Last Lecture: Network Layer

1. *Design goals and issues*
2. *Basic Routing Algorithms & Protocols*
3. *Addressing, Fragmentation and reassembly*
4. *Internet Routing Protocols and Inter-networking* ✔
   - *Intra- and Inter-domain Routing Protocols* ✔
   - *Introduction to BGP* ✔
   - *Why is routing so hard to get right?*
5. *Router design*
6. *Congestion Control, Quality of Service*
7. *More on the Internet's Network Layer*

This Lecture: Network Layer

1. *Design goals and issues*
2. *Basic Routing Algorithms & Protocols*
3. *Addressing, Fragmentation and reassembly*
4. *Internet Routing Protocols and Inter-networking*
   - *Intra- and Inter-domain Routing Protocols*
   - *Introduction to BGP*
   - *Why is routing so hard to get right?* ✔
   - *Credits: slides from Jennifer Rexford, Nick Feamster, Hari Balakrishnan, Timothy Griffin ICNP'02 Tutorial, Xin Hu & Z. Morley Mao*
5. *Router design*
6. *Congestion Control, Quality of Service*
7. *More on the Internet's Network Layer*

BGP is a Headache! (And Thus Opportunity!)

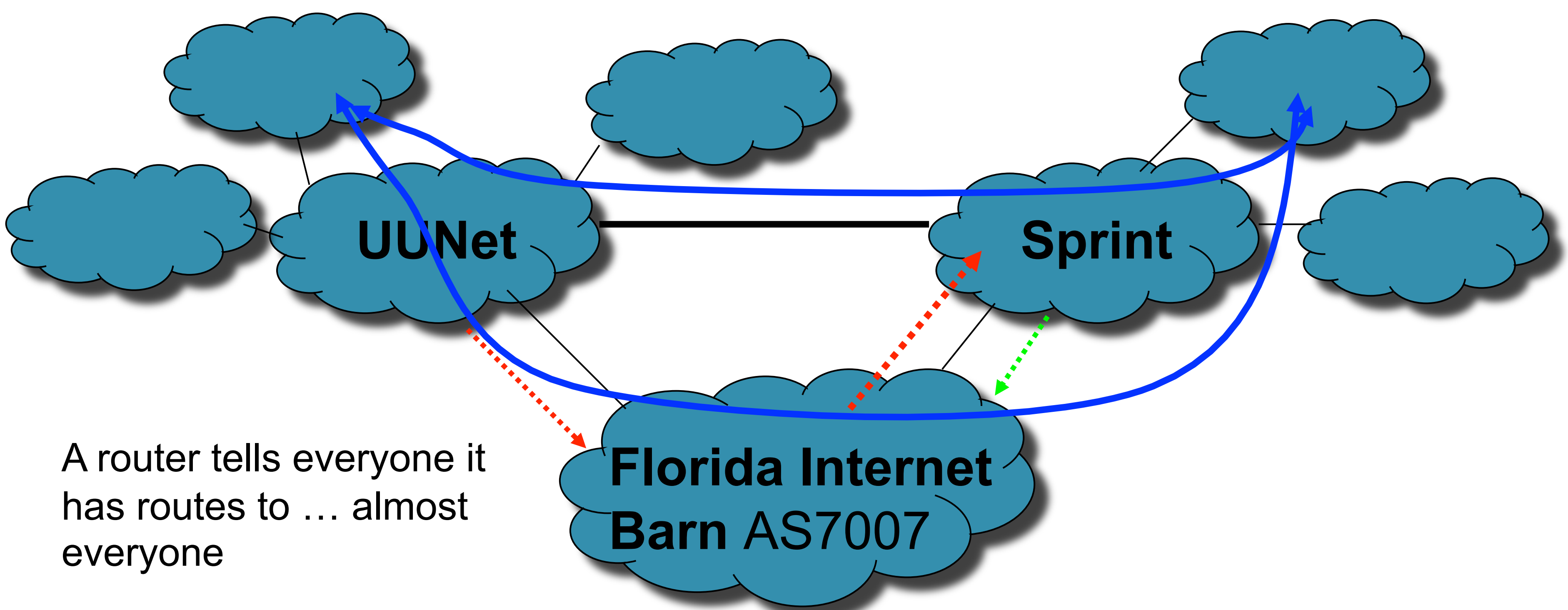
1. *Security:* *e.g.*, prefix hijacking
2. May take a long time to converge
   - *May never converge!*
   - The problem of determining if current policies lead to convergence is *NP-Hard*!
3. *Route oscillations*
4. *Forwarding loops*
5. *Black holes, partition*

- *Broken business model*
  - Depeering can lead to disconnectivity

These Problems Are Real

"…a glitch at a small ISP… triggered a major outage in Internet access across the country. The problem started when MAI Network Services...passed bad router information from one of its customers onto Sprint."

-- *news.com*, April 25, 1997



A router tells everyone it has routes to … almost everyone

These Problems Are Real

- *Apr 2001,* AS3561 propagated > 5000 improper

18:47:00 uninterrupted videos of exploding jello
18:47:45 first evidence of hijacked route propagating in Asia, AS path 3491 17557
18:48:00 several big trans-Pacific providers carrying hijacked route (9 ASNs)
18:48:30 several DFZ providers now carrying the bad route (and 47 ASNs)
18:49:00 most of the DFZ now carrying the bad route (and 93 ASNs)
18:49:30 all providers who will carry the hijacked route have it (total 97 ASNs)
20:07:25 YouTube, AS 36561 advertises the /24 that has been hijacked to its providers
20:07:30 several DFZ providers stop carrying the erroneous route
20:08:00 many downstream providers also drop the bad route
20:08:30 and a total of 40 some-odd providers have stopped using the hijacked route
20:18:43 and now, two more specific /25 routes are first seen from 36561
20:19:37 25 more providers prefer the /25 routes from 36561
20:28:12 peers of 36561 start seeing the routes that were advertised to transit at 20:07
20:50:59 evidence of attempted prepending, AS path was 3491 17557 17557
20:59:39 hijacked prefix is withdrawn by 3491, who disconnect 17557
21:00:00 the world rejoices; Leeroy Jenkins online again.

These Problems Are Real

"…a glitch at a small ISP… triggered a major outage in Internet access across the country. The problem started when MAI Network Services...passed bad router information from one of its customers onto Sprint."

-- *news.com*, April 25, 1997

"Microsoft's websites were offline for up to 23 hours...because of a [router] misconfiguration…it took nearly a day to determine what was wrong and undo the changes."

-- *wired.com*, January 25, 2001

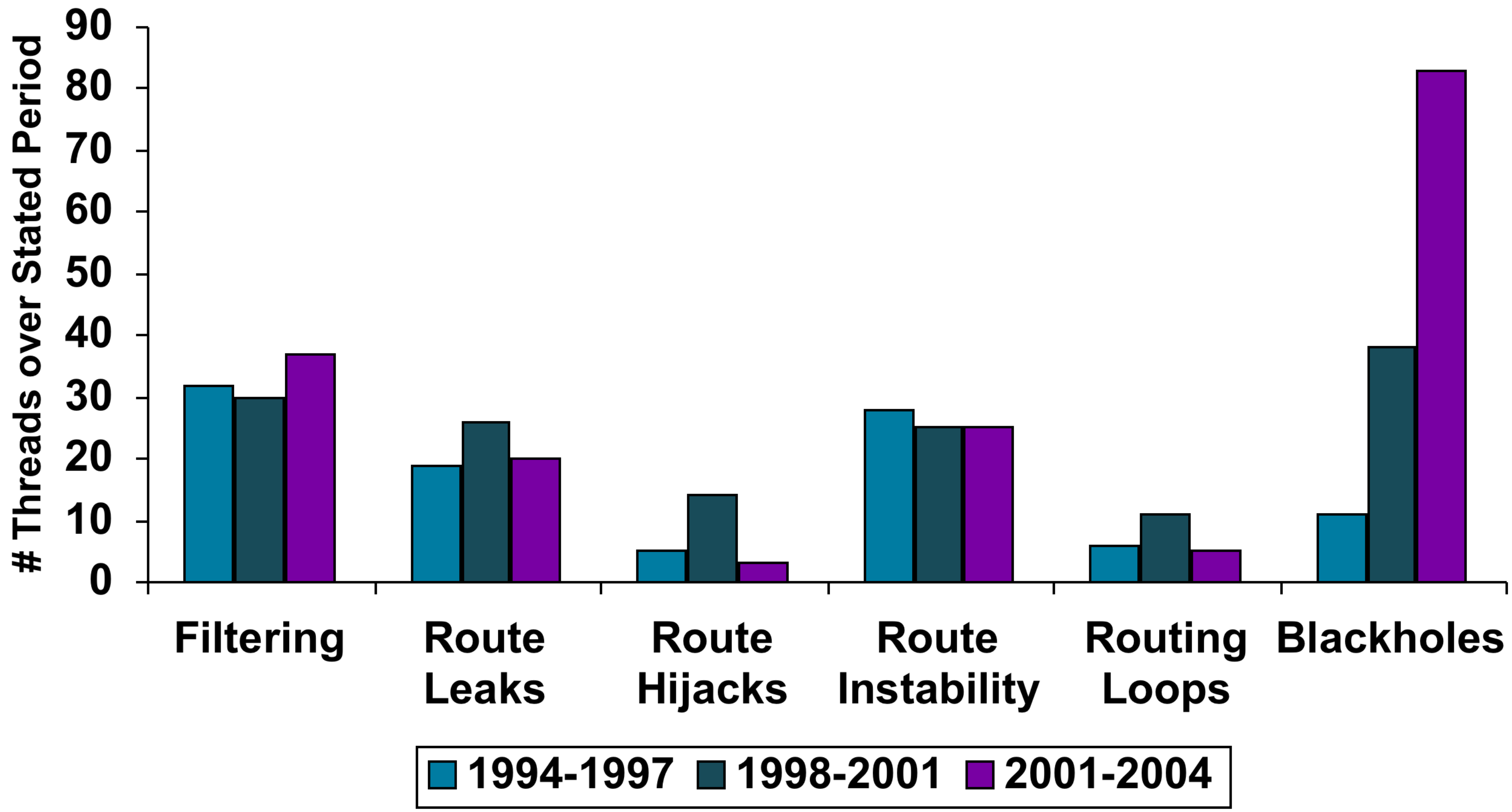
"WorldCom Inc…suffered a widespread outage on its Internet backbone that affected roughly 20 percent of its U.S. customer base. The network problems…affected millions of computer users worldwide. A spokeswoman attributed the outage to "a route table issue.""

-- *cnn.com*, October 3, 2002

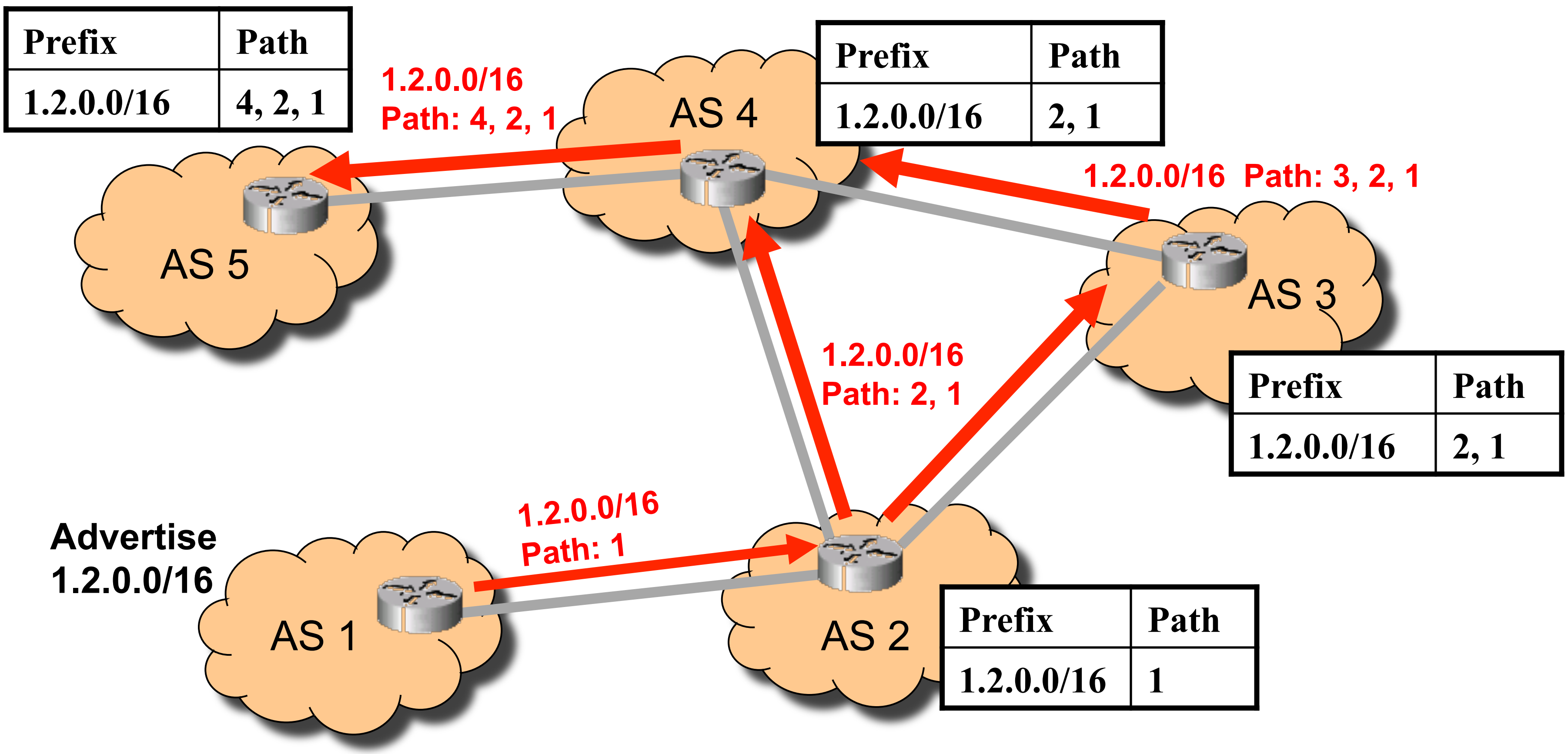
"A number of Covad customers went out from 5pm today due to, supposedly, a DDOS (distributed denial of service attack) on a key Level3 data center, which later was described as a route leak (misconfiguration)."

