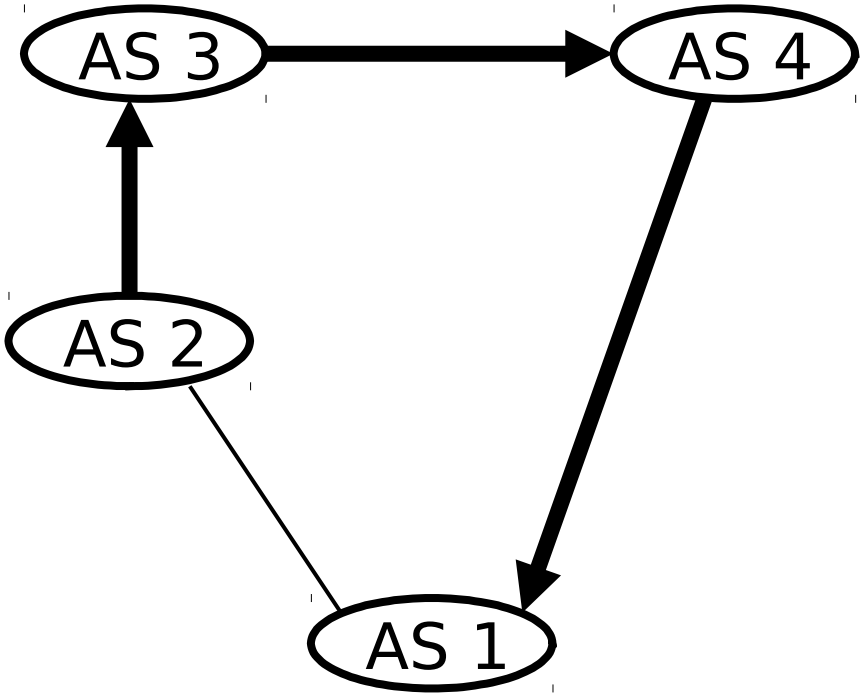
- BGP policies make sense locally
- Interaction of local policies may allow **multiple** stable routings
- Some routings may be consistent with intended policies, others not
- BGP is **wedged** when an unintended routing is installed
- **Manual intervention** is required to change to an intended routing
- Worst case : an unintended routing is installed and no single group of network operators has enough knowledge to debug the problem!

Simple Example

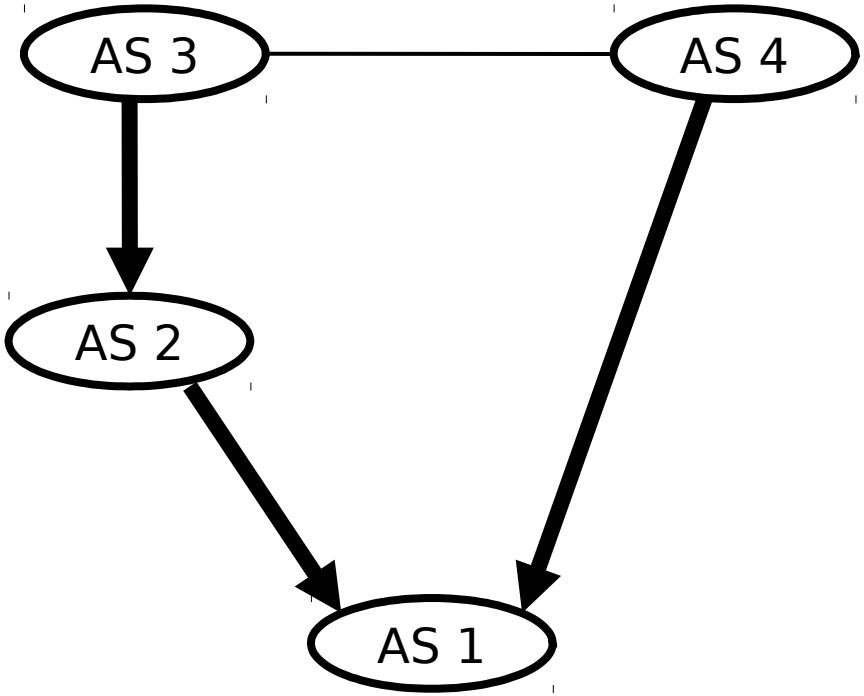


- AS 1 implements backup link by sending AS 2 a **depref-me** community.
- AS 2 implements this community so that the resulting local pref is below that of routes from it's upstream provider (AS 3 routes)

