Inter-Domain Routing and Policy

Richard T. B. Ma
School of Computing
National University of Singapore

CS 4226: Internet Architecture
Inter-Domain Routing

- Internet is a “network of networks”
- Roughly a hierarchy of Autonomous Systems
  - large, tier-1 provider with a nationwide backbone and international connections
  - medium-sized regional provider with smaller backbone
  - small network run by a single company or university
- How do ASes interact with each other?
What is an Autonomous System?

- A network of interconnected routers
- Identified by a globally unique AS Number (ASN)
- Controlled by a single administrative domain (a company can have several ASNs)
- Use common routing protocol and policy
Example: Singtel

<table>
<thead>
<tr>
<th>Company Name</th>
<th>ASN</th>
<th>General Policy</th>
</tr>
</thead>
<tbody>
<tr>
<td>HousingCenter</td>
<td>28707</td>
<td>Open</td>
</tr>
<tr>
<td>Institute of Molecular and Cell Biology, Singapore</td>
<td>38181</td>
<td>Open</td>
</tr>
<tr>
<td>Level3 formerly Global Crossing</td>
<td>3549</td>
<td>Restrictive</td>
</tr>
<tr>
<td>NTT Communications - NTT Singapore</td>
<td>17645</td>
<td>Selective</td>
</tr>
<tr>
<td>RisingNet.com</td>
<td>33211</td>
<td>Open</td>
</tr>
<tr>
<td>Singapore Internet Exchange (SGIX)</td>
<td>55518</td>
<td>Open</td>
</tr>
<tr>
<td>Singapore Telecommunications Limited</td>
<td>7473</td>
<td>Selective</td>
</tr>
<tr>
<td>Single Host Internet Brasil</td>
<td>52919</td>
<td>Open</td>
</tr>
</tbody>
</table>
AS Topology: UUNET (AS701)
AS Topology: Viatel, Renater

http://www.viatel.com/
https://www.renater.fr/
Challenges for Inter-domain Routing

- **Scale**
  - millions of routers and 200,000+ prefixes
  - 35,000+ self-operated networks and 50K+ ASes

- **Privacy**
  - ASes don’t want to expose internal topologies or their business relationships with neighbors

- **Policy**
  - no Internet-wide notion of a link cost metric
  - need control over where you send traffic and who can send traffic through you
Two classes of routing algorithm

Link state algorithm
- all routers have complete topology, link cost info
- Global or centralized
  - Dijsktra’s algorithm
- Open Shortest Path First (OSPF)

Distance vector algorithm
- router knows connected neighbors, link costs
- iterative process of computation, exchange of info with neighbors
- Decentralized algorithm
  - Bellman-Ford algorithm
- Routing Information Protocol (RIP)
Limitation of Link-State Routing

- Topology information is flooded
  - high bandwidth and storage overhead
  - nodes divulge sensitive information

- Entire path computed locally per node
  - high processing overhead in a large network

- Minimize some notion of total distance
  - works only if policy is shared and uniform
Distance Vector (DV) approach

- **Advantages**
  - hide details of the network topology
  - only next hop is determined per node

- **Disadvantages**
  - minimizes some notion of total distance, which is difficult in an inter-domain setting
  - slow convergence due to the counting-to-infinity problem

- **Solution:** extend the notion of a DV
Path-Vector Routing

- Extension of distance-vector routing
  - support flexible routing policies
  - avoid count-to-infinity problem

- Key ideas: advertise the entire path
  - DV: send distance metric per destination d
  - PV: send the entire path for each destination d
Faster Loop Detection

- Node can easily detect a loop
  - check if itself is in the path
- Node can simply discard paths with loops
  - e.g., node 1 simply discards the advertisement
Border Gateway Protocol (BGP)

- **BGP**: the de facto inter-domain routing protocol
  - prefix-based path-vector protocol
  - BGP4 described in RFC 4271 (104 pages)
  - RFC 4276 gives an implementation report on BGP
  - RFC 4277 describes operational experiences using BGP
  - enable policy-based routing based on AS Paths

- Allows subnet to advertise its existence to rest of Internet: "I am here"

- Allows ASes to determine “good” routes to other networks based on reachability info and policy
BGP operations

- **BGP session**: two BGP routers (or peers or speakers) exchange messages:
  - advertise *paths* to different destination network prefixes

