CSCI-1680
Network Layer:
Inter-domain Routing

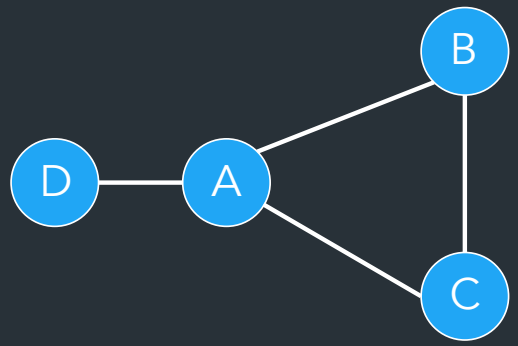
Nick DeMarinis

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…

- (1) When using standard RIP?
- (2) When using split horizon + poison reverse?

(1) (A, 1)
(C, 1)
(D, 2)

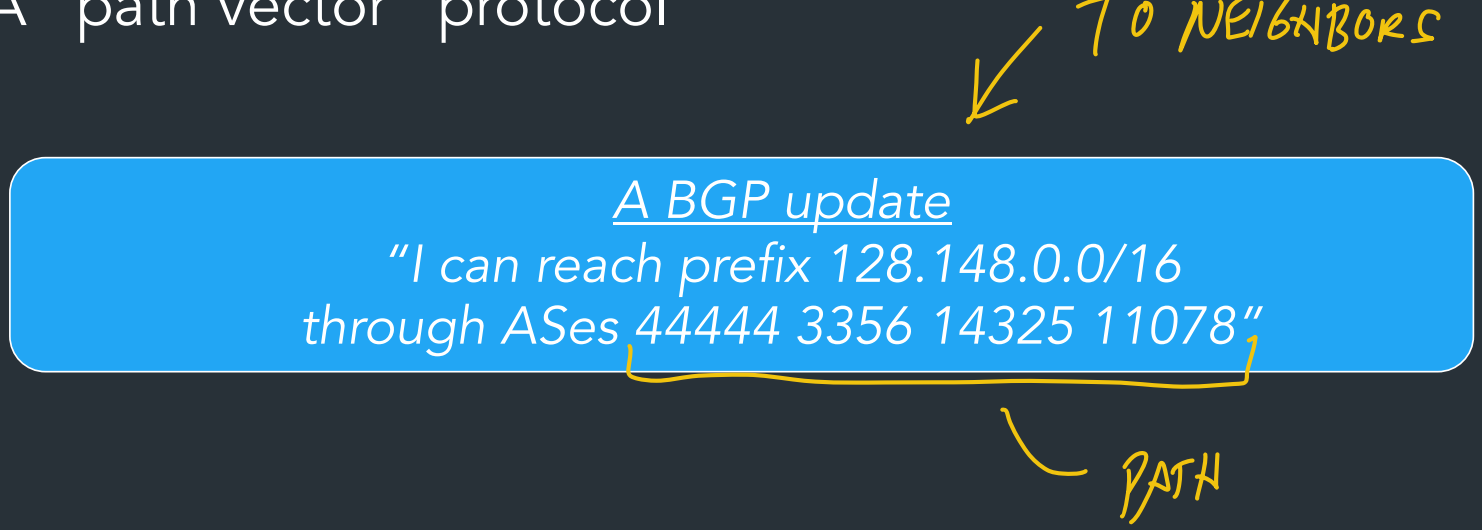
(2) (A, ∞)
(C, 1)
(D, ∞)

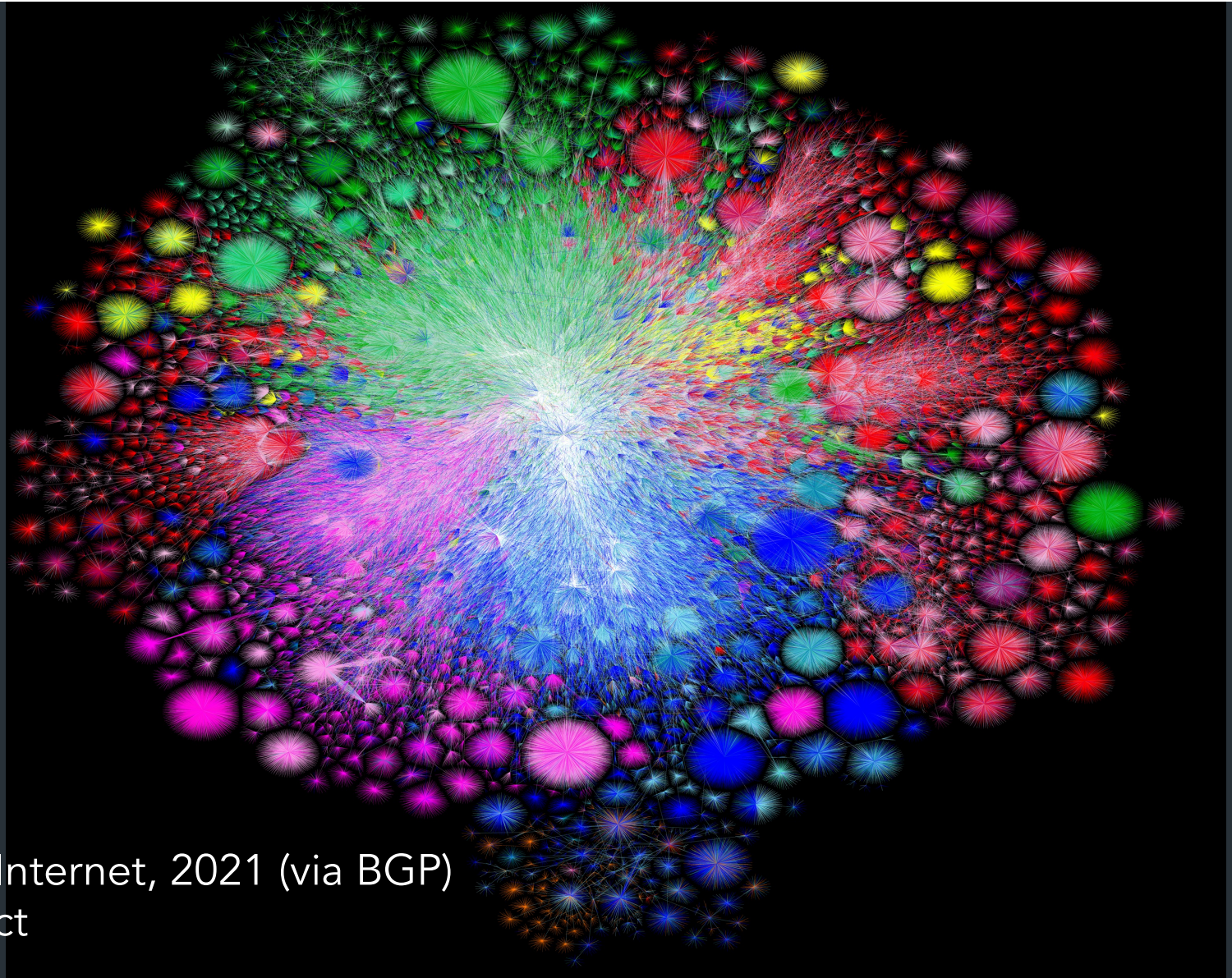
Recall: BGP

Exterior routing: between Autonomous Systems (ASes)

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

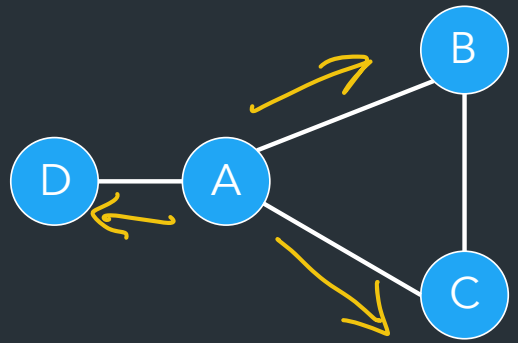
=> A "path vector" protocol

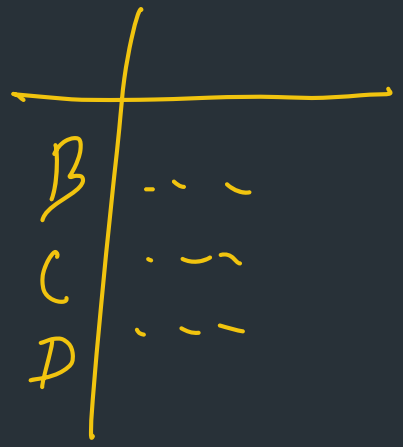




Map of the Internet, 2021 (via BGP)
OPTE project

Before: Interior routing

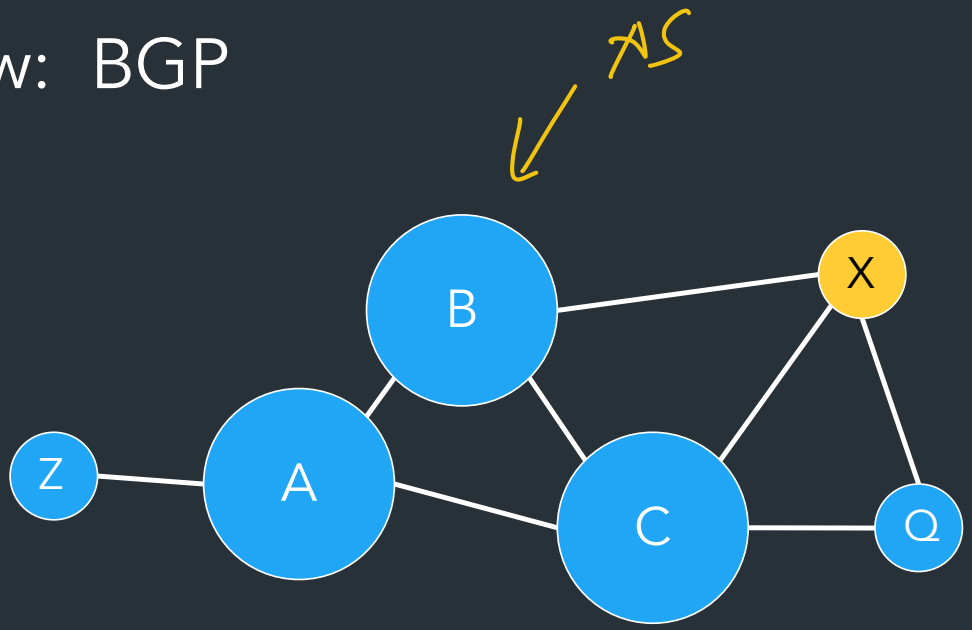




All nodes advertise their routes to all other nodes:

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

Now: BGP



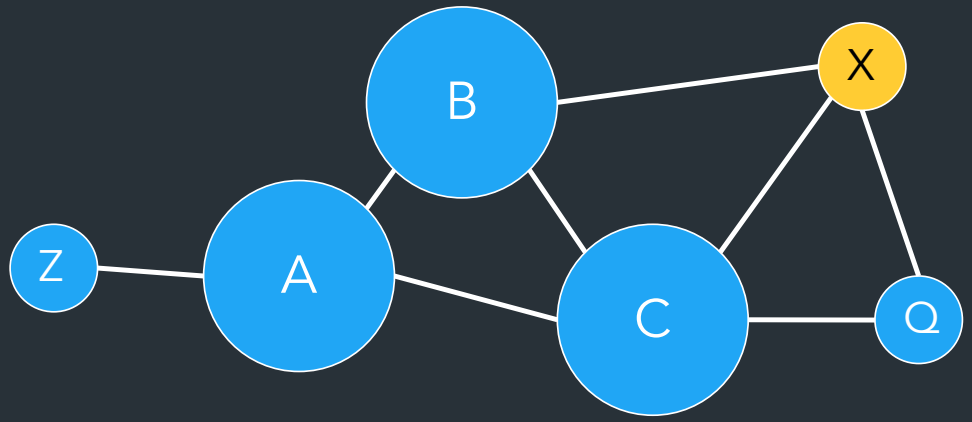
X's table (subset):

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

"Origin": prefixes assigned to X that it wants to advertise to the Internet

"X originates prefix 1.0.0.0/8"

Now: BGP



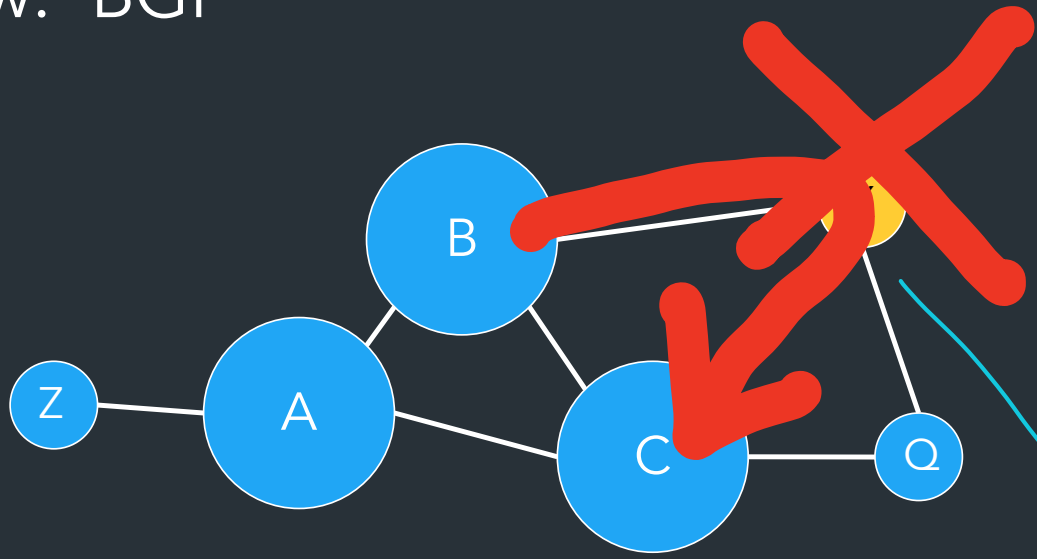
X's table (subset):

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?

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.

Difference between:
- What routes you add to YOUR forwarding table
- What routes you tell your neighbors about

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

If X tells B it has a route to C, B will start sending traffic to X to get to C! If B is a big network, this probably isn't what we want...

Key policy questions

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

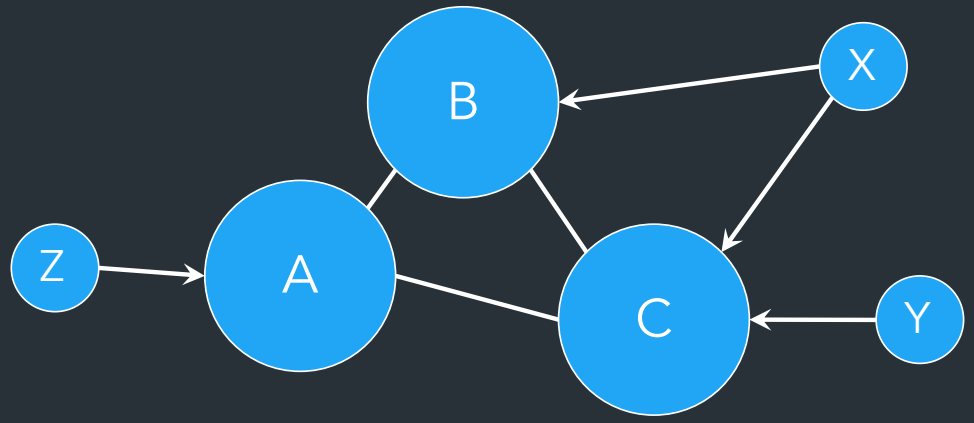
"How to use route info to update forwarding tables?"

⇒ "SELECTION POLICY"

≠

"What routing info to send to neighbors?"

⇒ "EXPORT POLICY"



Relationships between AS drive policy:

- Customer: Customer pays provider to advertise its routes

⇒Y pays C

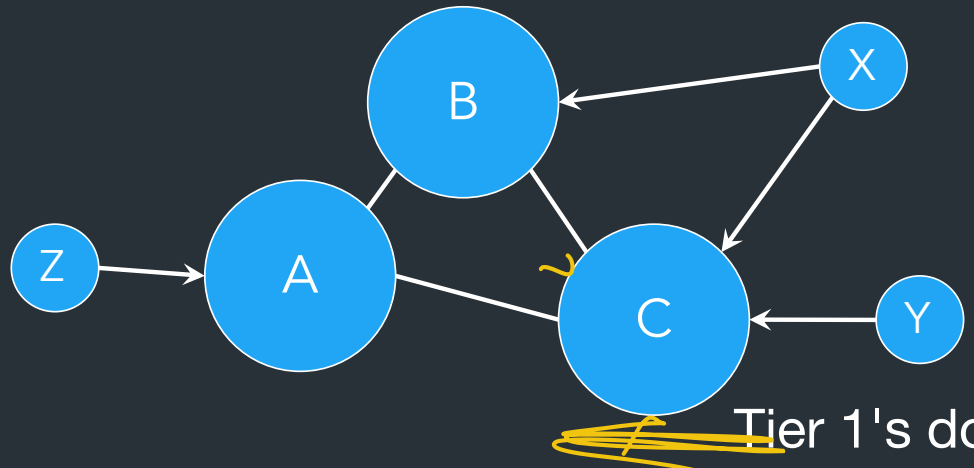
⇒X pays B, C (multihomed)

⇒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



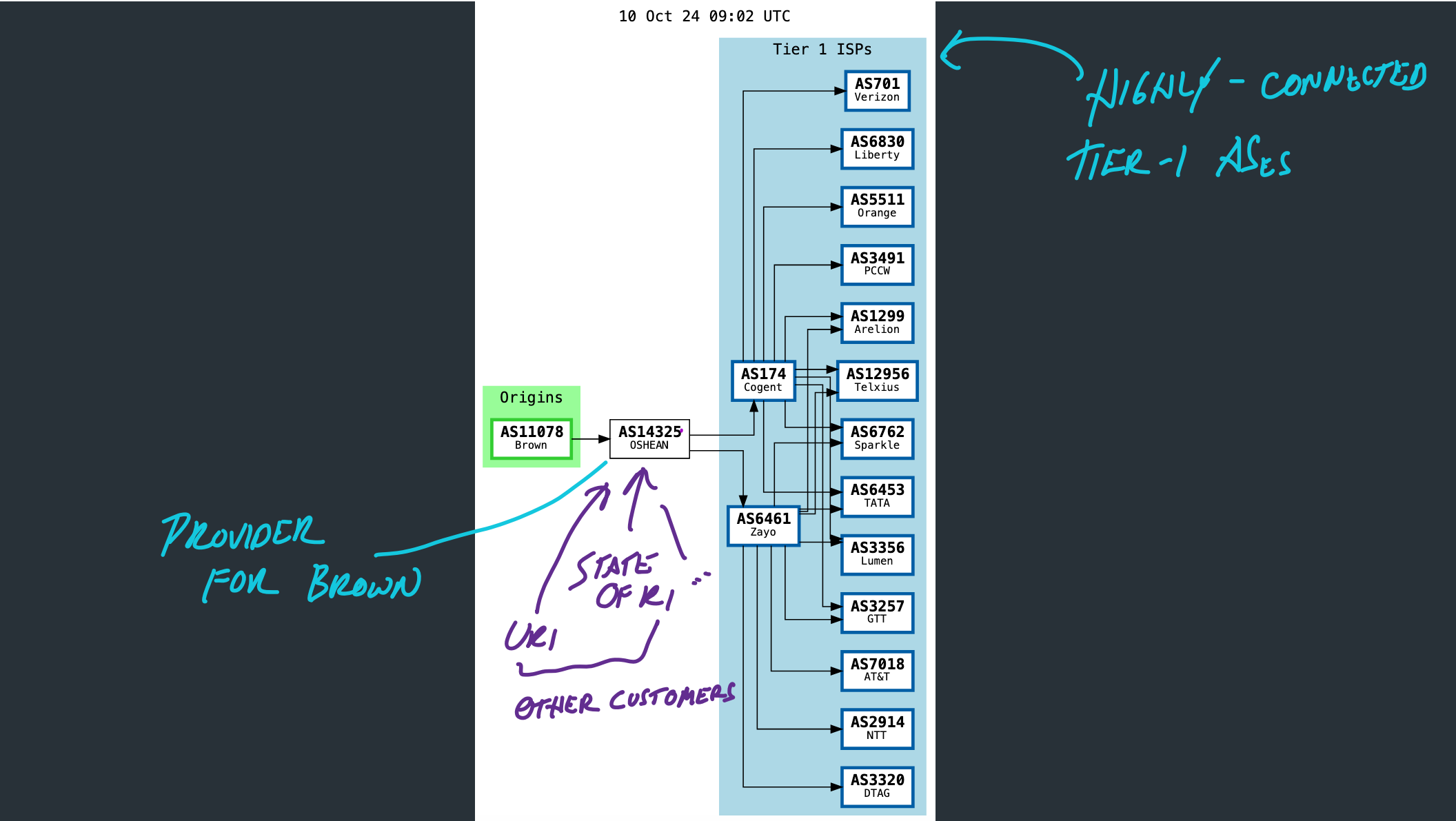
Tier 1's don't charge each other for traffic, because traffic between each other is equal