And the Routings are…



Intended Routing

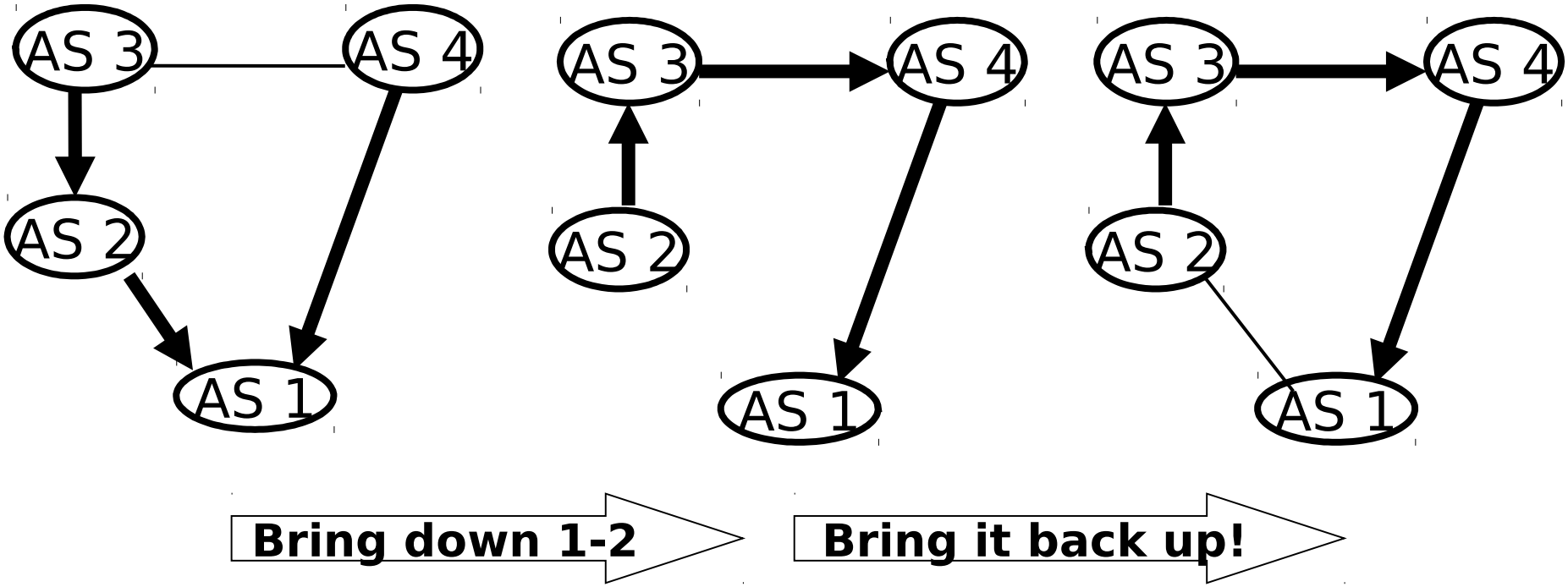


Unintended Routing

Note: this would be the ONLY routing if AS2 translated its "depref me" community to a "depref me" community of AS 3

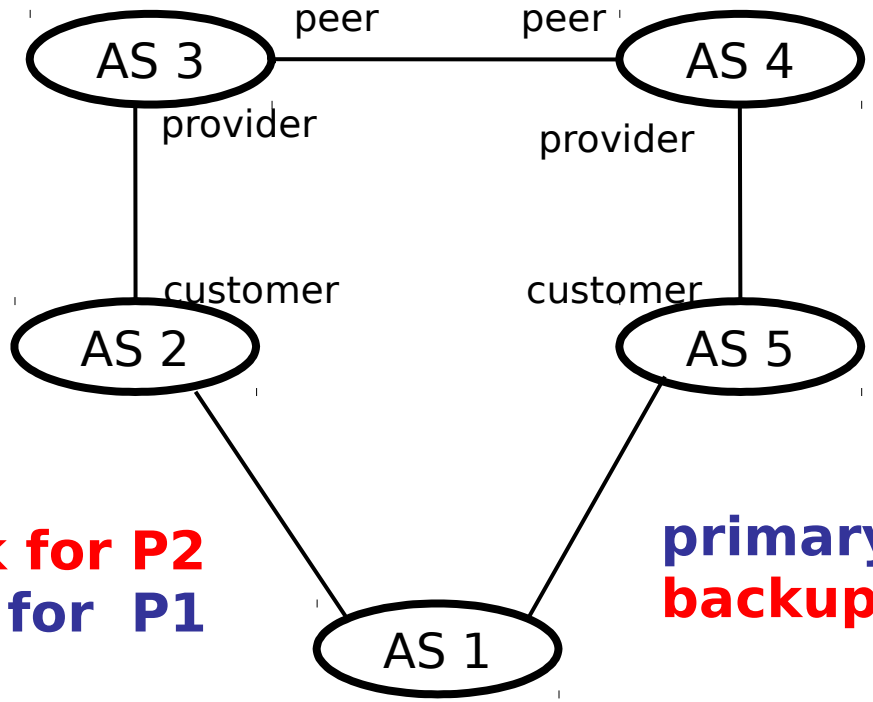
Note: This is easy to reach from the intended routing just by "bouncing" the BGP session on the primary link.

Recovery?



- Requires **manual intervention**
- Can be done in AS 1 or AS 2

Load balancing example



primary link for P2
backup link for P1

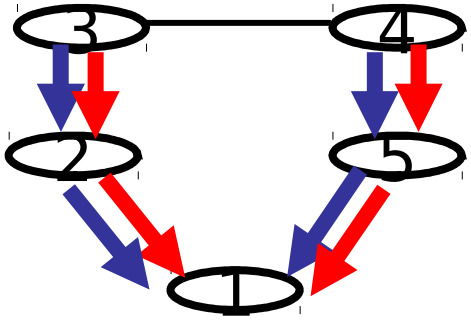
primary link for P1
backup link for P2

Originates prefixes
P1 and P2

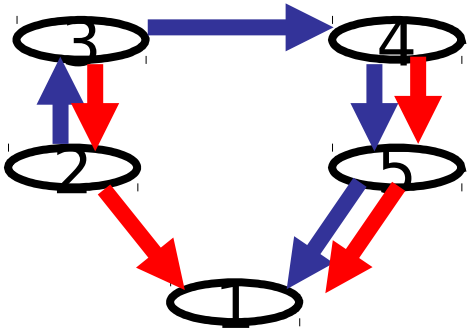
Simple session reset my not work!!

4 stable routings!

