CSCI-1680
Network Layer:
Inter-domain Routing

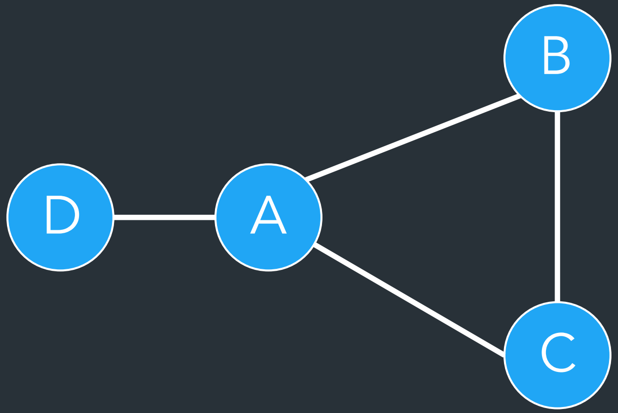
Nick DeMarinis

Based partly on lecture notes by Rachit Agarwal, Rodrigo Fonseca, Jennifer Rexford, Rob Sherwood, David Mazières, Phil Levis, John Jannotti

Administrivia

- IP: Due next Thursday (10/17)
- HW2: As soon as I can get there
- Long weekend: no hours on Monday (10/14), responses on Ed delayed

Warmup



B's routing table

Dest.	Cost	Next Hop
A	1	A
C	1	C
D	2	A

Routers A,B,C,D use RIP. When B sends a periodic update to A, what does it send…

- When using standard RIP?
- When using split horizon + poison reverse?

Recall: BGP

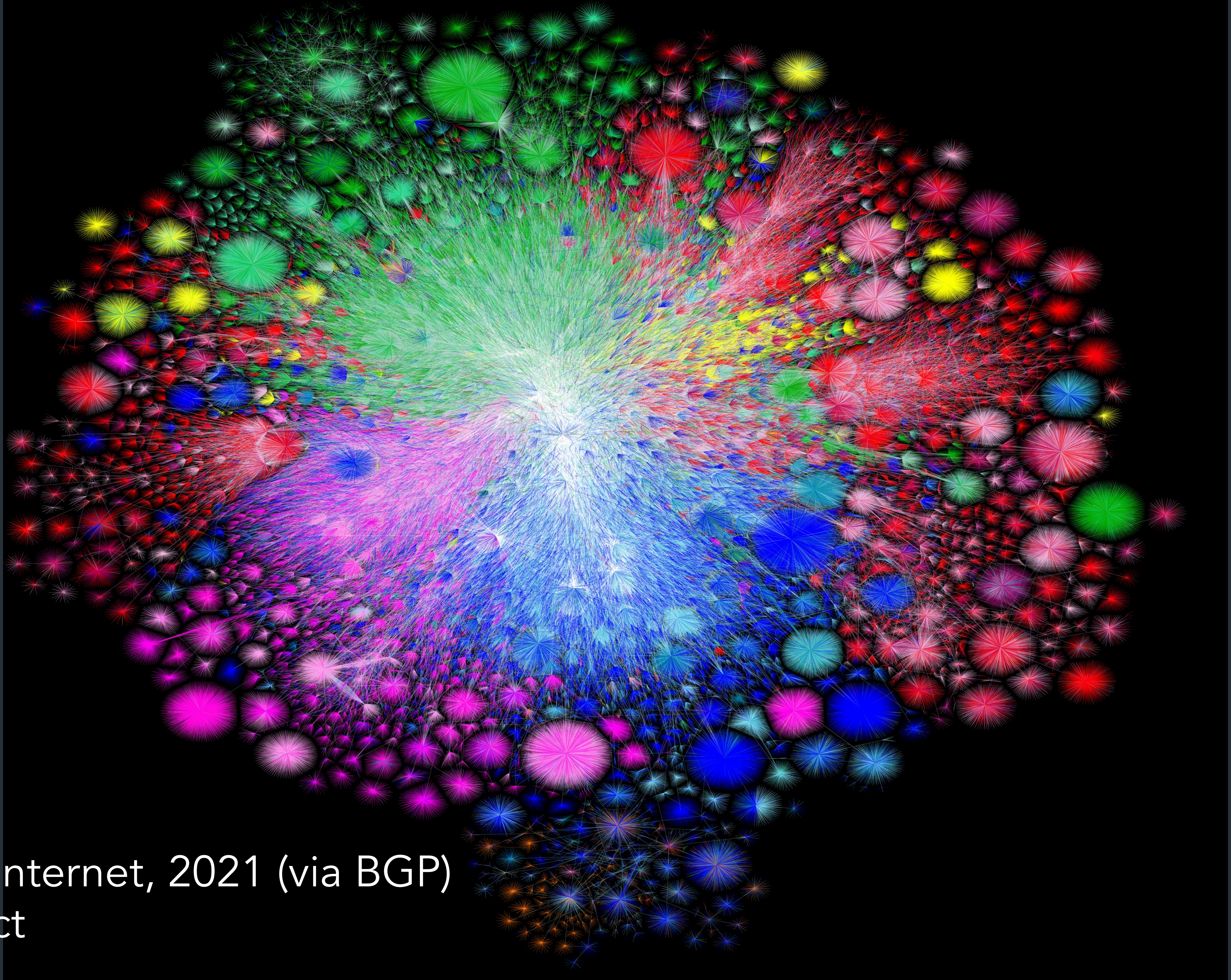
Exterior routing: between Autonomous Systems (ASes)

=> How networks with different goals/policies/incentives connect to each other (or don't)

=> A "path vector" protocol

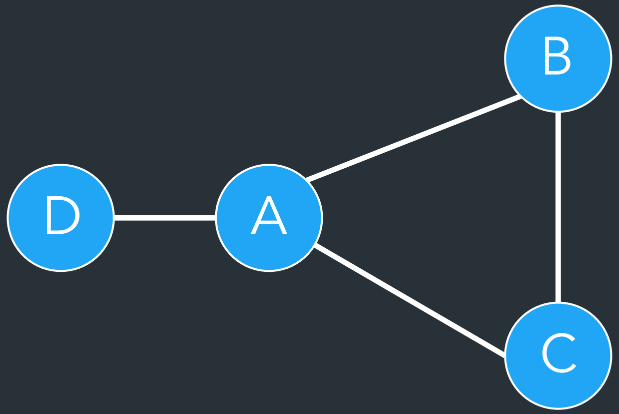
A BGP update

"I can reach prefix 128.148.0.0/16
through ASes 44444 3356 14325 11078"

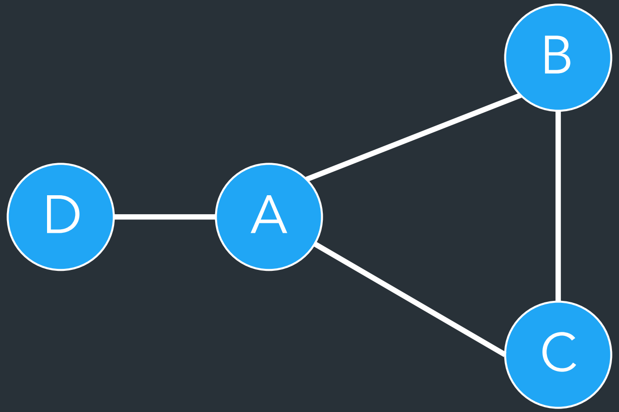


Map of the Internet, 2021 (via BGP)
OPTE project

Before: Interior routing



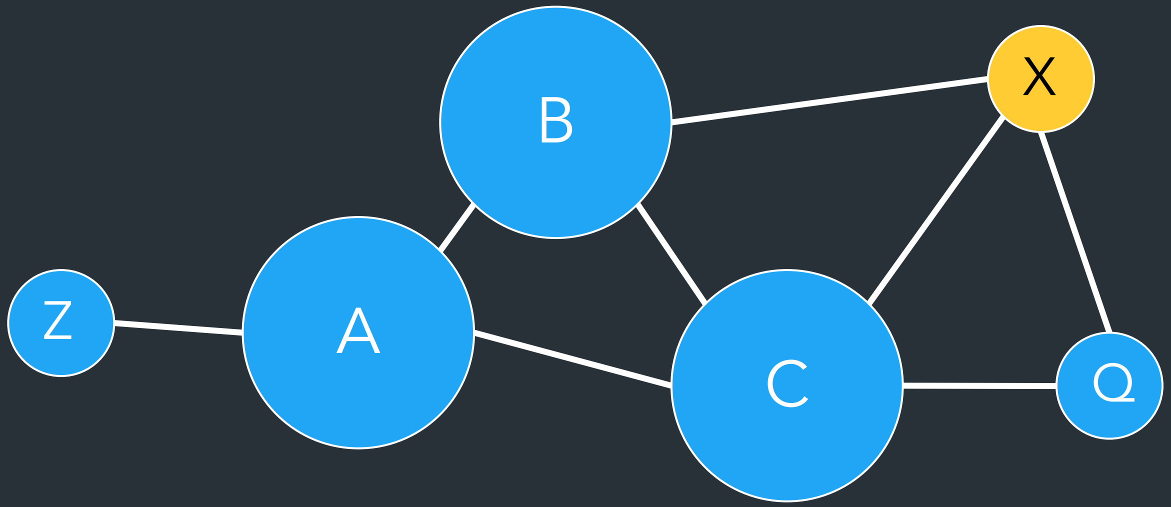
Before: Interior routing



All nodes advertise their routes to all other nodes:

- Goal: connect everything to everything
- One administrative domain
- Find optimal path

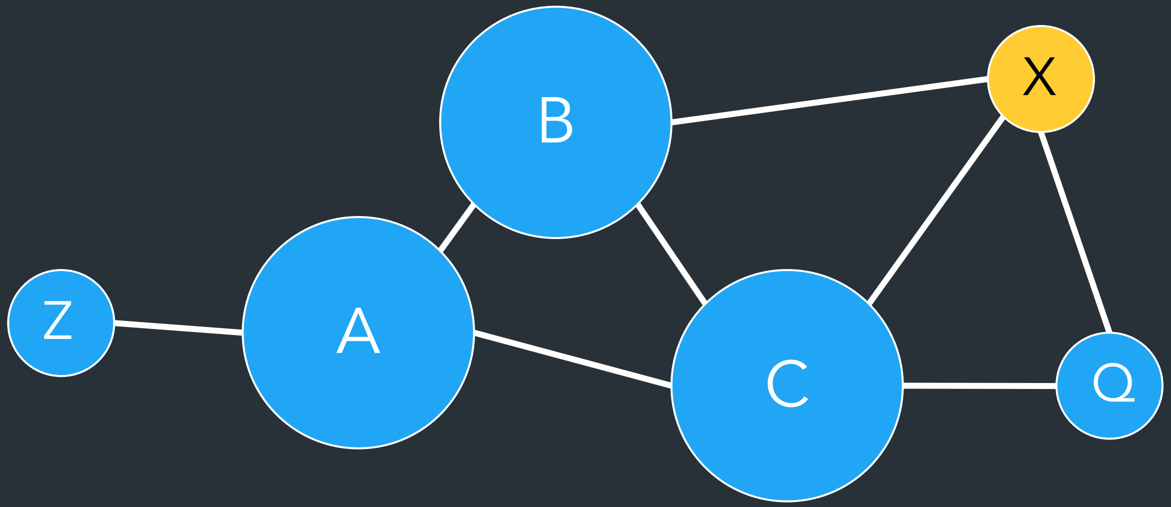
Now: BGP



X's table (subset):

Network	Next Hop	Path
X	--	(Origin)
B	B	B
C	C	C
Q	Q	Q
A	B	B A
…	…	…

Now: BGP



X's table (subset):

Network	Next Hop	Path
X	--	(Origin)
B	B	B
C	C	C
Q	Q	Q
A	B	B A
…	…	…

X has neighbors B, C, Q.

What routes might X NOT want to tell B? Why?

Key policy questions

A BGP update

*"I can reach prefix 128.148.0.0/16
through ASes 44444 3356 14325 11078"*

"How to use route info to update forwarding tables?"

"What routing info to send to neighbors?"

Key policy questions

A BGP update

"I can reach prefix 128.148.0.0/16 through ASes 44444 3356 14325 11078"

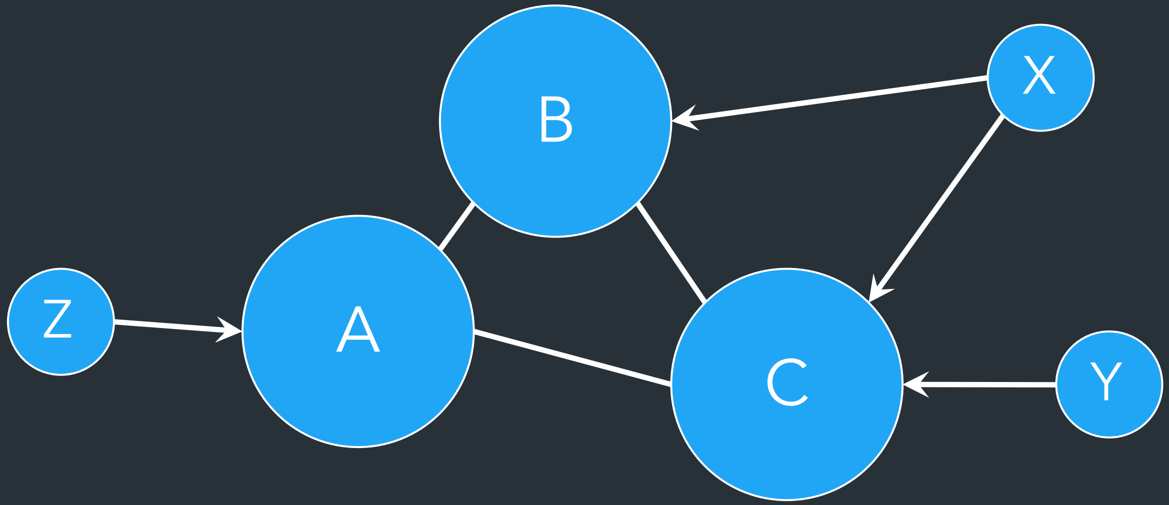
"How to use route info to update forwarding tables?"

=> Local routing policy ("Selection policy")

"What routing info to send to neighbors?"

=> Export policy

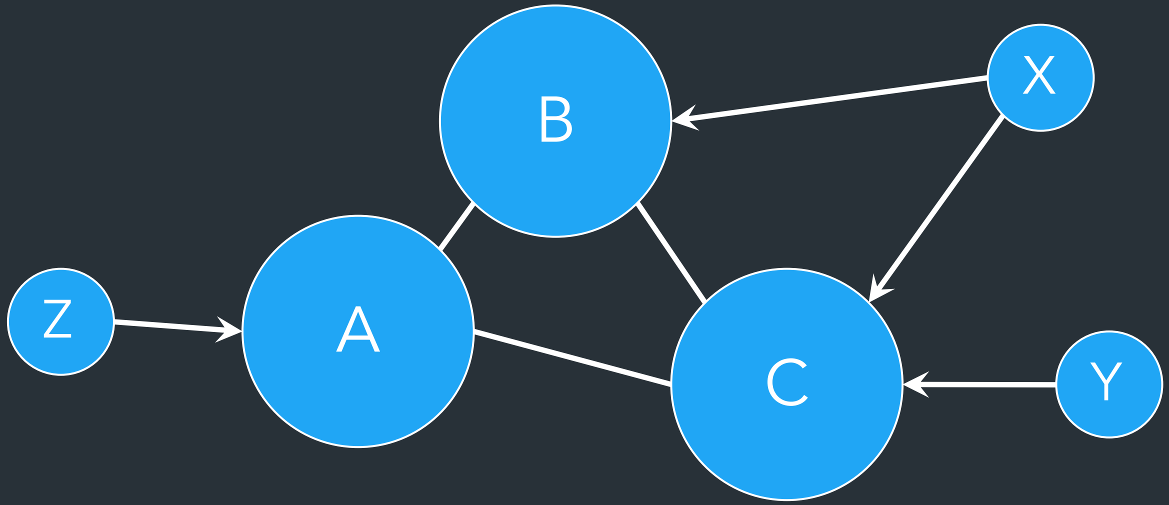
=> Policy Implications? What can go wrong?