-- *dslreports.com*, February 23, 2004

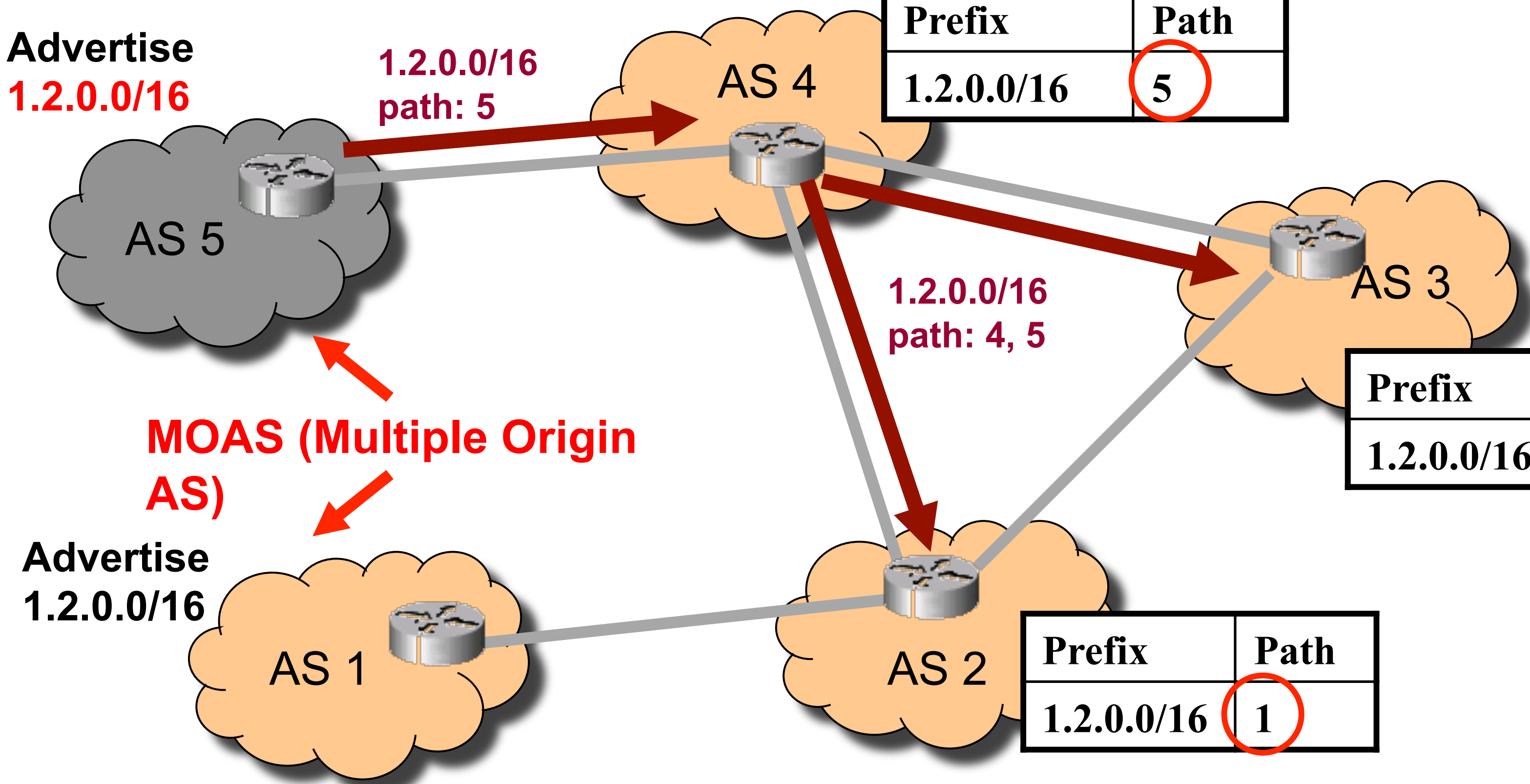
Several "Big" Problems a Week



Reminder: Normal Operations



Type 1: Hijack A Prefix

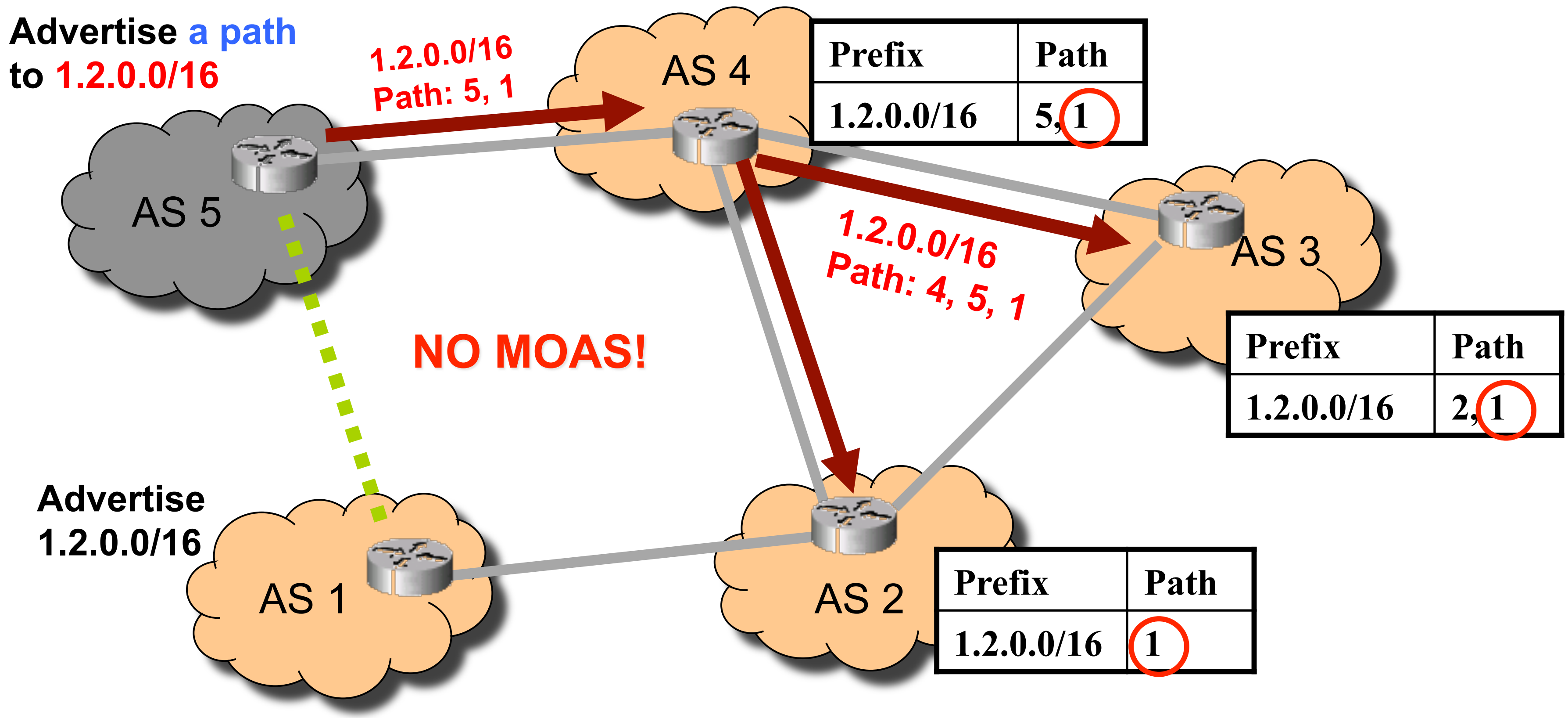


Prefix	Path
1.2.0.0/16	5

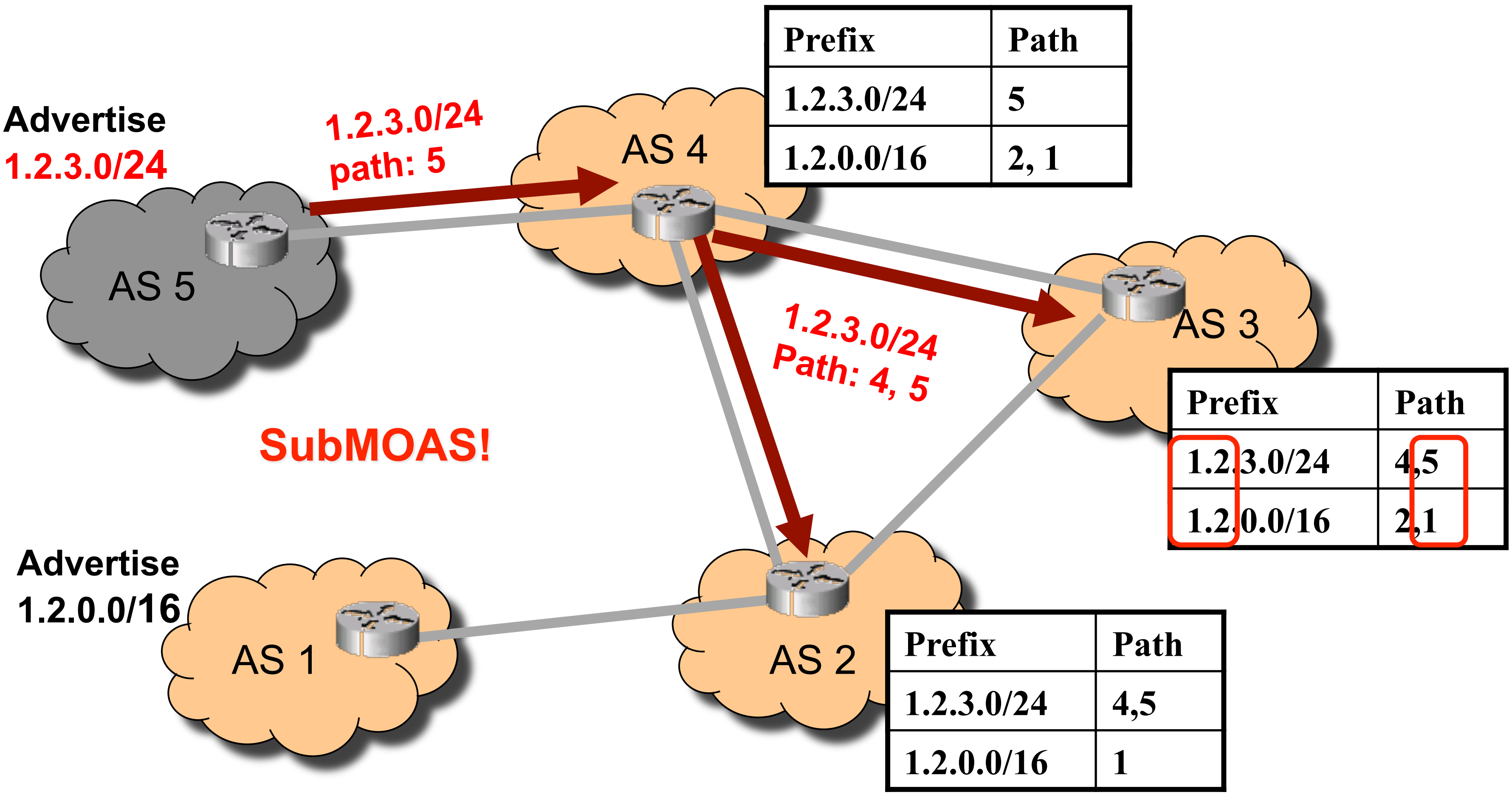
Prefix	Path
1.2.0.0/16	4, 5

Prefix	Path
1.2.0.0/16	1

Type 2: Hijack a Prefix & Its AS Number



Type 3: Hijack a Subnet of the Prefix



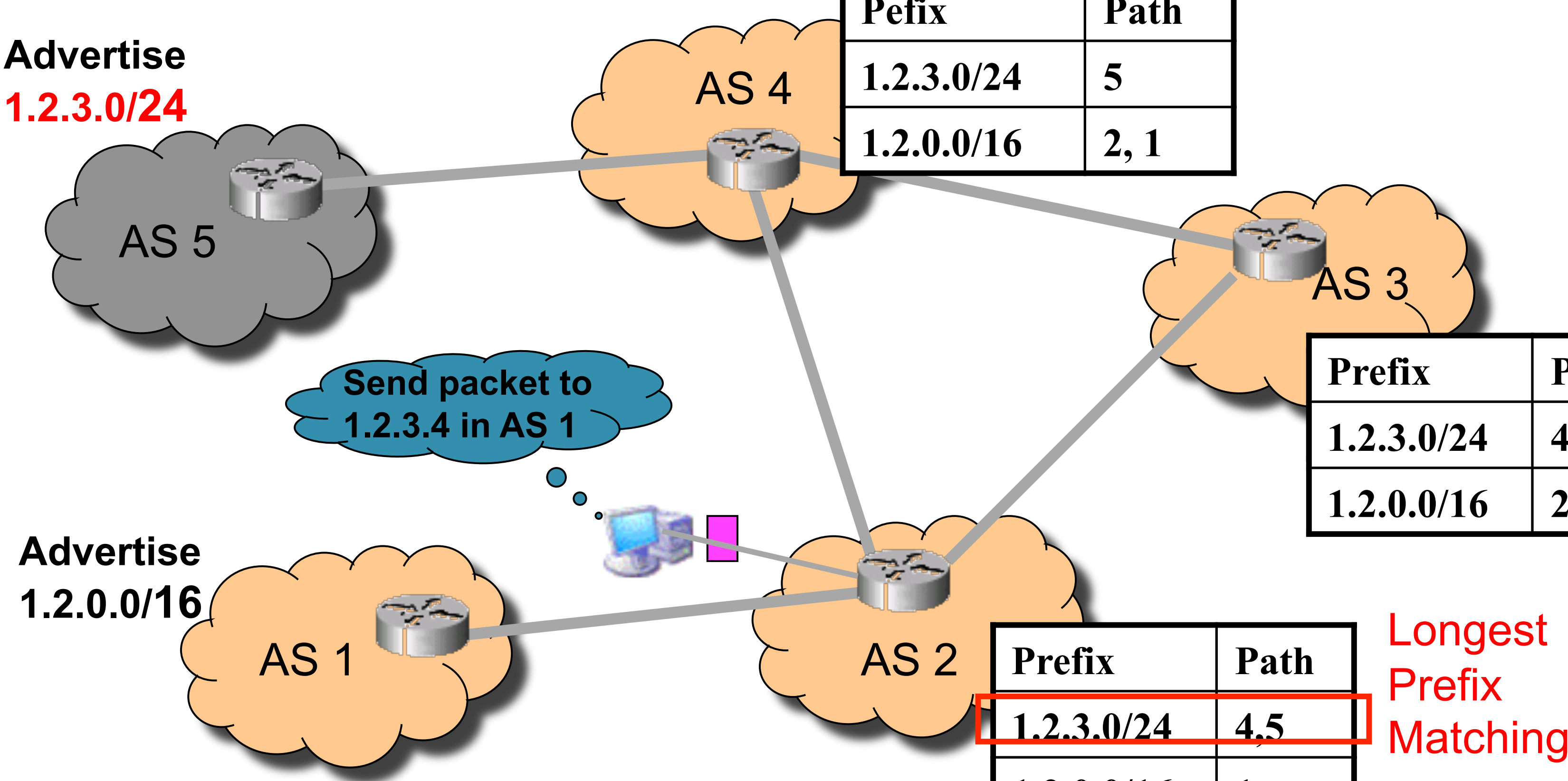
Prefix	Path
1.2.3.0/24	5
1.2.0.0/16	2, 1

Prefix	Path
1.2.3.0/24	4,5
1.2.0.0/16	2,1

Prefix	Path
1.2.3.0/24	4,5
1.2.0.0/16	1

Longest Prefix Match

- Attacker is able to attract all traffic

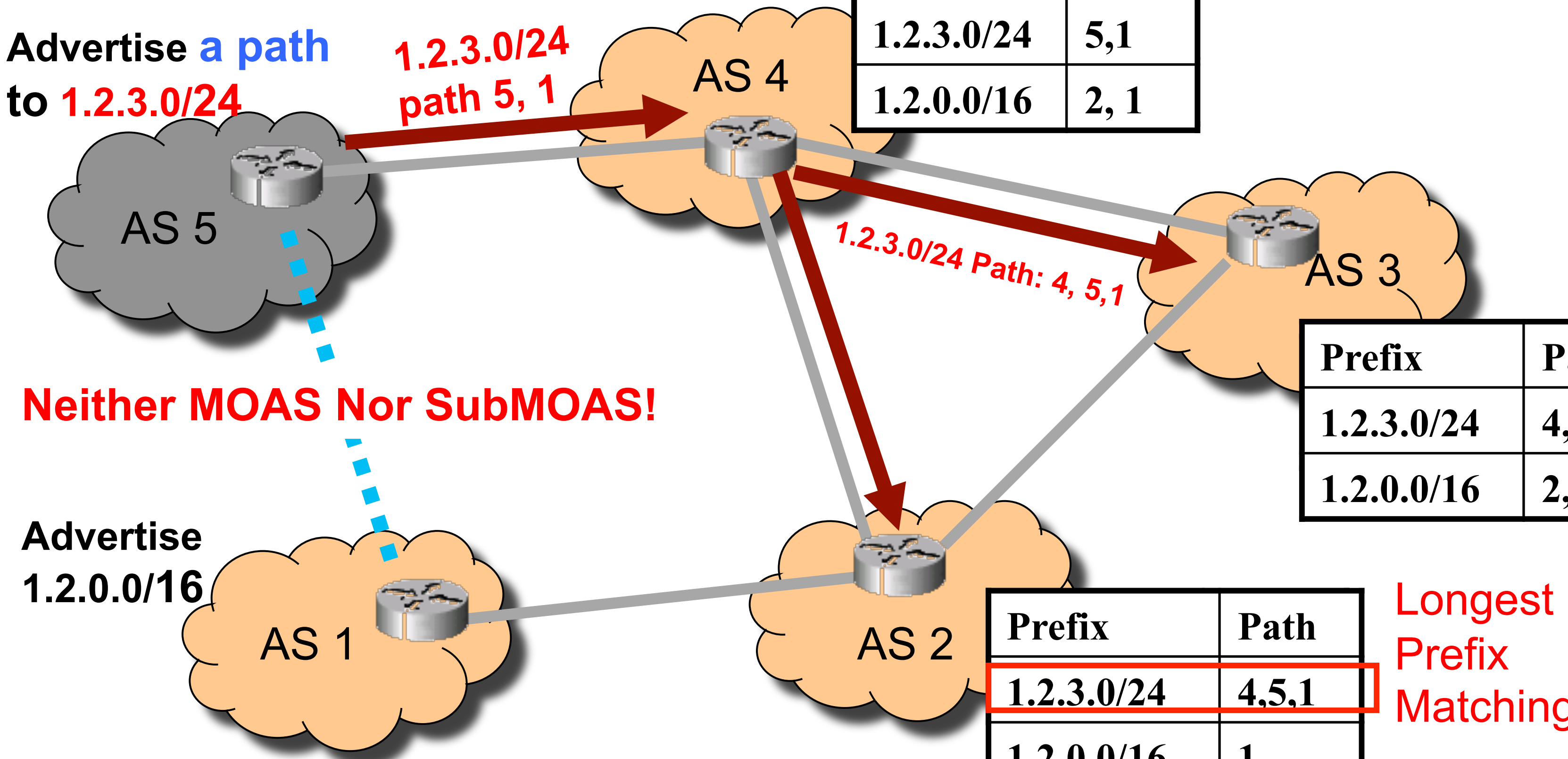


Pefix	Path
1.2.3.0/24	5
1.2.0.0/16	2, 1

Prefix	Path
1.2.3.0/24	4,5
1.2.0.0/16	2, 1

Prefix	Path
1.2.3.0/24	4,5
1.2.0.0/16	1

Type 4: Hijack Subnet & AS Number



Prefix	Path
1.2.3.0/24	5,1
1.2.0.0/16	2, 1

Prefix	Path
1.2.3.0/24	4,5,1
1.2.0.0/16	2, 1

Prefix	Path
1.2.3.0/24	4,5,1
1.2.0.0/16	1

Some Proposed Solutions

- *Prevention*
  - S-BGP,SO-BGP,SPV
- *Mitigation*
  - Wang et.al: PG-BGP,
  - Zhang et al.: AnycastRouting
- *Detect & Alert*
  - myASN, IAR, Phas->Cyclops, BGPmon.net
- *Detect & Recover*
  - Probabilistic IP Prefix Hijacking(PIPA)

- None satisfactory

Frankenstein's Monster: Convergence

- **BGP <u>is not guaranteed</u> to converge on a stable routing. Policy interactions could lead to "livelock" protocol oscillations.**
  See "Persistent Route Oscillations in Inter-domain Routing" by K. Varadhan, R. Govindan, and D. Estrin. ISI report, 1996
- **Corollary: BGP <u>is not guaranteed</u> to recover from network failures.**

Need a theoretical framework to discuss BGP

Griffin, Shepherd, Wilfong – *Transactions on Networking 2002* – gave us an answer

What Problem is BGP solving?