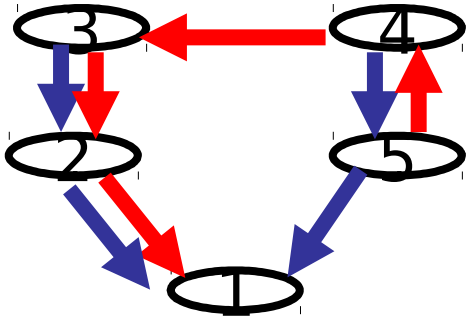
P1 and P2 wedged



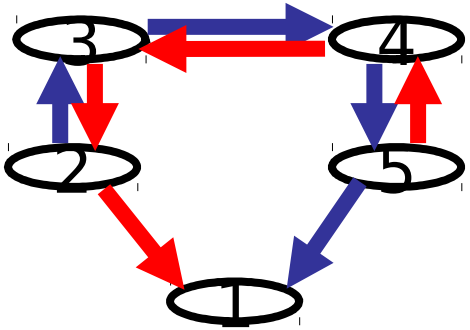
P2 wedged



P1 wedged



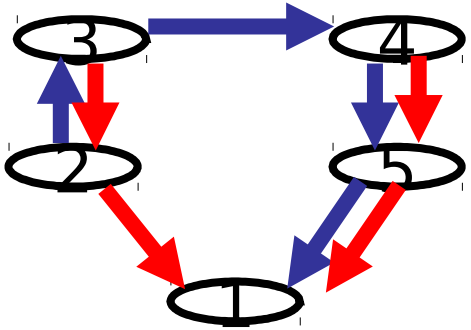
Intended



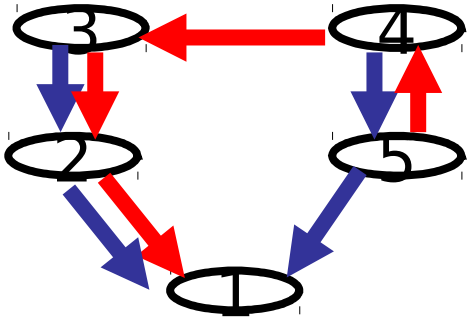