Relationships between AS drive policy:

- Provider
- Customer
- Peers

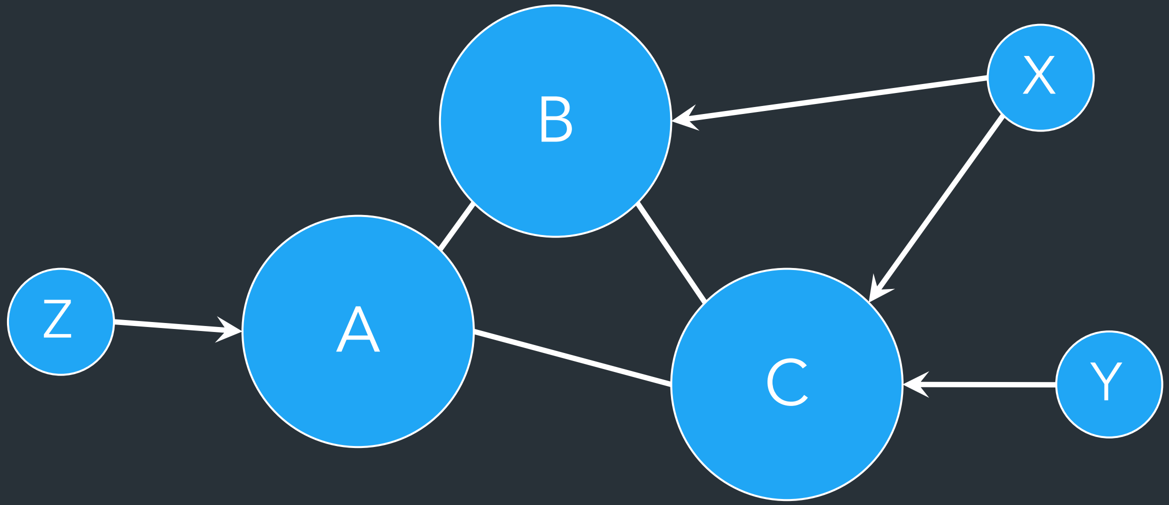
Example from Kurose and Ross, 5th Ed



Relationships between AS drive policy:

- Customer->Provider: Customer pays provider to advertise its routes, send it traffic

Example from Kurose and Ross, 5th Ed



Relationships between AS drive policy:

- Customer: Pays *provider* to advertise its routes, send it traffic

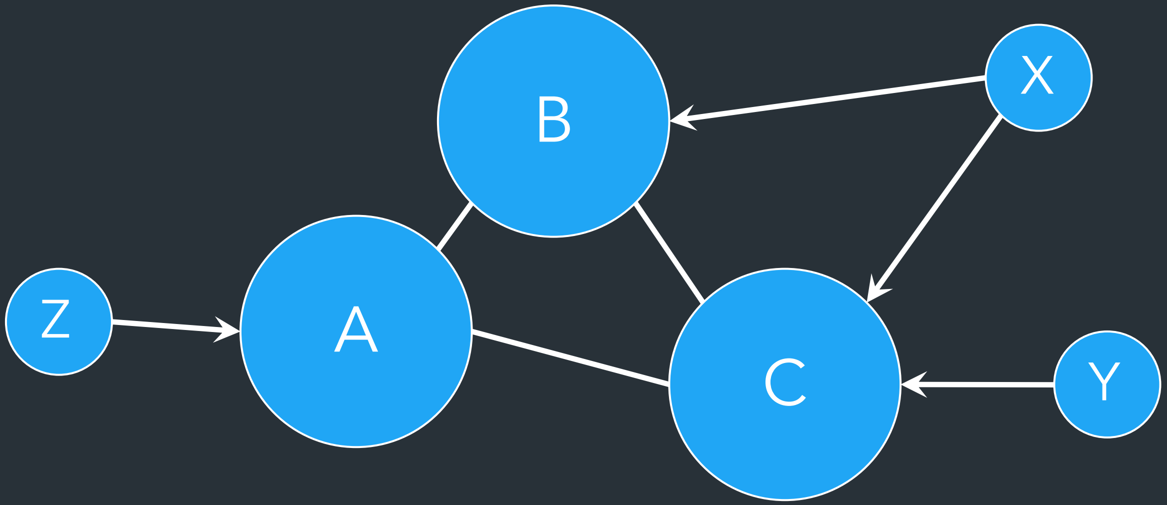
⇒Y pays C

⇒X pays B, C (multihomed)

⇒ B *is transit [provider] for* X: Traffic destined for X goes through B

⇒ X is not transit for B, C: Traffic from B->C must not go through X!

Example from Kurose and Ross, 5th Ed



Relationships between AS drive policy:

- Customer: Customer pays provider to advertise its routes

⇒Y pays C

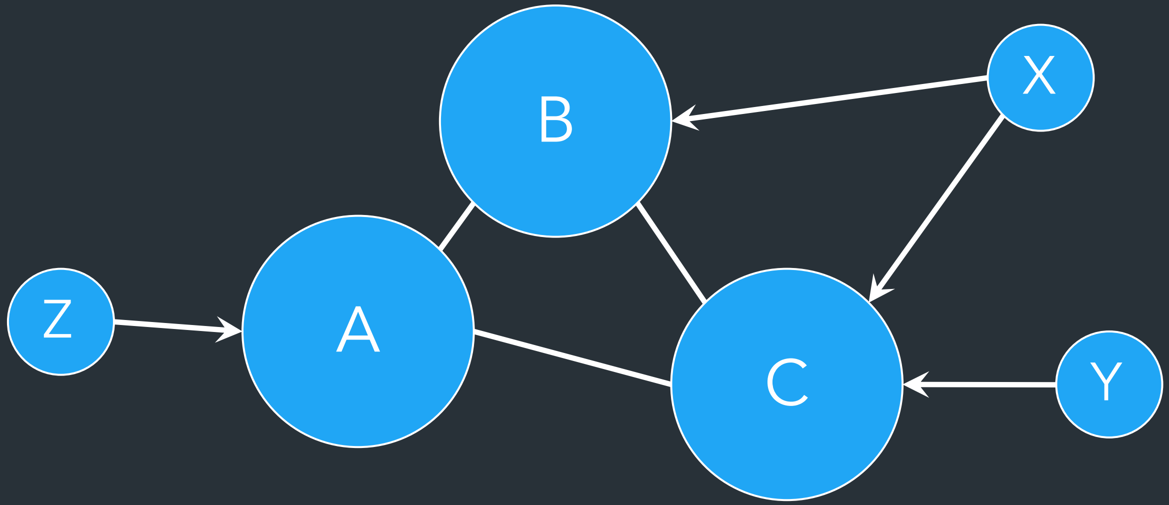
⇒X pays B, C (multihomed)

⇒B "*is transit [provider] for*" X: Traffic destined for X goes through B

⇒X is not transit for B, C: Traffic from B->C must not go through X!

=> Why not? X gains nothing!

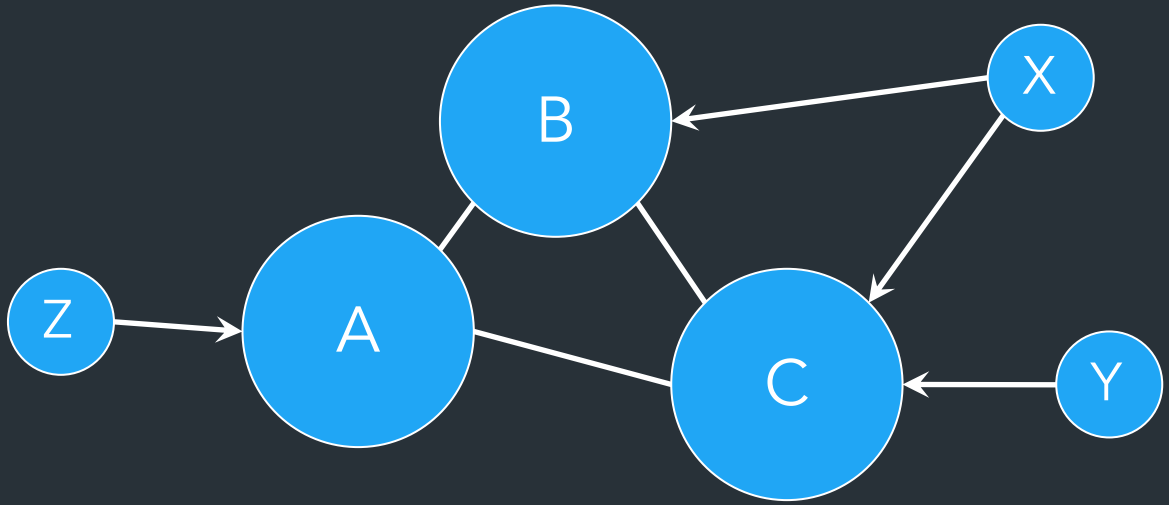
Example from Kurose and Ross, 5th Ed



Relationships between AS drive policy:

- Customer: Customer pays provider to advertise its routes
- Peers: Providers may share routes at no cost for mutual benefit

Example from Kurose and Ross, 5th Ed



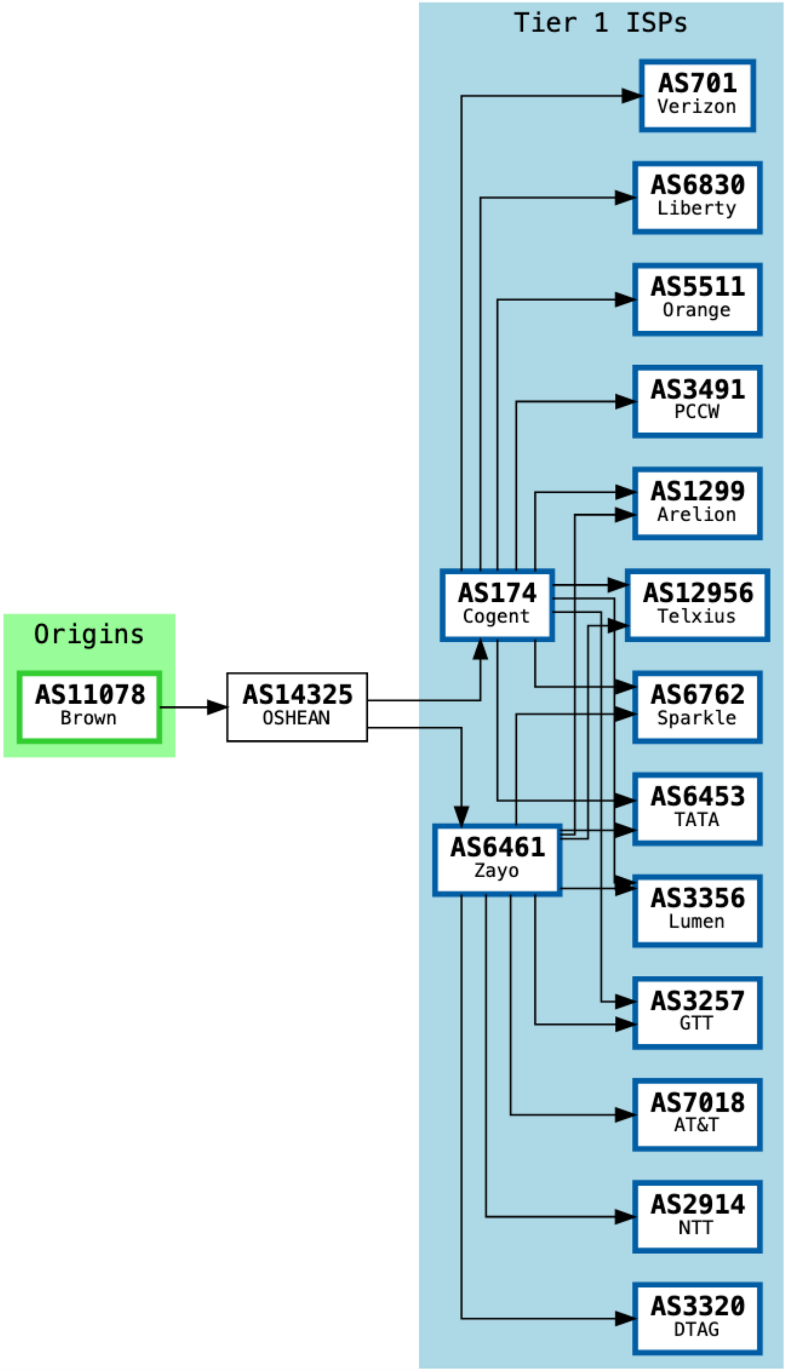
- Providers: highly connected ISPs
  - Most connected (“Tier 1”) have no default route!
  - Tier 2 is customer of Tier 1, …

- Peers: Providers may share routes at no cost for mutual benefit

  => A peers with B
  => A peers with C
  . . .

Example from Kurose and Ross, 5th Ed

10 Oct 24 09:02 UTC



How to think about policies

Update processing

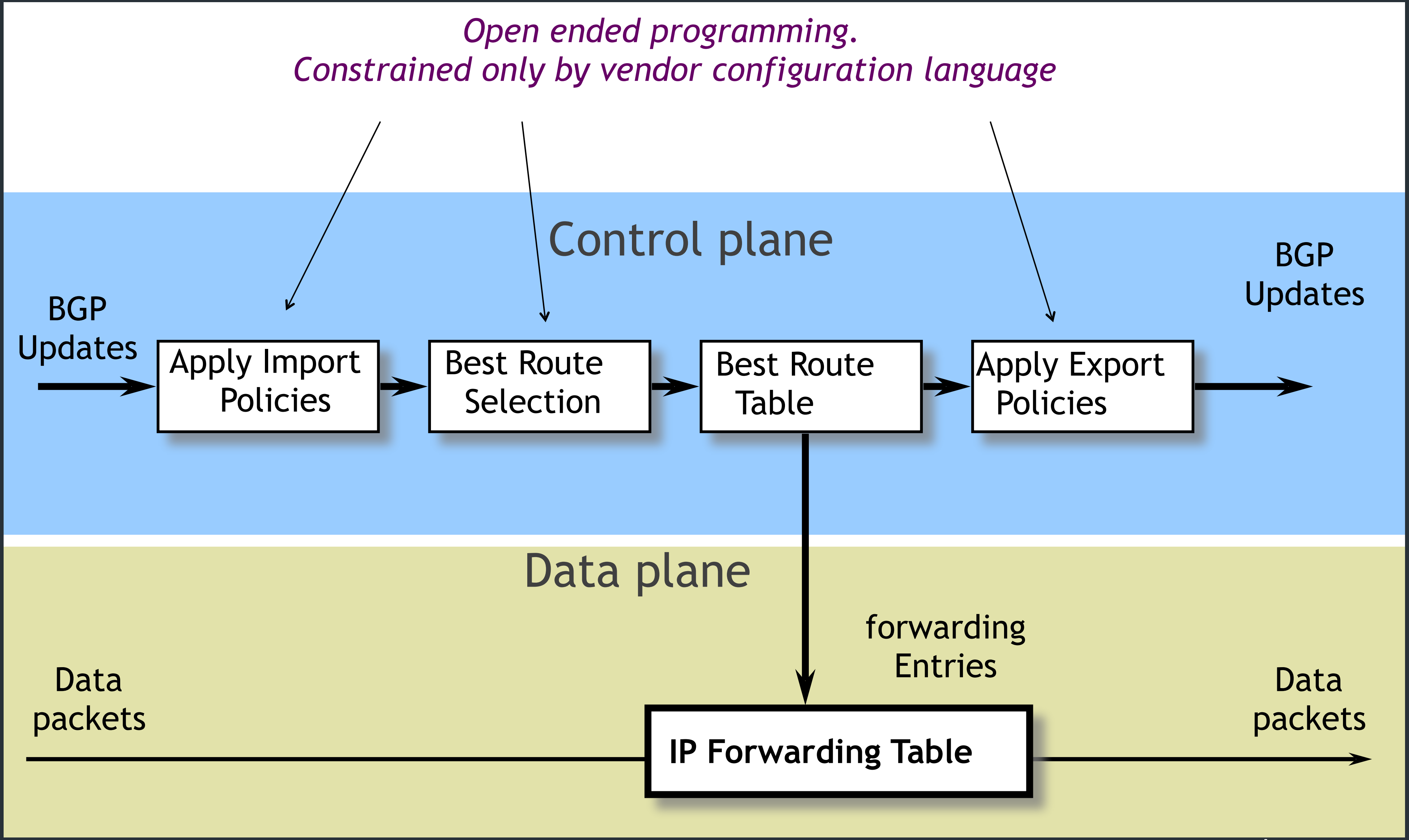


Image credit Rachit Agarwal

Typical route selection policy

In decreasing priority order:

1. Make or save money (send to customer > peer > provider)
2. Try to maximize performance (smallest AS path length)
3. Minimize use of my network bandwidth ("hot potato routing"
4. …

How to turn this into a policy?

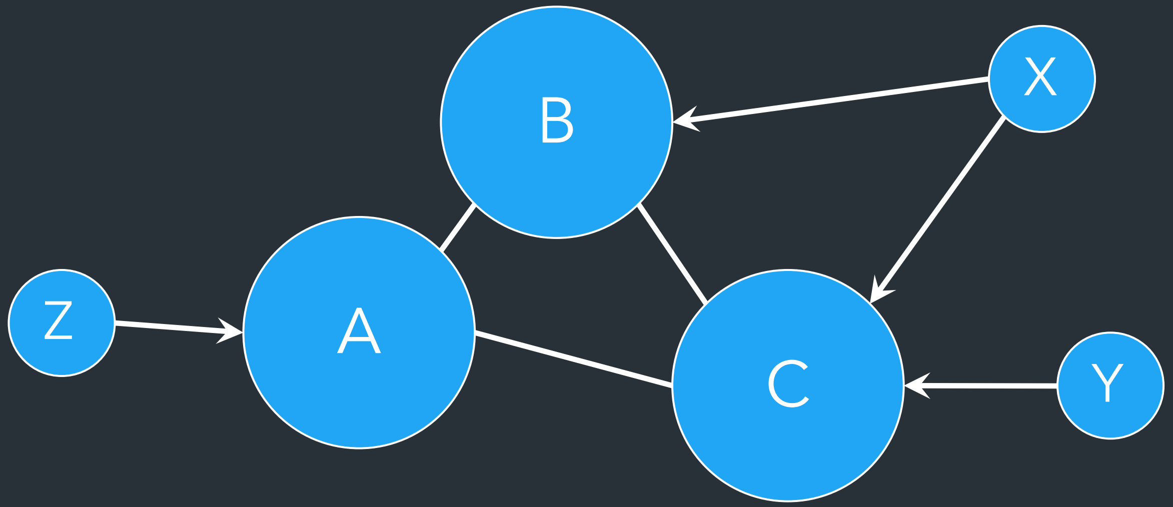
- Selection Policy: which path to use in your network

- Export Policy: which path to advertise

Typical Export Policy

Destination prefix advertised by…	Export route to…
Customer	Everyone (providers, peers, other customers…)
Peer	Customers only
Provider	Customers only

Known as Gao-Rexford principles: define common practices for AS relationships



How to prevent X from forwarding transit between B and C?

How to avoid transit between CBA ?

Example from Kurose and Ross, 5th Ed

What can go wrong?

How to advertise *your* prefixes?