Underlying problem	Distributed means of computing a solution.
Shortest Paths	RIP, OSPF, IS-IS
X?	BGP

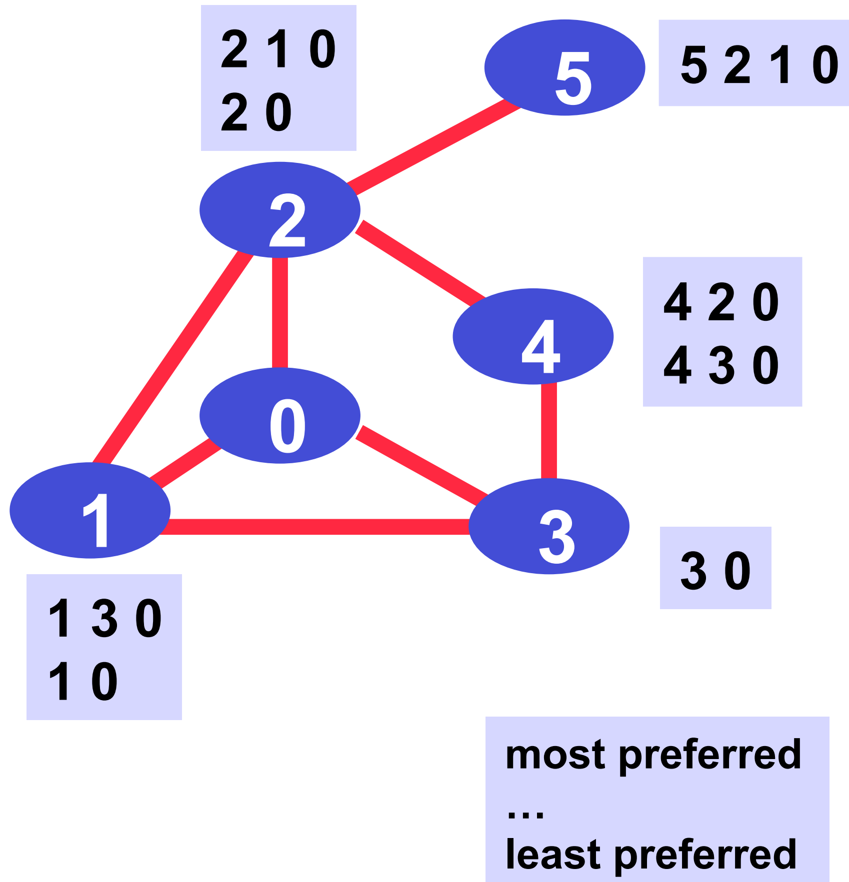
Having an X can

- aid in the design of policy analysis algorithms and heuristics,
- aid in the analysis and design of BGP and extensions,
- help explain some BGP routing anomalies,
- provide a fun way of thinking about the protocol

Our focus

Candidate for X : Stable Paths Problem (SPP)

- A graph of nodes and edges,
- Node 0, called *the origin*,
- For each non-zero node, a set or *permitted paths* to the origin. This set always contains the "*null path*".
- A *ranking* of permitted paths at each node. Null path is always least preferred. (Not shown in diagram)



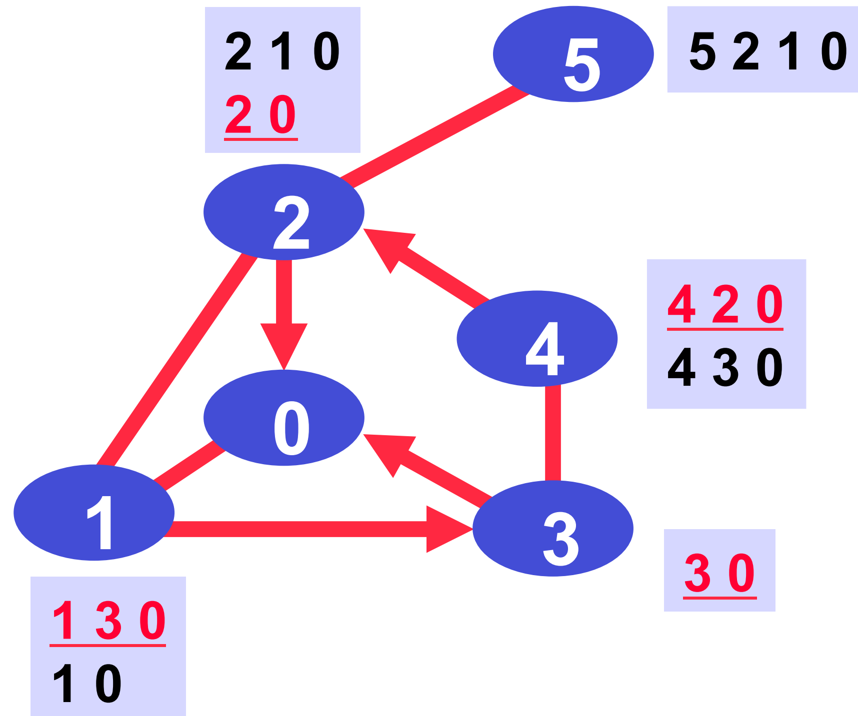
When modeling BGP : nodes represent BGP speaking routers, and 0 represents a node originating some address block

Yes, the translation gets messy!

A Solution to the SPP Problem

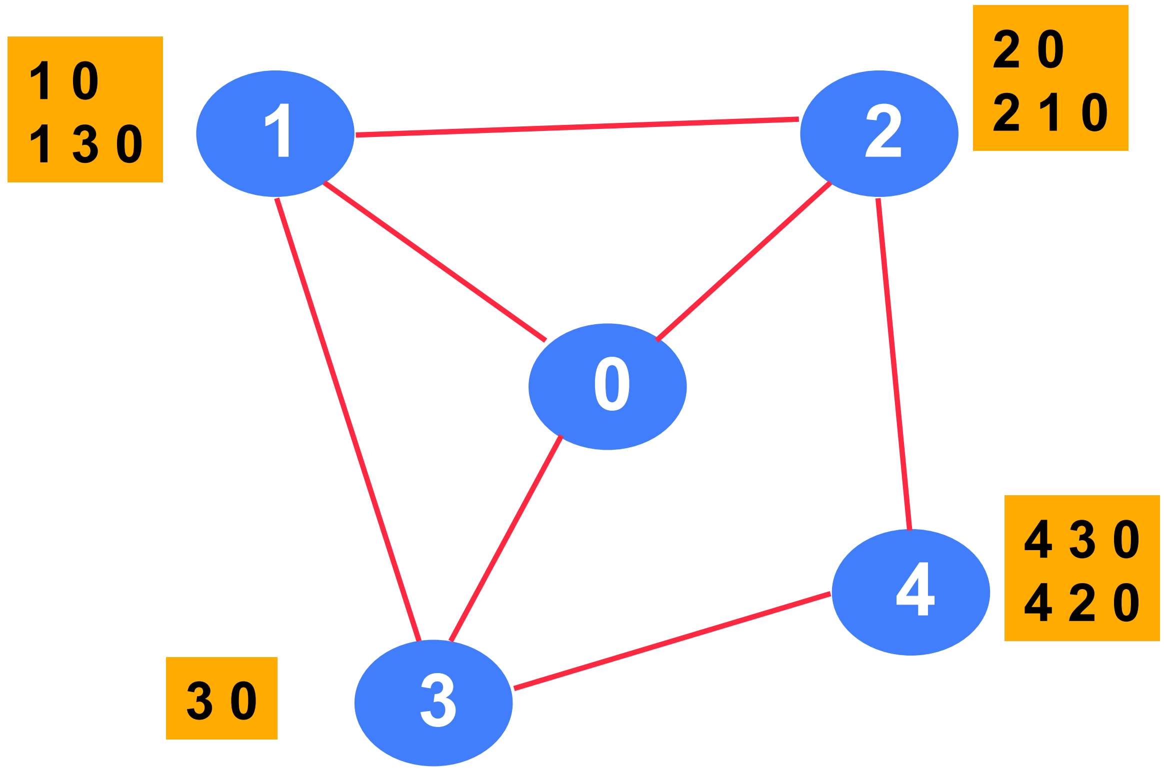
A *solution* is an assignment of permitted paths to each node such that

- node u's assigned path is either the null path or is a path uwP, where wP is assigned to node w and uw is an edge in the graph,
- each node is assigned the highest ranked path among those consistent with the paths assigned to its neighbors.

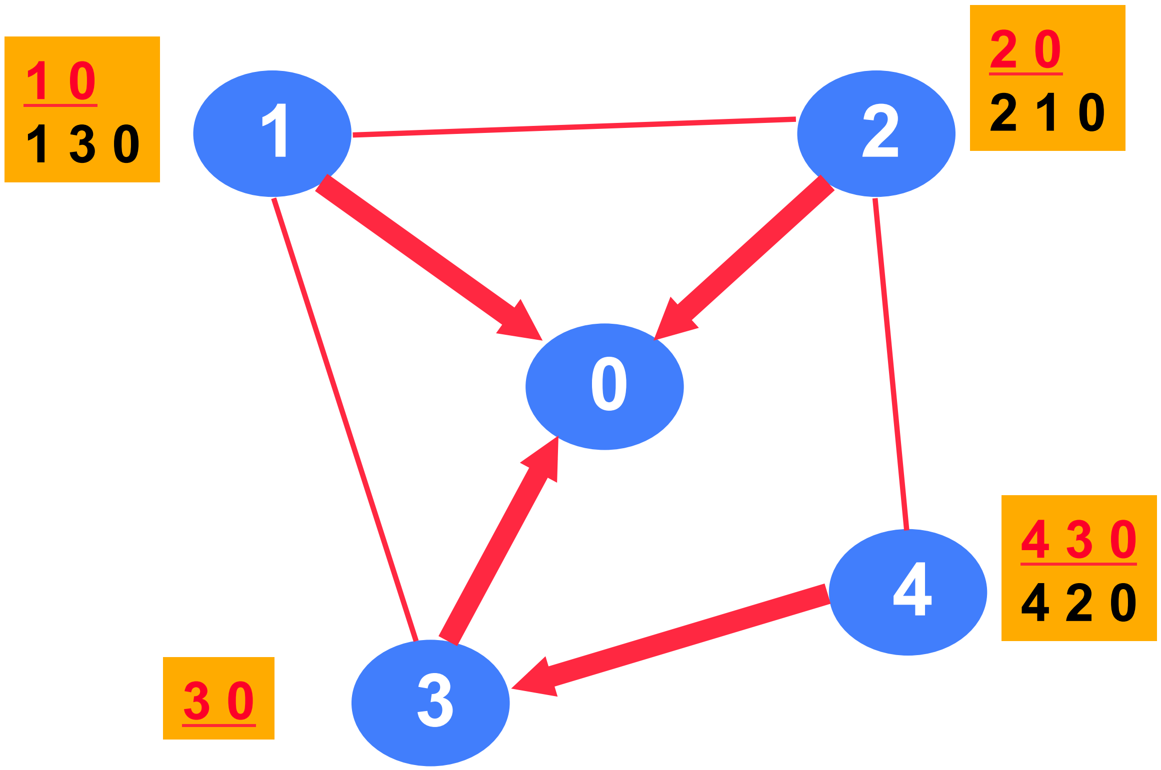


A Solution need not represent
a shortest path tree, or
a spanning tree.

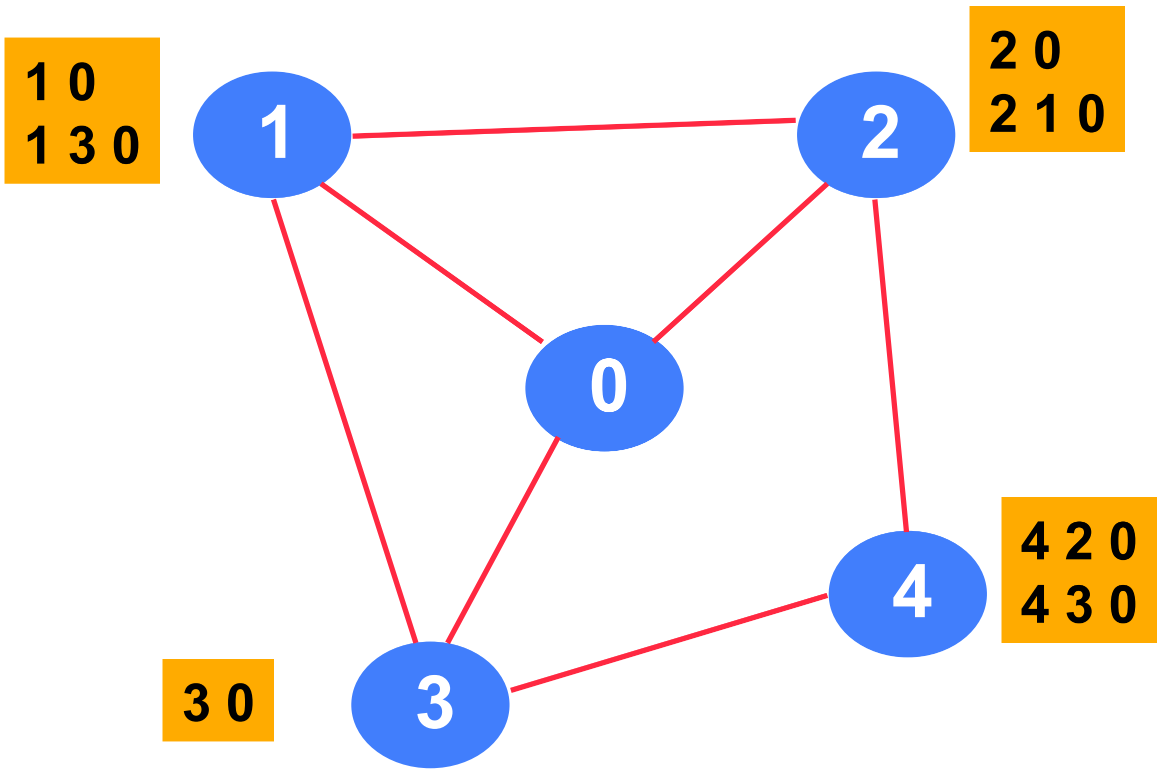
Example 1: Ranking by Shortest Path Length



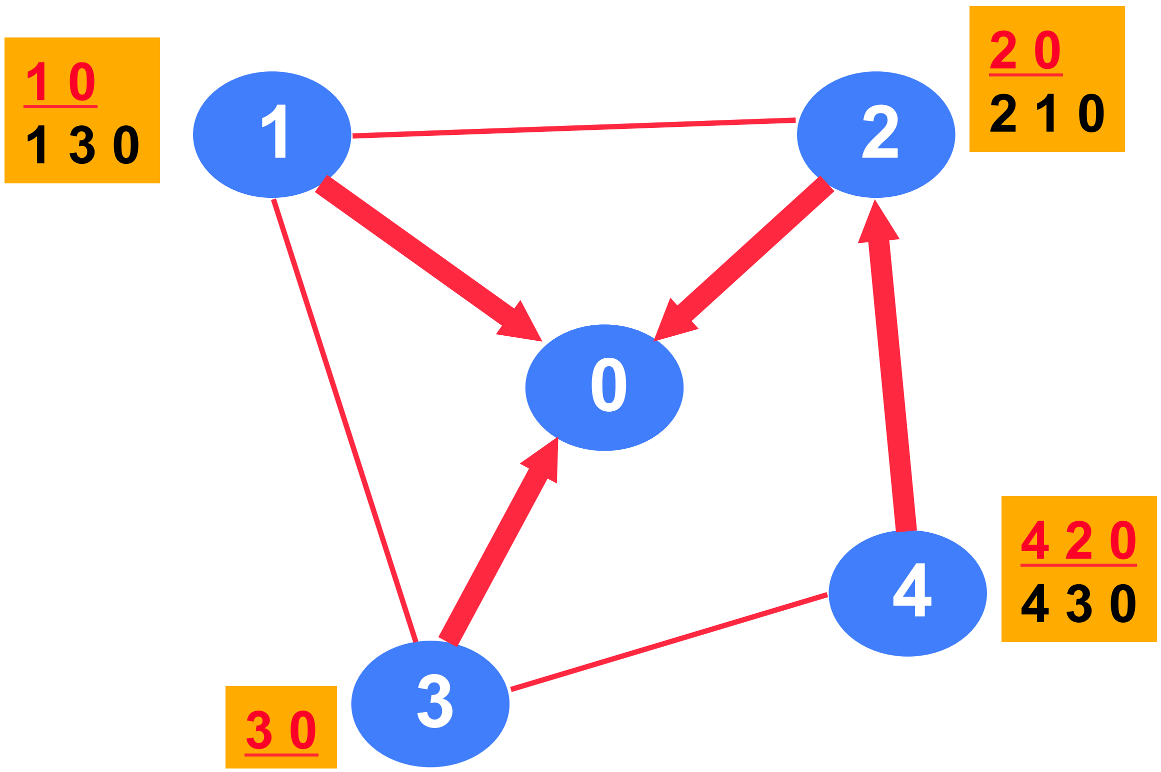
Example 1: Solution



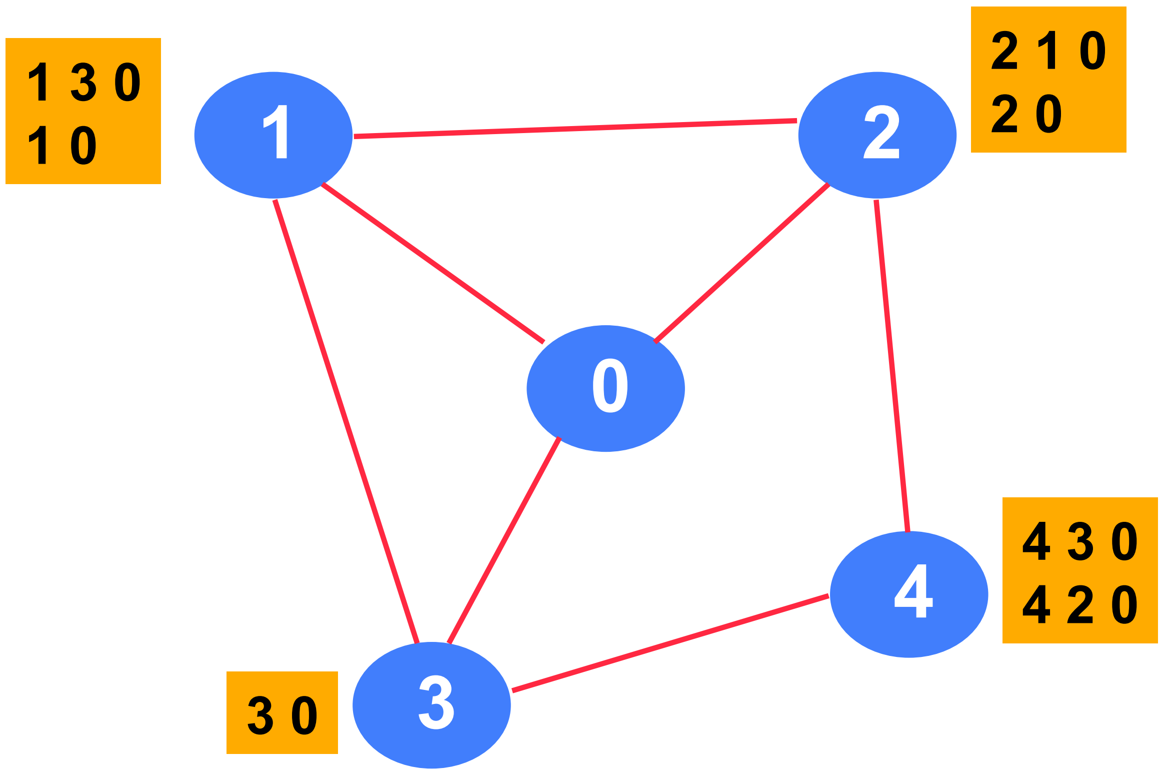
Example 2: Ranking by Shortest Path Length