One reset will not help

Reset 1-5

P2 wedged

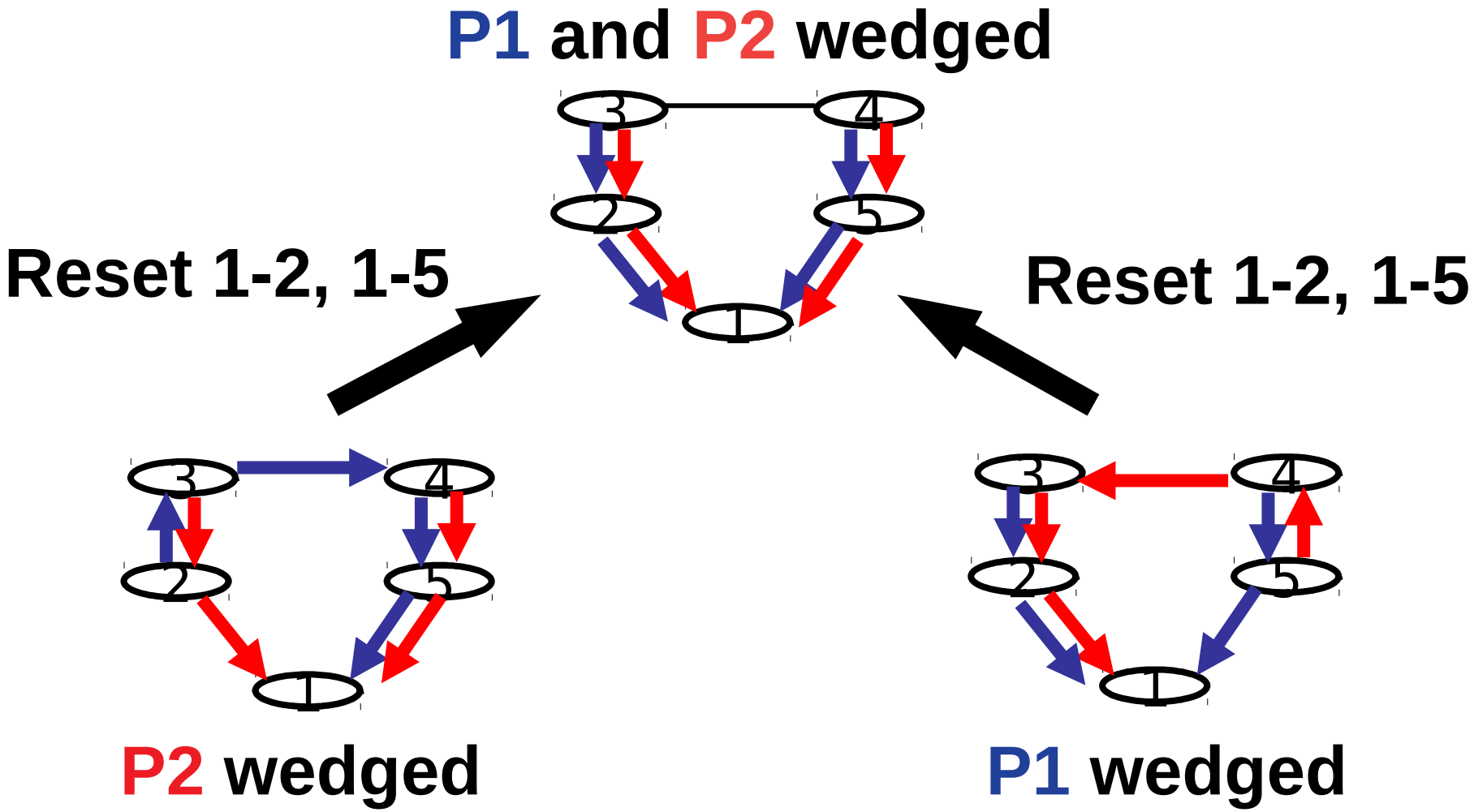


P1 wedged

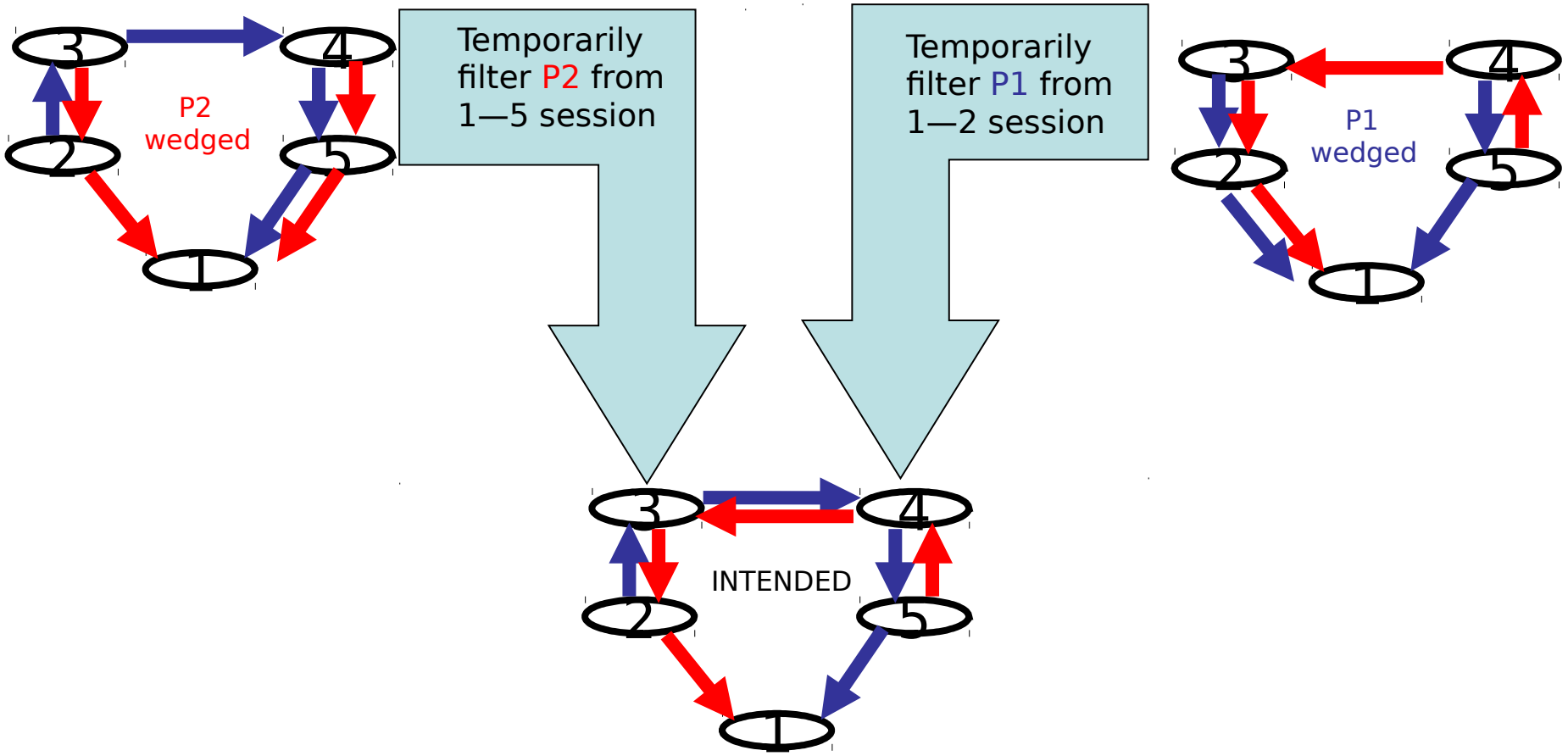


Reset 1-2

Can get double trouble by resetting both sessions!