- 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

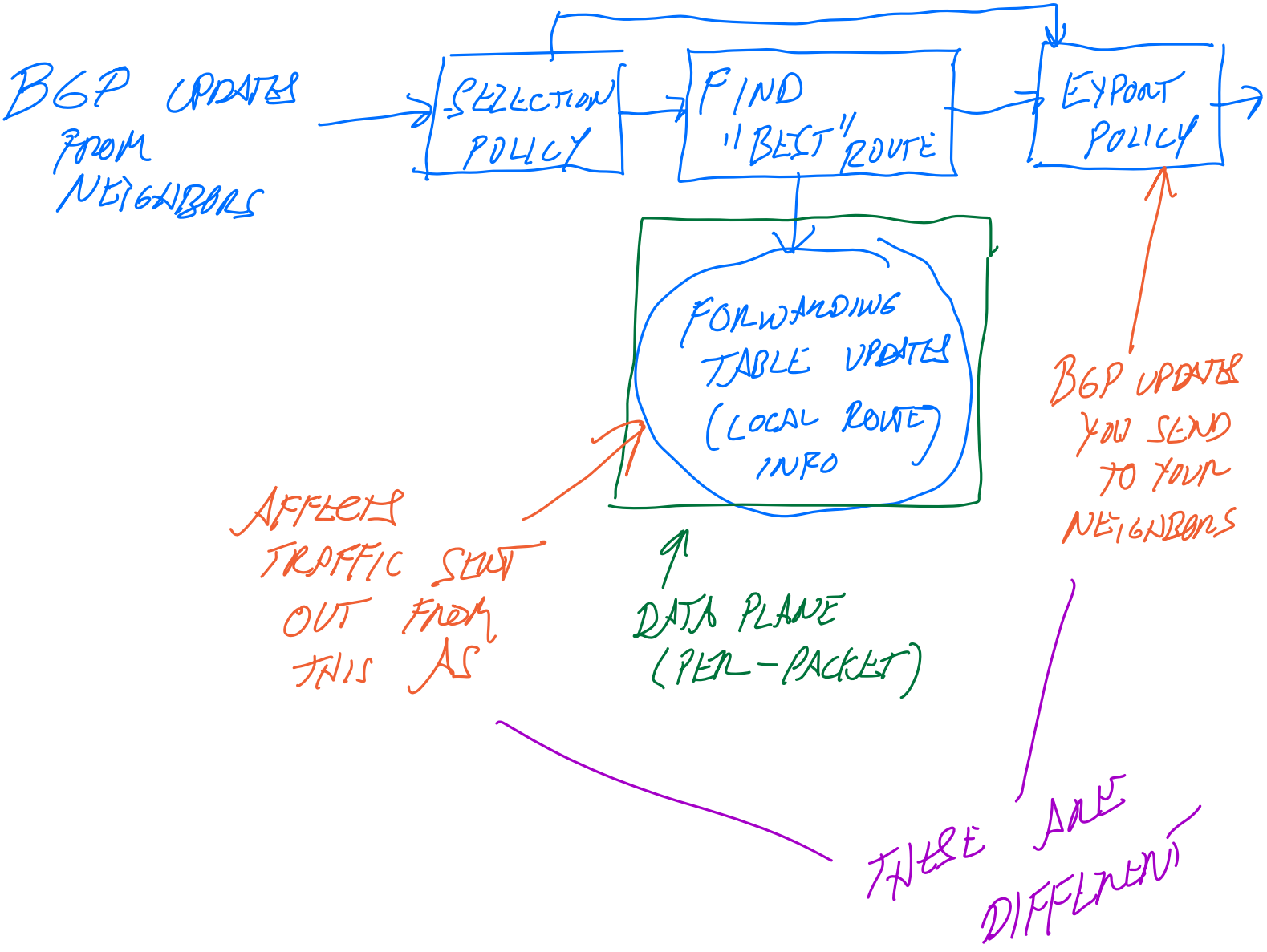
=> A peers with B
=> A peers with C
. . .

Example from Kurose and Ross, 5th Ed


10 Oct 24 09:02 UTC
Tier 1 ISPs
AS701 Verizon
AS6830 Liberty
AS5511 Orange
AS3491 PCCW
AS1299 Arelion
AS12956 Telxius
AS174 Cogent
AS6762 Sparkle
AS6453 TATA
AS6461 Zayo
AS3356 Lumen
AS3257 GTT
AS7018 AT&T
AS2914 NTT
AS3320 DTAG
Origins
AS11078 Brown
AS14325 OSHEAN
HIGHLY - CONNECTED TIER-1 ASES
PROVIDER FOR BROWN
STATE OF R1 ...
R1
OTHER CUSTOMERS

NOW TO THINK ABOUT POLICIES:

⇒ CONTROL PLANE:



Update processing

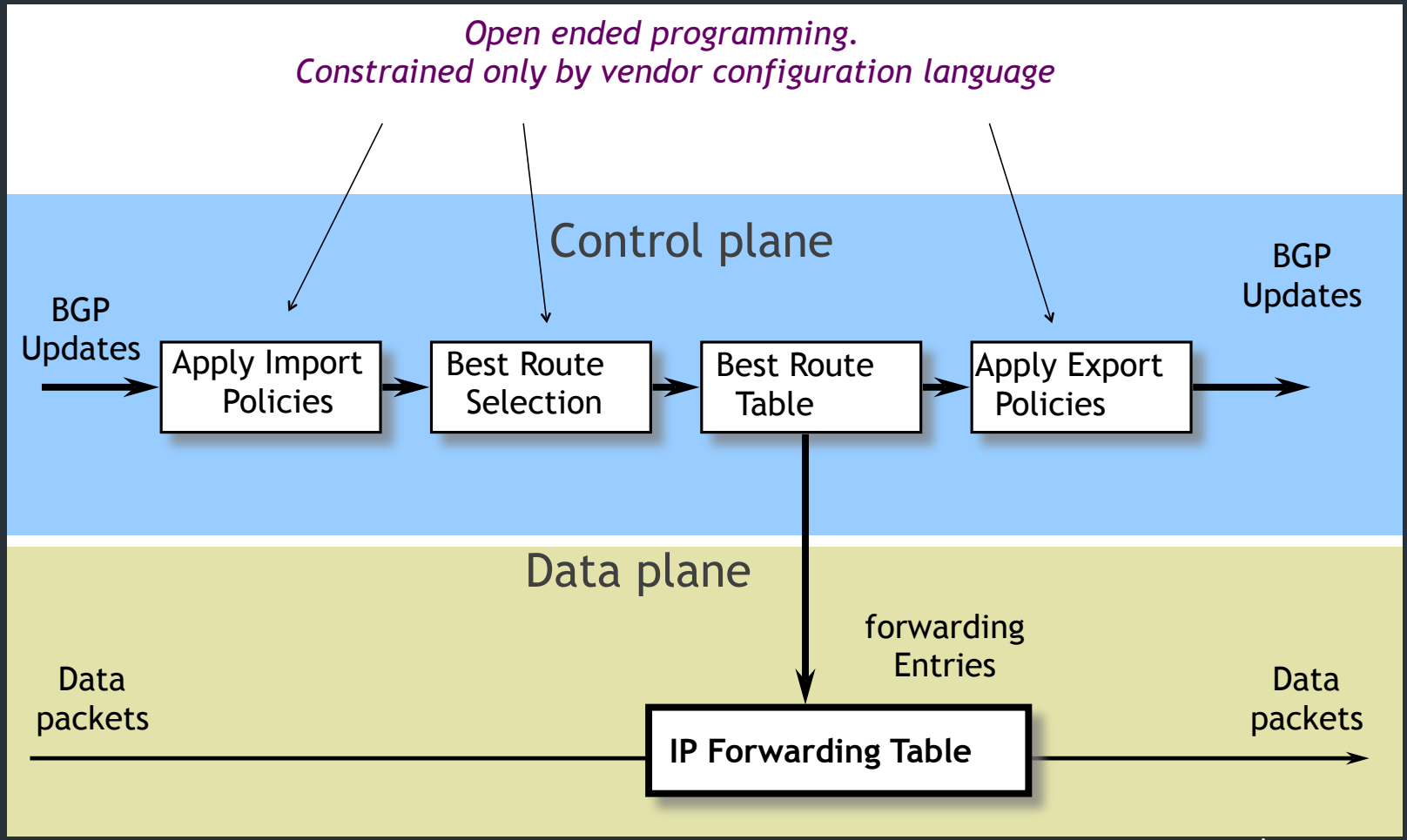
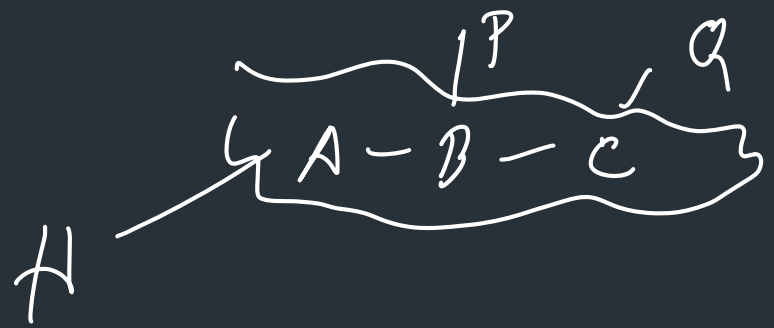


Image credit Rachit Agarwal

Typical route selection policy

In decreasing priority order:

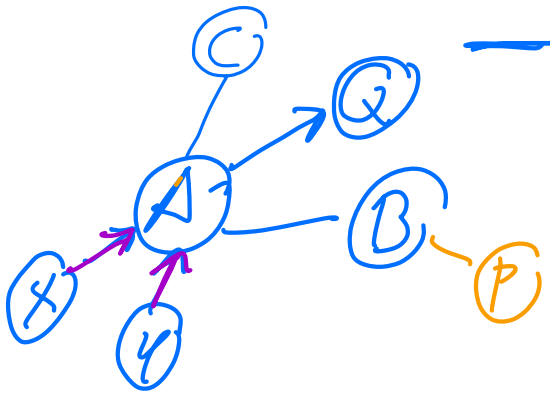
1. Make or save money (send to customer > peer > provider)
2. Try to maximize performance (smallest AS path length)
3. Minimize use of my network bandwidth ("hot potato routing"
4. …