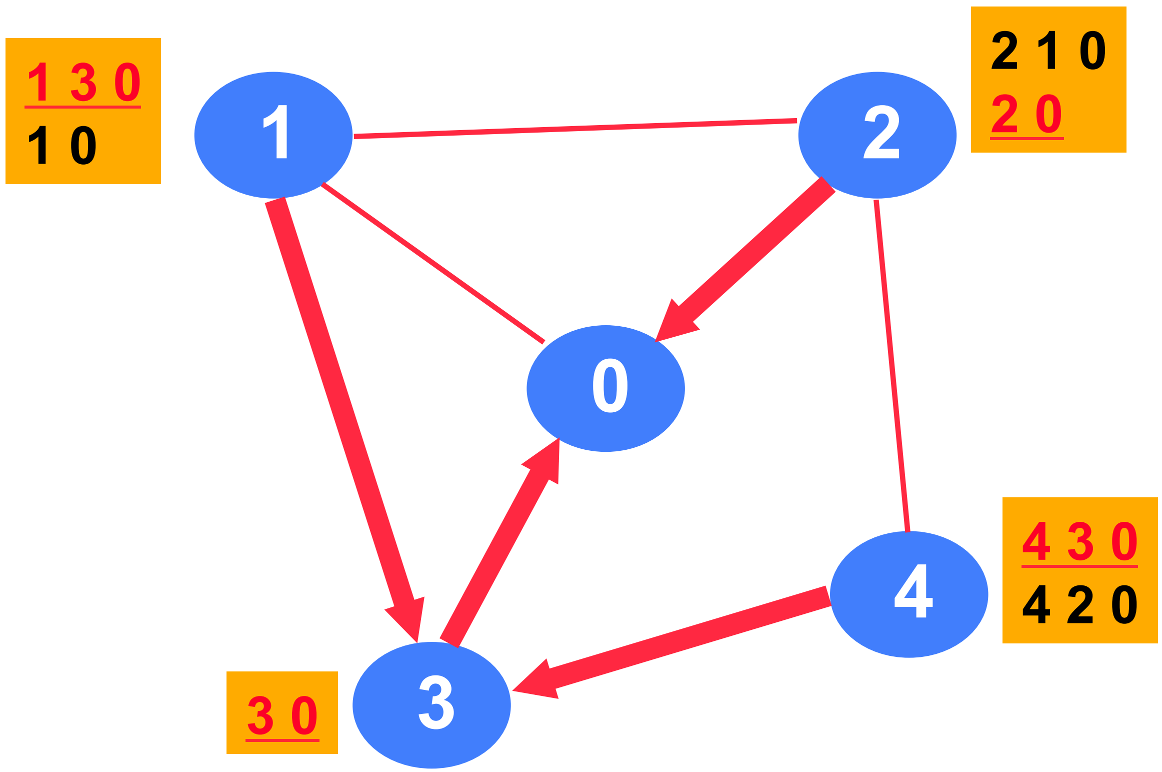
Example 2: Solution



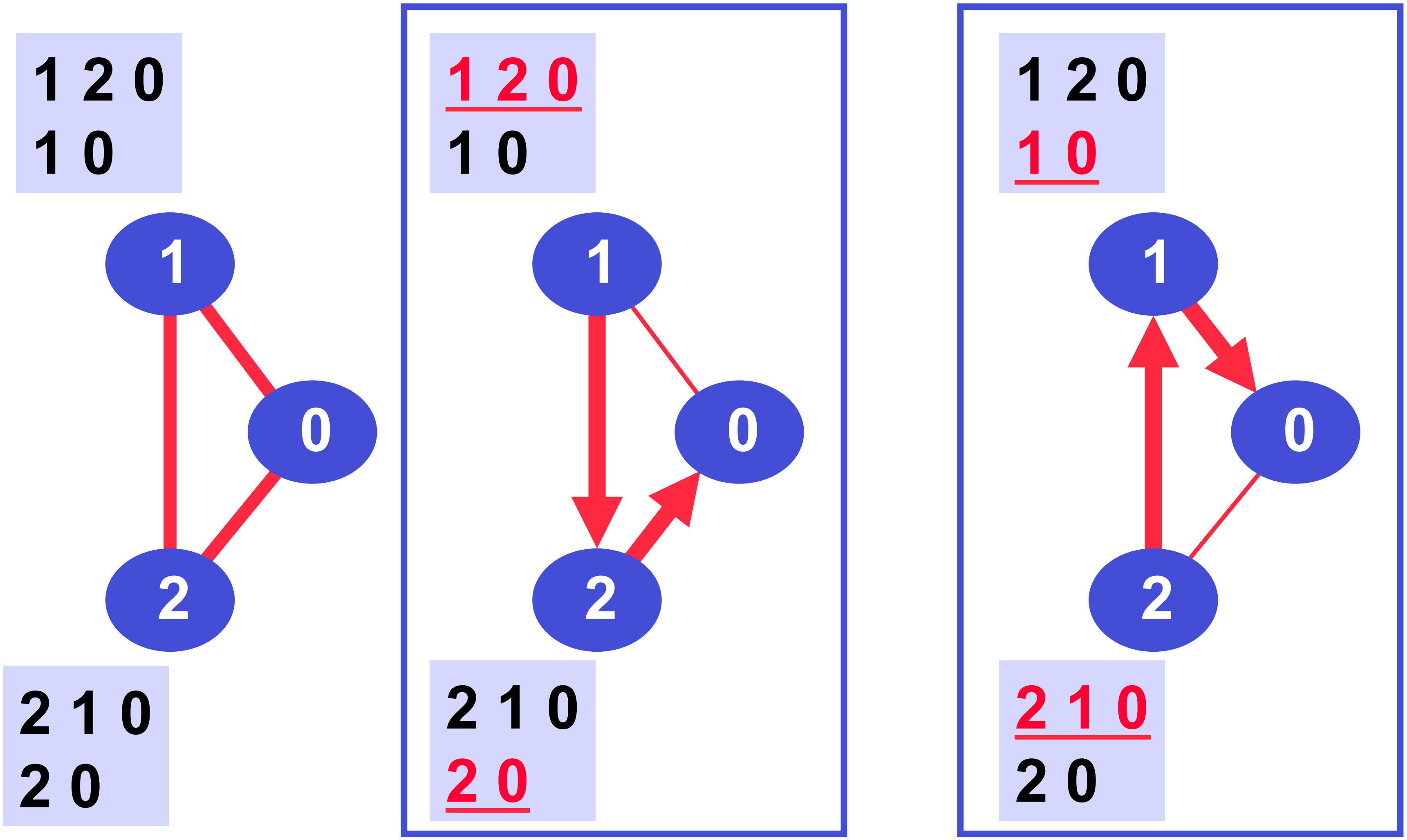
Example 3: (Another) Good Gadget



Example 3: Good Gadget's Solution



Example 4: Multiple Solutions

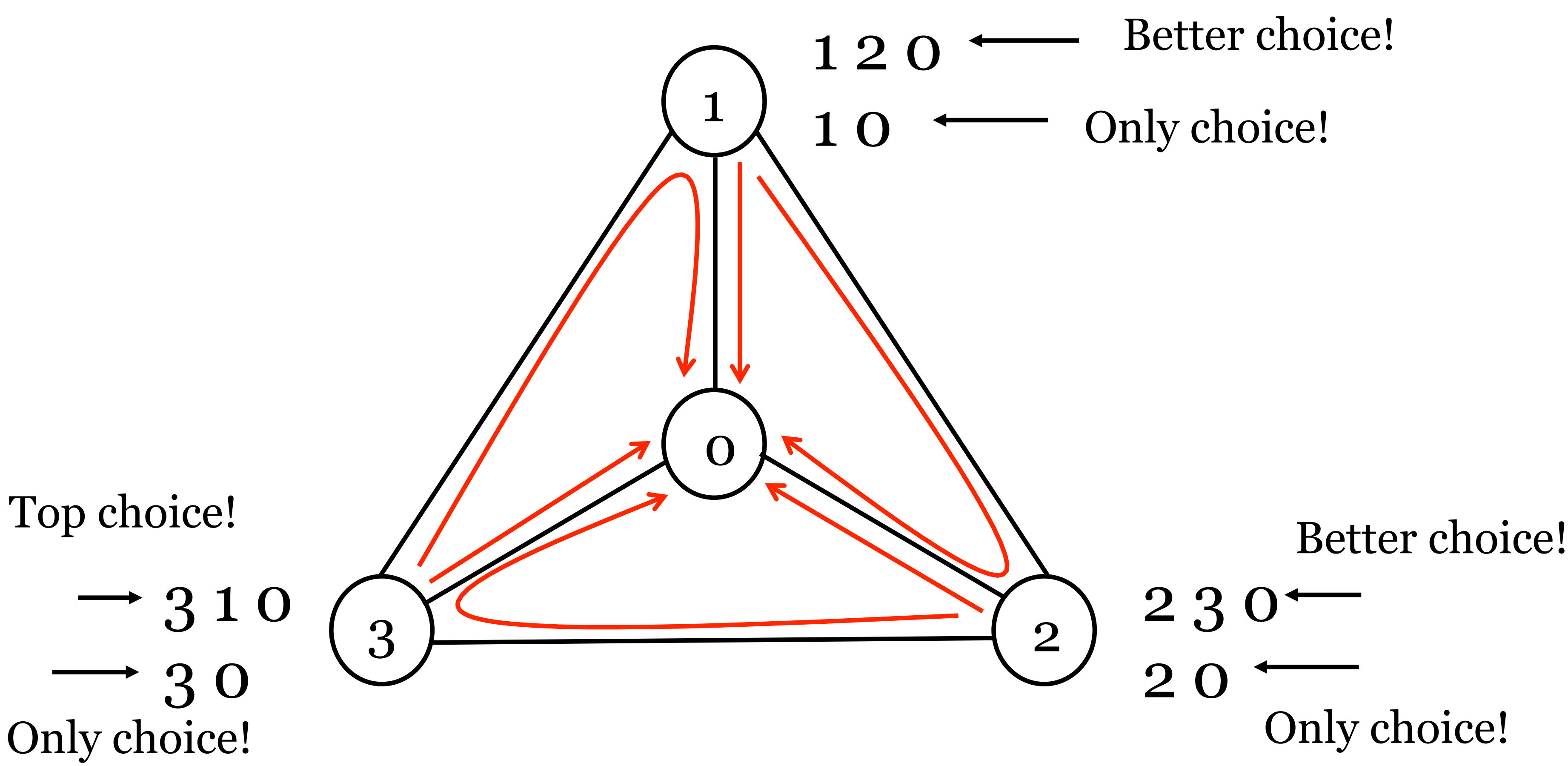


DISAGREE

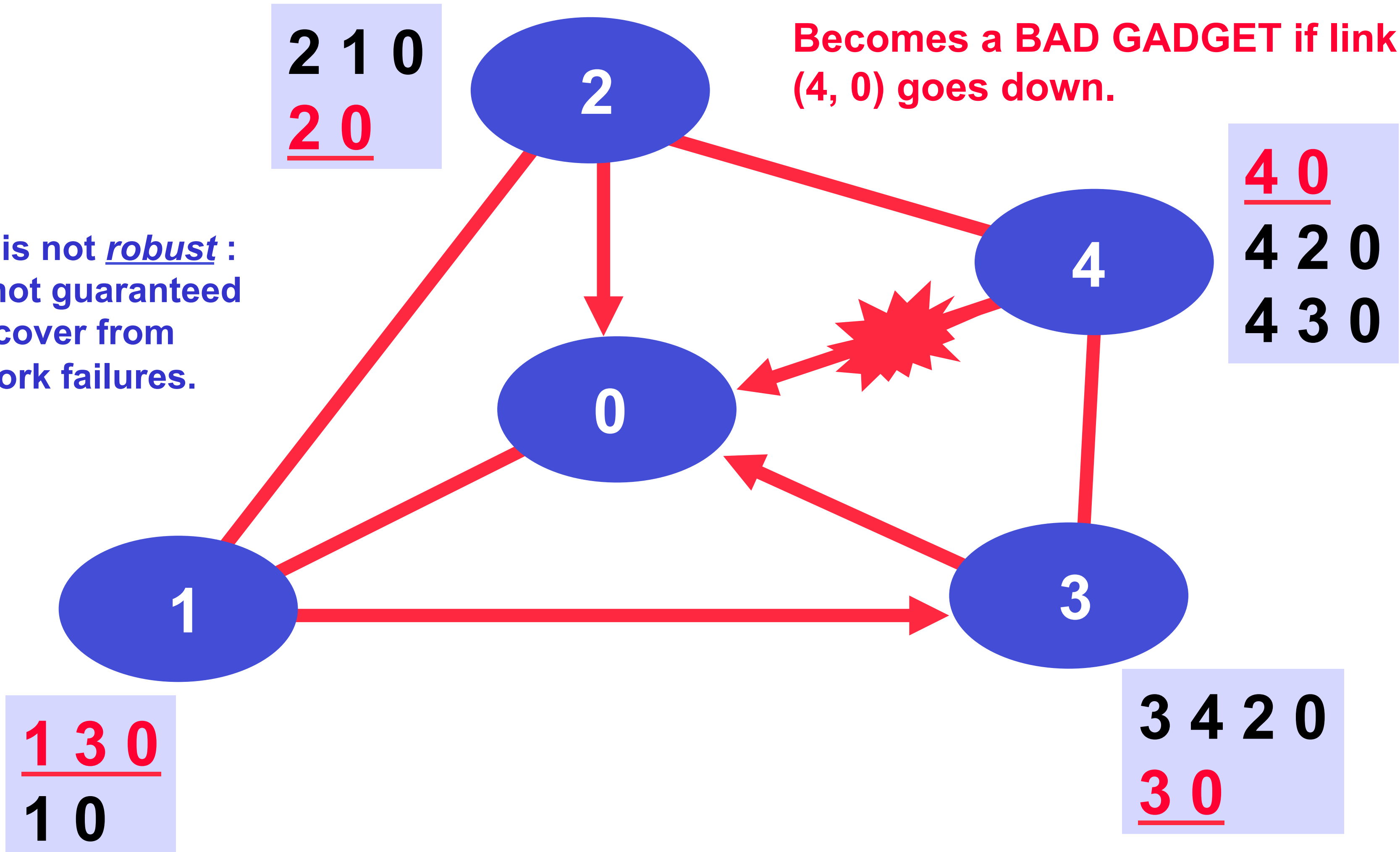
First solution

Second solution

Example 5: No Solution



Link Down, Good Gadget May Become Bad

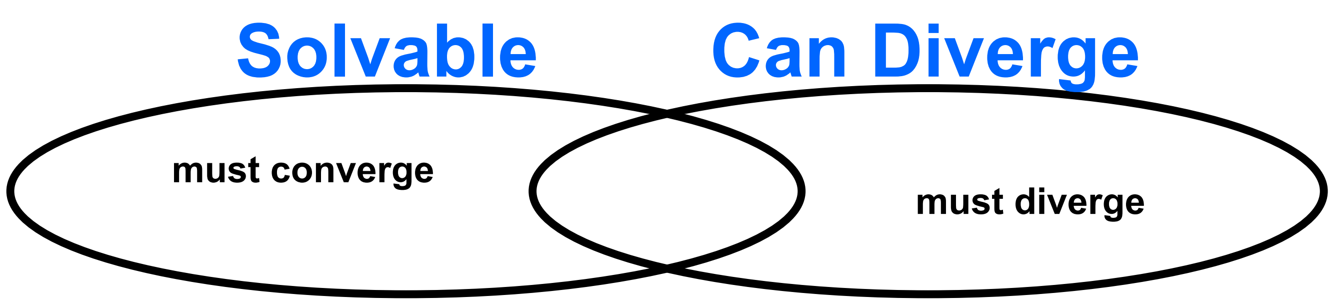


BGP is not *robust* : it is not guaranteed to recover from network failures.