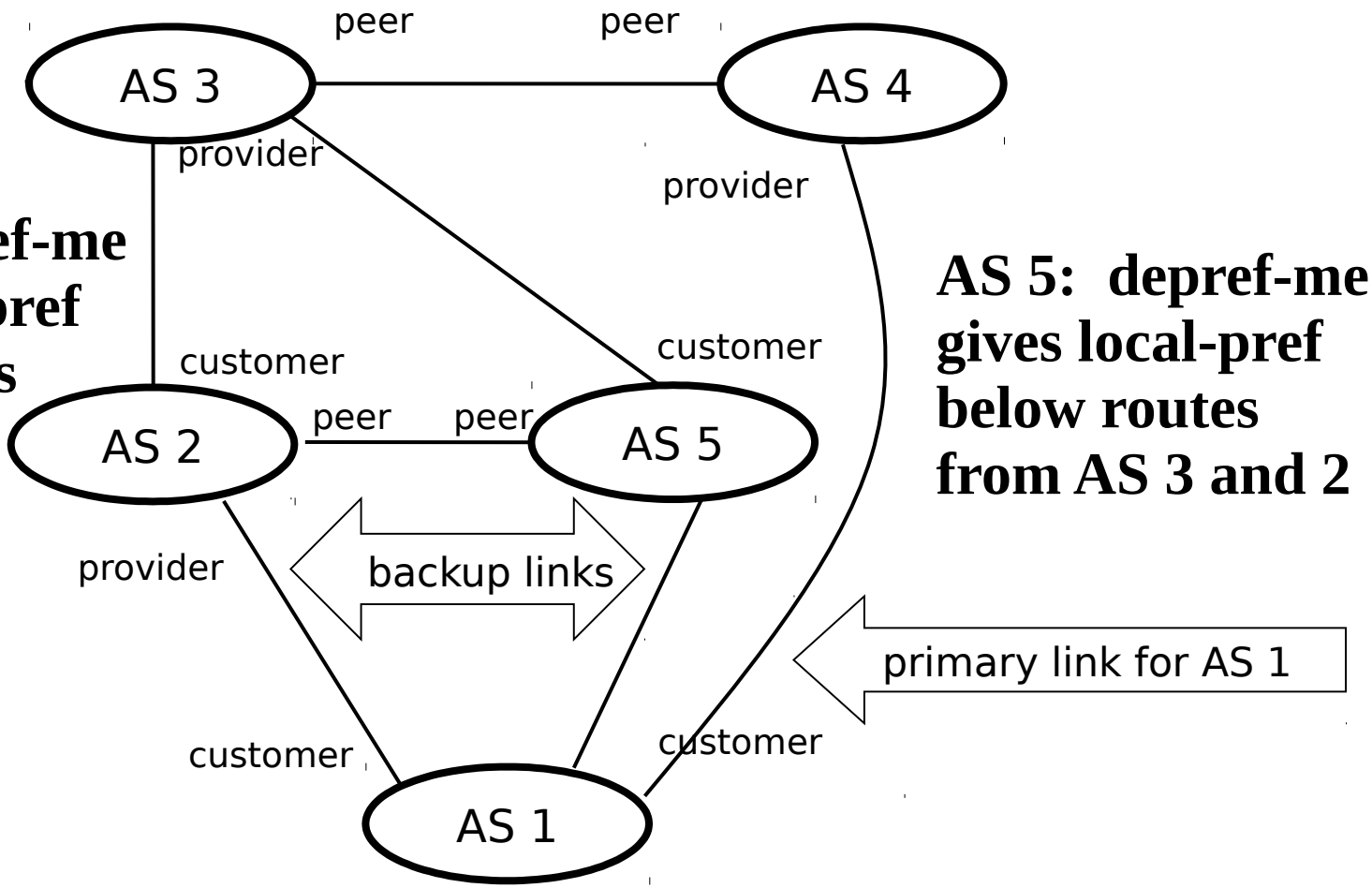


Recovery?



Who among us could do this when 1-2 is in Moscow and 1-5 is in Tokyo?

The Full Wedgie

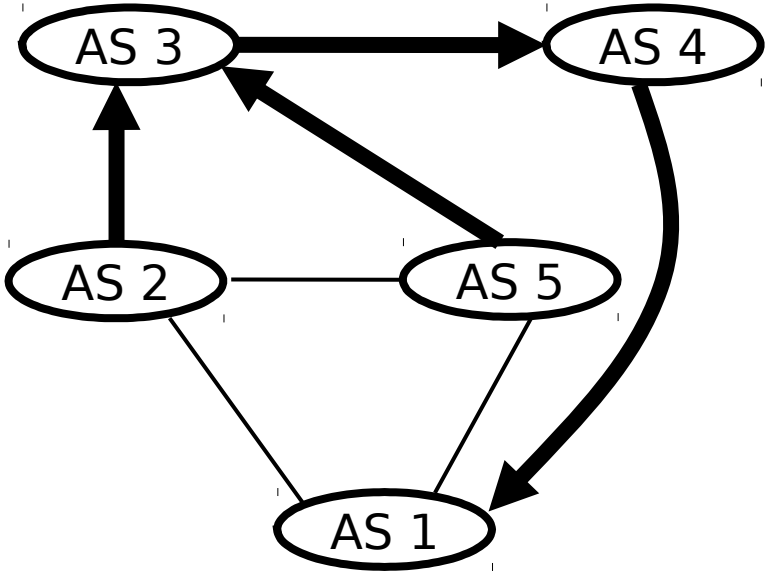


AS 2: depref-me gives local-pref below routes from AS 3 and 5

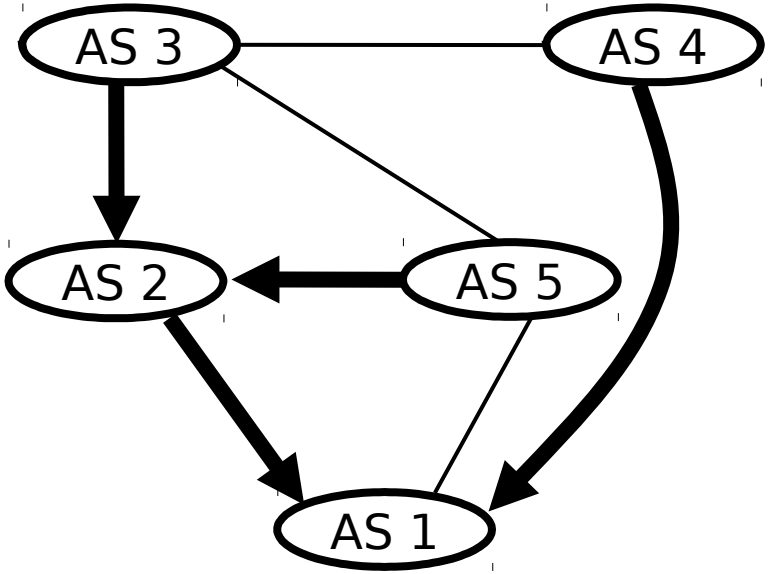
AS 5: depref-me gives local-pref below routes from AS 3 and 2

AS 1 sends AS 2 and AS 5 depref-me communities.