How to Think About Export Policies

Gao-Rexford Principles

Given: ISP A has:
- Customers: X, Y
- Peer with B, C
- Customer of Q



→ Customer

— Peer

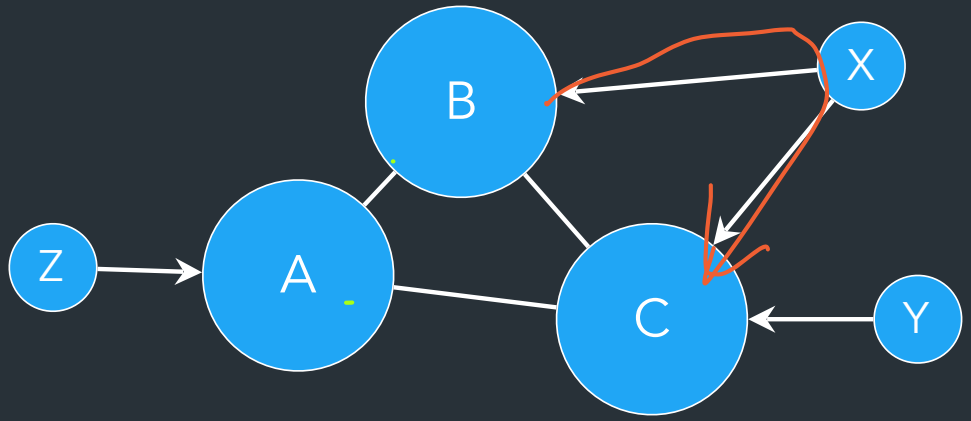
If prefix is advertised by...	Export prefix to...
Customer (eg. X, Y)	Everyone! (X, Y, C, B, Q)
Peer (eg. B)	Customers only (X, Y) (not C, Q)
Provider (Q)	Customers only (X, Y)

Goal: don't become transit if no gain!

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?

X NEVER TELLS B ABOUT C (OR VICE VERSA)

How to avoid transit between CBA ?

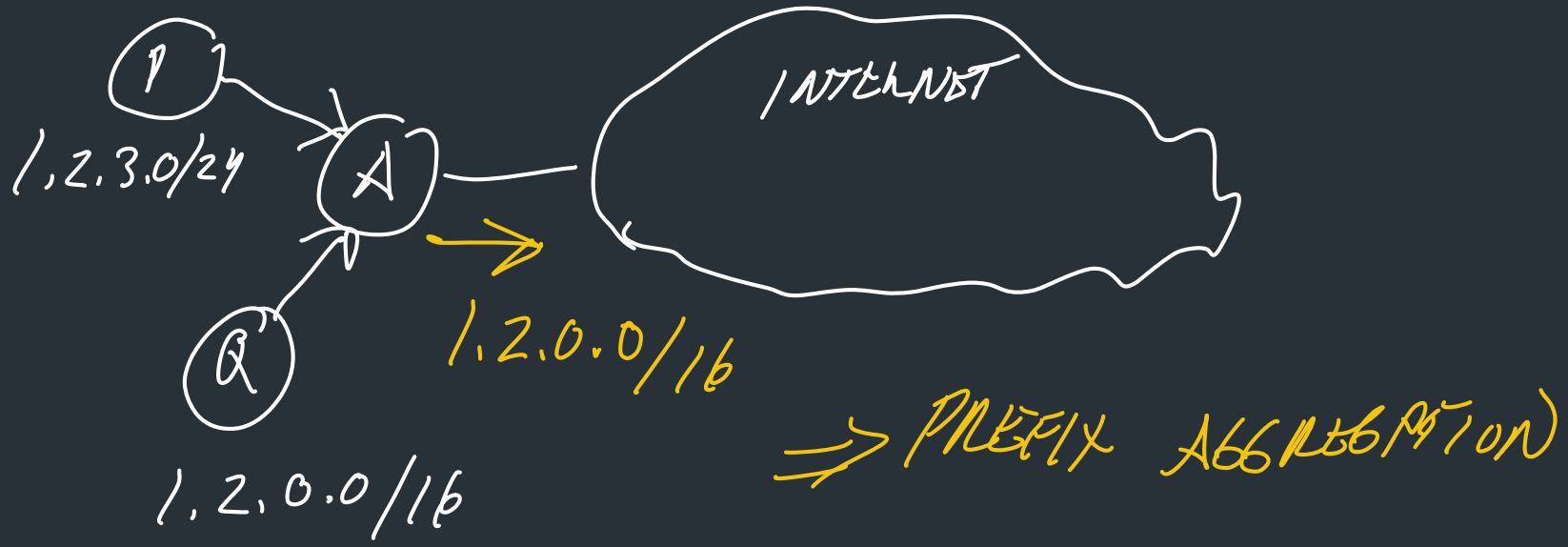
B NEVER TELLS A ABOUT C

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



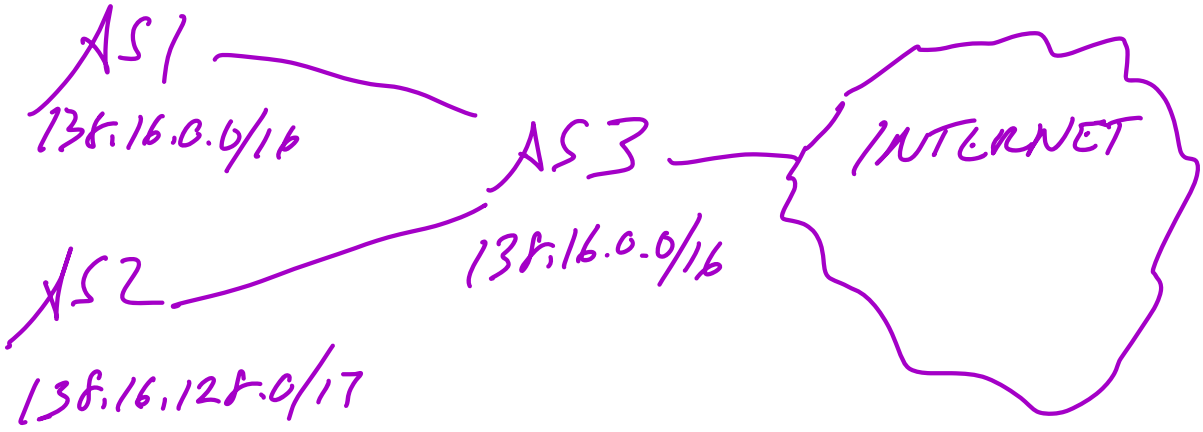
IP PREFIXES / ROUTE AGGREGATION

138.16.0.0/16 138.16.X.X

IDEA: ALLOCATE SMALLER NETWORKS FROM ONE PREFIX

EX. COULD DIVIDE INTO TWO NETWORKS

① 138.16.0.0/17 0000 0000
② 138.16.128.0/17 1000 0000



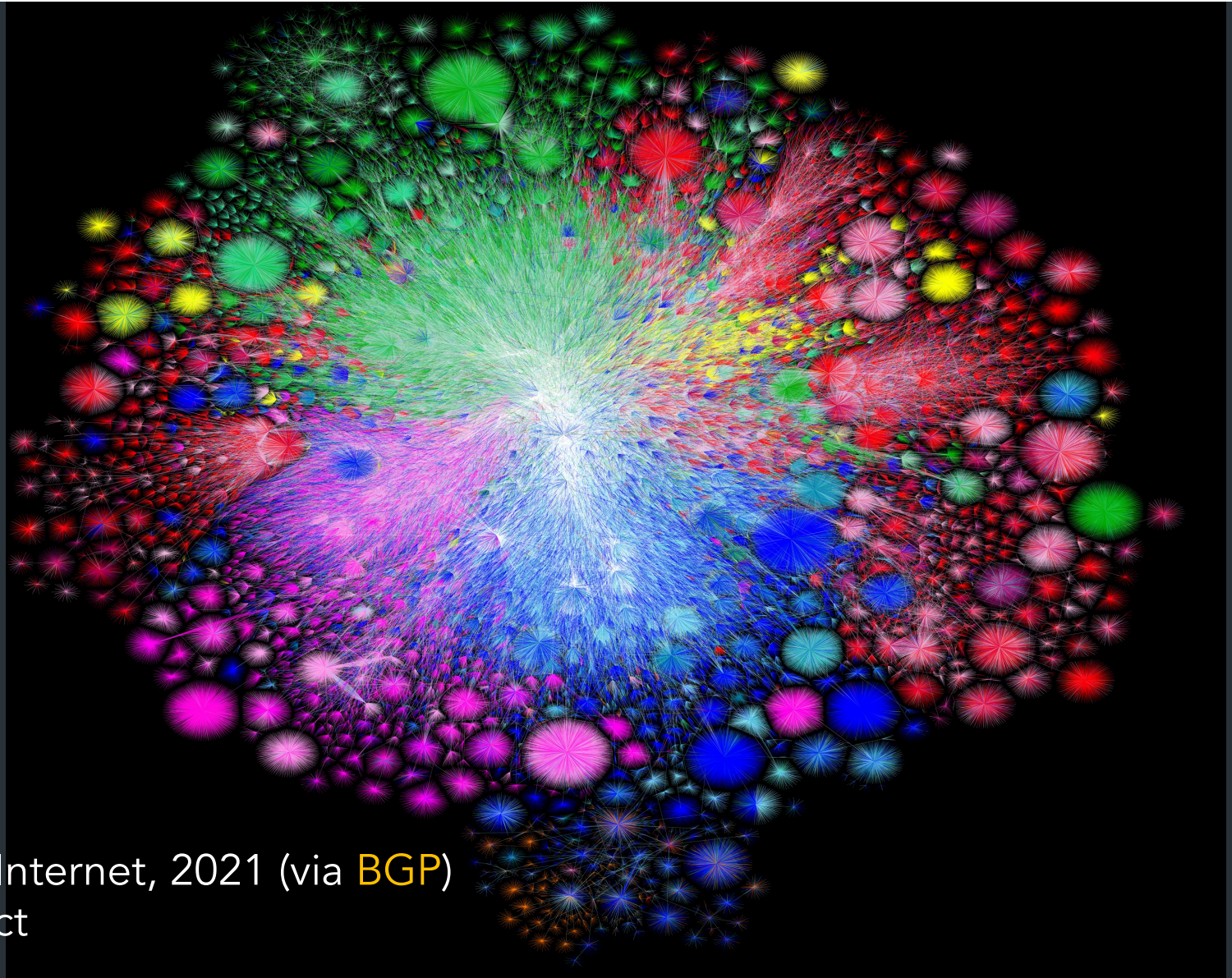
IDEA: AS3 COMBINES, OR AGGREGATES PREFIXES FOR ITS CUSTOMERS
⇒ LEVERAGE HIERARCHY OF ADDRESSES!