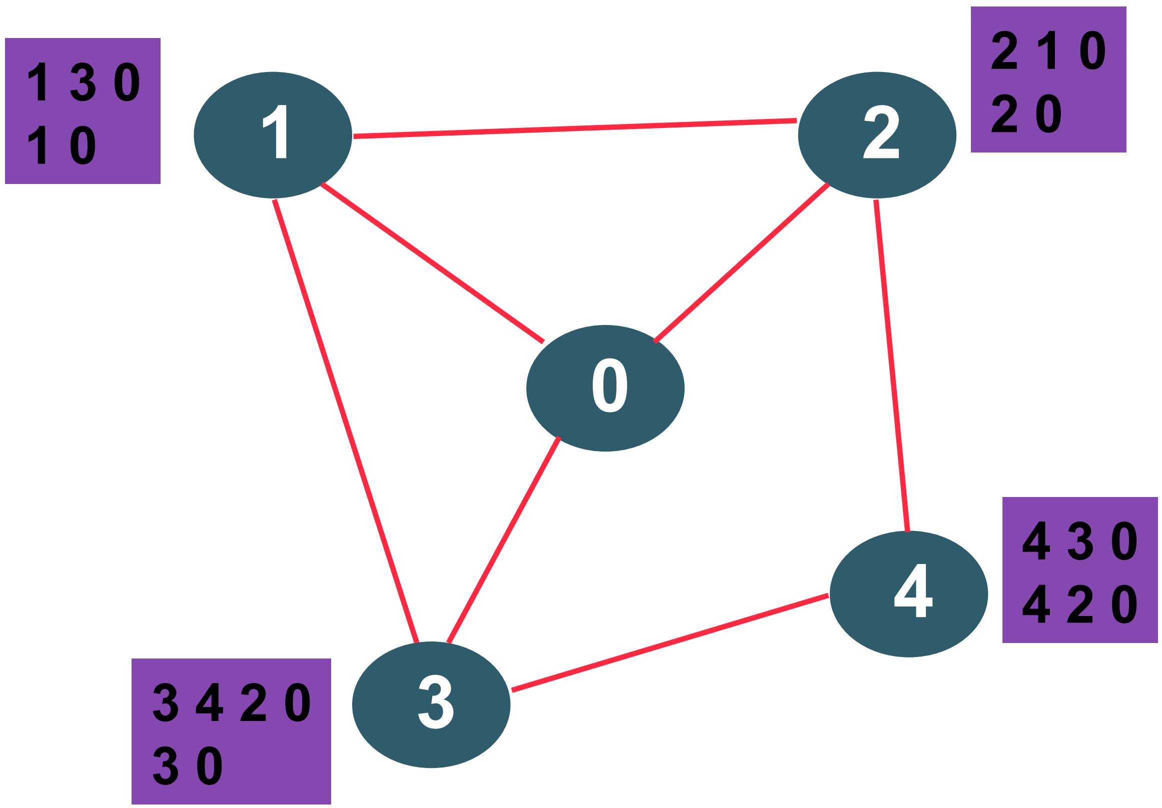
SPP Explains Possibility of BGP Divergence

- BGP is not guaranteed to converge to a stable routing. Policy inconsistencies can lead to "livelock" protocol oscillations.
- See "*Persistent Route Oscillations in Inter-domain Routing*" by K. Varadhan, R. Govindan, and D. Estrin. ISI report, 1996

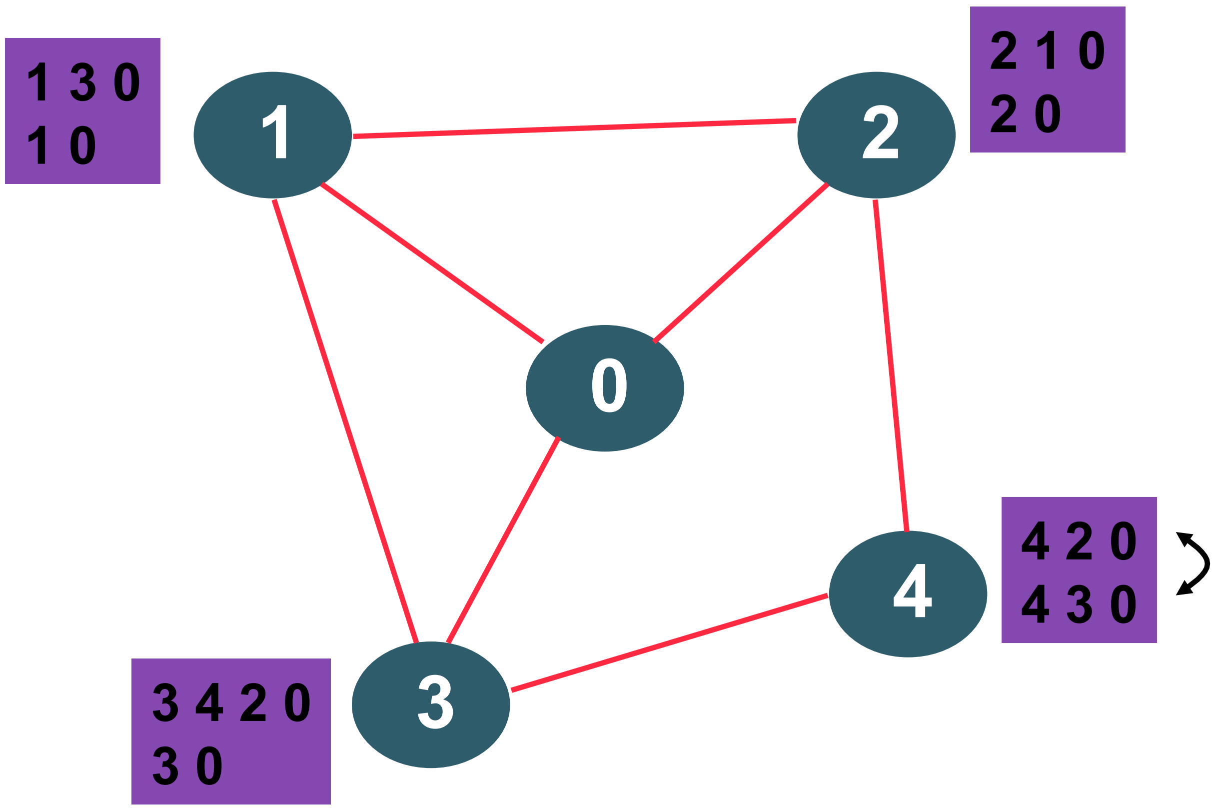
The SPP view :



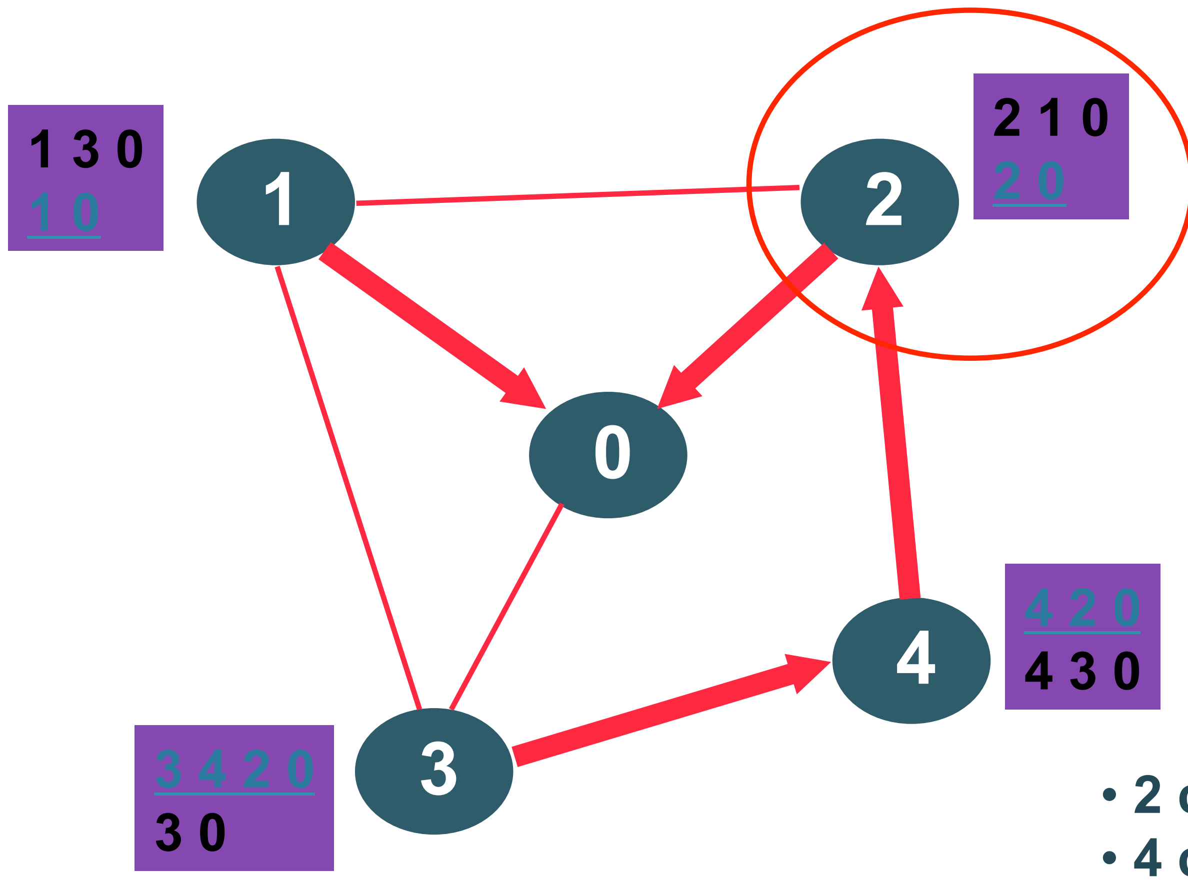
Example: NAUGHTY GADGET



Example: BAD GADGET

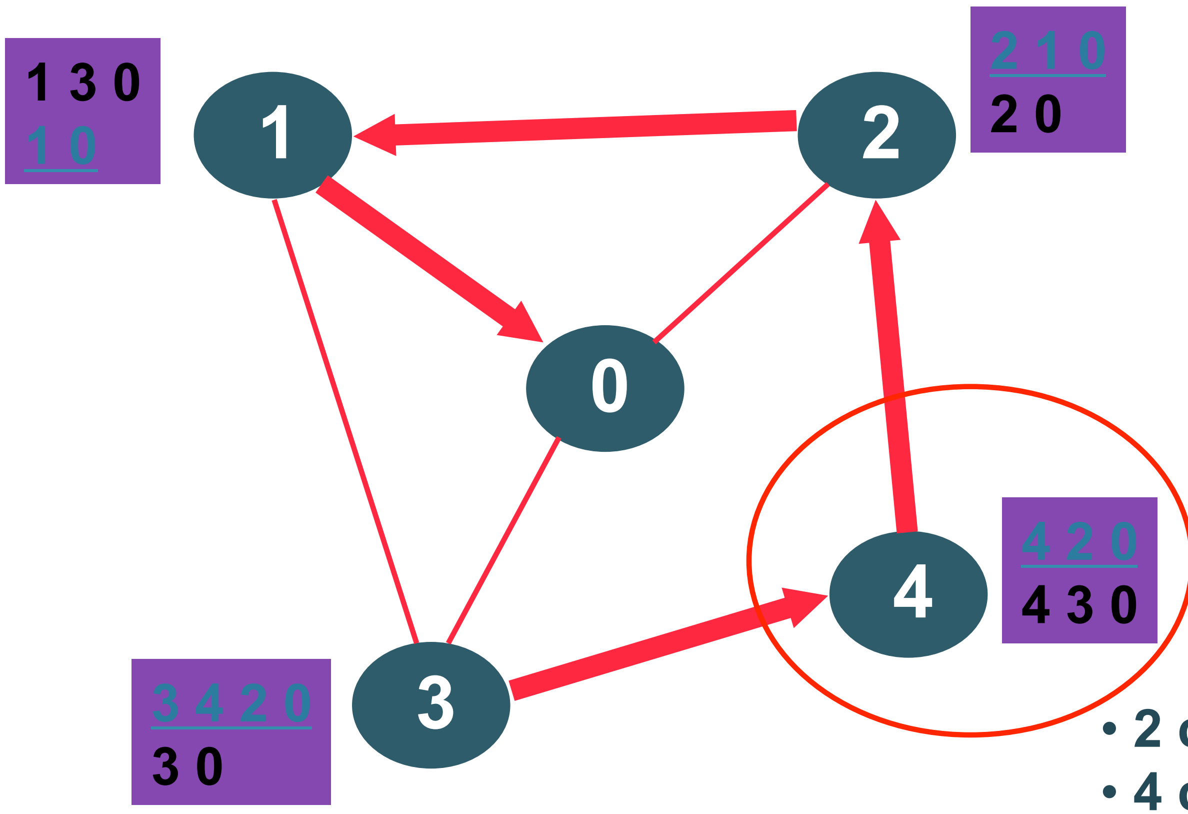


Example: BAD GADGET



- 2 chooses (2 0)
- 4 chooses (4 2 0)
- 3 chooses (3 4 2 0)
- 1 chooses (1 0)

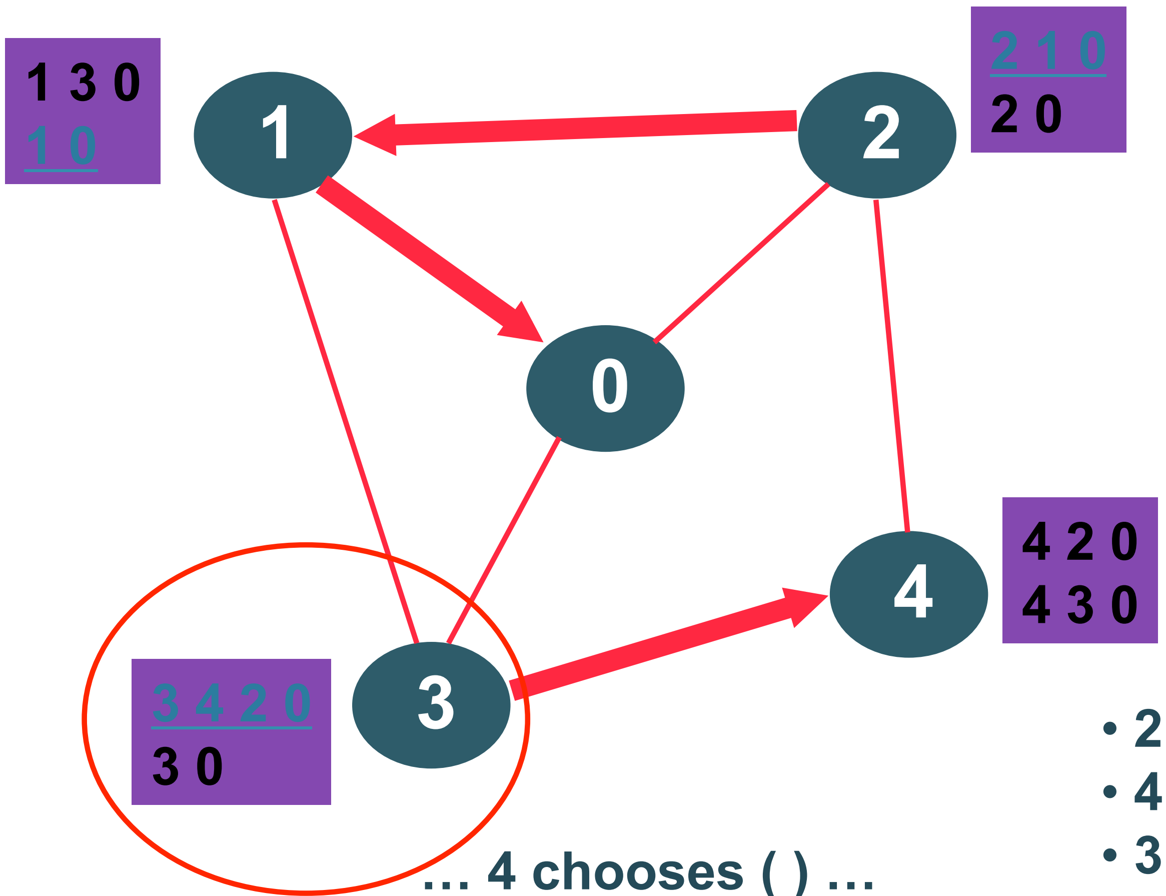
Example: BAD GADGET



… 2 chooses (2 1 0) …

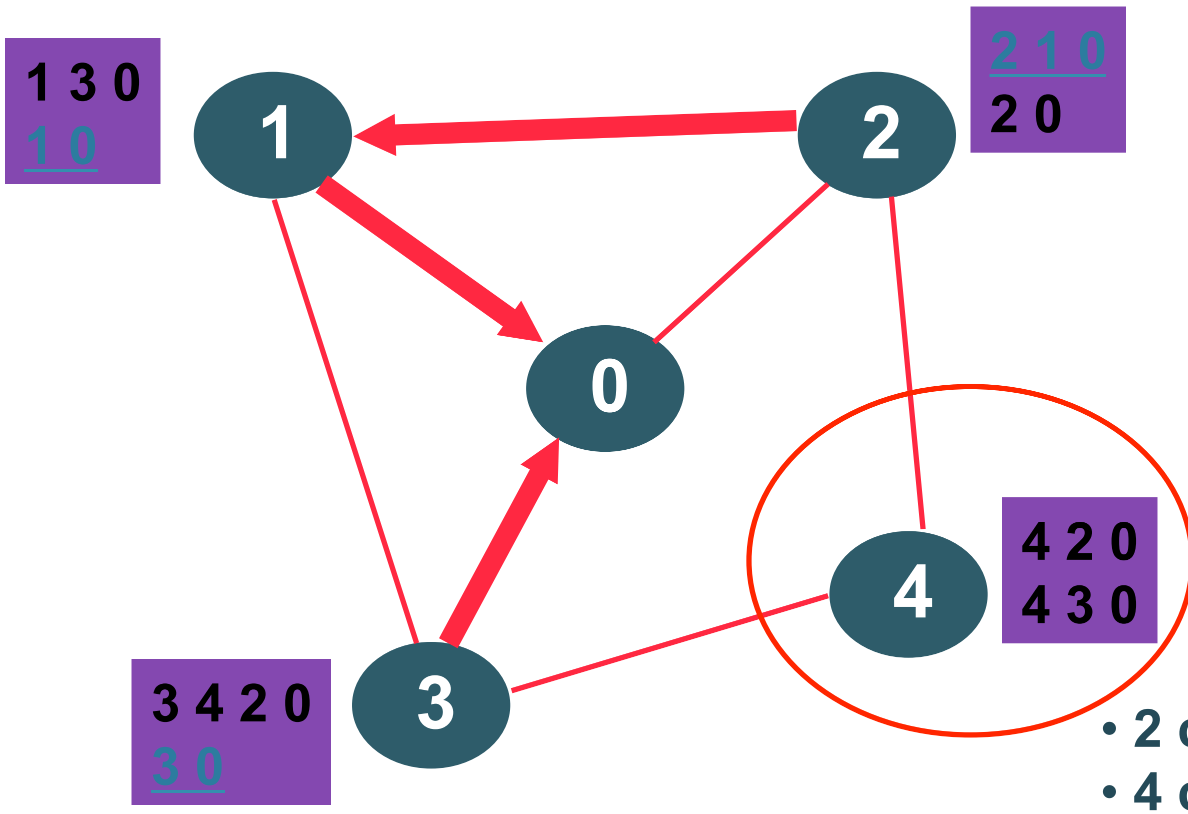
- 2 chooses (2 0)
- 4 chooses (4 2 0)
- 3 chooses (3 4 2 0)
- 1 chooses (1 0)

Example: BAD GADGET



- 2 chooses (2 0)
- 4 chooses (4 2 0)
- 3 chooses (3 4 2 0)
- 1 chooses (1 0)