Try to aggregate (summarize) prefixes for networks you own, but not always possible

More specific prefix => More preferred
=> Can have policy, security implications…

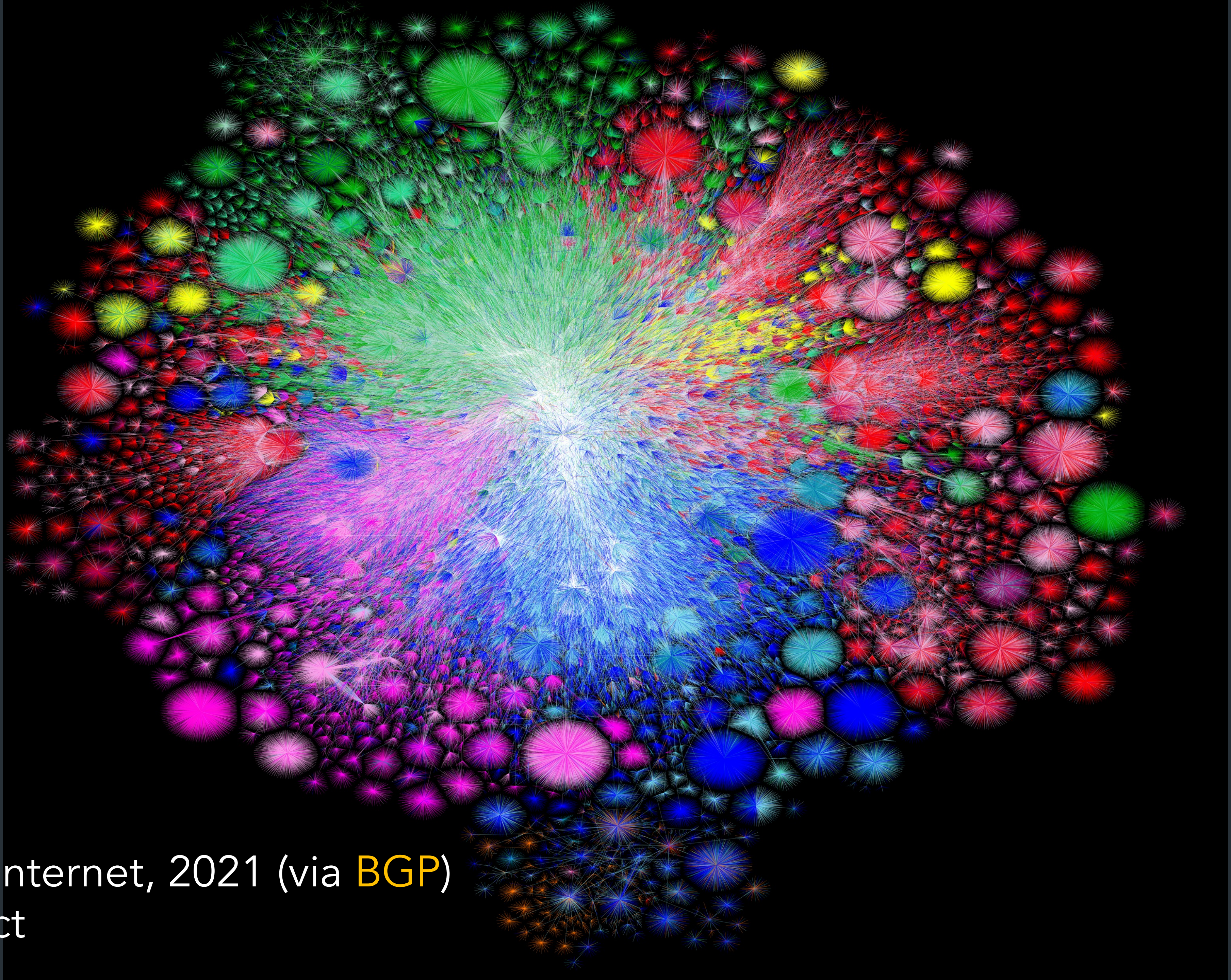
How to advertise *your* prefixes?

Try to aggregate (summarize) prefixes for networks you own, but not always possible

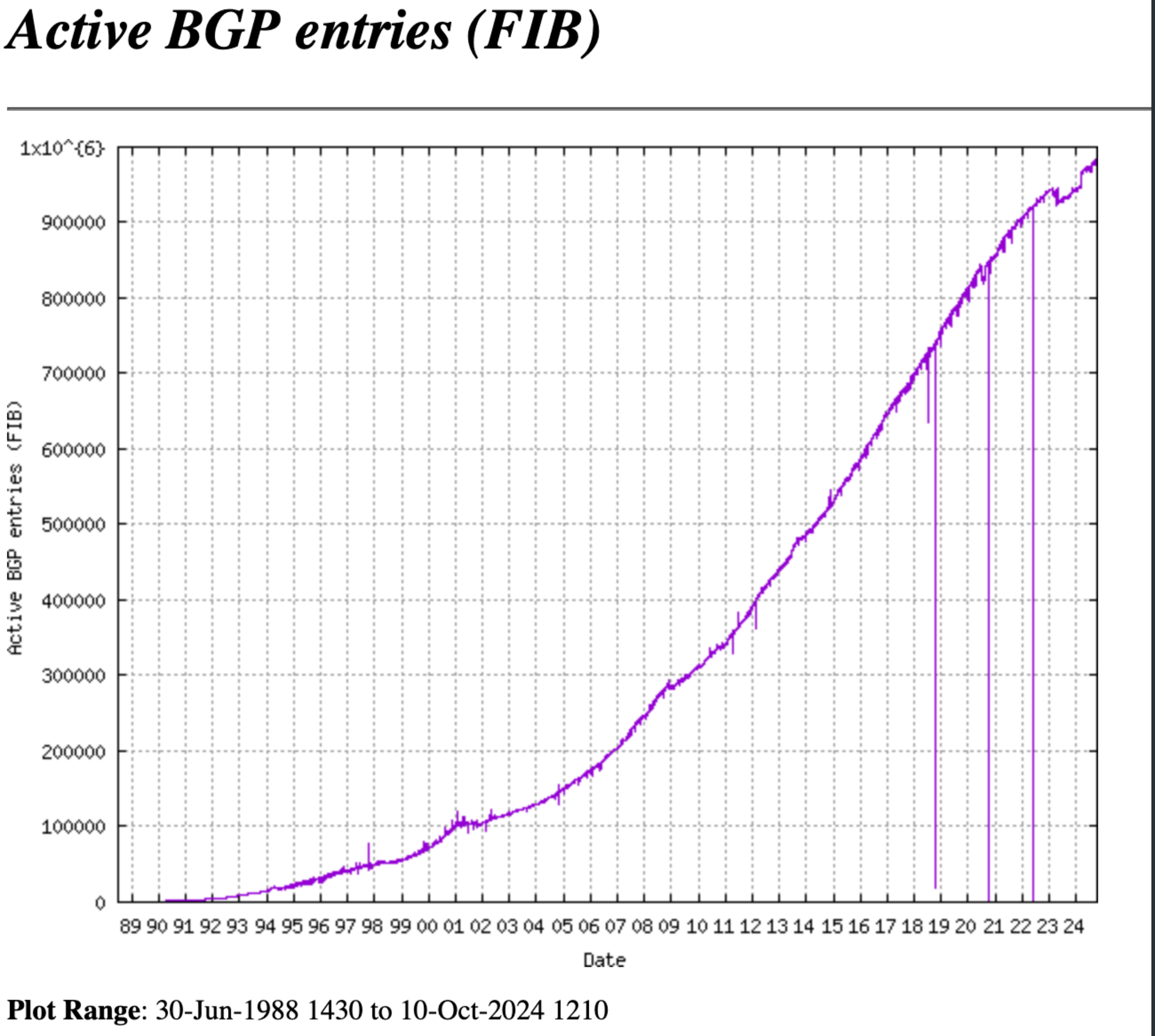
Problem: smaller allocations => more prefixes in table
=> Forwarding table size limited by fast memory (TCAM) inside routers

What can lead to table growth?

- More addresses being allocated
- Fragmentation
  - Multihoming
  - Change of ISPs
  - Address re-selling



Map of the Internet, 2021 (via BGP)
OPTE project



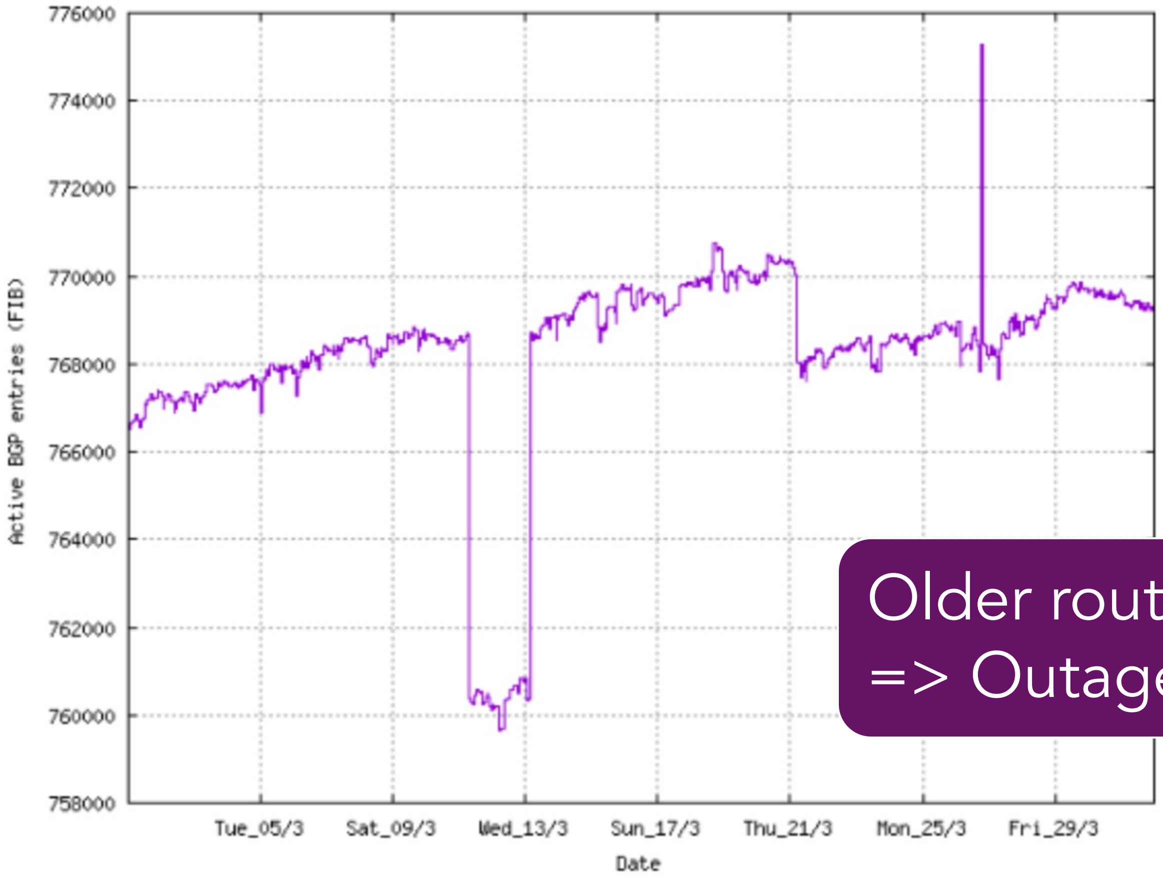
Active BGP entries (FIB)

Plot Range: 30-Jun-1988 1430 to 10-Oct-2024 1210

CIDR Report

How big can the table get?

- August 12, 2014: the full IPv4 BGP table reached 512k prefixes
- March 5, 2019: 768k prefixes



Plot Range: 01-Mar-2019 0001 to 01-Apr-2019 0001

Older routers run out of space => Outages

Peering Drama

- Cogent vs. Level3 were peers
- In 2003, Level3 decided to start charging Cogent
- Cogent said no
- Internet partition: Cogent's customers couldn't get to Level3's customers and vice-versa
  - Other ISPs were affected as well
- Took 3 weeks to reach an undisclosed agreement

BGP can be fragile!

- Individual router configurations and policy can affect whole network

- Consequences sometimes disastrous…

BGP Problems and Security Issues

Who owns a prefix?

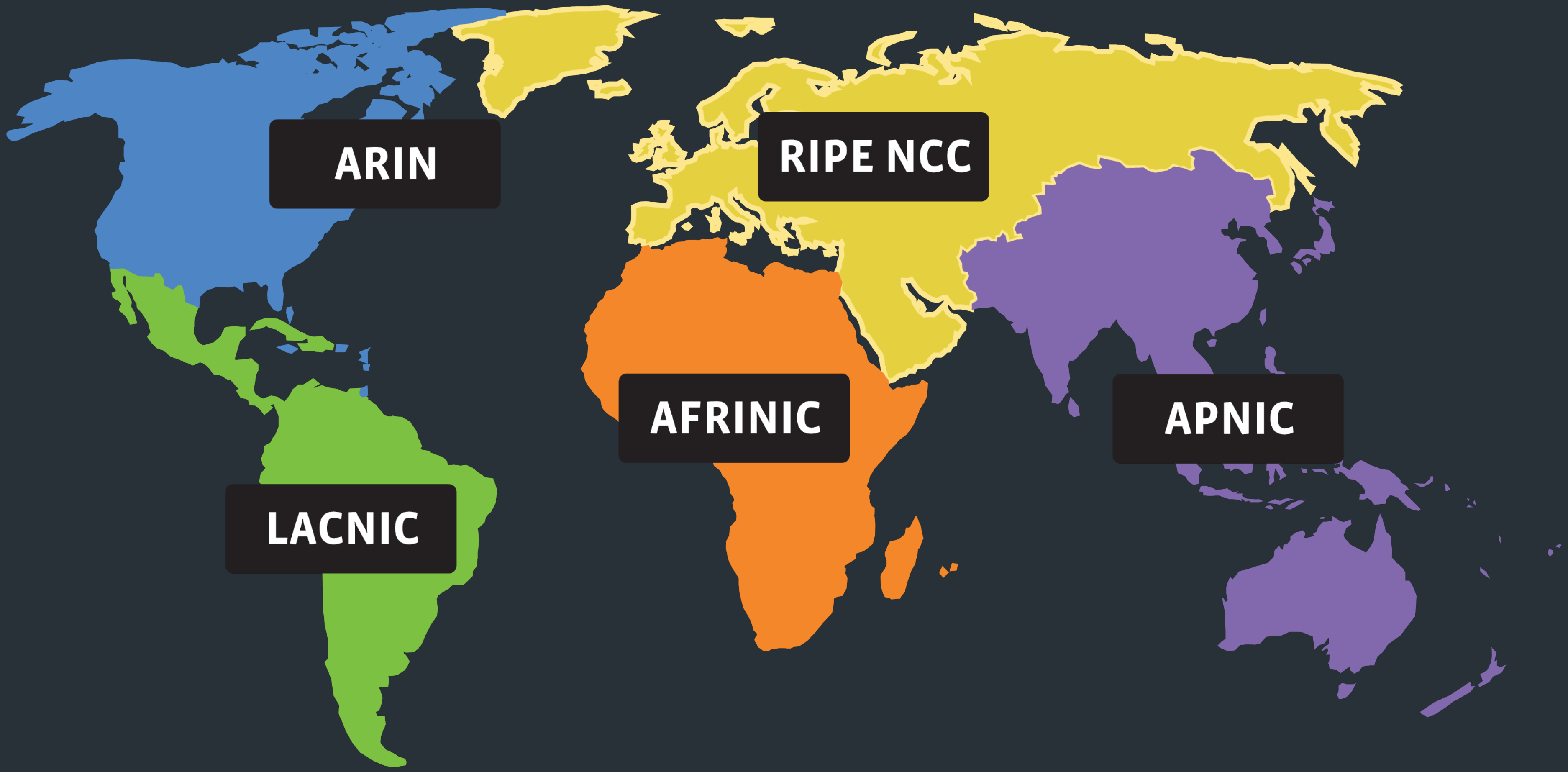
- Allocated by Internet authorities
  - Regional Internet Registries (ARIN, RIPE, APNIC)
  - Internet Service Providers

- Ideally, AS who owns prefix (or its providers) should advertise it
- However: BGP does not verify this

Who owns a prefix?

- Allocated by Internet authorities
  - Regional Internet Registries (ARIN, RIPE, APNIC)
  - Internet Service Providers

The Five RIRs

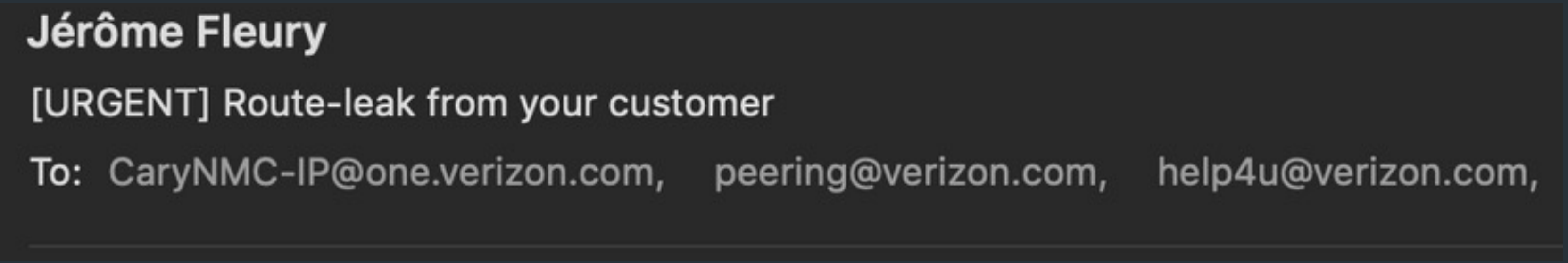


What can go wrong?

Prefix hijacking

Some Notable incidents

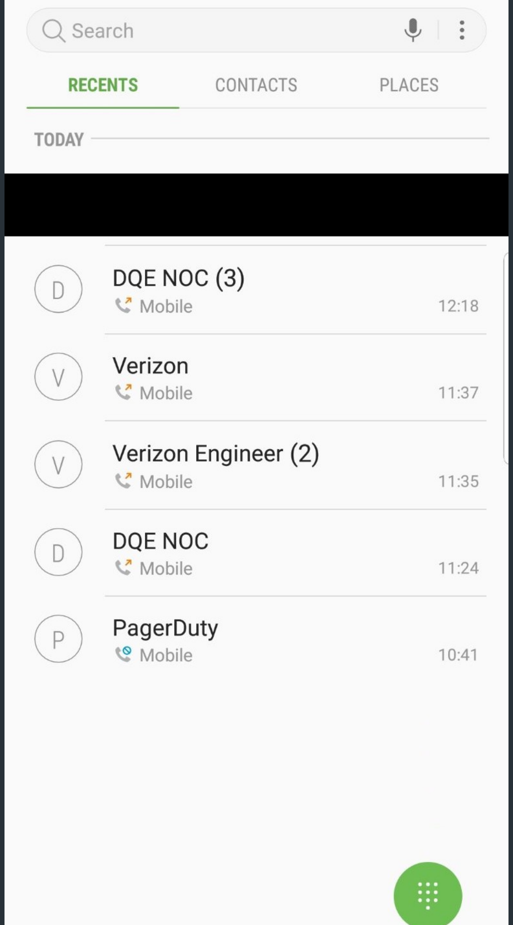
June 24, 2019: Misconfigured small customer router accepted lots of transit traffic



Jérôme Fleury

[URGENT] Route-leak from your customer

To: CaryNMC-IP@one.verizon.com, peering@verizon.com, help4u@verizon.com,



[redacted]

D	DQE NOC (3) Mobile	12:18
V	Verizon Mobile	11:37
V	Verizon Engineer (2) Mobile	11:35
D	DQE NOC Mobile	11:24
P	PagerDuty Mobile	10:41

At this level, solving problems involves a lot of human expertise!