And the Routings are…

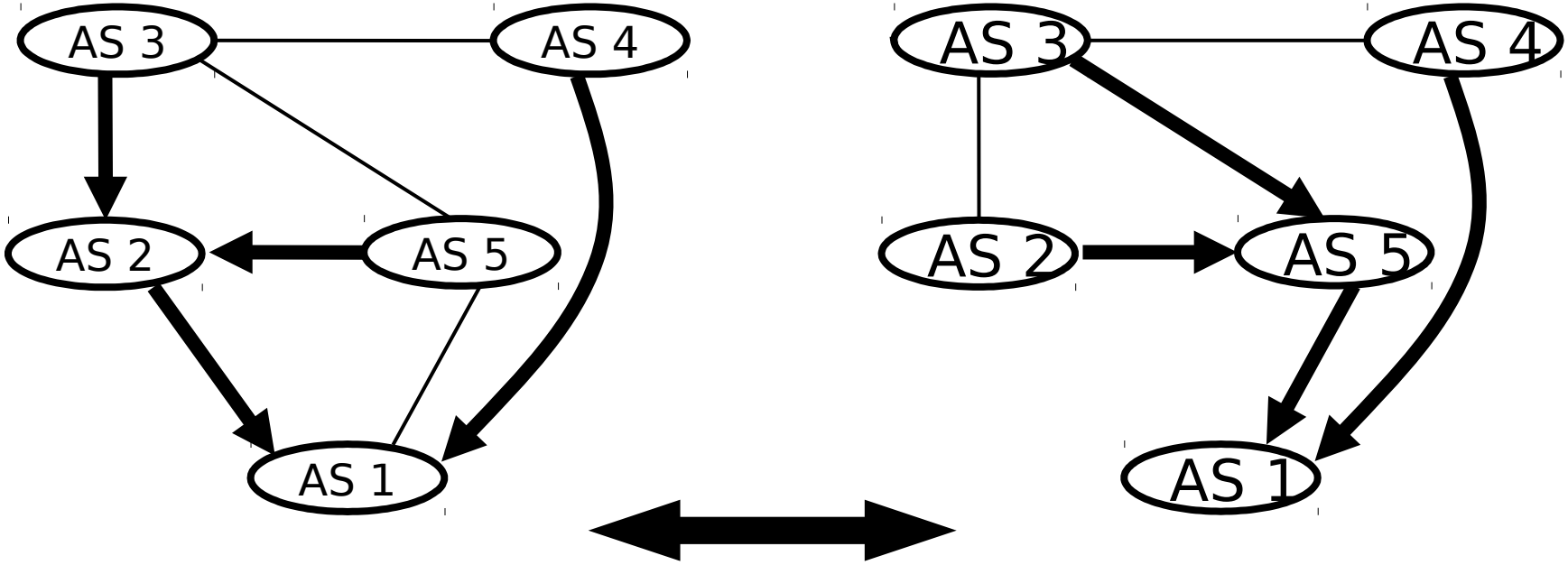


Intended Routing

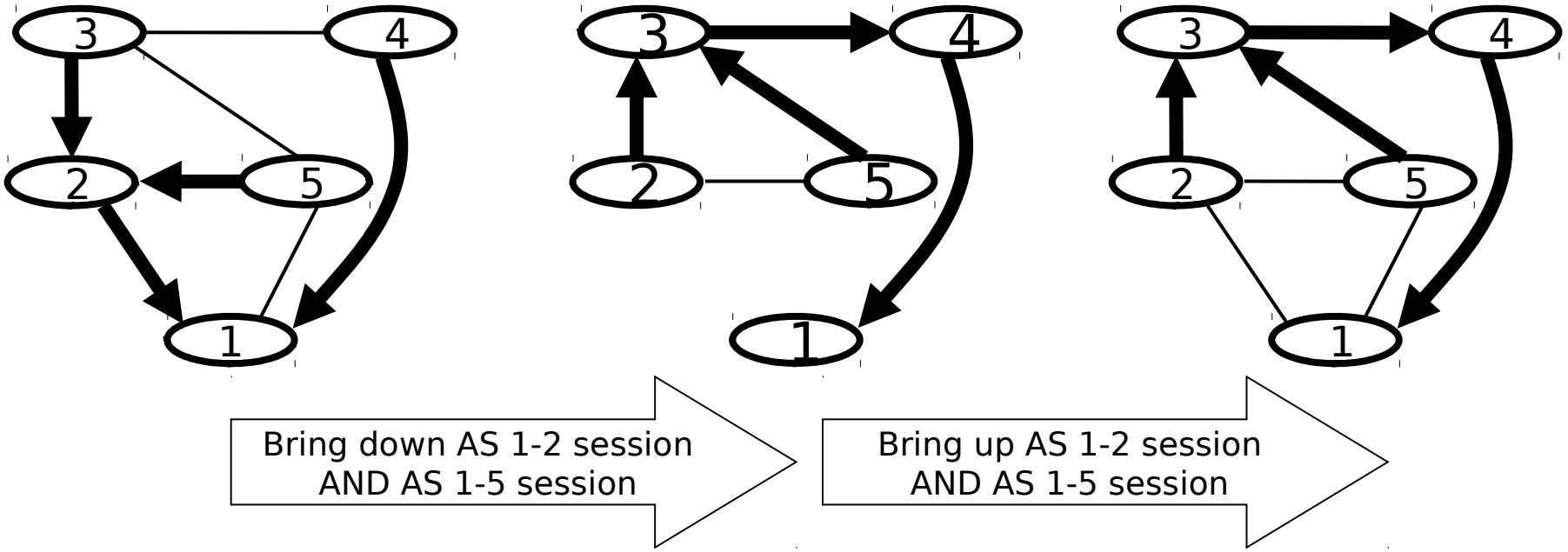