HOWEVER, NOT SO EASY IN PRACTICE...

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

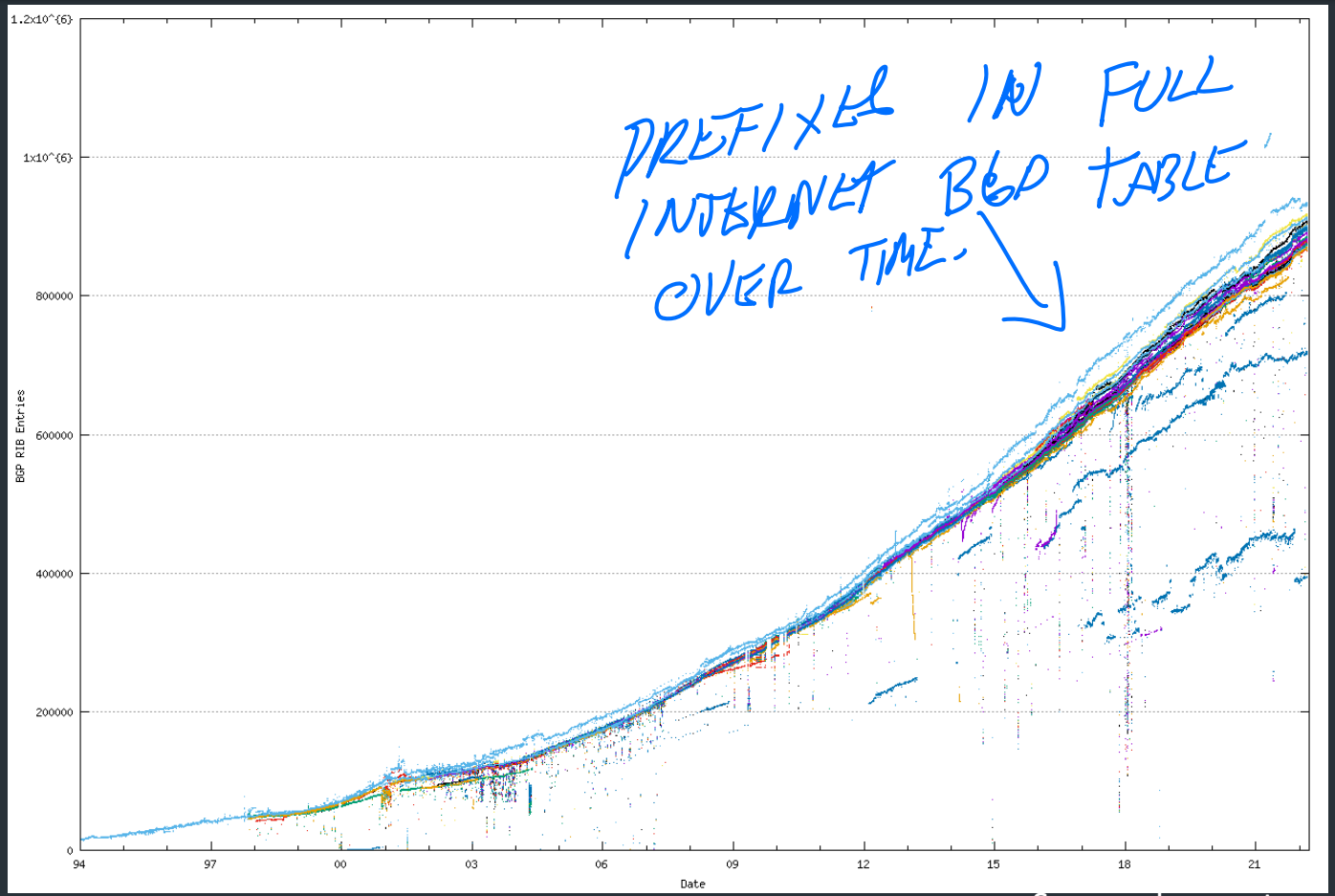


Map of the Internet, 2021 (via BGP)
OPTE project

What can lead to table growth?

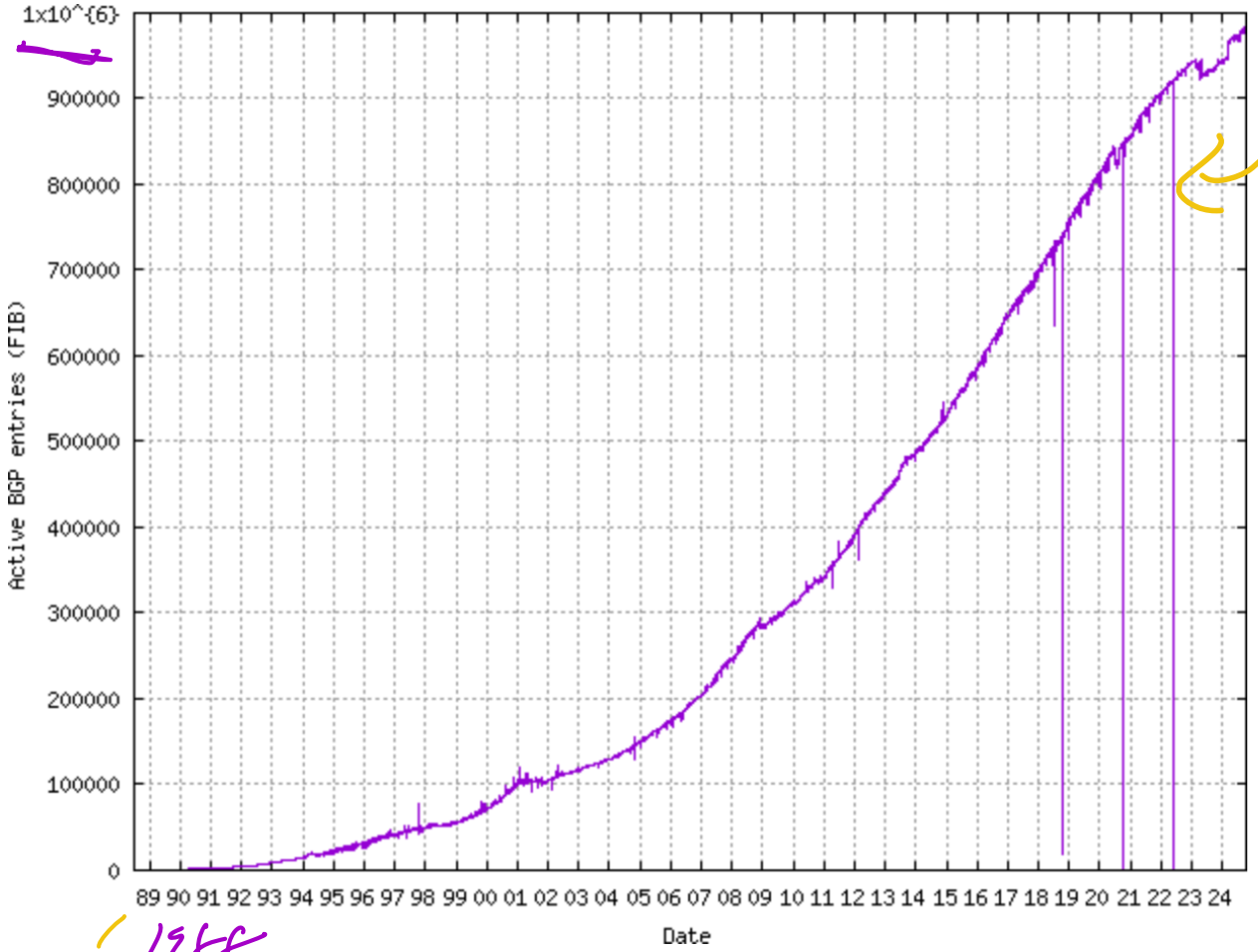
- More addresses being allocated
- Fragmentation
  - Multihoming
  - Change of ISPs
  - Address re-selling

BGP Table Growth



Source: bgp.potaroo.net

Active BGP entries (FIB)



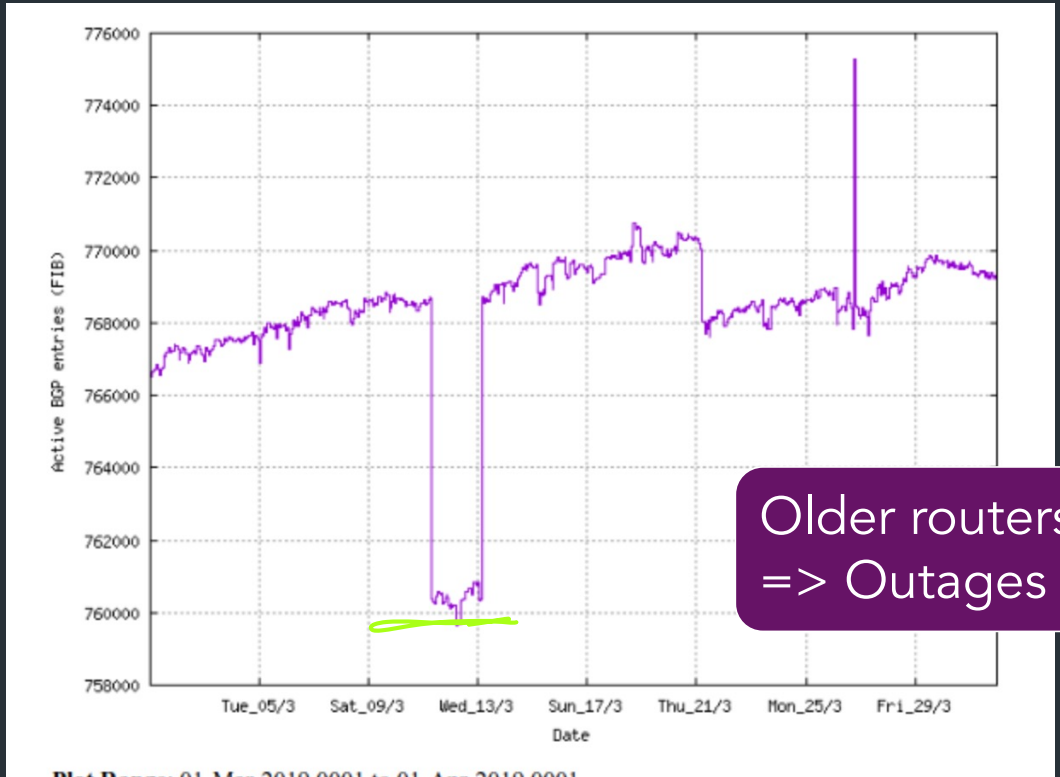
PREFIXES IN FULL-INTERNET BGP TABLE

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



Older routers run out of space => Outages

BGP can be fragile!