Facebook DNS outage

- October 2021: Misconfiguration causes Facebook to withdraw routes for its DNS servers

- DNS: core service that translates domain names to Ips

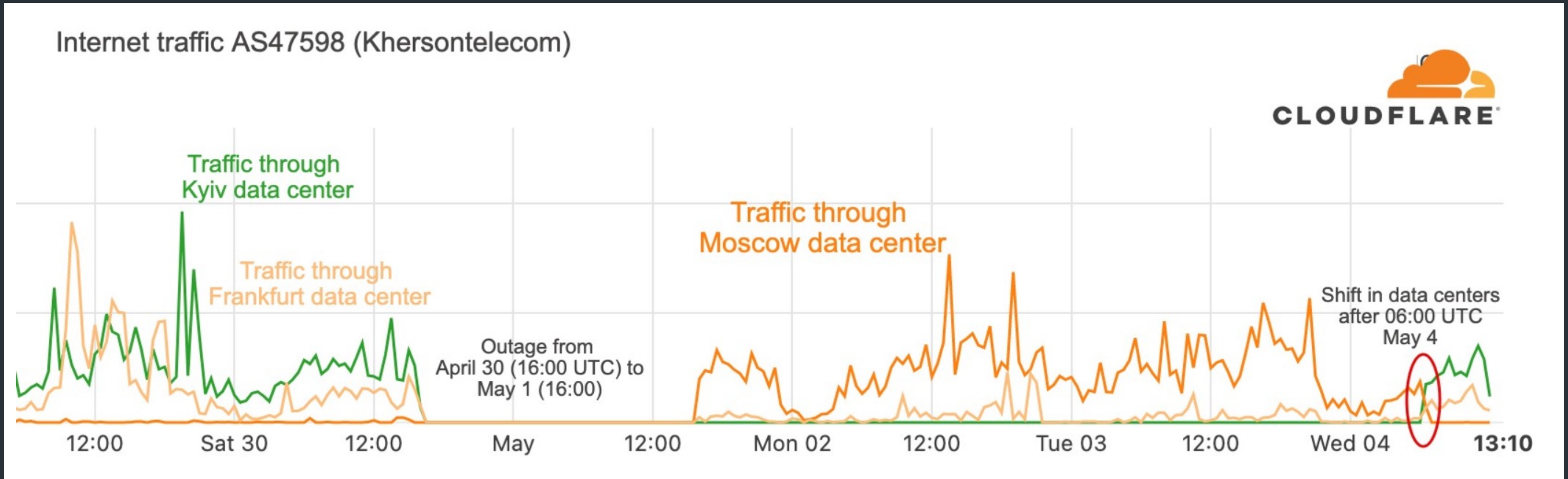
  facebook.com => 1.2.3.6

- All services dependent on Facebook services go offline

Pakistan Youtube incident

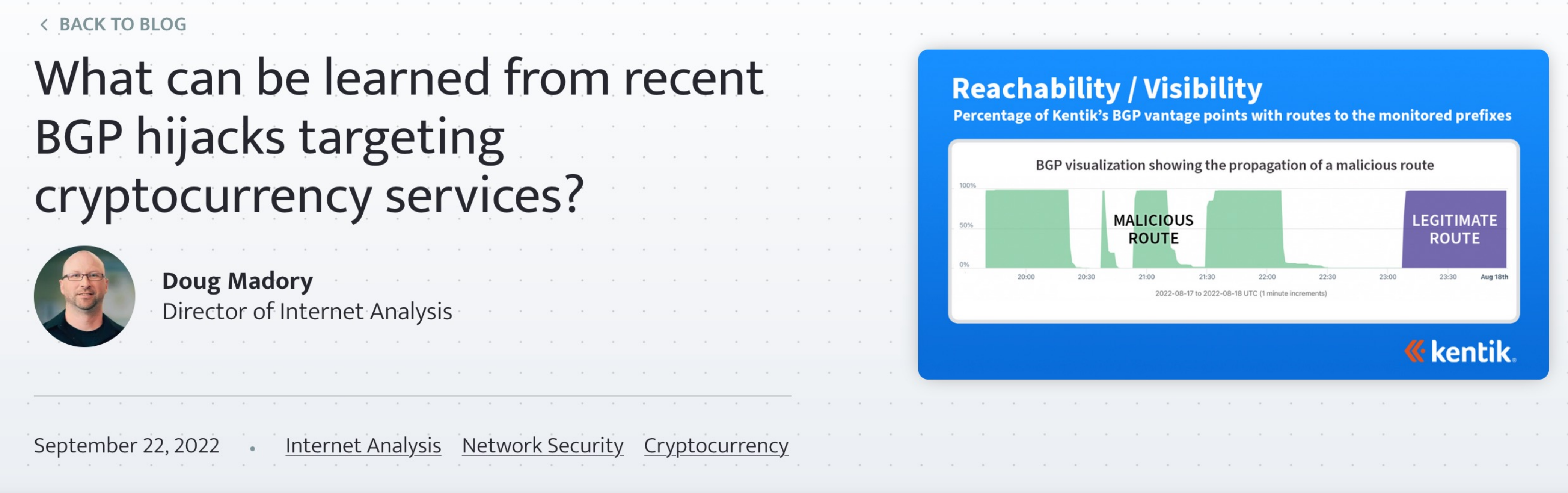
- Youtube's has prefix 208.65.152.0/22
- Pakistan's government order Youtube blocked
- Pakistan Telecom (AS 17557) announces 208.65.153.0/24 in the wrong direction (outwards!)
- Longest prefix match caused worldwide outage
- http://www.youtube.com/watch?v=IzLPKuAOe50

- ISP outage in Russian-occupied city of Kherson, Ukraine
- Comes back several days later… with traffic routed through a Russian ISP



https://blog.cloudflare.com/tracking-shifts-in-internet-connectivity-in-kherson-ukraine/

Prefix Hijacking in the wild



‹ BACK TO BLOG

What can be learned from recent BGP hijacks targeting cryptocurrency services?

Doug Madory
Director of Internet Analysis

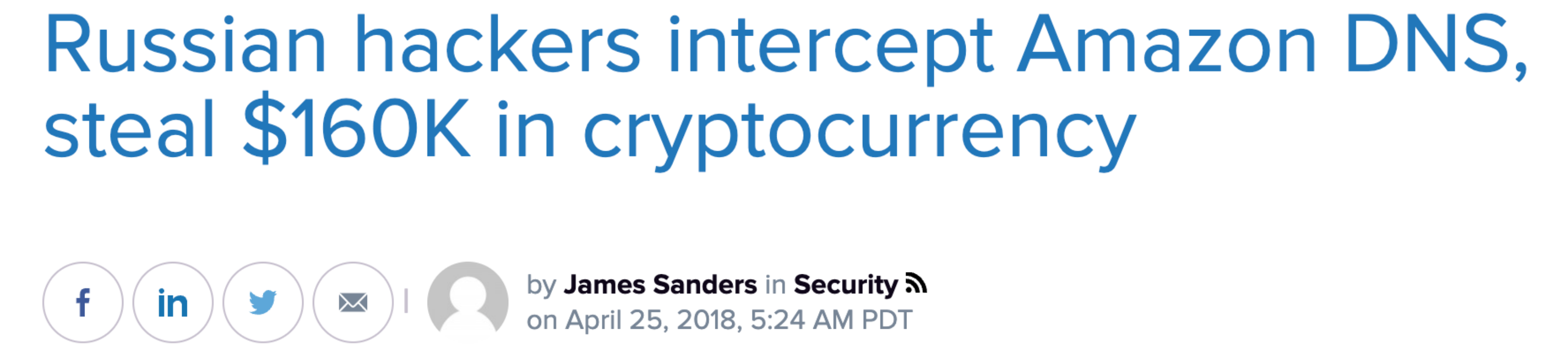
September 22, 2022 • Internet Analysis Network Security Cryptocurrency

Reachability / Visibility
Percentage of Kentik's BGP vantage points with routes to the monitored prefixes
BGP visualization showing the propagation of a malicious route
100%
50%
0%
MALICIOUS ROUTE
LEGITIMATE ROUTE
20:00
20:30
21:00
21:30
22:00
22:30
23:00
23:30
Aug 18th
2022-08-17 to 2022-08-18 UTC (1 minute increments)
kentik.

Writeup (more)

Many other incidents

- China incident, April 8th 2010
  - China Telecom's AS23724 generally announces 40 prefixes
  - On April 8th, announced ~37,000 prefixes
  - About 10% leaked outside of China
  - Suddenly, going to www.dell.com might have you routing through AS23724!

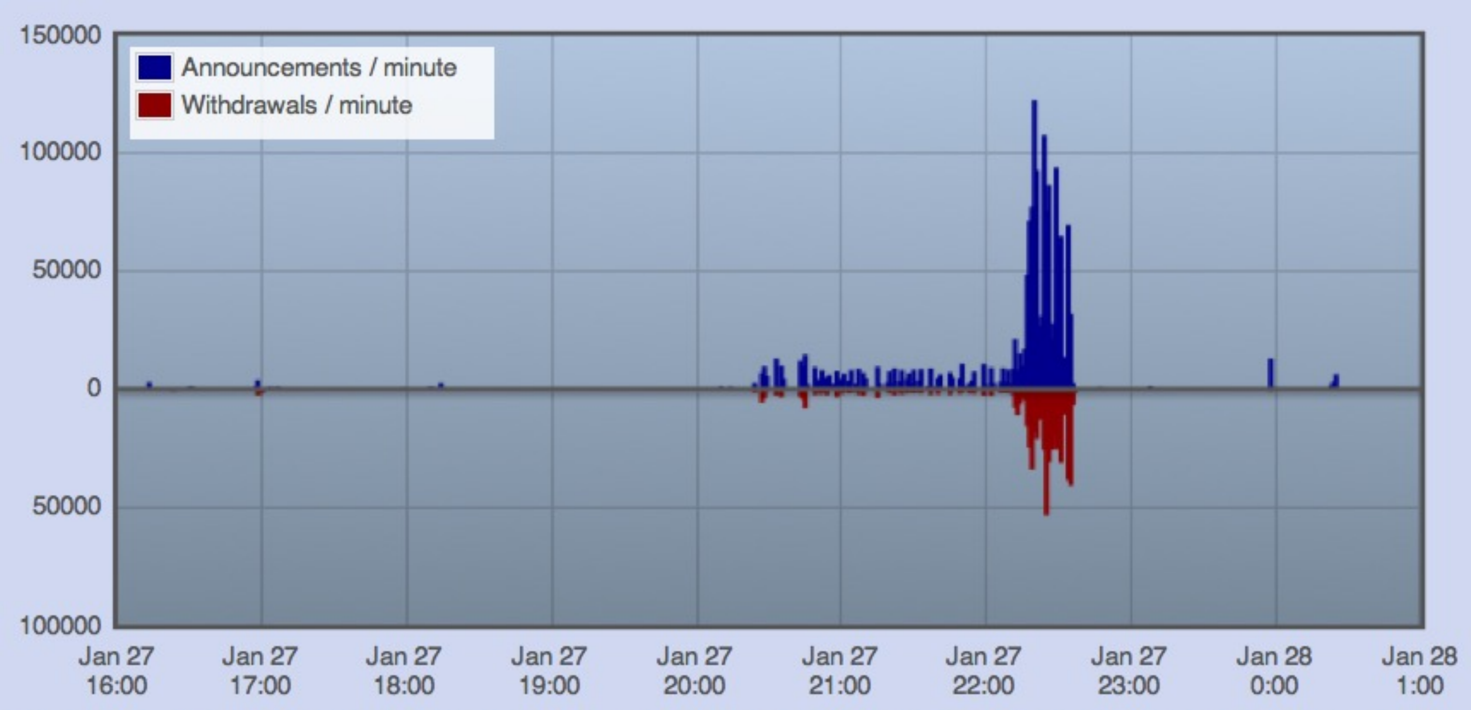


Russian hackers intercept Amazon DNS, steal $160K in cryptocurrency

by **James Sanders** in **Security**
on April 25, 2018, 5:24 AM PDT

“Shutting off” the Internet

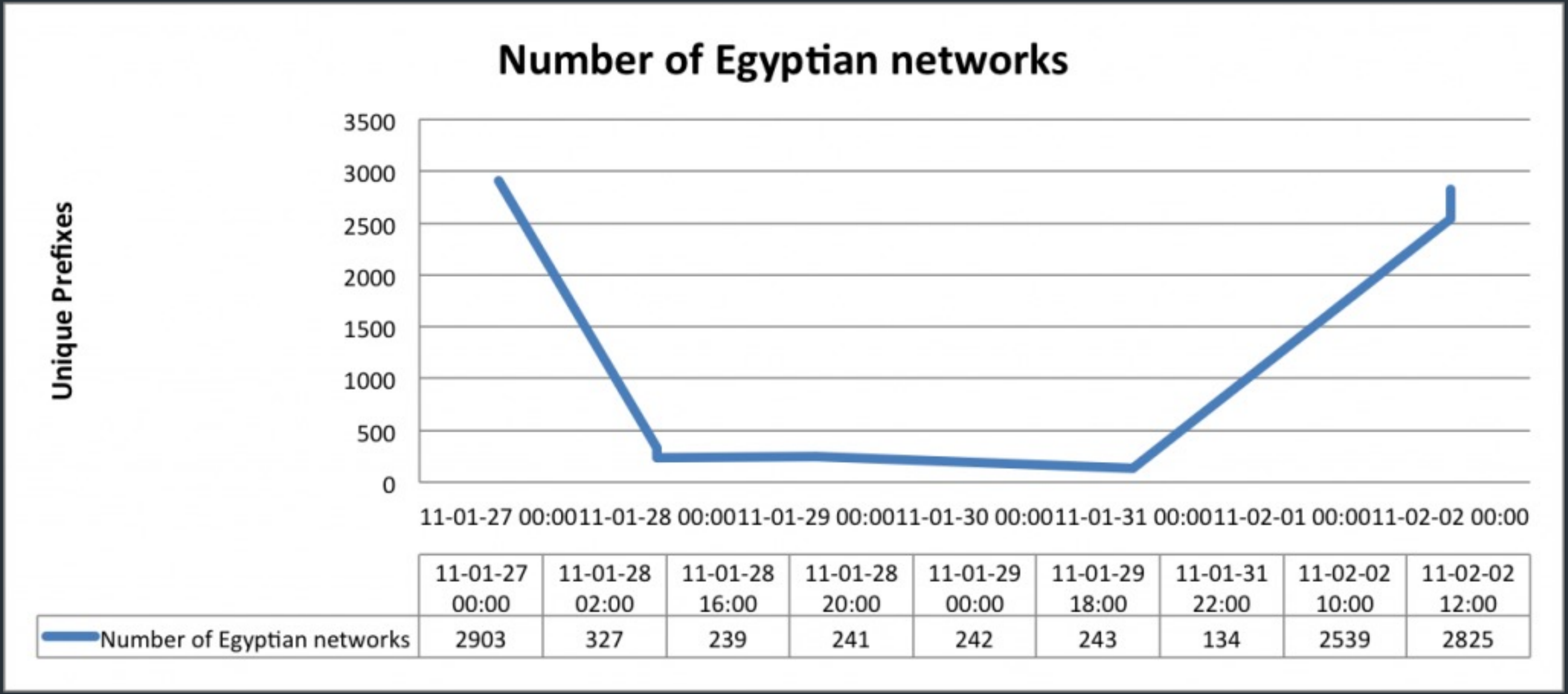
- Starting from Jan 27th, 2011, Egypt was disconnected from the Internet
  - 2769/2903 networks withdrawn from BGP (95%)!



Static view on BGP activity for prefixes originating from Egyptian organisations between 27 Jan 16:00 UTC and 28 Jan 01:00 UTC

Source: RIPEStat - http://stat.ripe.net/egypt/

Egypt Incident



	11-01-27 00:00	11-01-28 02:00	11-01-28 16:00	11-01-28 20:00	11-01-29 00:00	11-01-29 18:00	11-01-31 22:00	11-02-02 10:00	11-02-02 12:00
Number of Egyptian networks	2903	327	239	241	242	243	134	2539	2825

Source: BGPMon (http://bgpmon.net/blog/?p=480)

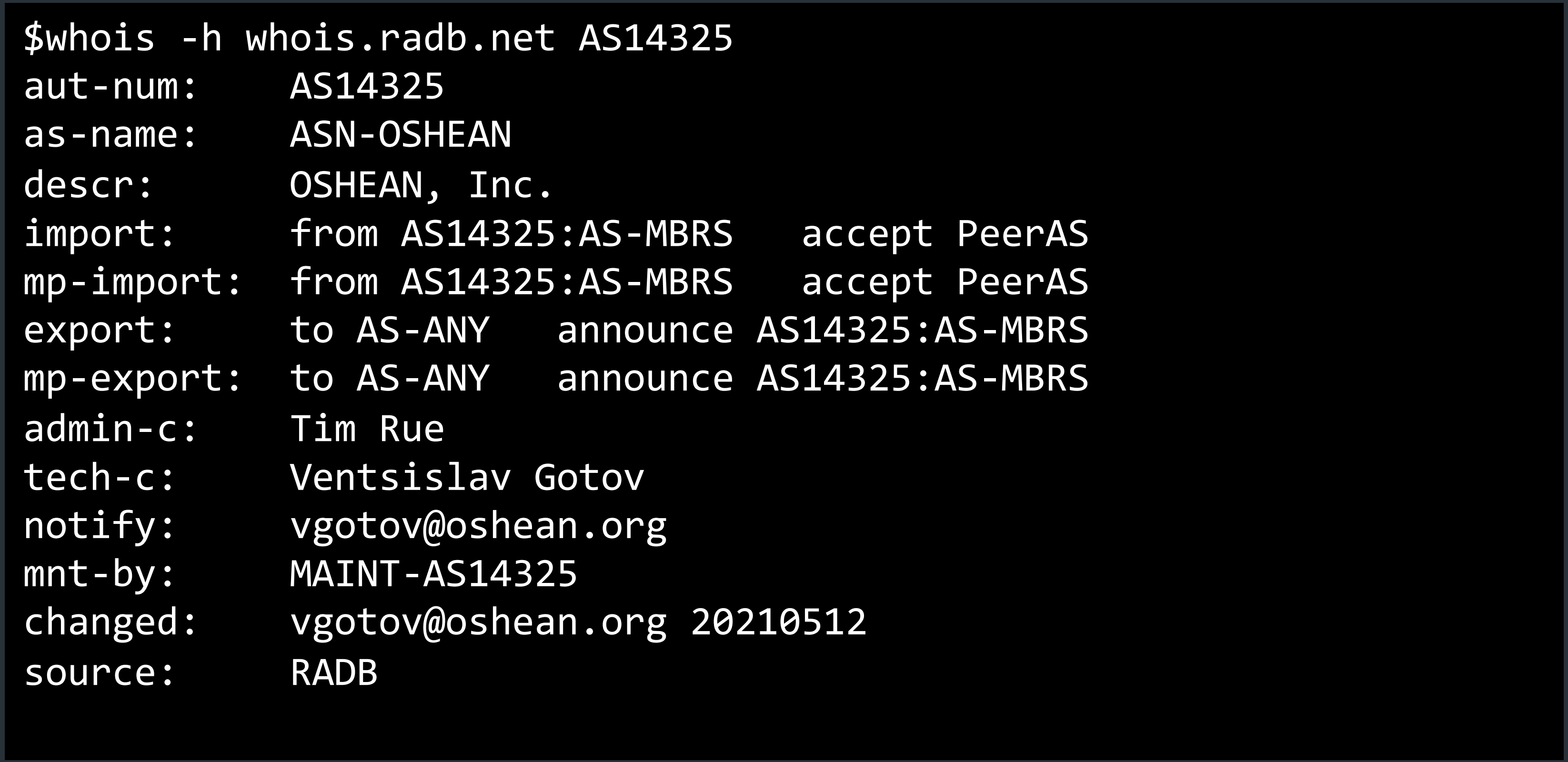
What can be done?