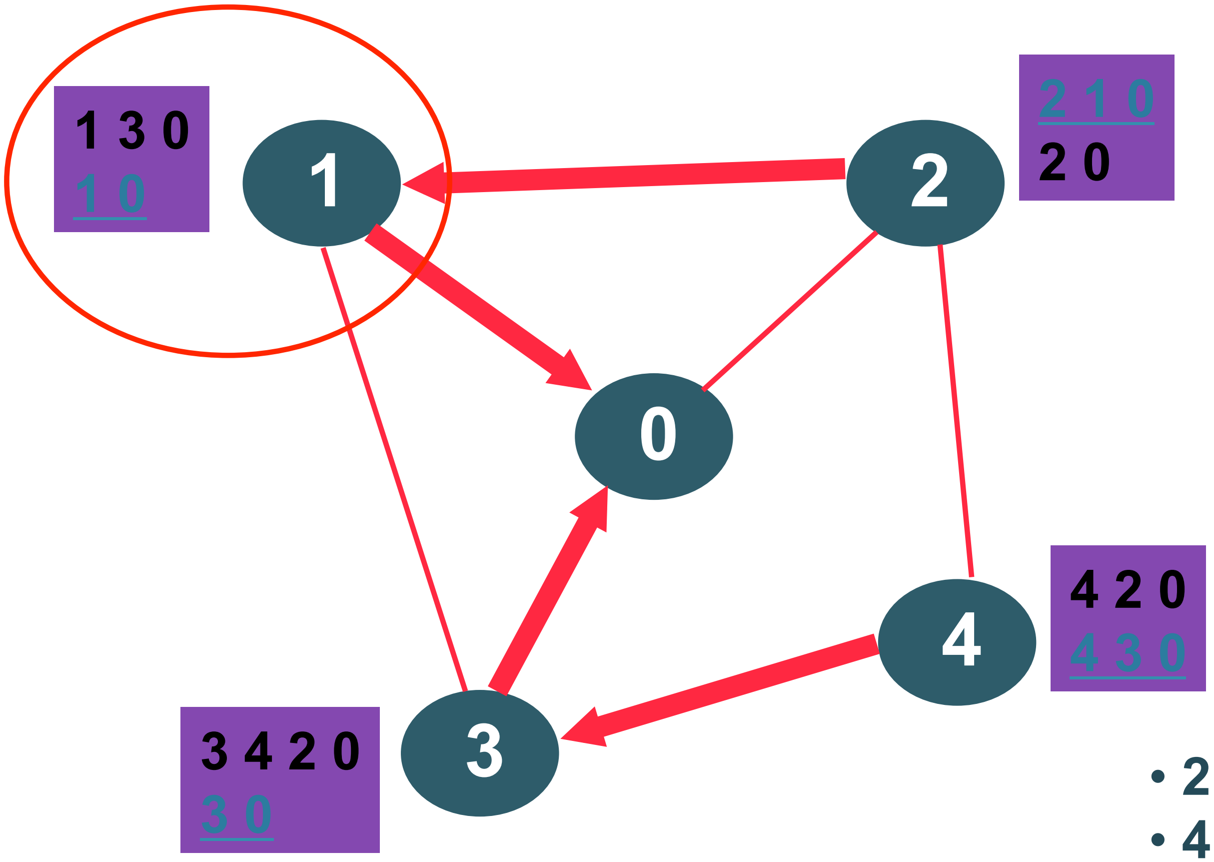
Example: BAD GADGET



… 3 chooses (3 0) …

- 2 chooses (2 0)
- 4 chooses (4 2 0)
- 3 chooses (3 4 2 0)
- 1 chooses (1 0)

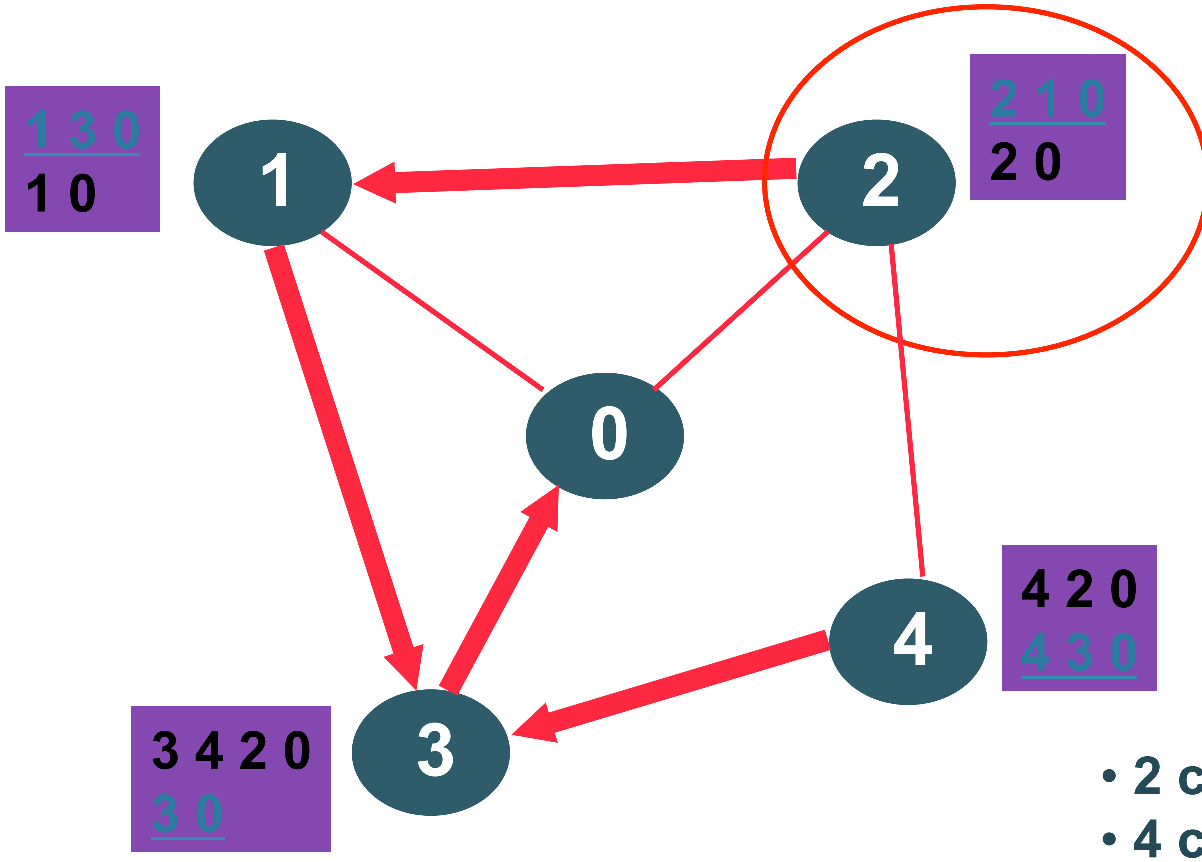
Example: BAD GADGET



… 4 chooses (4 3 0) …

- 2 chooses (2 0)
- 4 chooses (4 2 0)
- 3 chooses (3 4 2 0)
- 1 chooses (1 0)

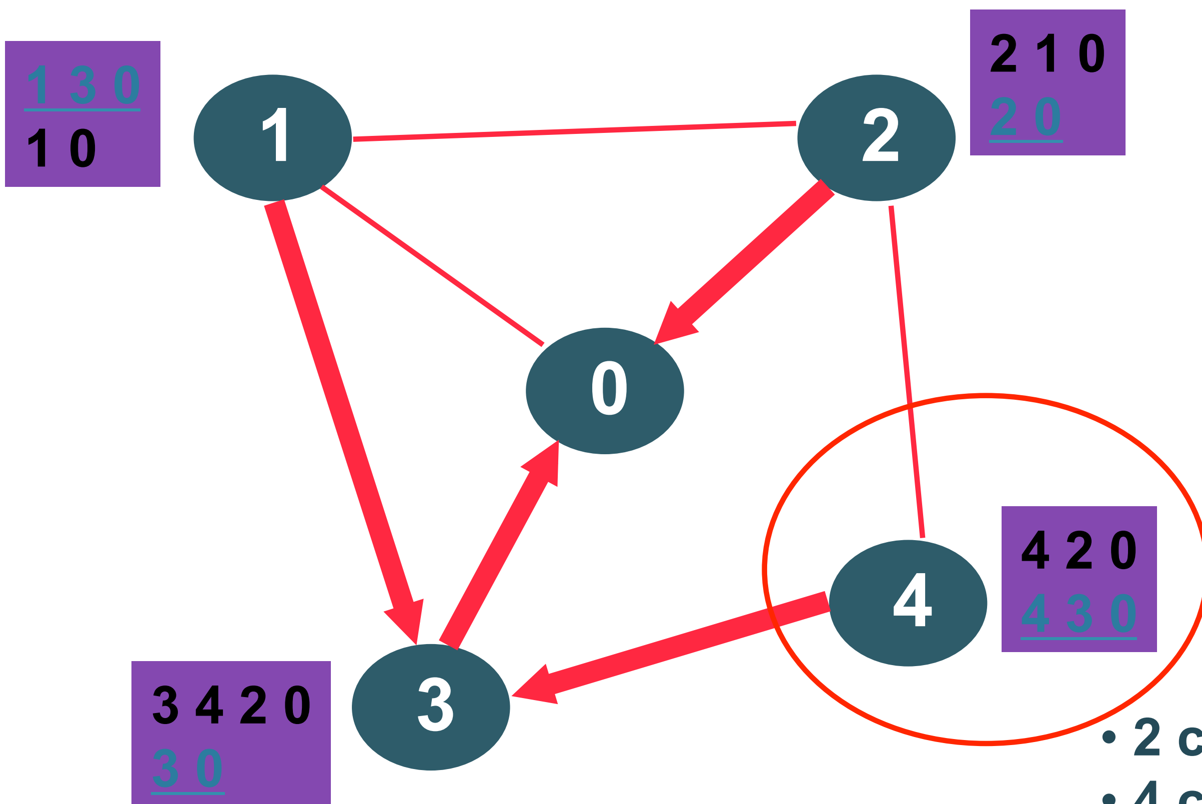
Example: BAD GADGET



… 1 chooses (1 3 0) …

- 2 chooses (2 0)
- 4 chooses (4 2 0)
- 3 chooses (3 4 2 0)
- 1 chooses (1 0)

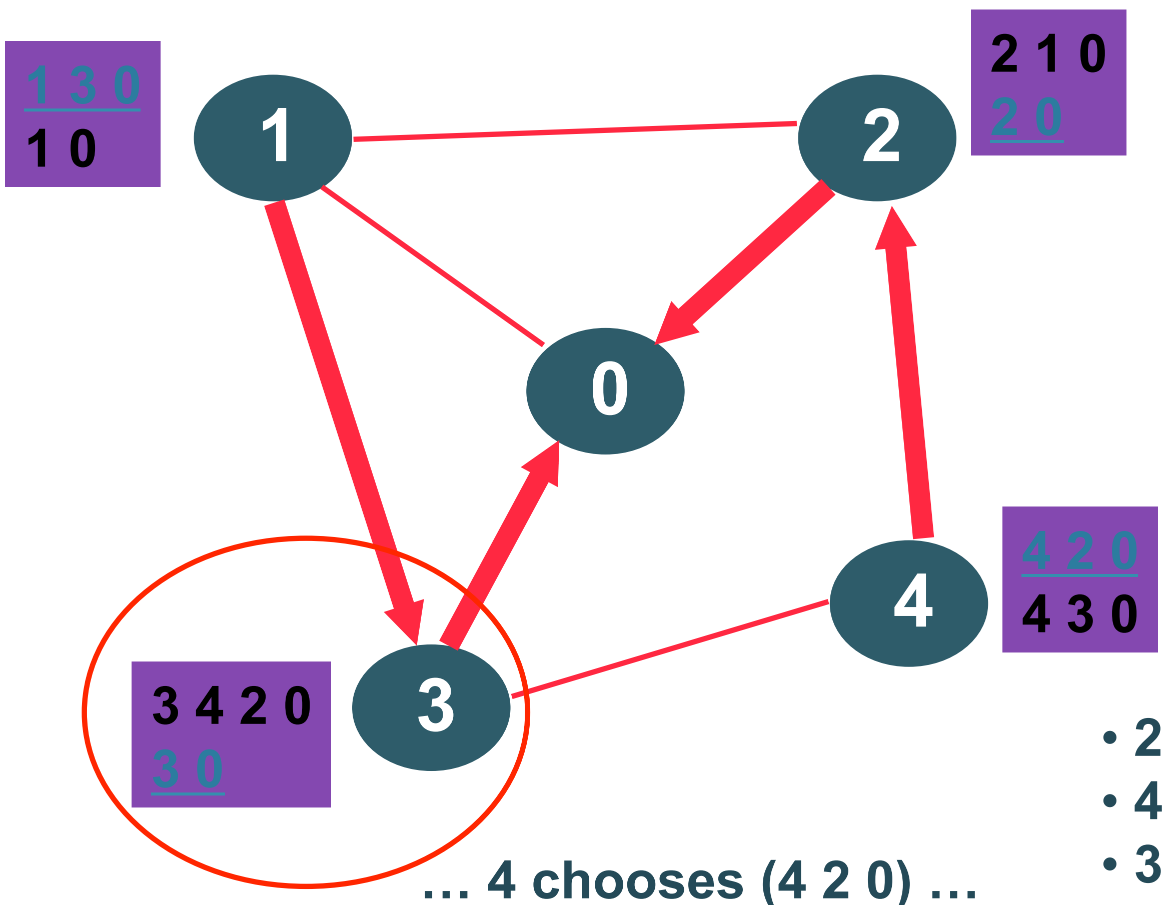
Example: BAD GADGET



… 2 chooses (2 0) …

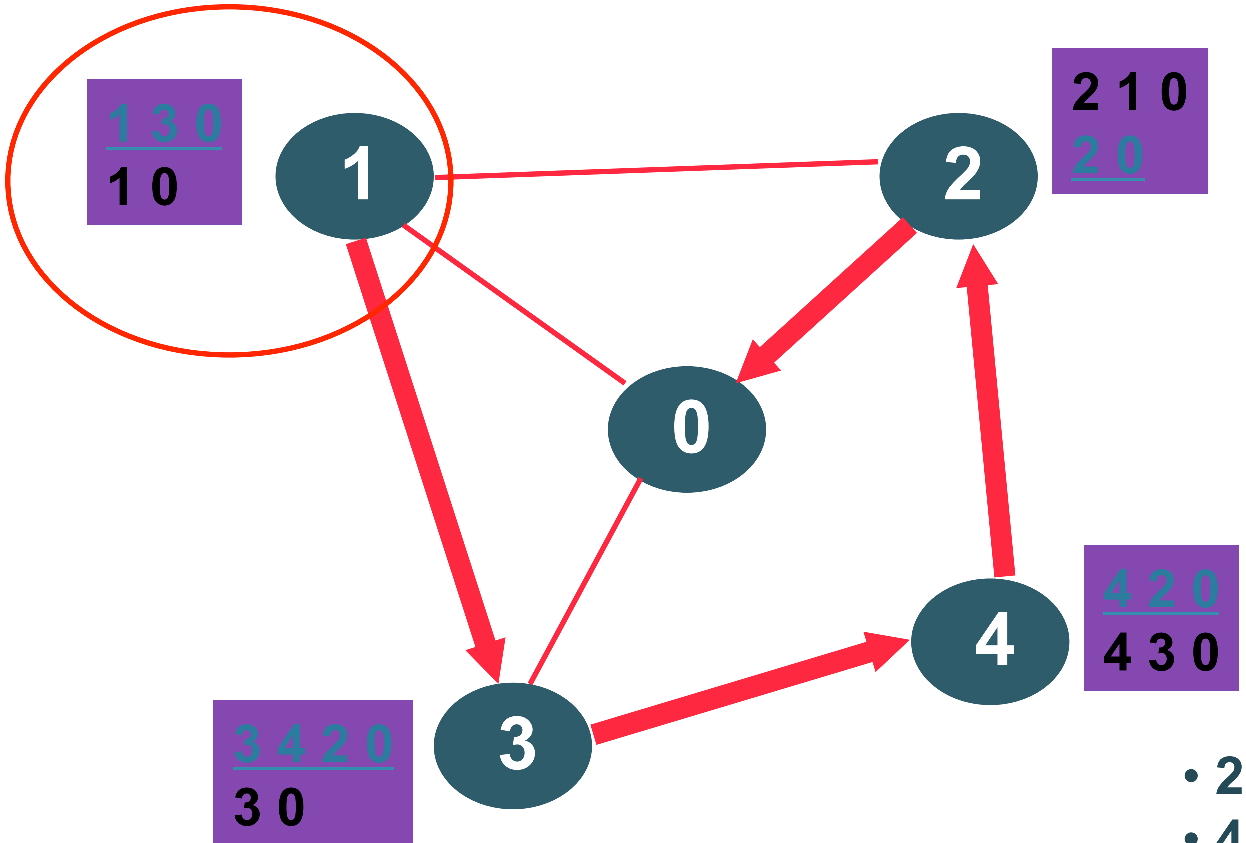
- 2 chooses (2 0)
- 4 chooses (4 2 0)
- 3 chooses (3 4 2 0)
- 1 chooses (1 0)

Example: BAD GADGET



- 2 chooses (2 0)
- 4 chooses (4 2 0)
- 3 chooses (3 4 2 0)
- 1 chooses (1 0)

Example: BAD GADGET

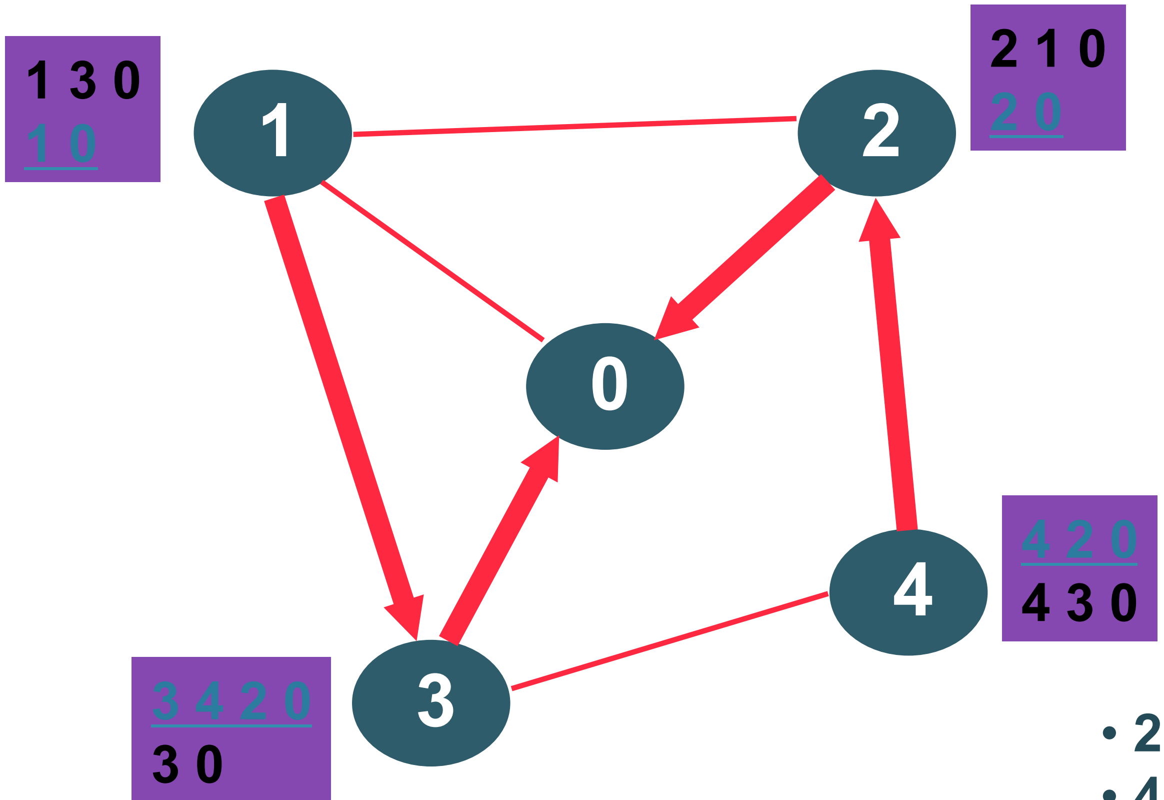


… 3 chooses (3 4 2 0) …

- 2 chooses (2 0)
- 4 chooses (4 2 0)
- 3 chooses (3 4 2 0)
- 1 chooses (1 0)

Example: BAD GADGET

That was one round of oscillation!