One Unintended Routin
(there are others)

Resetting 1—2 does not help!



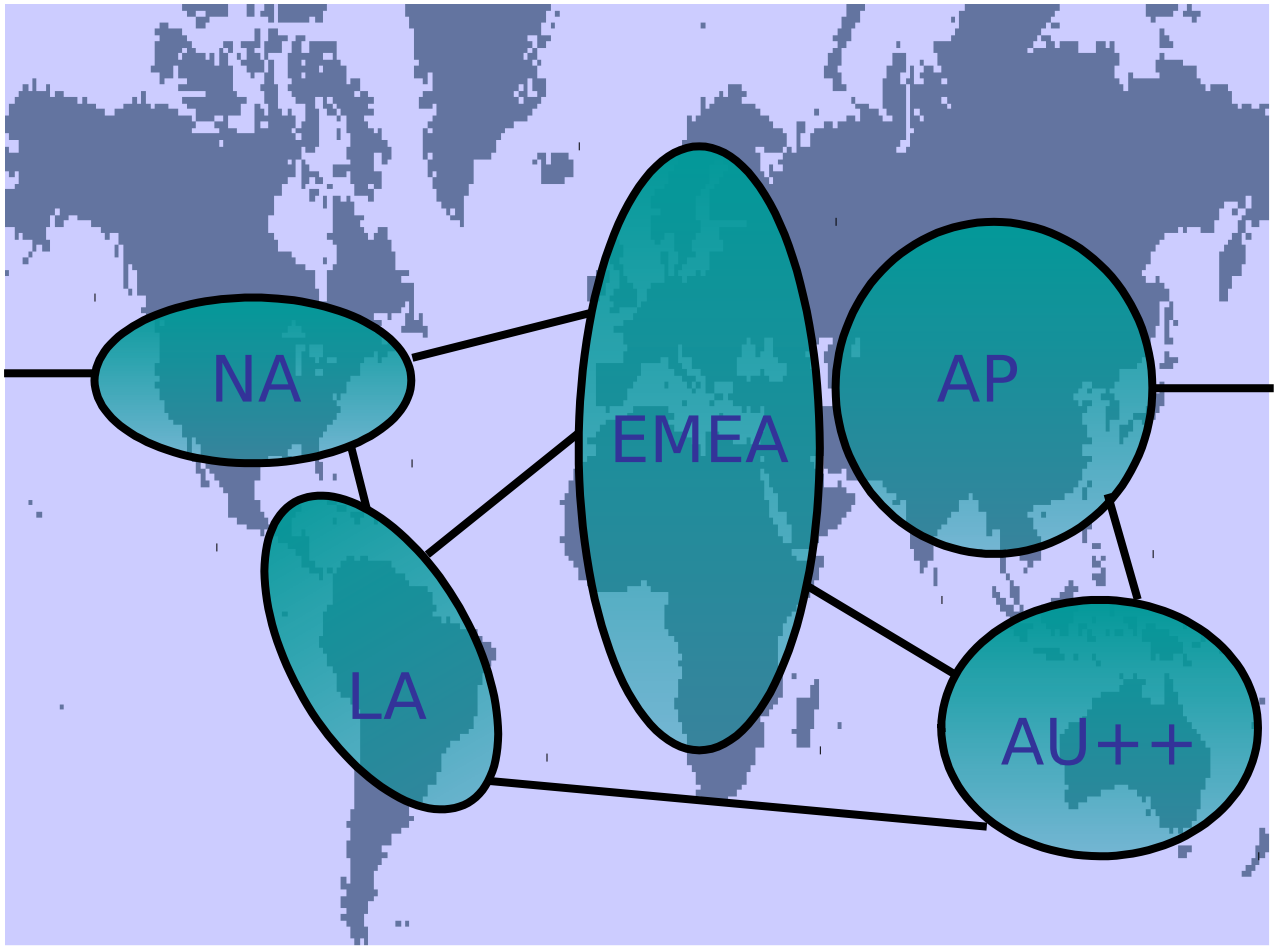
Recovery?