Originally: Internet Routing Registries (IRRs): public database listing IP allocations

```
route: 10.0.0.0/8
descr: University of Blogging
descr: Anytown, USA
origin: AS65099
mnt-by: MNT-UNIVERSITY
notify: person@example.com
changed: person@example.com 20180101
source: RADB
```

But, database not verified and often incomplete/wrong

What can be done?



```
$whois -h whois.radb.net AS14325
aut-num:    AS14325
as-name:    ASN-OSHEAN
descr:      OSHEAN, Inc.
import:     from AS14325:AS-MBRS   accept PeerAS
mp-import:  from AS14325:AS-MBRS   accept PeerAS
export:     to AS-ANY   announce AS14325:AS-MBRS
mp-export:  to AS-ANY   announce AS14325:AS-MBRS
admin-c:    Tim Rue
tech-c:     Ventsislav Gotov
notify:     vgotov@oshean.org
mnt-by:     MAINT-AS14325
changed:    vgotov@oshean.org 20210512
source:     RADB
```

Proposed Solution: RPKI

- Based on a public key infrastructure
- Address attestations
  - Claims the right to originate a prefix
  - Signed and distributed out of band, checked on BGP updates
  - Checked through delegation chain from ICANN
- Can avoid
  - Prefix hijacking
  - Addition, removal, or reordering of intermediate ASes

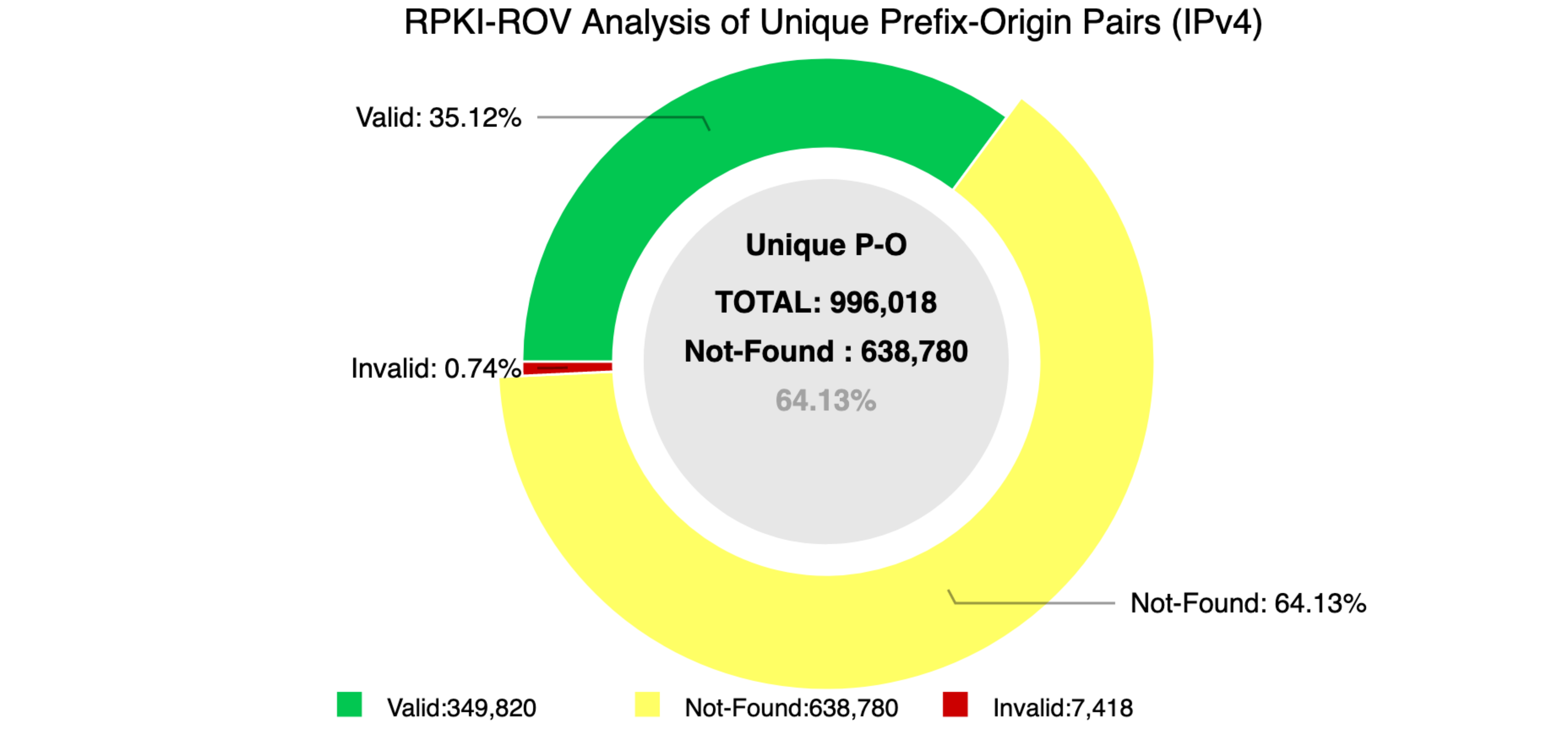
Proposed Solution: RPKI

- Every AS adds *signature* of its route info in database
  - Max prefix size, etc.

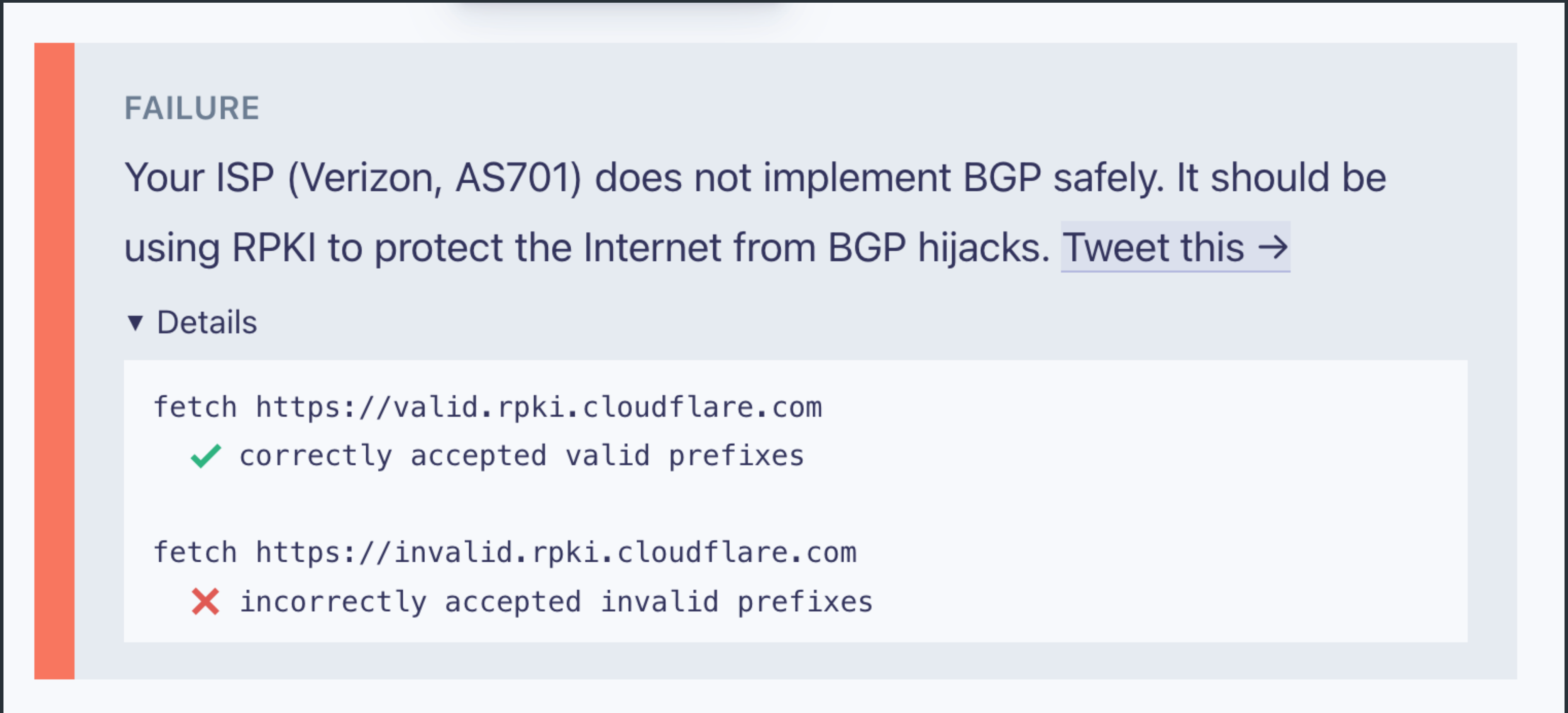
- Other ASes using routes can cryptographically verify advertised routes against signature

- Can avoid
  - Prefix hijacking
  - Addition, removal, or reordering of intermediate ASes

RPKI deployment



RPKI at Brown?



FAILURE

Your ISP (Verizon, AS701) does not implement BGP safely. It should be using RPKI to protect the Internet from BGP hijacks. Tweet this →

▼ Details

```
fetch https://valid.rpki.cloudflare.com
   ✓ correctly accepted valid prefixes

fetch https://invalid.rpki.cloudflare.com
   ✗ incorrectly accepted invalid prefixes
```

Following slides not covered, but interesting

BGP Protocol Details

- *BGP speakers*: nodes that communicates with other ASes over BGP

- Speakers connect over TCP on port 179

- Exact protocol details are out of scope for this class; most important messages have type UPDATE

Prefixes

- Nodes in local network share prefix
  - Key to decide whether to send message locally
- Prefixes can also aggregate multiple networks
  - E.g., 100.20.33.128/25, 100.20.33.0/25 -> 100.20.33.0/24
- If networks connected hierarchically, can have significant aggregation
- But allocations aren't so hierarchical… what does this mean?

Anatomy of an UPDATE

- Withdrawn routes: list of withdrawn IP prefixes
- Network Layer Reachability Information (NLRI)
  - List of prefixes to which path attributes apply
- Path attributes
  - ORIGIN, AS_PATH, NEXT_HOP, MULTI-EXIT-DISC, LOCAL_PREF, ATOMIC_AGGREGATE, AGGREGATOR, …
  - Extensible: can add new types of attributes

Example

- NLRI: 128.148.0.0/16
- AS-Path: ASN 44444 3356 14325 11078
- Next Hop IP
- Various knobs for traffic engineering:
  - Metric, weight, LocalPath, MED, Communities
  - Lots of voodoo

Demo: AS11078

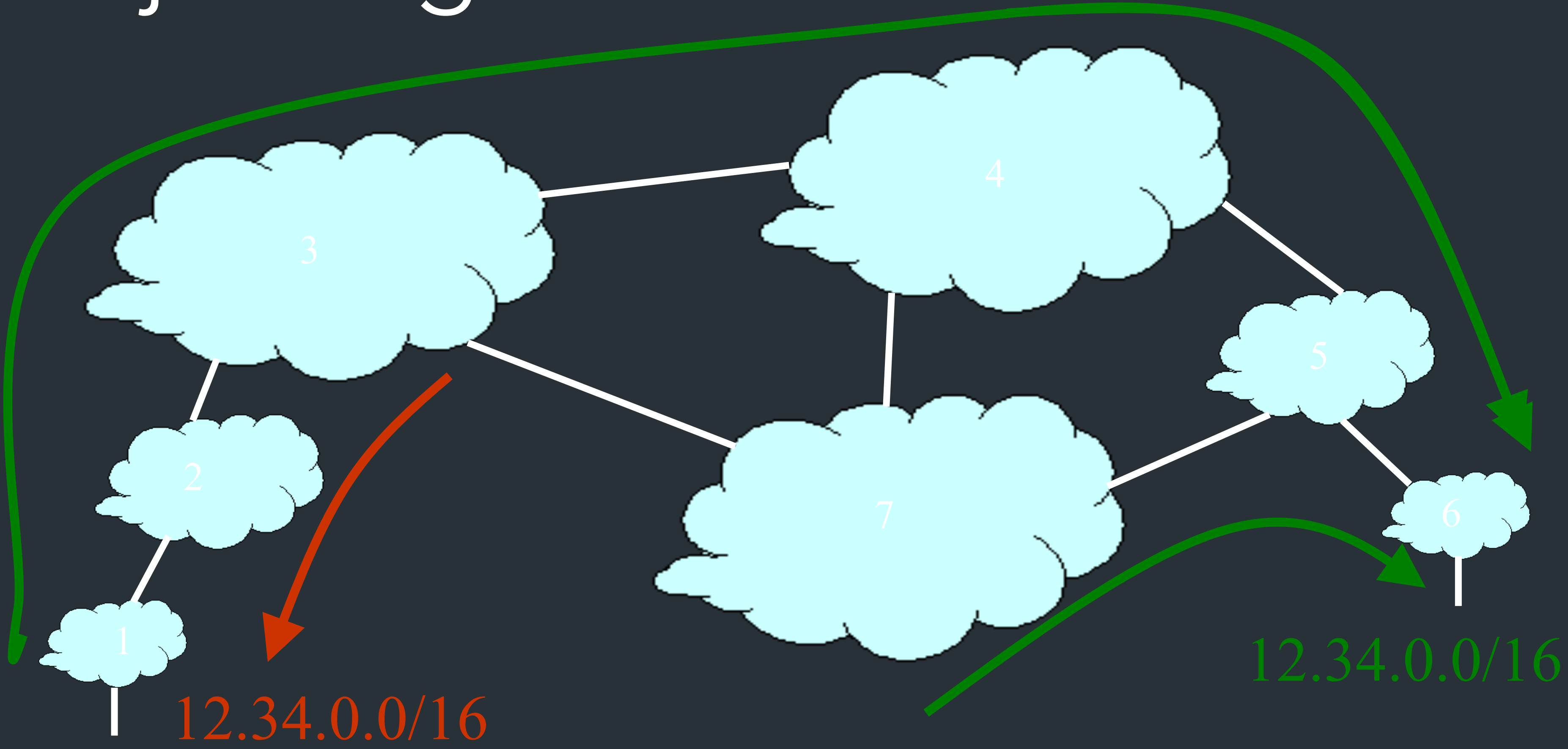
BGP Security Goals

- Confidential message exchange between neighbors
- Validity of routing information
  - Origin, Path, Policy
- Correspondence to the data path

Origin: IP Address Ownership and Hijacking

- IP address block assignment
  - Regional Internet Registries (ARIN, RIPE, APNIC)
  - Internet Service Providers
- Proper origination of a prefix into BGP
  - By the AS who owns the prefix
  - … or, by its upstream provider(s) in its behalf
- However, what's to stop someone else?
  - Prefix hijacking: another AS originates the prefix
  - BGP does not verify that the AS is authorized
  - Registries of prefix ownership are inaccurate