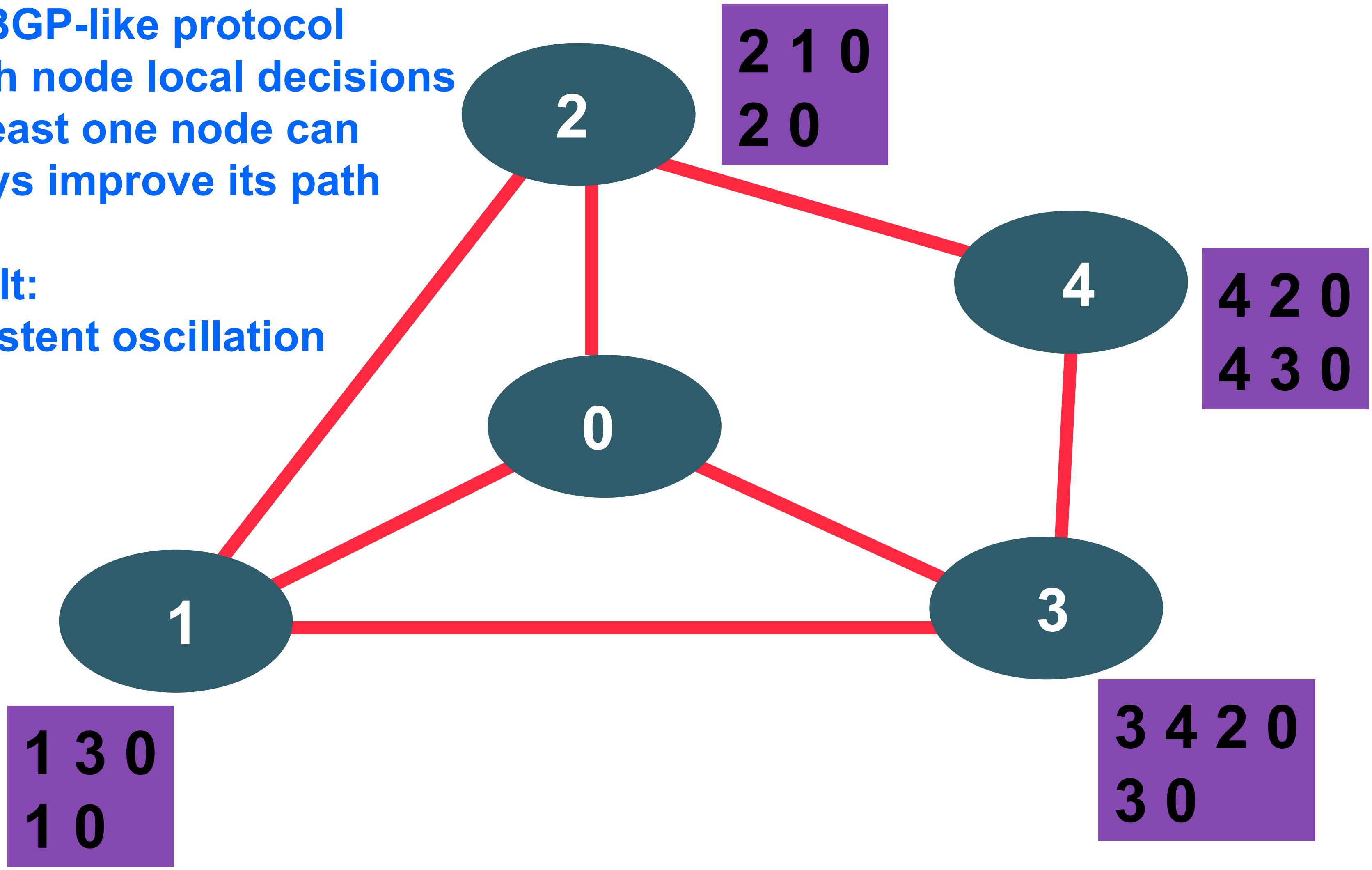
… 1 chooses (1 0) …

- 2 chooses (2 0)
- 4 chooses (4 2 0)
- 3 chooses (3 4 2 0)
- 1 chooses (1 0)

BAD GADGET : No Solution

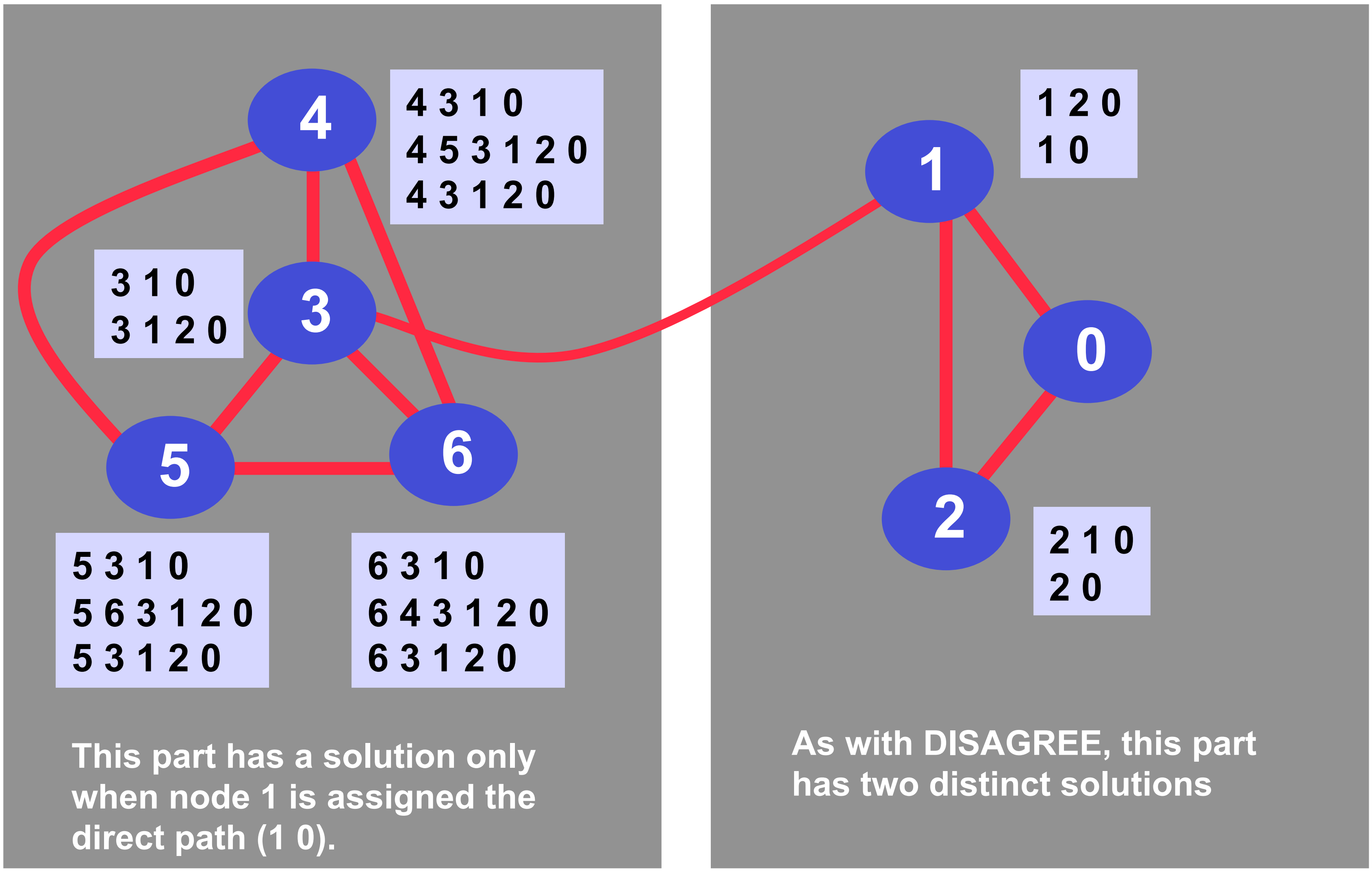
In a BGP-like protocol
- **each node local decisions**
- **at least one node can always improve its path**

Result:
persistent oscillation



Precarious

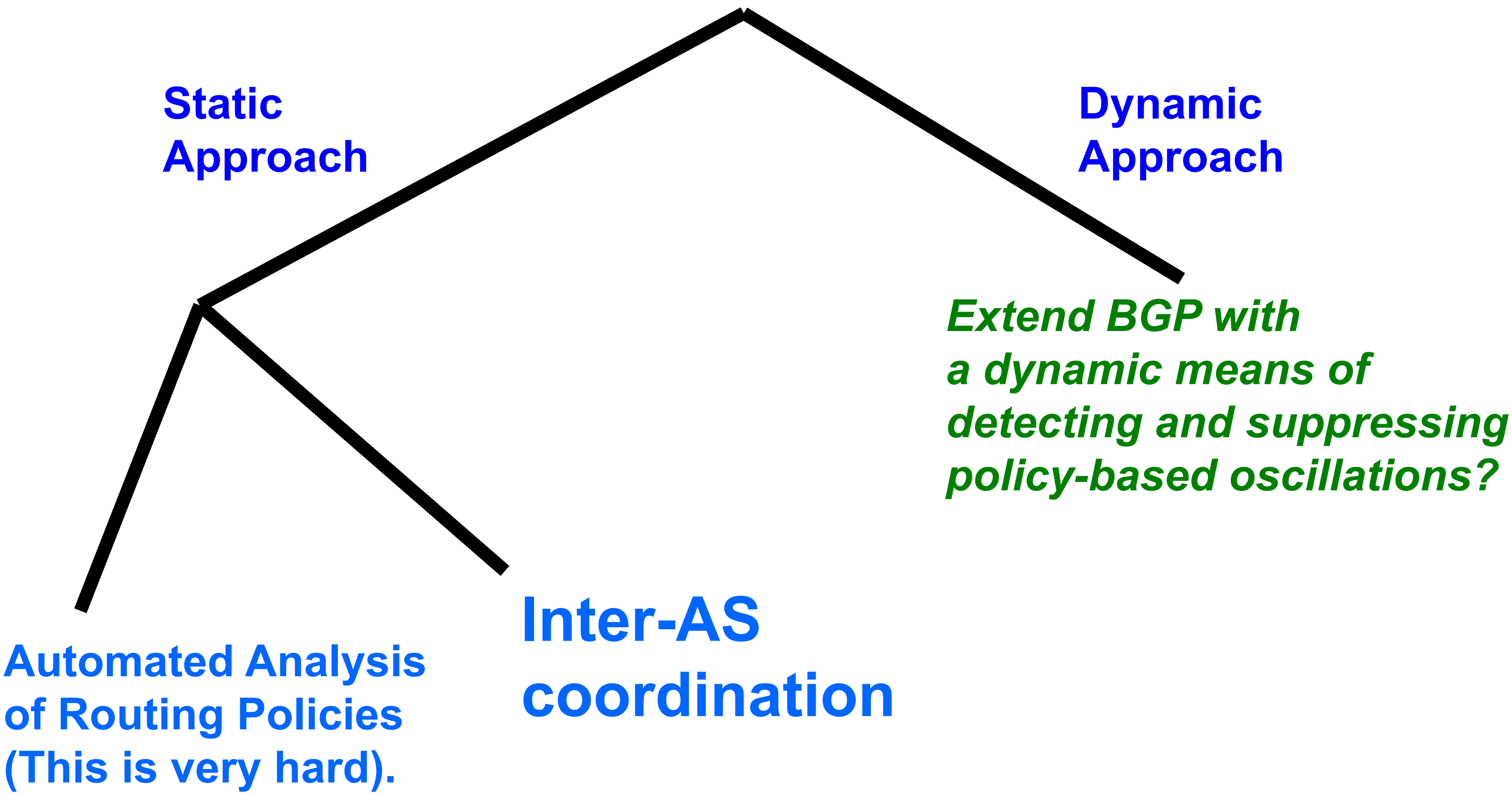
Has a solution, but can get "trapped"



Theoretical Results

- The problem of determining whether an instance of *stable paths problem* is solvable is NP-complete
- *Shortest path route* selection is provably safe

What is to be done?



These approaches are <u>complementary</u>

Strawman: Global Policy Check

- Require each AS to publish its policies
- Detect and resolve conflicts

Problems:

- ASes typically unwilling to reveal policies
- Checking for convergence is NP-complete
- Failures may still cause oscillations

Think Globally, Act Locally

- Key features of a good solution
  - *Safety*: guaranteed convergence
  - *Expressiveness*: allow diverse policies for each AS
  - *Autonomy*: do not require revelation/coordination
  - *Backwards-compatibility*: no changes to BGP

- *Local* restrictions on configuration semantics
  - Ranking
  - Filtering

Main Idea of Gao-Rexford (2001)

- Permit only two business arrangements
  - Customer-provider
  - Peering
- Constrain both filtering and ranking based on these arrangements to guarantee safety
  - *These are still restrictive, newer results relax them*
- Surprising result: these arrangements correspond to today's (common) behavior

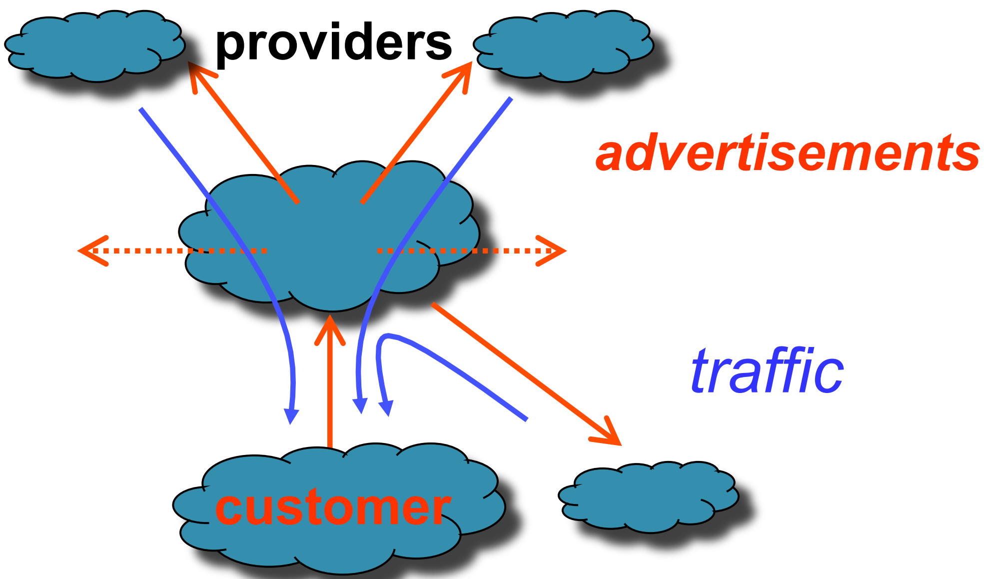
Gao & Rexford, "Stable Internet Routing without Global Coordination", *IEEE/ACM ToN*, 2001

Relationship #1: Customer-Provider

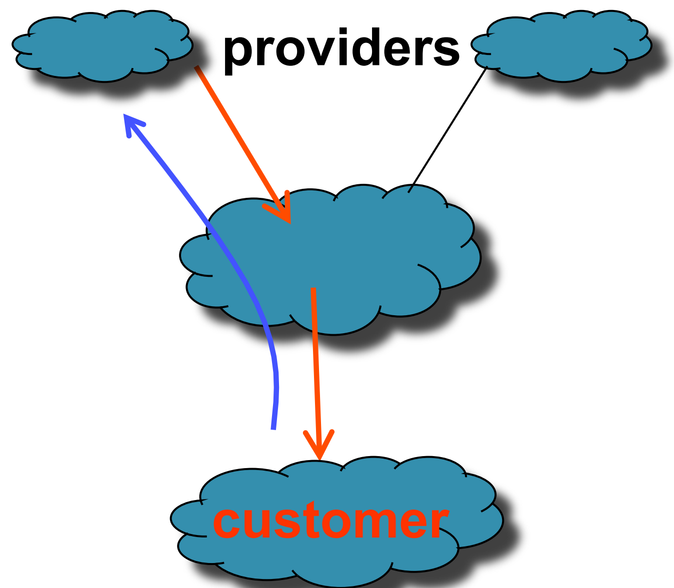
Filtering

- Routes from customer: to *everyone*
- Routes from provider: only to *customers*