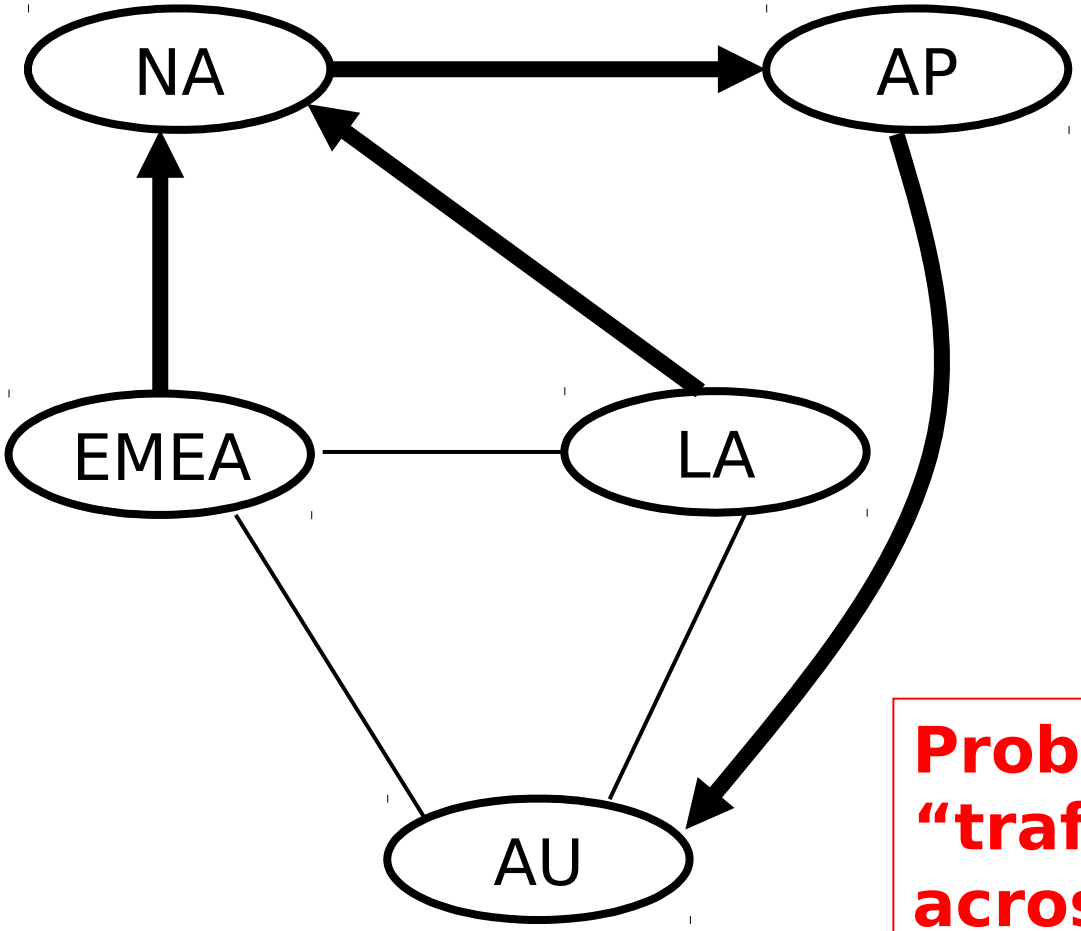
A lot of “non-local” knowledge is required to arrive at this recovery strategy!

Try to convince AS 5 and AS 1 that their session has be reset (or filtered) even though it is not associated with an active route!

That Can’t happen in MY network!



Does this look familiar?



Intended Routing for some prefixes in AU, implemented with communities.

Problems can arise with “traffic engineering” across regional networks.

A wee bit of theory

- Theoretical work has shown that policy-rich routing (such as in BGP) requires a new mathematical framework.
- Why? The theory of shortest-path routing (and closely related approaches) relies on the equation **p(best(a, b)) = best(p(a), p(b)),** where **a** and **b** are routes and **p(_)** represents the application of some policy **p** to a route.
- I'll call a protocol **policy-rich** if that equation can fail.
- Policy-rich protocols can fail to converge (BGP oscillations) and have multiple disctinct solutions (BGP wedgies).
- **How can we fix this?** New equation: **best(a, p(a)) = a**. That is, a route **a** never becomes more preferred after policy is applied.
- New results: if the new equation holds, then a distributed Bellman-Ford computation will always find a unique solution.
- The solution will not be globally optimal --- only **locally optimal** (you get the best routes you can get given what your neighbours give you).

How to ensure **best(x, p(x)) = x**

- The problem : the BGP decision procedure for route **x** is run at one router while the decision procedure for **p(x)** is run at a neighbour.
- Enforcing the equation **within an AS** is not too difficult.
- Enforcing the equation **between Ases** is much more difficult …
- … and requires some additional cooperation between network operators.
- Since the interpretation of **local preference** is local to an AS, the equation cannot be read in absolute terms, but only in relative terms (for example, customers might get loc pref of 100 in one AS and 50 in a neighbouring AS).
- Most important : translate **depref me** communities from one neighbor into **depref me** communities for all of your other neighbors.

Thank you!

Questions?

Some References

“Classical” algebraic theory of Shortest-path-like protocols:

- Baras, J.S., Theodorakopoulos, G.: Path problems in networks. Synthesis Lectures on Communication Networks (2010)
- A model of Internet routing using semi-modules. John N. Billings and Timothy G. Griffin. RelMiCS 2009. http://www.cl.cam.ac.uk/~tgg22/publications/relmics2009.pdf

Algebraic theory of BGP-like (policy-rich) protocols:

- RFC 4264
- João Luís Sobrinho. 2005. An algebraic theory of dynamic network routing. IEEE/ACM Transactions on Networking (TON) 13, 5 (2005), 1160–1173.
- Daggitt, M.L., Gurney, A.J.T., Griffin, T.G.: Asynchronous convergence of policy-rich distributed bellman-ford routing protocols. In: SIGCOMM proceedings. ACM (2018) --- to appear in September 2018