Prefix Hijacking

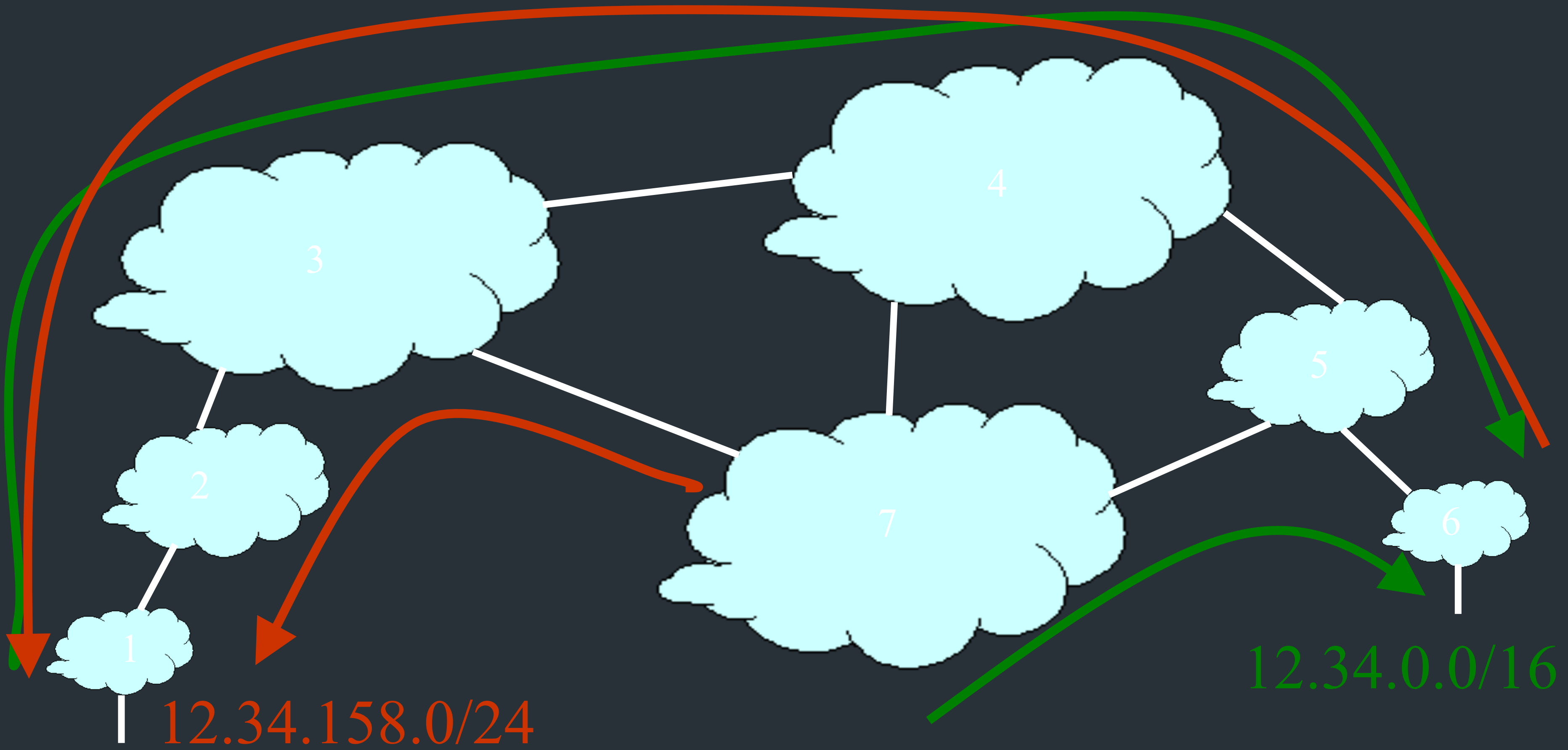


- Consequences for the affected ASes
  - Blackhole: data traffic is discarded
  - Snooping: data traffic is inspected, and then redirected
  - Impersonation: data traffic is sent to bogus destinations

Hijacking is Hard to Debug

- Real origin AS doesn’t see the problem
  - Picks its own route
  - Might not even learn the bogus route
- May not cause loss of connectivity
  - E.g., if the bogus AS snoops and redirects
  - … may only cause performance degradation
- Or, loss of connectivity is isolated
  - E.g., only for sources in parts of the Internet
- Diagnosing prefix hijacking
  - Analyzing updates from many vantage points
  - Launching traceroute from many vantage points

Sub-Prefix Hijacking



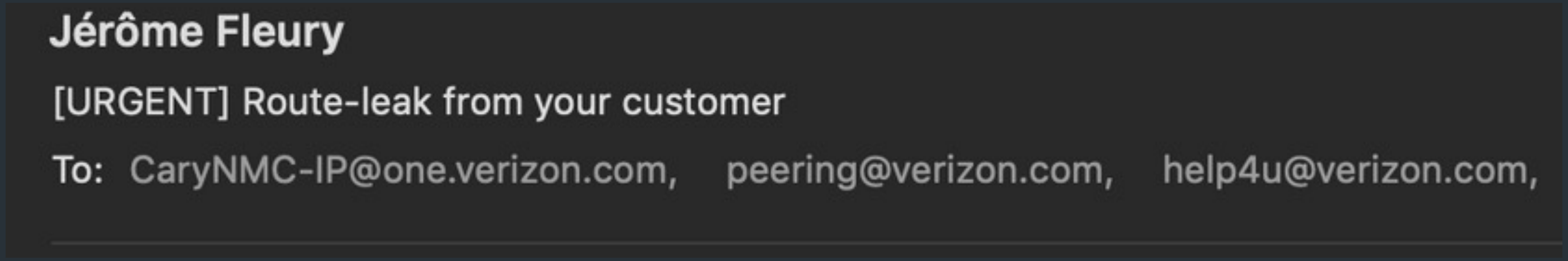
- Originating a more-specific prefix
  - Every AS picks the bogus route for that prefix
  - Traffic follows the longest matching prefix

How to Hijack a Prefix

- The hijacking AS has
  - Router with eBGP session(s)
  - Configured to originate the prefix
- Getting access to the router
  - Network operator makes configuration mistake
  - Disgruntled operator launches an attack
  - Outsider breaks into the router and reconfigures
- Getting other ASes to believe bogus route
  - Neighbor ASes not filtering the routes
  - … e.g., by allowing only expected prefixes
  - But, specifying filters on *peering* links is hard

Recent Notable incidents

- October 4 2021: Facebook accidentally removed routes for its DNS servers
  - Outside world couldn't resolve facebook.com, and neither could Facebook!
- June 24, 2019: Misconfigured router accepted lots of transit traffic



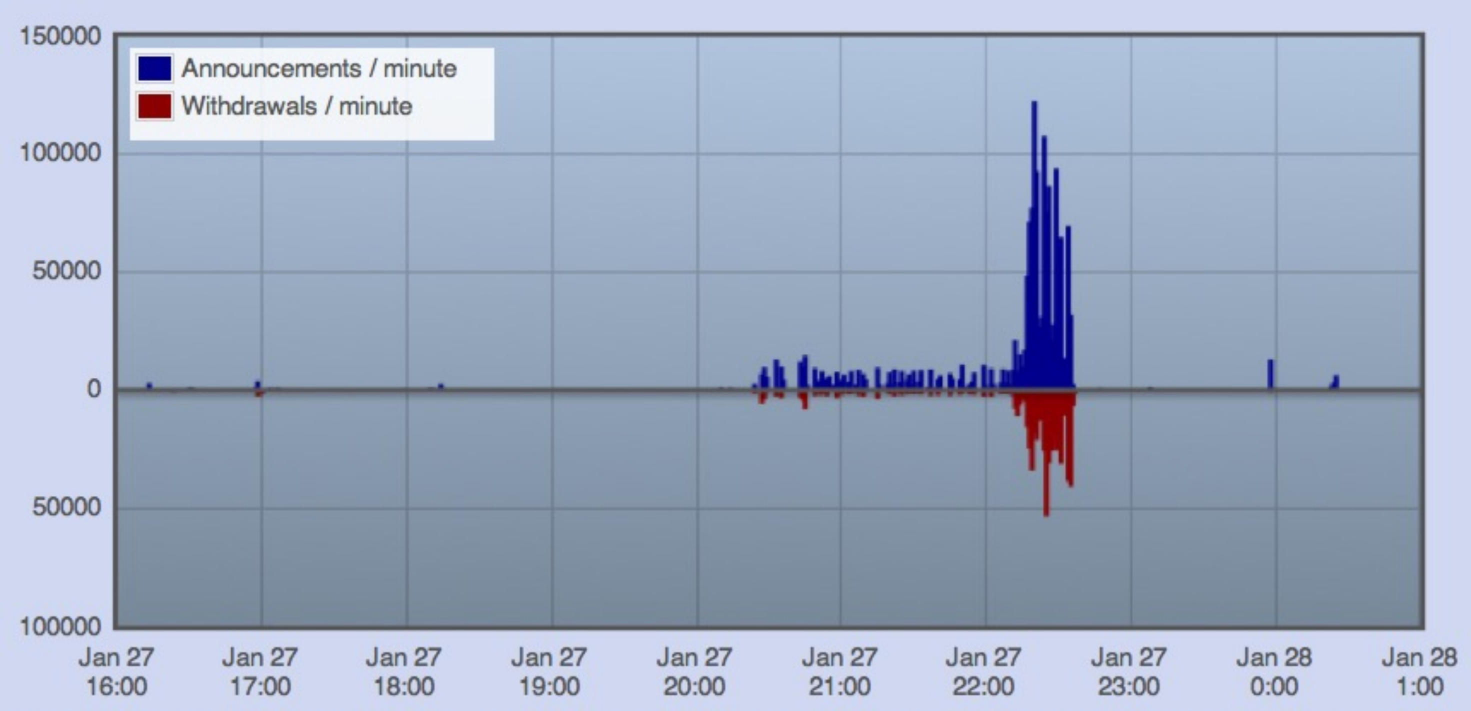
Jérôme Fleury

[URGENT] Route-leak from your customer

To: CaryNMC-IP@one.verizon.com, peering@verizon.com, help4u@verizon.com,

“Shutting off” the Internet

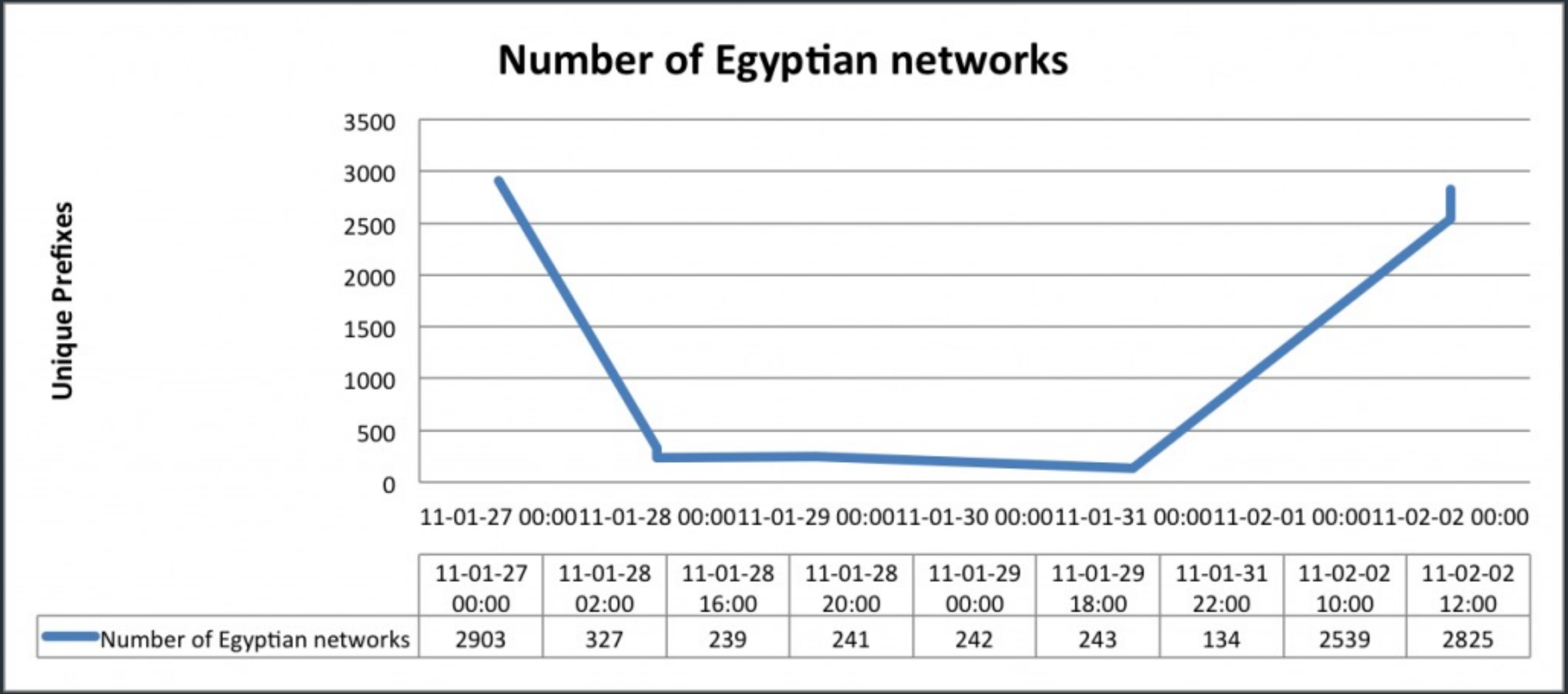
- Starting from Jan 27th, 2011, Egypt was disconnected from the Internet
  - 2769/2903 networks withdrawn from BGP (95%)!



Static view on BGP activity for prefixes originating from Egyptian organisations between 27 Jan 16:00 UTC and 28 Jan 01:00 UTC

Source: RIPEStat - http://stat.ripe.net/egypt/

Egypt Incident



Number of Egyptian networks

	11-01-27 00:00	11-01-28 02:00	11-01-28 16:00	11-01-28 20:00	11-01-29 00:00	11-01-29 18:00	11-01-31 22:00	11-02-02 10:00	11-02-02 12:00
Number of Egyptian networks	2903	327	239	241	242	243	134	2539	2825

Source: BGPMon (http://bgpmon.net/blog/?p=480)

Pakistan Youtube incident

- Youtube's has prefix 208.65.152.0/22
- Pakistan's government order Youtube blocked
- Pakistan Telecom (AS 17557) announces 208.65.153.0/24 in the wrong direction (outwards!)
- Longest prefix match caused worldwide outage
- http://www.youtube.com/watch?v=IzLPKuAOe50

Many other incidents

- Spammers steal unused IP space to hide
  - Announce very short prefixes (e.g., /8). Why?
  - For a short amount of time
- China incident, April 8th 2010
  - China Telecom's AS23724 generally announces 40 prefixes
  - On April 8th, announced ~37,000 prefixes
  - About 10% leaked outside of China
  - Suddenly, going to www.dell.com might have you routing through AS23724!

Attacks on BGP Paths

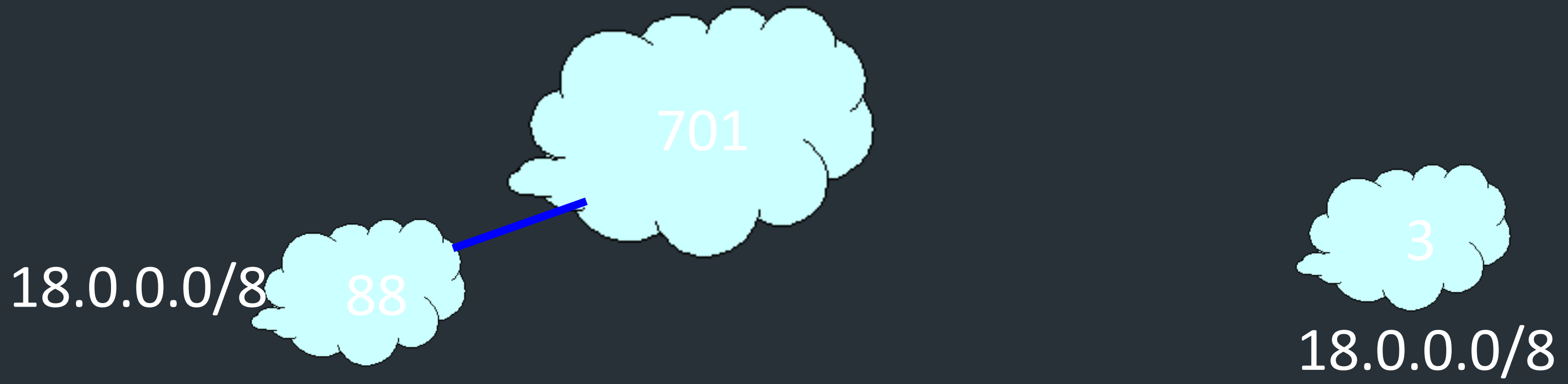
- Remove an AS from the path
  - E.g., 701 3715 88 -> 701 88
- Why?
  - Attract sources that would normally avoid AS 3715
  - Make path through you look more attractive
  - Make AS 88 look like it is closer to the core
  - Can fool loop detection!
- May be hard to tell whether this is a lie
  - 88 could indeed connect directly to 701!

Attacks on BGP Paths

- Adding ASes to the path
  - E.g., 701 88 -> 701 3715 88
- Why?
  - Trigger loop detection in AS 3715
    - This would block unwanted traffic from AS 3715!
  - Make your AS look more connected
- Who can tell this is a lie?
  - AS 3715 could, if it could see the route
  - AS 88 could, but would it really care?

Attacks on BGP Paths

- Adding ASes at the end of the path
  - E.g., 701 88 into 701 88 3
- Why?
  - Evade detection for a bogus route (if added AS is legitimate owner of a prefix)
- Hard to tell that the path is bogus!

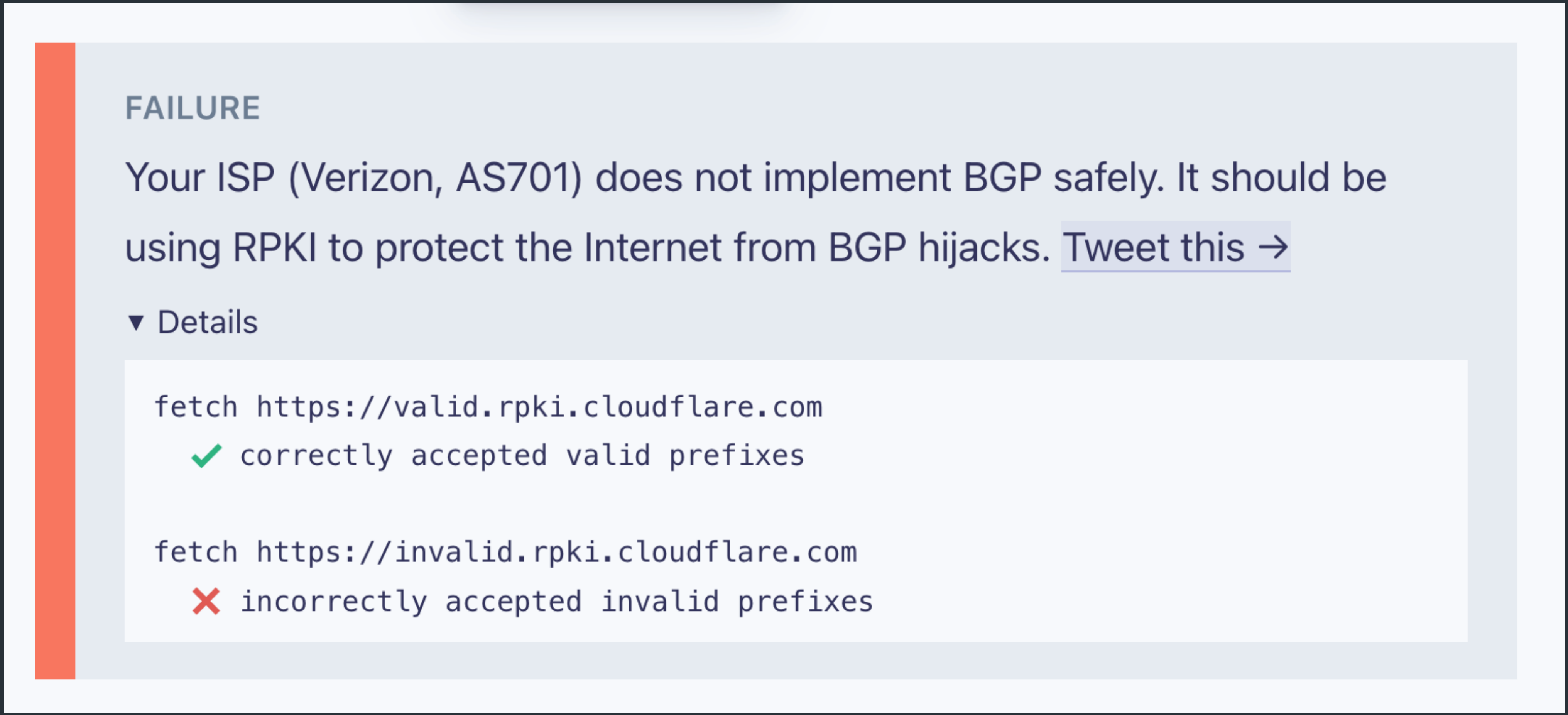


Proposed Solution: S-BGP

- Based on a public key infrastructure
- Address attestations
  - Claims the right to originate a prefix
  - Signed and distributed out of band
  - Checked through delegation chain from ICANN
- Route attestations
  - Attribute in BGP update message
  - Signed by each AS as route along path
- S-BGP can avoid
  - Prefix hijacking
  - Addition, removal, or reordering of intermediate ASes

S-BGP Deployment

- Very challenging
  - PKI (RPKI)
  - Accurate address registries
  - Need to perform cryptographic operations on all path operations
  - Flag day almost impossible
  - Incremental deployment offers little incentive
- But there is hope! [Goldberg et al, 2011]
  - Road to incremental deployment
  - Change rules to break ties for secure paths
  - If a few top Tier-1 ISPs
    - Plus their respective stub clients deploy simplified version (just sign, not validate)
    - Gains in traffic => $ => adoption!