- Individual router configurations and policy can affect whole network
- Consequences sometimes disastrous…

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

Who owns a prefix?

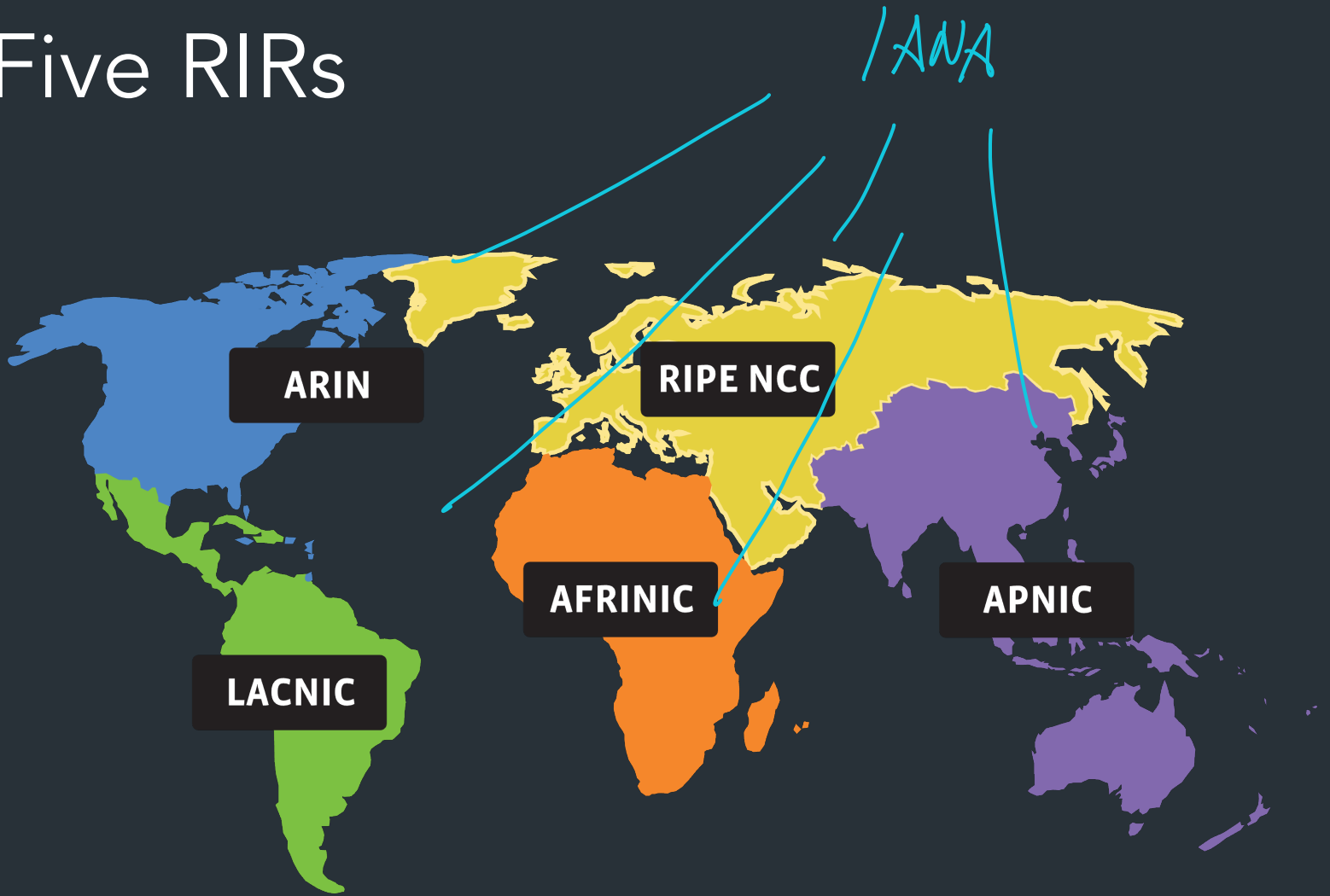
"I own 1.2.3.0/24"

- 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

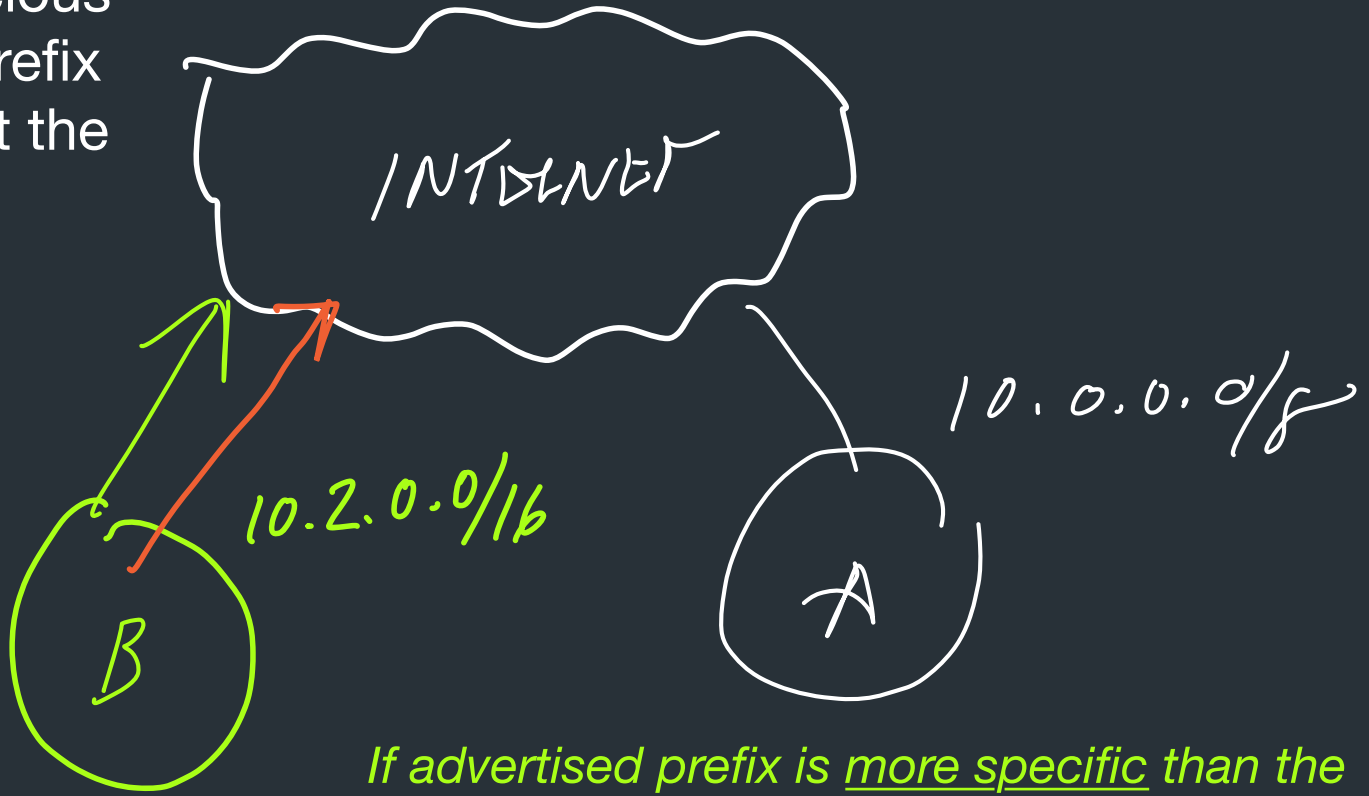
No built-in way to verify ownership, but modern standards like RPKI offer some hope (more on this later)

The Five RIRs



What can go wrong?

Prefix hijacking: malicious router can advertise prefix it does not own => get the traffic for that prefix



If advertised prefix is more specific than the original, other routers will prefer the more specific prefix!

Prefix Hijacking

- Problem: Who "owns" a prefix? Who is allowed to *originate* a prefix?
  => BGP by default **does not verify** announce messages match the network that owns them.
  => ASes have their own security polices (and they are being more widely adopted), but they are not unified

If you can hijack a prefix, what can you do?
- Intercept or redirect packets for some IP range
- Snooping
- Modify/slow down traffic

=> Prefix is hard to debug, because it may only be visible from certain parts of a network. (Though this is easier to see for companies that have visibility from very large networks.)