From other destinations
To the customer



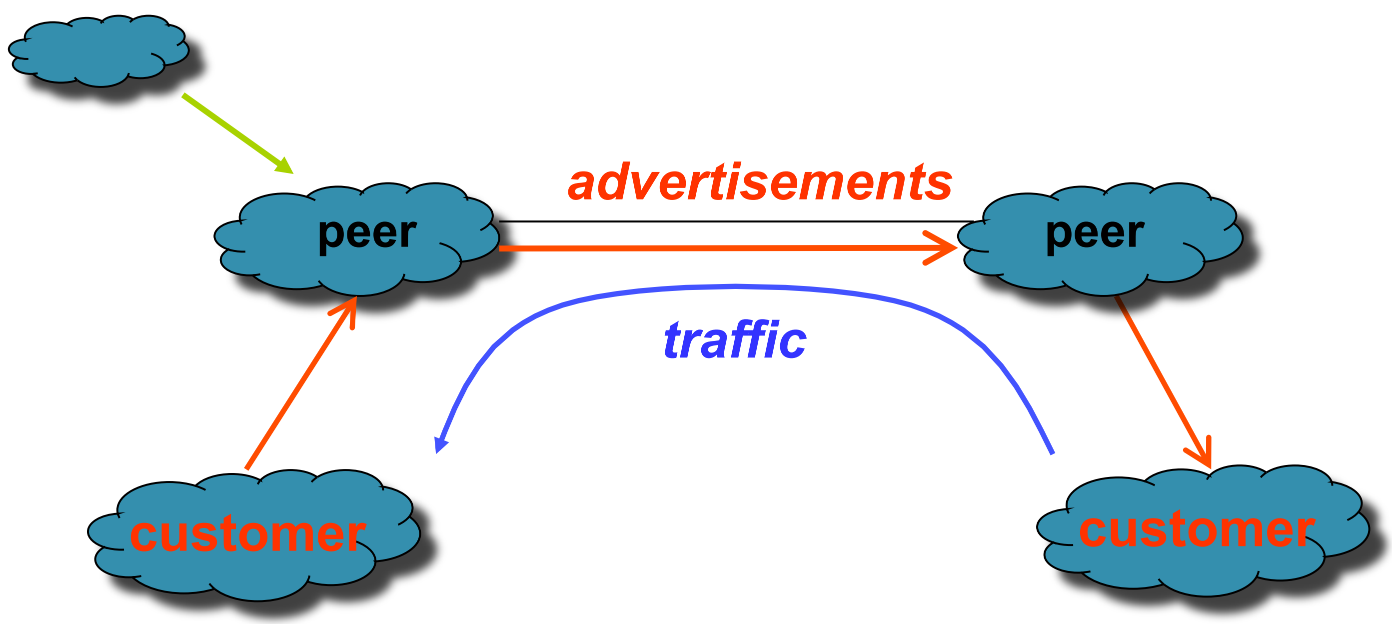
From the customer
To other destinations



Relationship #2: Peering

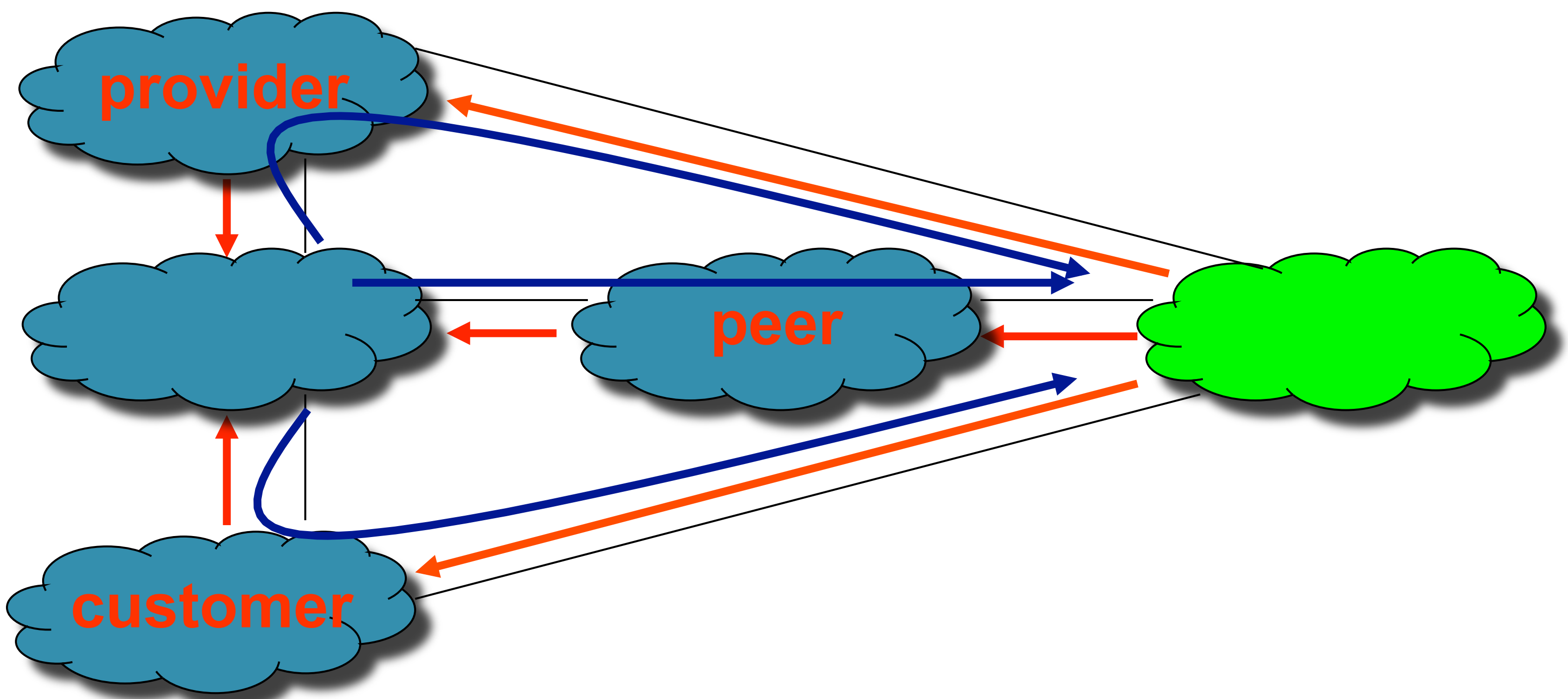
Filtering

- Routes from peer: only to customers
- No routes from other peers or providers

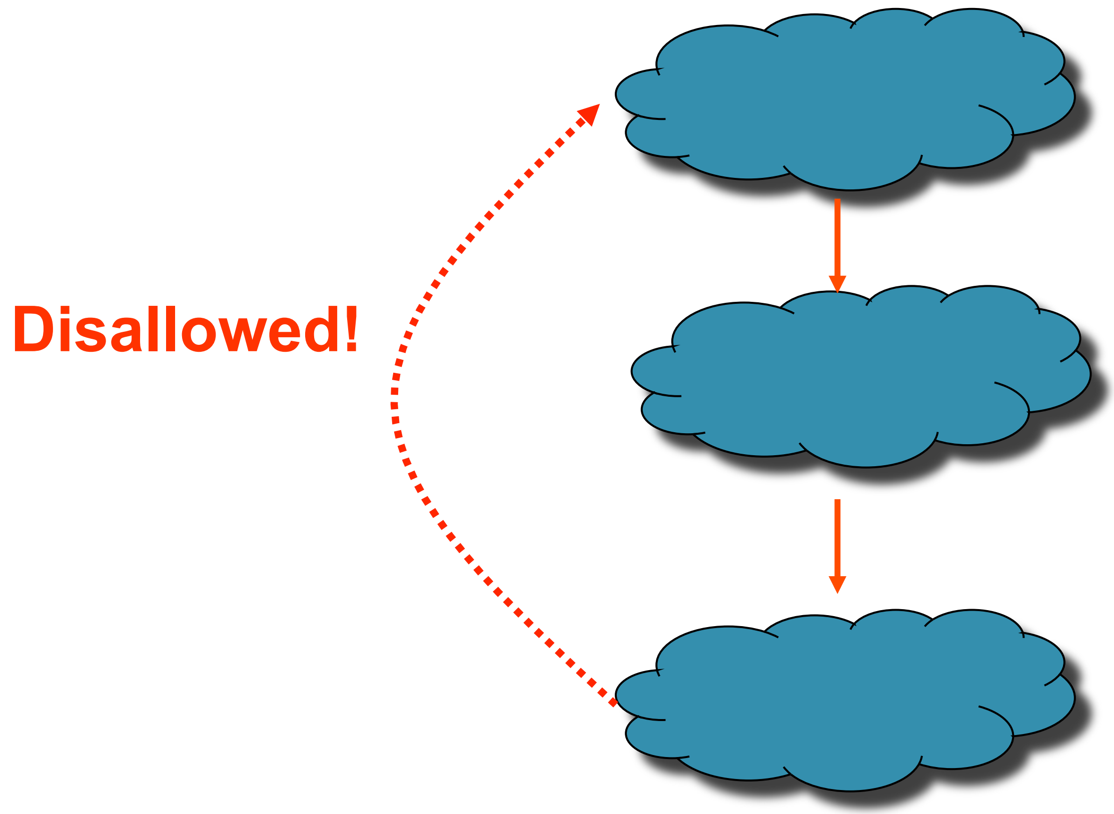


Rankings

- Routes from customers over routes from peers
- Routes from peers over routes from providers

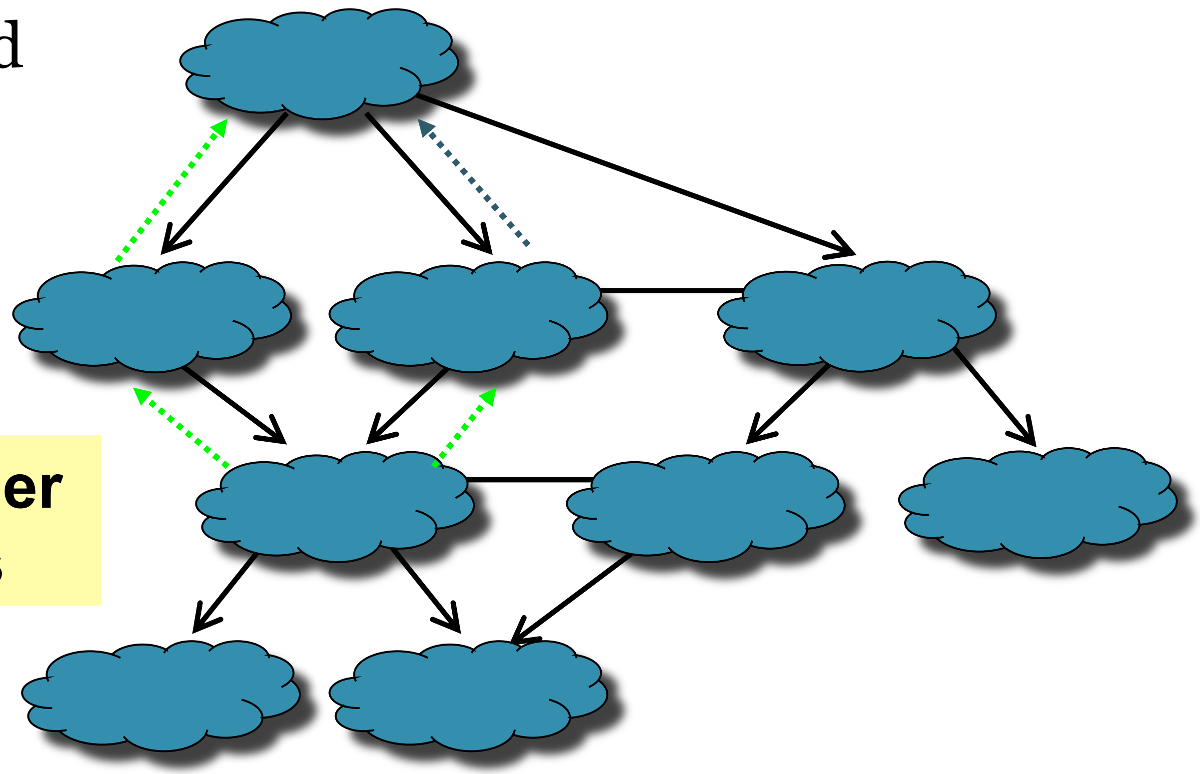


Additional Assumption: Hierarchy



Proof Sketch, Step 1: Customer Routes

- Activate ASes from customer to provider
  - AS picks a customer route if one exists
  - Decision of one AS cannot cause an earlier AS to change its mind

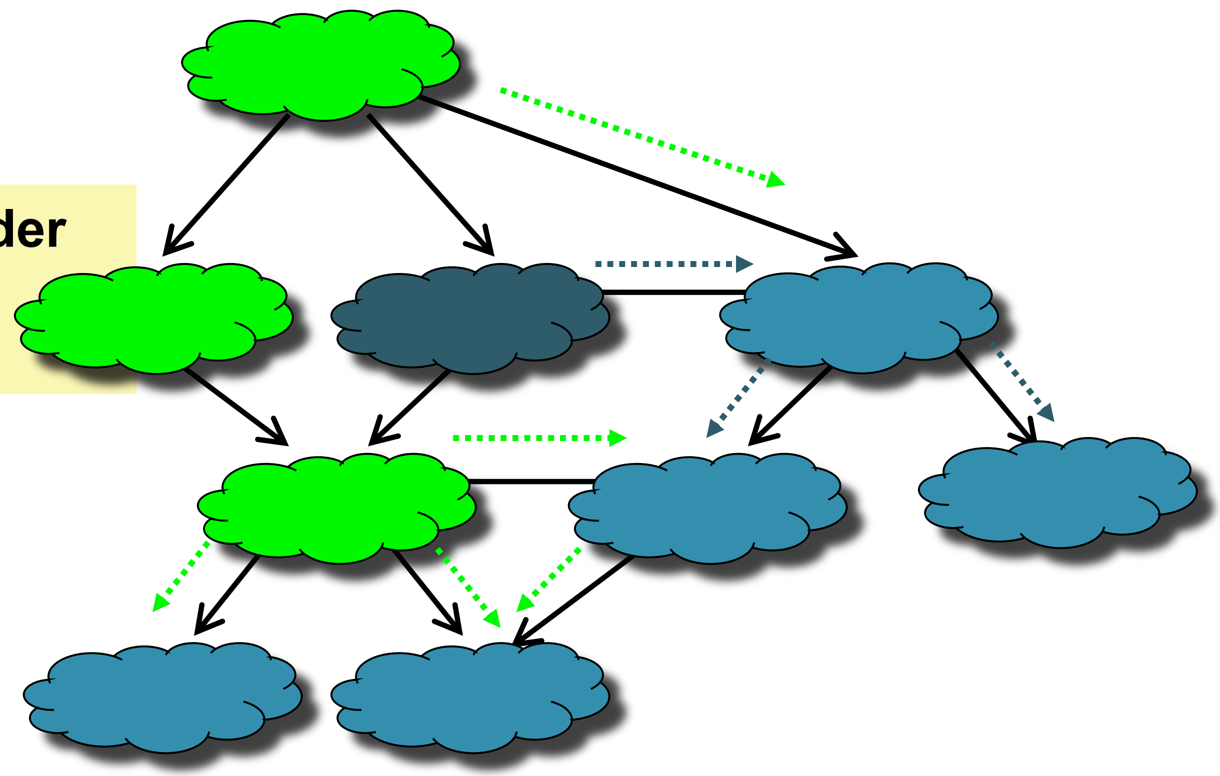


An AS picks a customer route when one exists

Proof Sketch, Step 2: Peer & Provider Routes

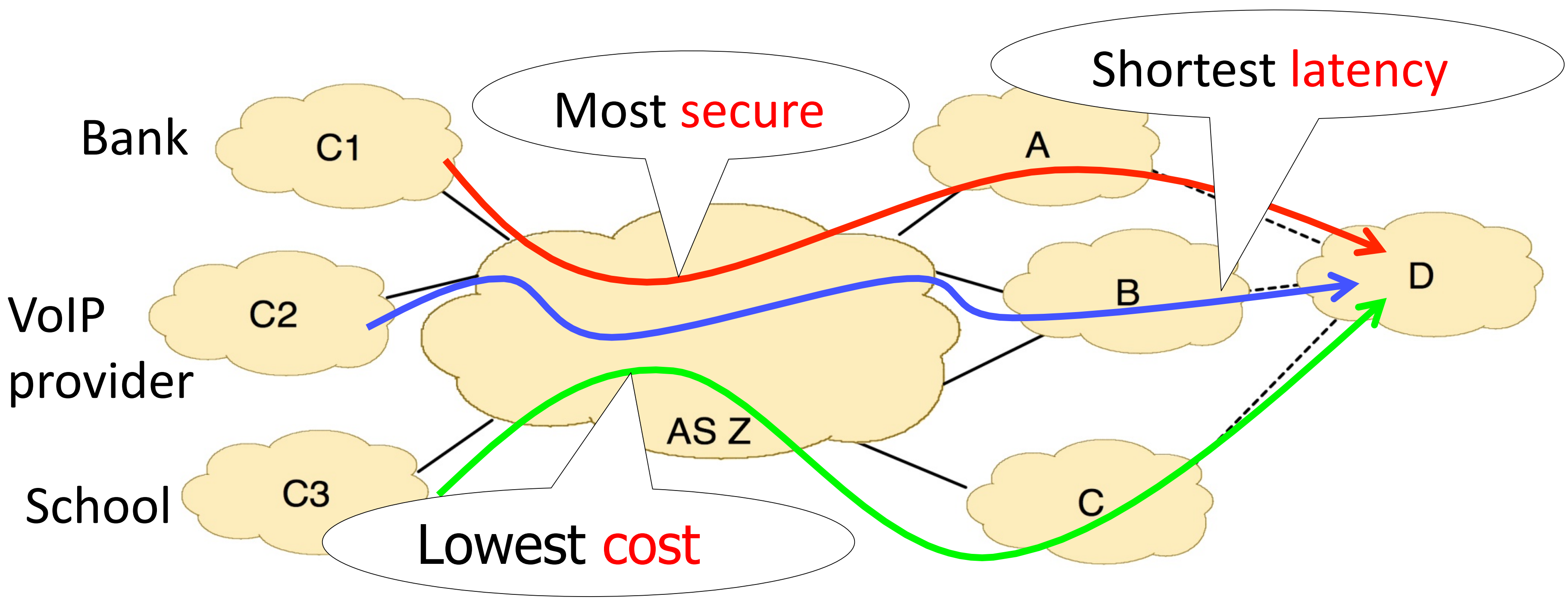
- Activate remaining ASes from provider to customer
  - Decision of one Step-2 AS cannot cause an earlier Step-2 AS to change its mind
  - Decision of Step-2 AS cannot affect a Step-1 AS

AS picks a peer or provider route when no customer route is available



SPP Might be too Restrictive

- ISPs usually have multiple paths to the destination
- Different paths have different properties
- Different neighbors may prefer different routes



Conclusions

- BGP is solving a hard problem
  - Routing protocol operating at a global scale
  - With tens of thousands of independent networks
  - That each have their own policy goals
  - And all want fast convergence
- Key features of BGP
  - Prefix-based path-vector protocol
  - Incremental updates (announcements and withdrawals)
  - Policies applied at import and export of routes
- Active research topic!