FAILURE

Your ISP (Verizon, AS701) does not implement BGP safely. It should be using RPKI to protect the Internet from BGP hijacks. Tweet this →

▼ Details

```
fetch https://valid.rpki.cloudflare.com
  ✓ correctly accepted valid prefixes

fetch https://invalid.rpki.cloudflare.com
  ✗ incorrectly accepted invalid prefixes
```

Data Plane Attacks

- Routers/ASes can advertise one route, but not necessarily follow it!
- May drop packets
  - Or a fraction of packets
  - What if you just slow down some traffic?
- Can send packets in a different direction
  - Impersonation attack
  - Snooping attack
- How to detect?
  - Congestion or an attack?
  - Can let ping/traceroute packets go through
  - End-to-end checks?
- Harder to pull off, as you need control of a router

BGP Recap

- Key protocol that holds Internet routing together
- Path Vector Protocol among Autonomous Systems
- Policy, feasibility first; non-optimal routes
- Important security problems

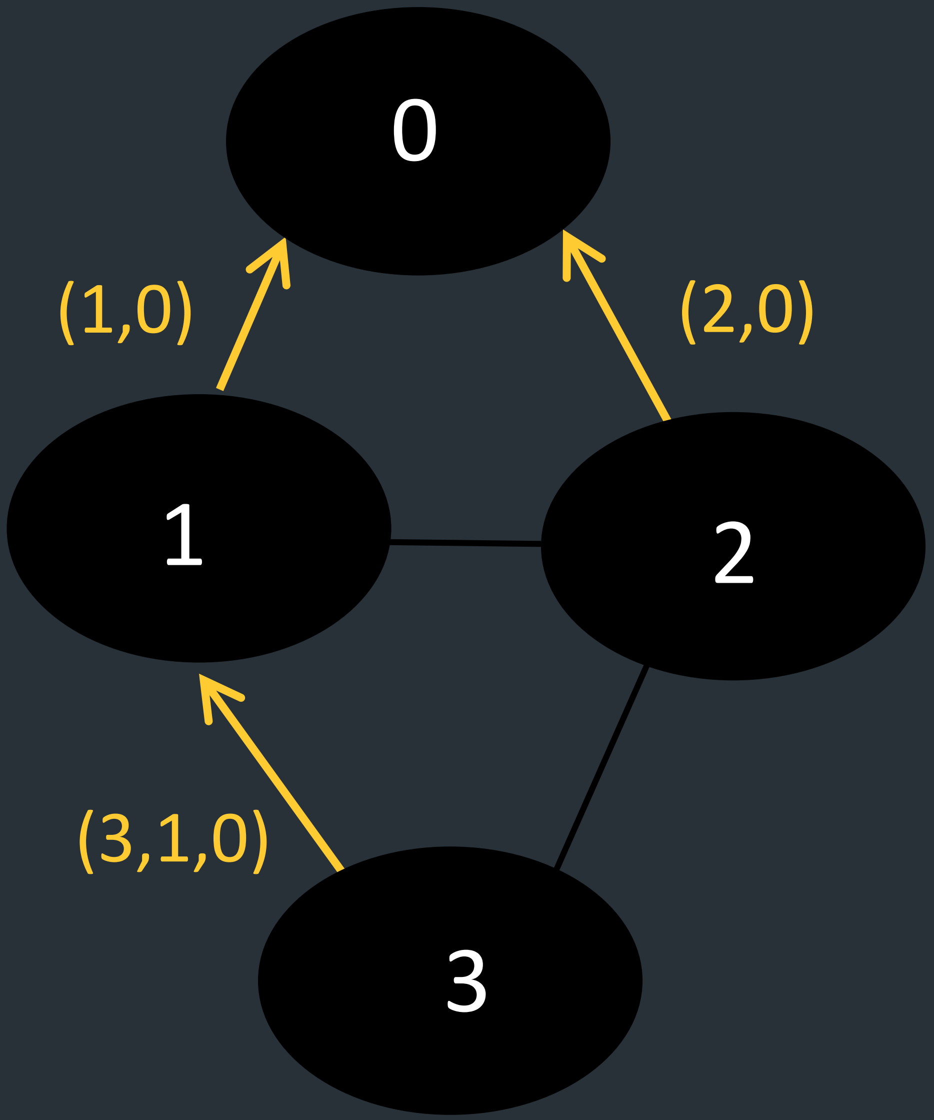
Next Class

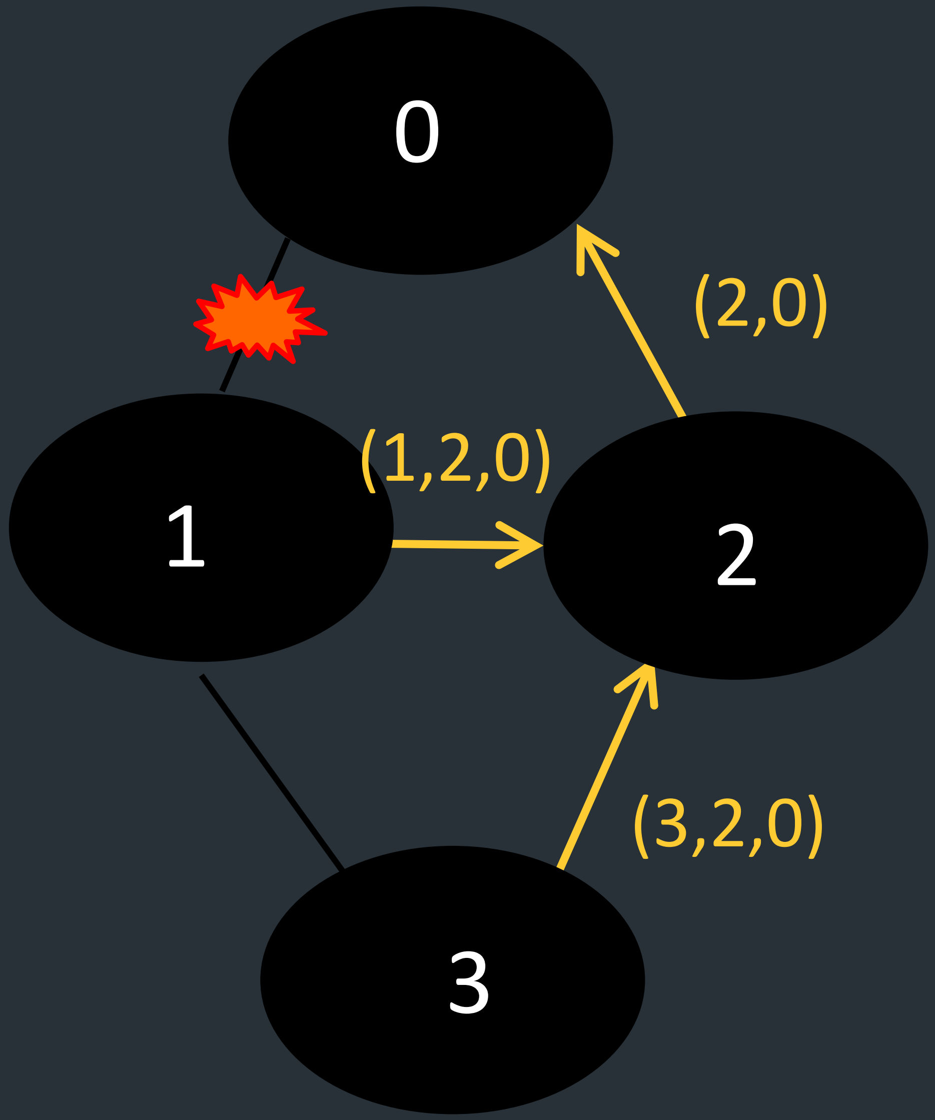
- Network layer wrap up

Convergence

- Given a change, how long until the network re-stabilizes?
  - Depends on change: sometimes never
  - Open research problem: “tweak and pray”
  - Distributed setting is challenging
- Some reasons for change
  - Topology changes
  - BGP session failures
  - Changes in policy
  - Conflicts between policies can cause oscillation

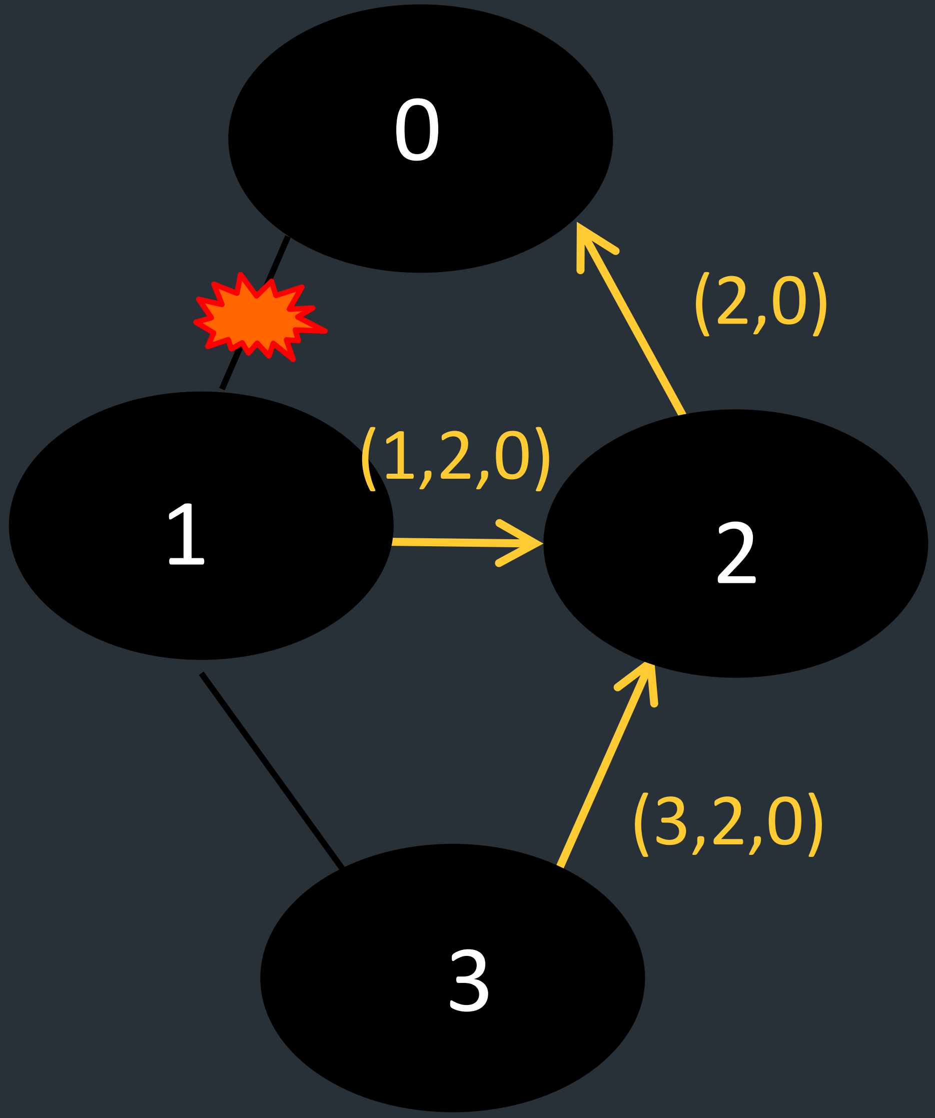
Routing Change: Before and After





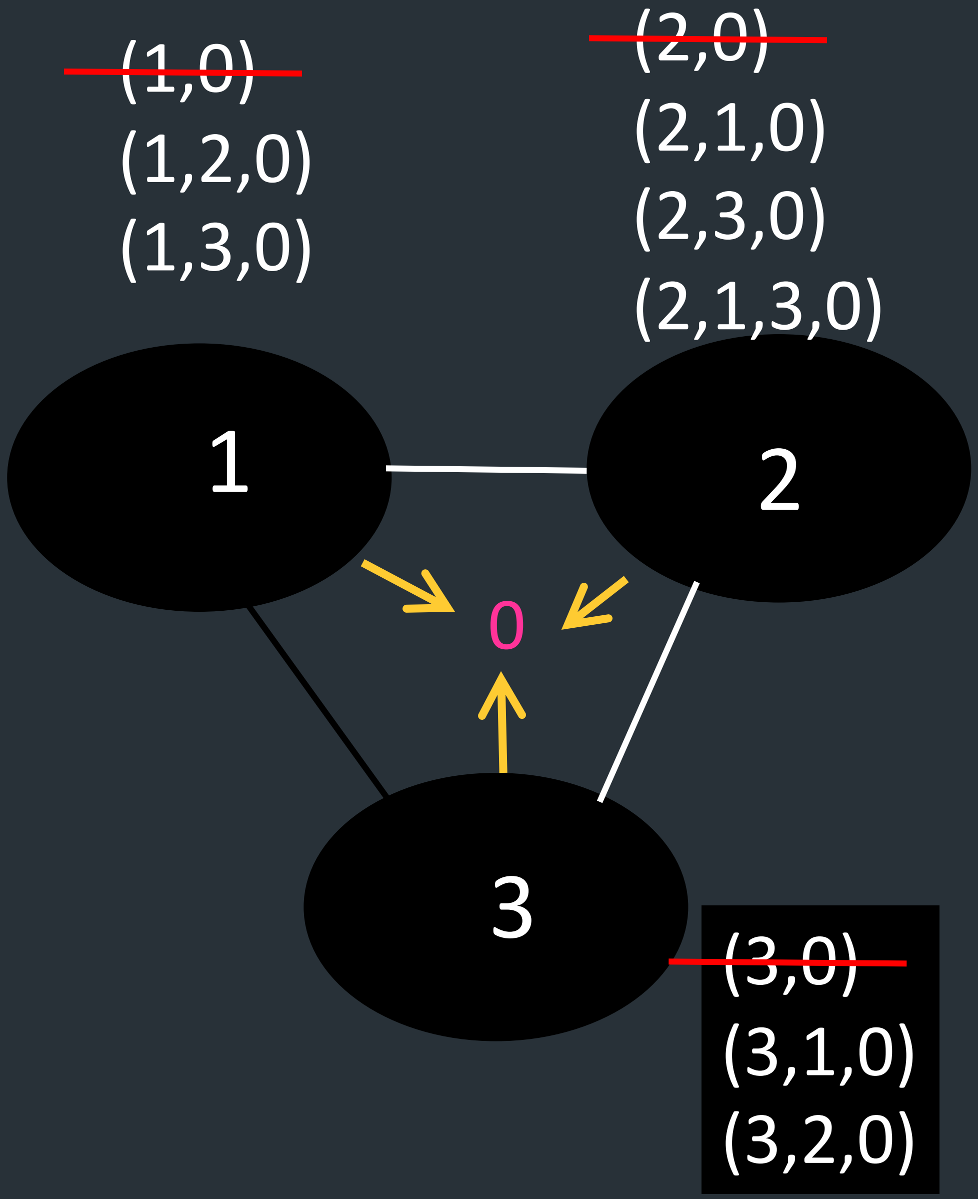
Routing Change: Path Exploration

- AS 1
  - Delete the route (1,0)
  - Switch to next route (1,2,0)
  - Send route (1,2,0) to AS 3
- AS 3
  - Sees (1,2,0) replace (1,0)
  - Compares to route (2,0)
  - Switches to using AS 2



Routing Change: Path Exploration

- Initial situation
  - Destination 0 is alive
  - All ASes use direct path
- When destination dies
  - All ASes lose direct path
  - All switch to longer paths
  - Eventually withdrawn
- E.g., AS 2
  - (2,0) → (2,1,0)
  - (2,1,0) → (2,3,0)
  - (2,3,0) → (2,1,3,0)
  - (2,1,3,0) → null
- Convergence may be slow!