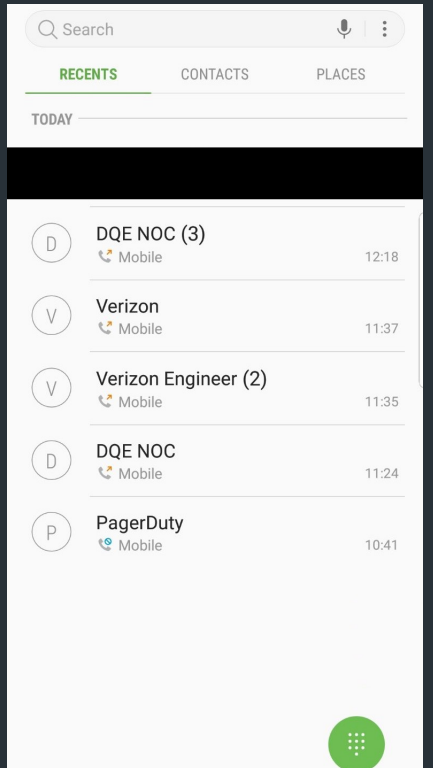
Some Notable incidents

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,



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

Facebook DNS outage

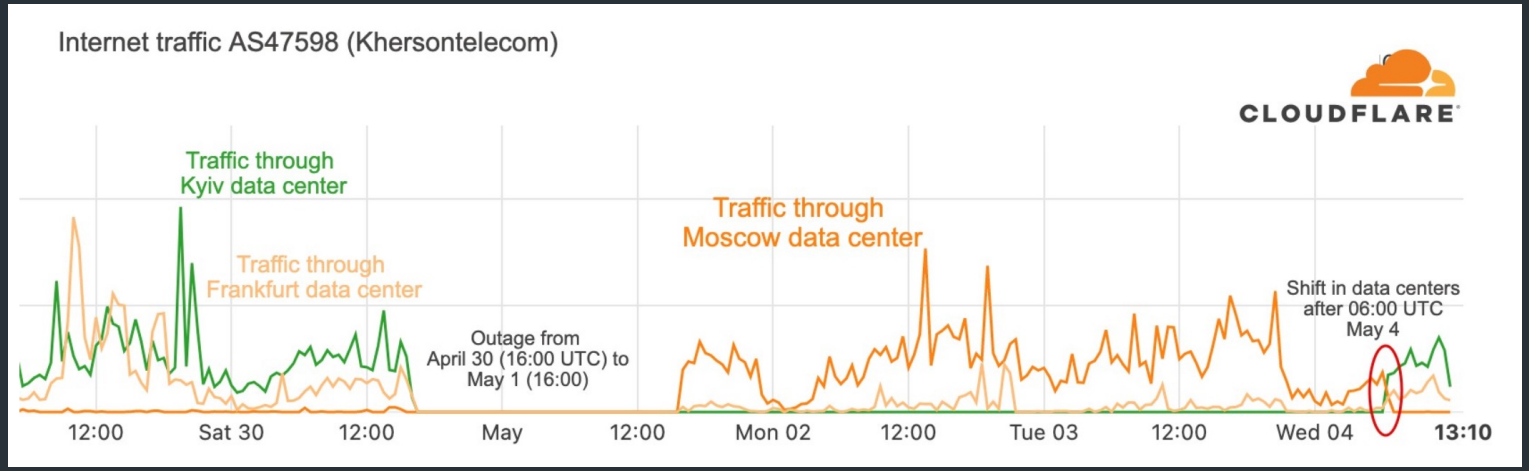
- October 2021: Misconfiguration causes Facebook to withdraw routes for its DNS servers
- DNS: core service that translates domain names to IPs

  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/

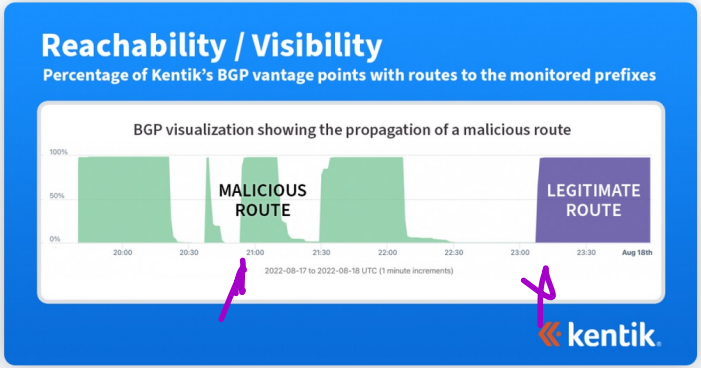
Prefix Hijacking in the wild

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



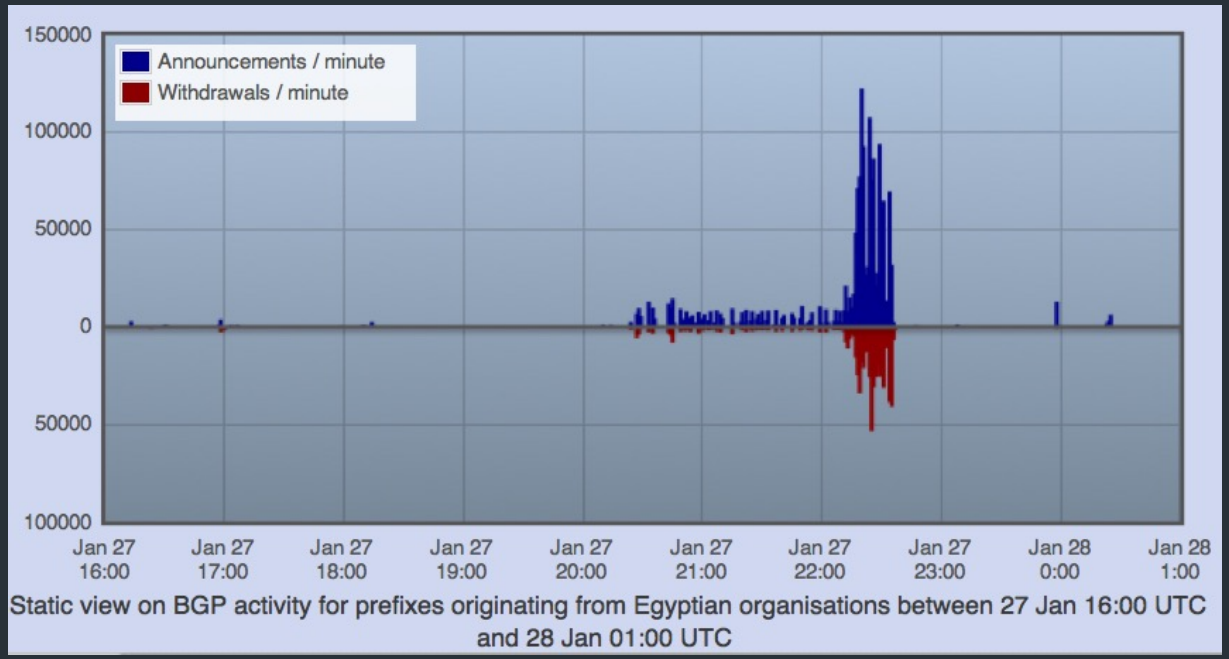
Writeup (more)

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!

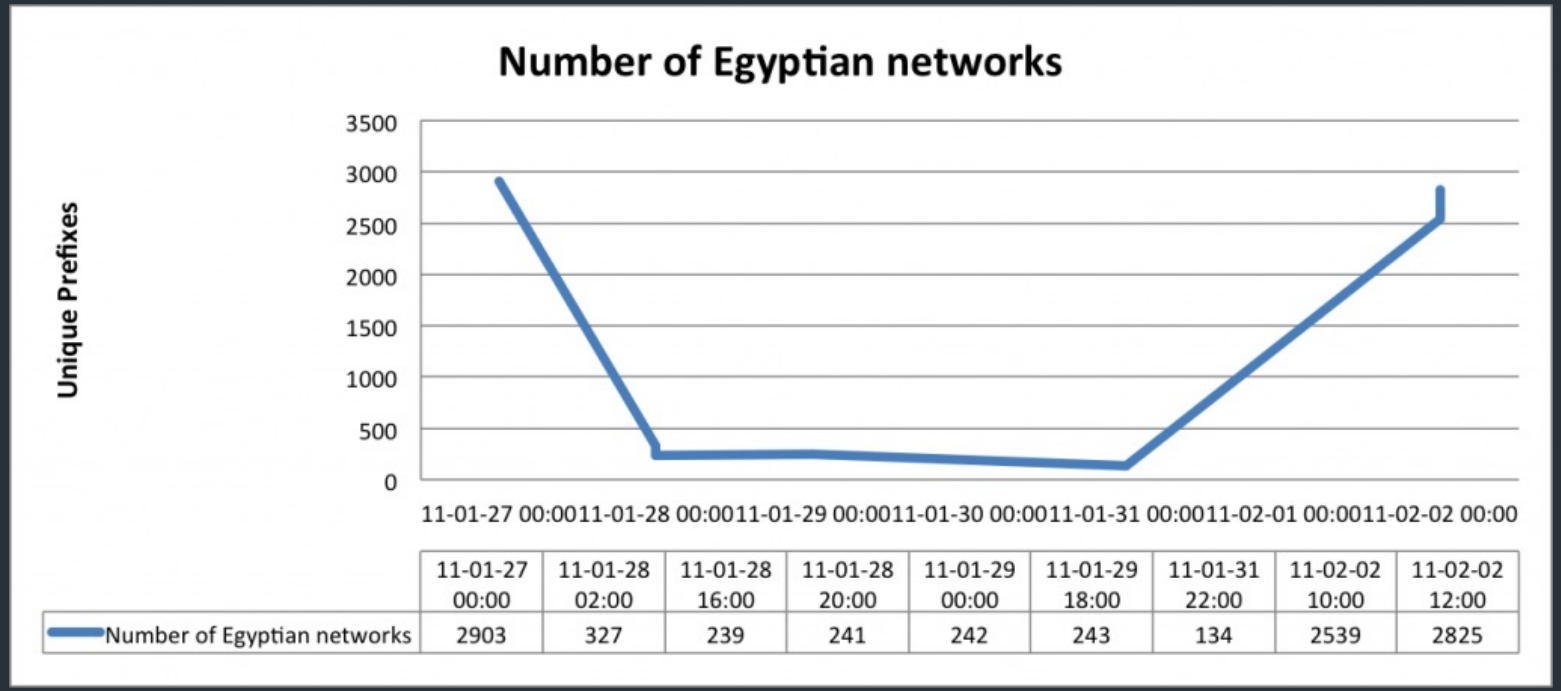
"Shutting off" the Internet

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



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)

— EXTRA CONTENT
WE DID NOT
COVER

What can be done?

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

SET OF ALLOWED PREFIXES.

```
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

- 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 CHECK ADVERTISEMENTS BEFORE INSTALLING THEM.

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

What can be done?

Brown's ISP