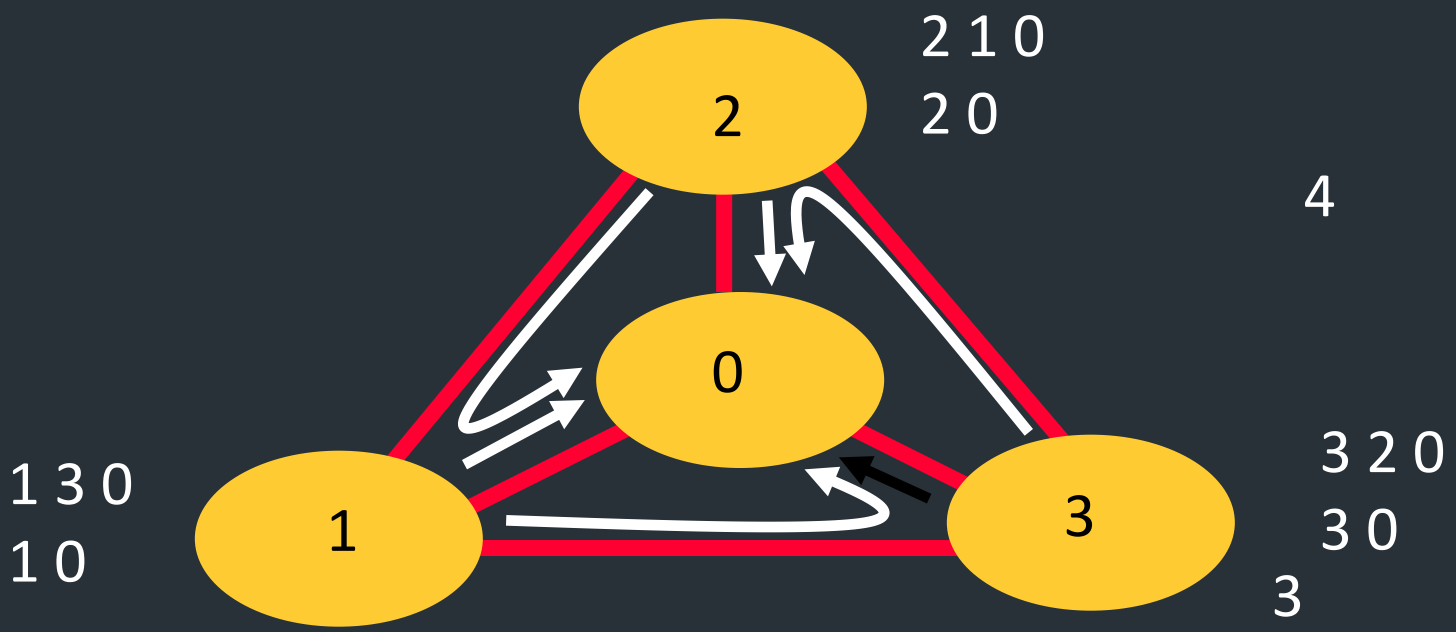


Route Engineering

- Route filtering
- Setting weights
- More specific routes: longest prefix
- AS prepending: “477 477 477 477”
- More of an art than science

Unstable Configurations

- Due to policy conflicts (Dispute Wheel)

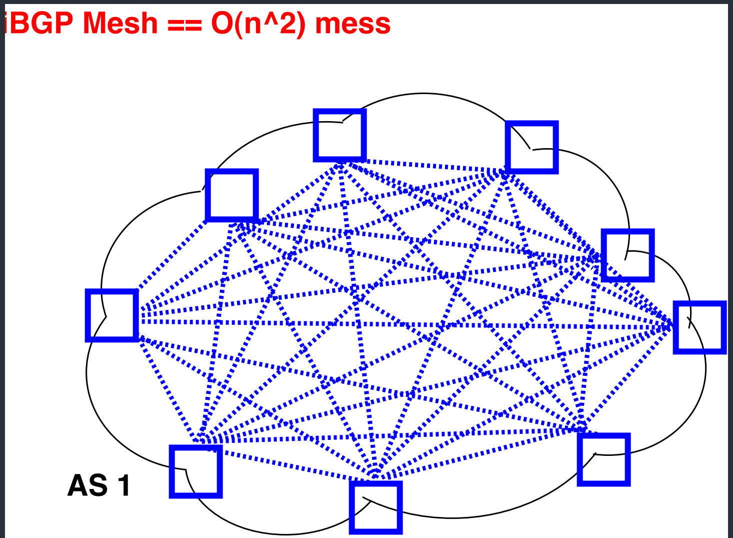


Avoiding BGP Instabilities

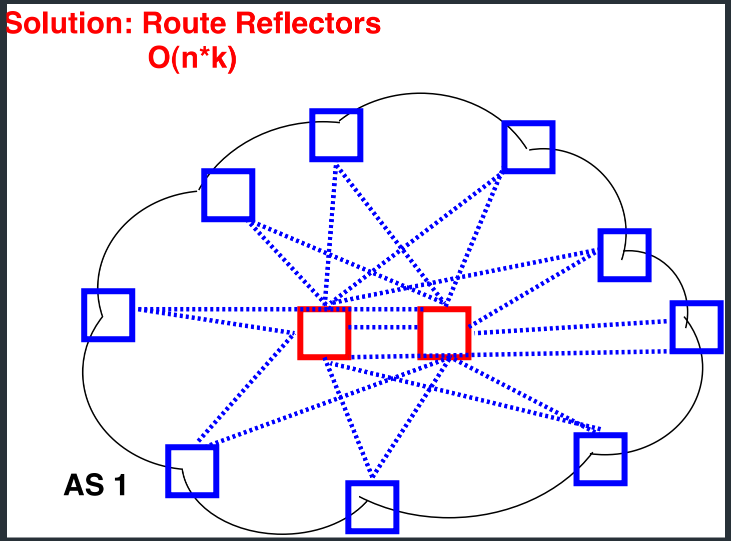
- Detecting conflicting policies
  - Centralized: NP-Complete problem!
  - Distributed: open research problem
  - Requires too much cooperation
- Detecting oscillations
  - Monitoring for repetitive BGP messages
- Restricted routing policies and topologies
  - Some topologies / policies proven to be safe*

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

Scaling iBGP: route reflectors



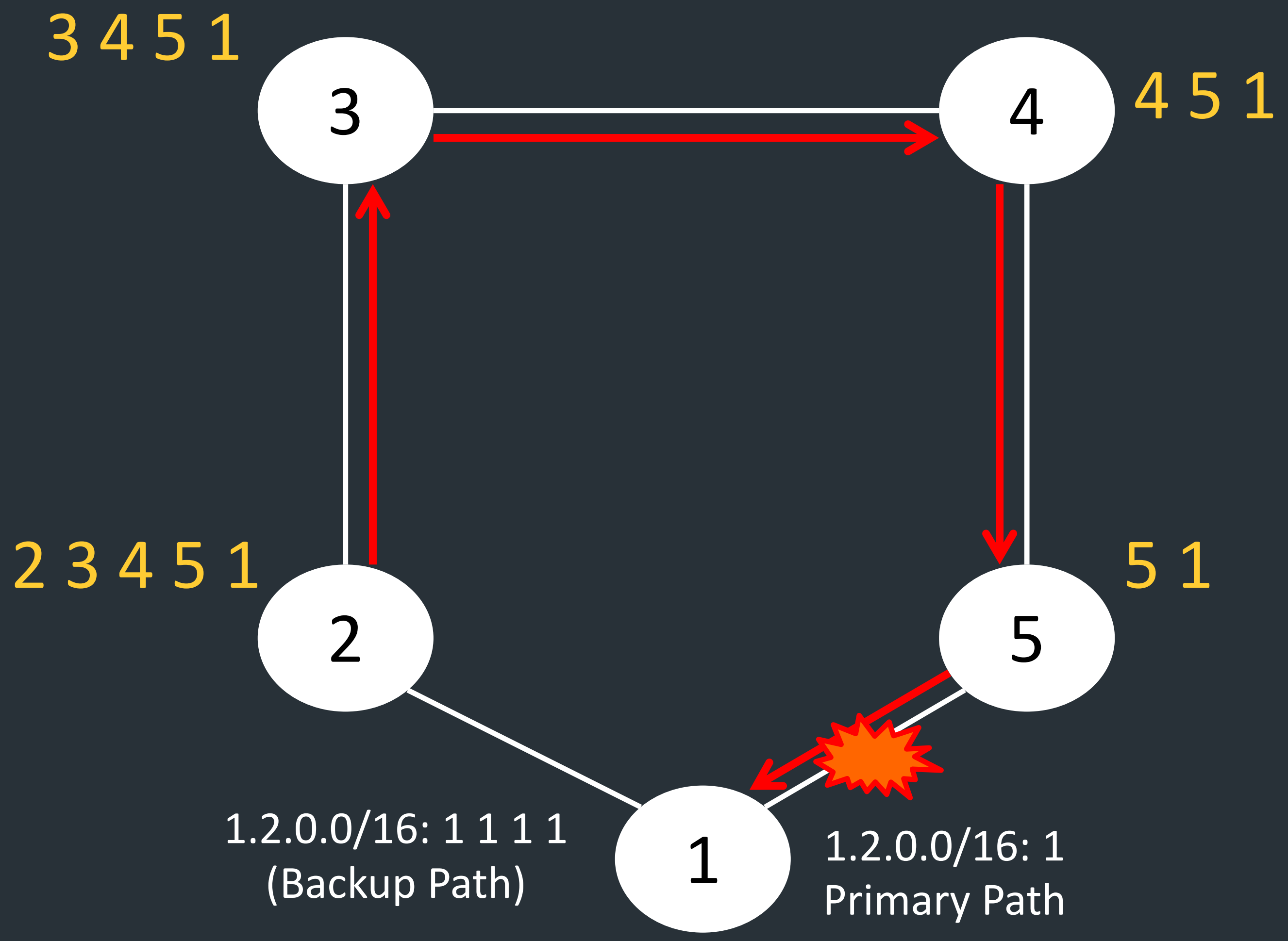
Scaling iBGP: route reflectors



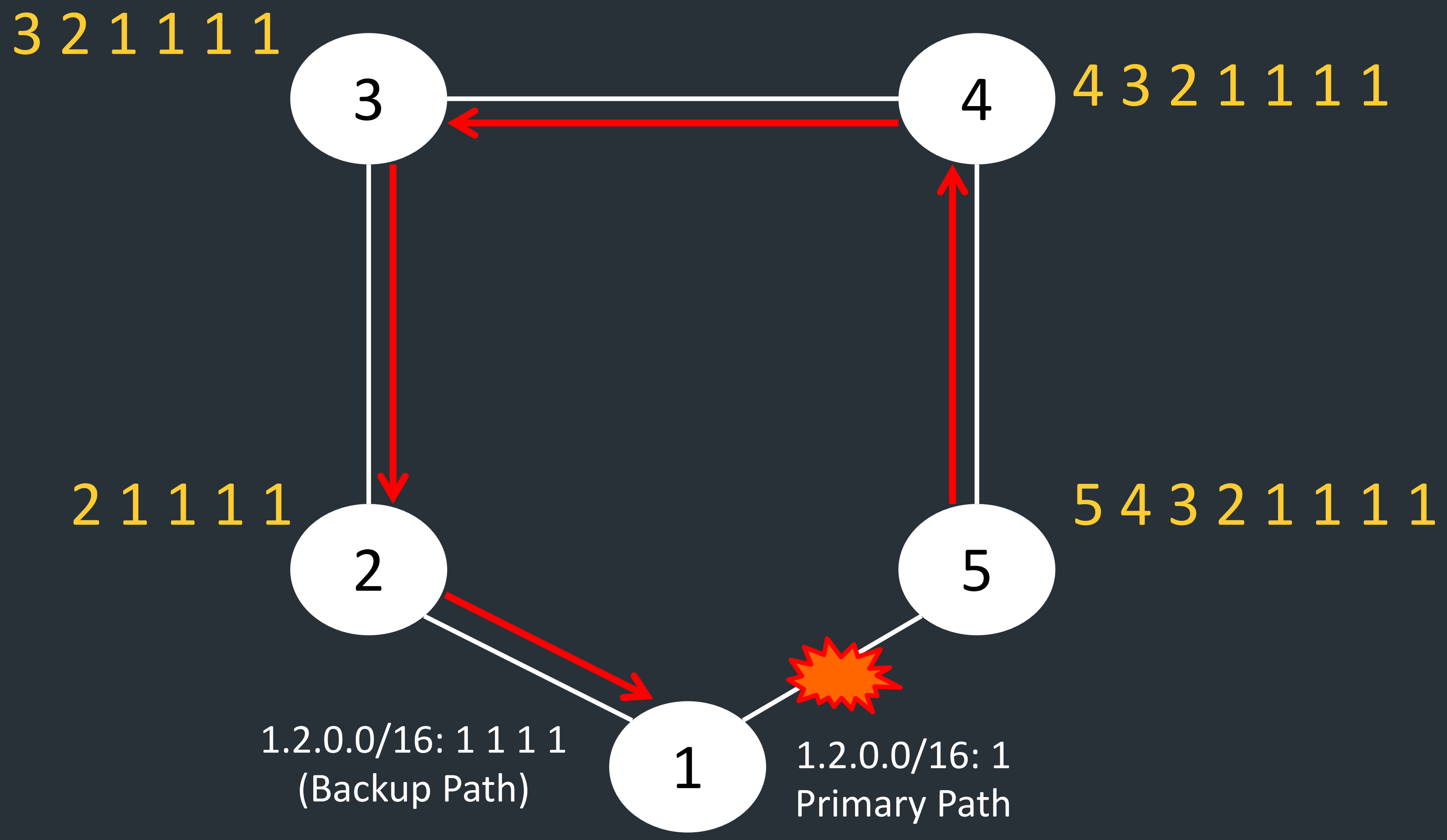
Multiple Stable Configurations BGP Wedgies [RFC 4264]

- Typical policy:
  - Prefer routes from customers
  - Then prefer shortest paths

BGP Wedgies

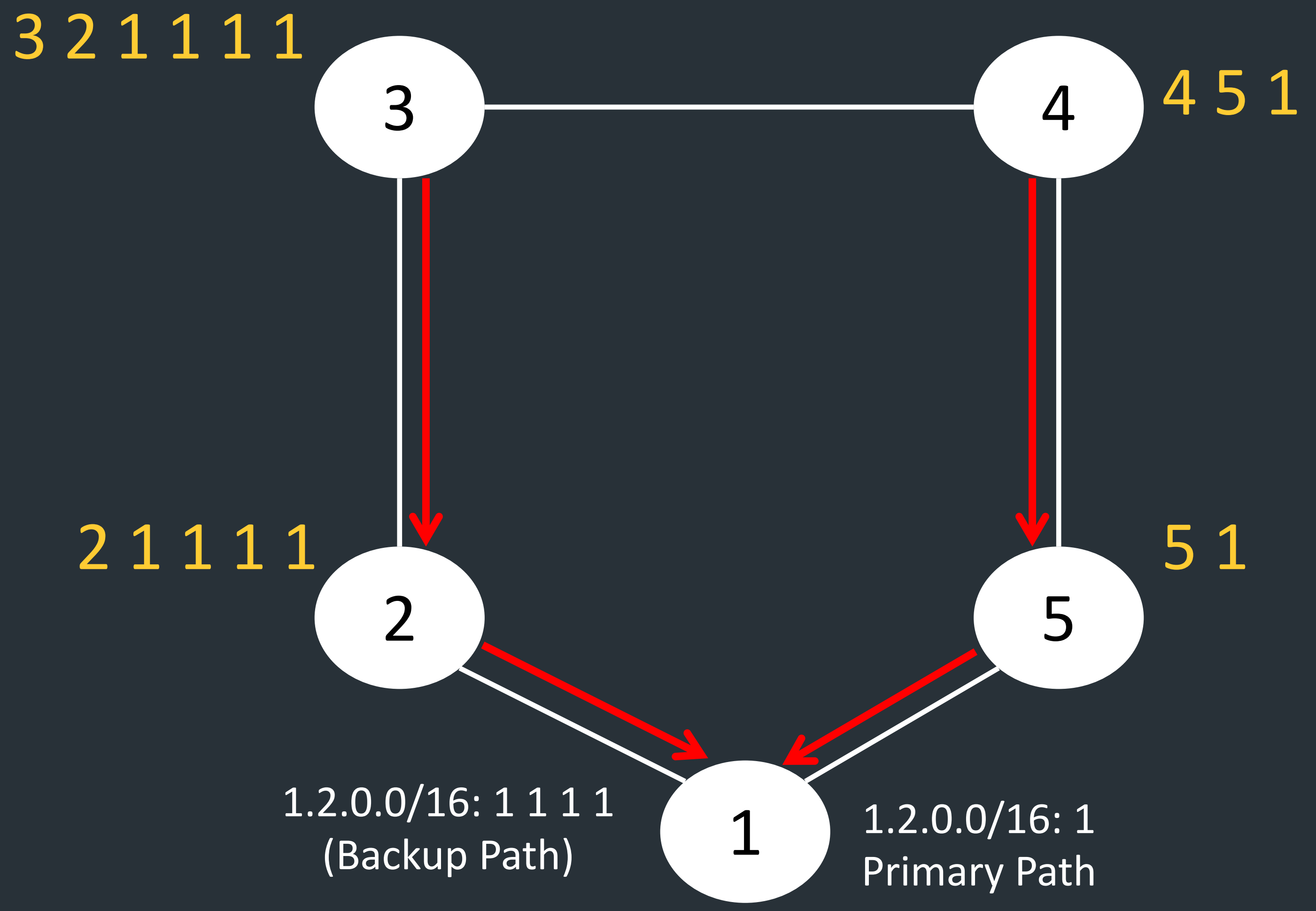

3 4 5 1
3
4
4 5 1
2 3 4 5 1
2
5
5 1
1.2.0.0/16: 1 1 1 1
(Backup Path)
1
1.2.0.0/16: 1
Primary Path

BGP Wedgies


3 2 1 1 1 1
3
4
4 3 2 1 1 1 1
2 1 1 1 1
2
5
5 4 3 2 1 1 1 1
1
1.2.0.0/16: 1 1 1 1
(Backup Path)
1.2.0.0/16: 1
Primary Path

BGP Wedgies

3 prefers customer route: stable configuration!



Warmup for discussion

Given this routing table, to which prefix would a router map each IP?

- 1.2.3.4
- 138.16.100.5
- 138.16.10.200
- 12.34.5.120
- 12.34.18.5

Prefix	Next Hop
`1.0.0.0/8`	...
`12.34.0.0/16`	...
`12.34.16.0/20`	...
`138.16.0.0/16`	...
`138.16.100.0/24`	...

Longest Prefix Match

When performing a forwarding table lookup, select the most specific prefix that matches an address

- Eg. 12.34.18.5

Prefix	Next Hop
1.0.0.0/8	…
12.34.0.0/16	…
12.34.16.0/20	…
138.16.0.0/16	…
138.16.100.0/24	…

Internet routers have specialized memory called TCAM (Ternary Content Addressable Memory) to do longest prefix match *fast* (one clock cycle!)
Goal: forward at *line rate* (as fast as link allows)