```
$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
```

CAN CONTAIN SOME INFO ON THIS AS'S POLICY.

IN THEORY, SHOULD REFLECT HOW BGP ANNOUNCEMENTS ARE SENT.

Proposed Solution: RPKI

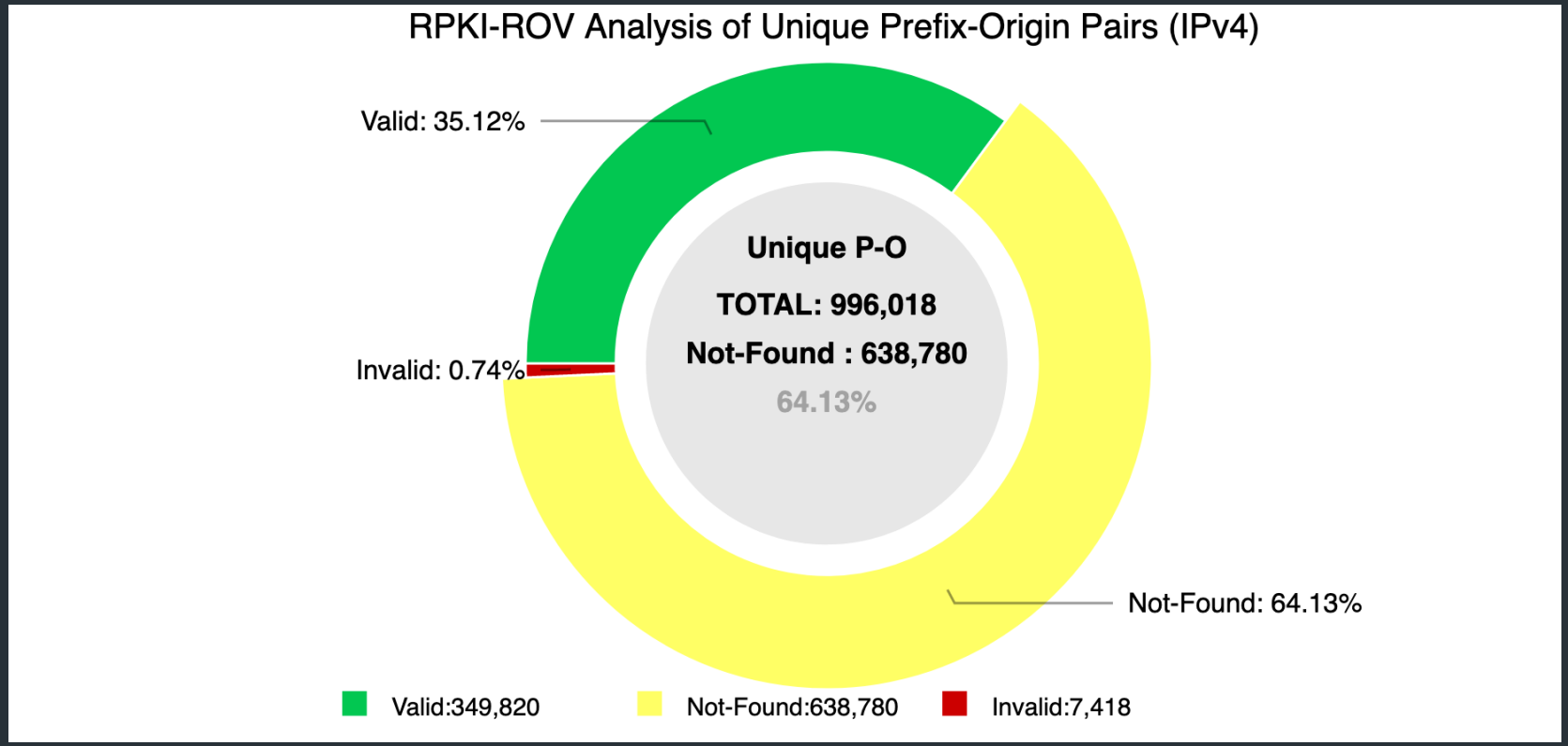
- Based on a public key infrastructure
- Address attestations
  - Claims the right to originate a prefix
  - Signed and distributed out of band, checked on BGP updates
  - Checked through delegation chain from ICANN
- Can avoid
  - Prefix hijacking
  - Addition, removal, or reordering of intermediate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

① EVERY AS ADDS A SIGNATURE OF ROUTE INFO TO DB,
- MAX PREFIX SIZE.
- PREVENT OTHERS FROM ADVERTISING A MORE SPECIFIC PREFIX.

② ASES ACCEPTING ROUTES SUPPOSED TO VALIDATE AGAINST DB

⇒ CAN WORK, IF EVERYONE COOPERATES.

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

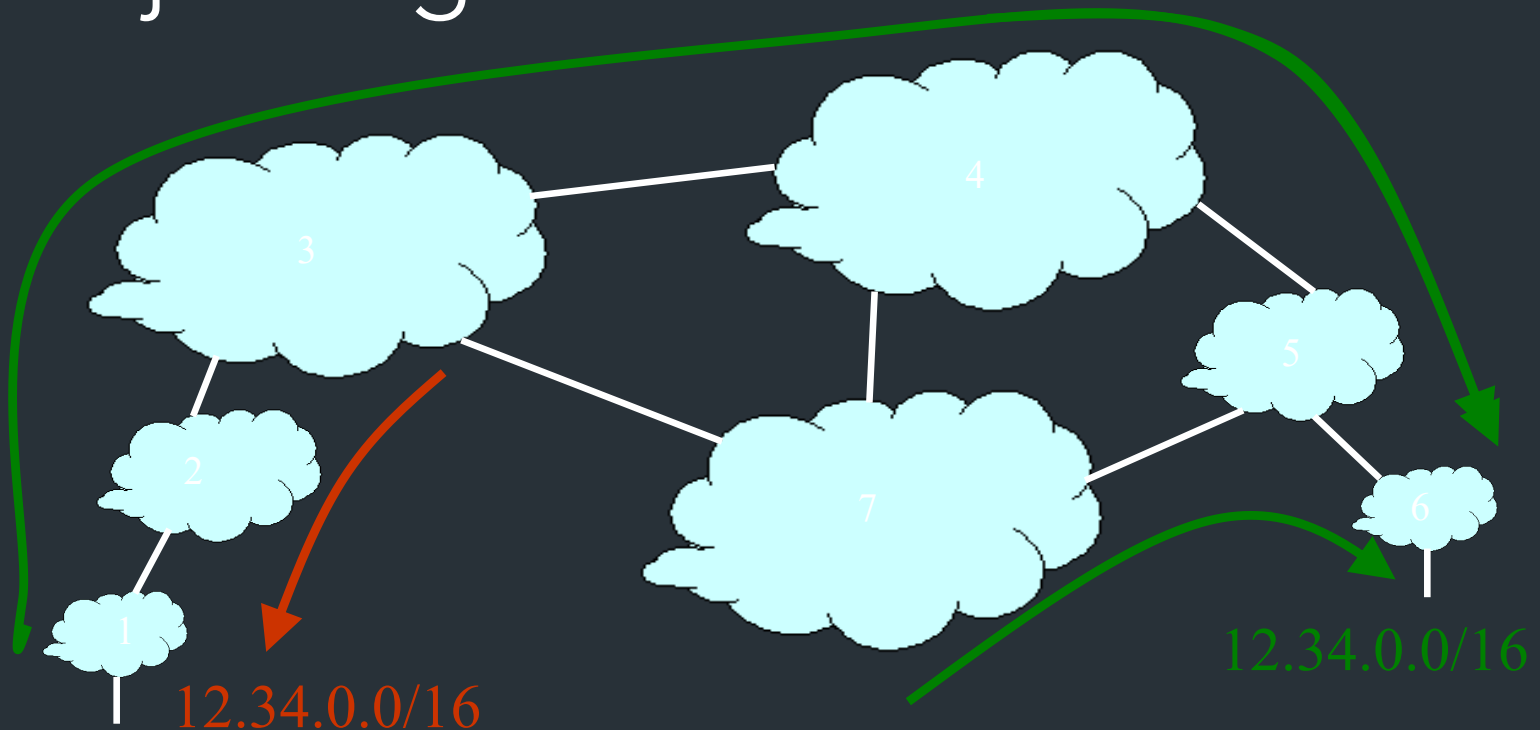
BGP Security Goals

- 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

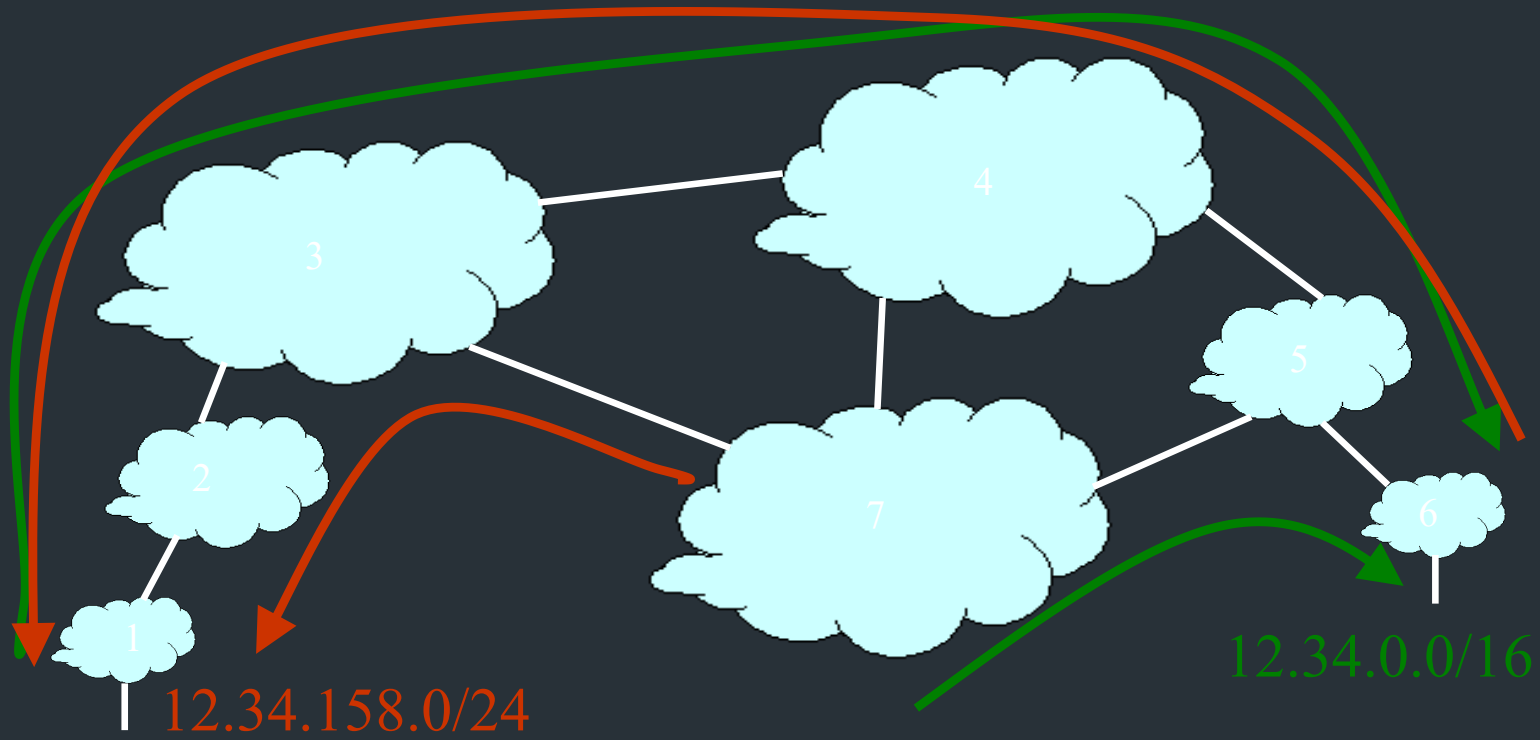


- Consequences for the affected ASes
  - Blackhole: data traffic is discarded
  - Snooping: data traffic is inspected, and then redirected
  - Impersonation: data traffic is sent to bogus destinations

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

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!