Diagram:

1. Establish session on TCP port 179
2. Exchange all active routes
3. Exchange incremental updates

While connection ALIVE, exchange route UPDATE messages
BGP/IGP model used in ISPs

- **eBGP**: exchange reachability info from neighbor ASes; implement routing policy
- **iBGP**: propagate reachability info across backbone; carry ISP’s own customer prefixes
**eBGP**

- **external BGP peering (eBGP)**
  - between BGP speakers in different ASes
  - should be directly connected
  - never run an IGP between eBGP peers

- **when AS3 advertises a prefix to AS1:**
  - AS3 *promises* it will forward datagrams towards that prefix
  - AS3 can aggregate prefixes in its advertisement
internal BGP peering (iBGP)

- peers within an AS; not required to be directly connected
  - IGP takes care of inter-BGP speaker connectivity
- iBGP peers must be fully meshed
  - they originate connected networks
  - pass on prefixes learned from outside the AS
  - do not pass on prefixes learned from other iBGP speakers

1c can use iBGP to distribute prefix info to all routers in AS1; 1b can re-advertise info to AS2 over eBGP
BGP messages

- **OPEN**: opens TCP connection to peer and authenticates sender

- **UPDATE**: advertises new paths (or withdraws old paths)

- **KEEPALIVE**: keeps connection alive in absence of UPDATES; also ACKs OPEN request

- **NOTIFICATION**: reports errors in previous messages; also used to close connection
**UPDATE Message Format**

<table>
<thead>
<tr>
<th>Marker (16)</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Length (2)</strong></td>
<td><strong>Type (1)</strong></td>
</tr>
<tr>
<td>Withdrawn Routes Length (2)</td>
<td>Withdrawn Routes (variable)</td>
</tr>
<tr>
<td>Path Attribute Length (2)</td>
<td>Path Attributes (variable)</td>
</tr>
<tr>
<td>Network Layer Reachability Information (variable)</td>
<td></td>
</tr>
</tbody>
</table>

- **Withdrawn Routes**: IP prefixes for the routes withdrawn
- **Can withdraw multiple routes in an UPDATE message**
- **Can only advertise one feasible route for the NLRI**
- **Network Layer Reachability Information (NLRI)**: IP prefixes that can be reached from the advertised route  
  - IP prefixes are coded more compactly (refer to RFC)
Withdrawn Routes

- No expiration timer for the routes like RIP
- Invalidate routes are actively withdrawn by the original advertiser
- Or use UPDATE message to replace the existing routes
- All routes from a peer become invalid when the peer goes down
BGP Path Attributes

- Fall into four separate categories:
  1. well-known mandatory
  2. well-known discretionary
  3. optional transitive
  4. optional non-transitive

- Some implementation rules:
  - must recognize all well-known attributes
  - mandatory attributes must be included in UPDATE messages that contain NLRI
  - once a BGP peer updates well-known attributes, it must pass them to its peers
# Common Path Attributes

<table>
<thead>
<tr>
<th>Attribute Name</th>
<th>Category</th>
</tr>
</thead>
<tbody>
<tr>
<td>ORIGIN</td>
<td>well-known mandatory</td>
</tr>
<tr>
<td>AS_PATH</td>
<td>well-known mandatory</td>
</tr>
<tr>
<td>NEXT_HOP</td>
<td>well-known mandatory</td>
</tr>
<tr>
<td>LOCAL_PREF</td>
<td>well-known discretionary</td>
</tr>
<tr>
<td>ATOMIC_AGGREGATE</td>
<td>well-known discretionary</td>
</tr>
<tr>
<td>AGGREGATOR</td>
<td>optional transitive</td>
</tr>
<tr>
<td>COMMUNITY</td>
<td>optional transitive</td>
</tr>
<tr>
<td>MULTI_EXIT_DISC (MED)</td>
<td>optional non-transitive</td>
</tr>
</tbody>
</table>
Well-Known mandatory attributes

- **ORIGIN:**
  - conveys the origin of the prefix
  - historical attribute used in transition from EGP to BGP

- **AS-PATH:**
  - contains ASes through which NLRI has passed
  - expressed as a sequence, e.g., AS 79, AS 11 ... , or a set

- **NEXT-HOP:**
  - indicates IP address of the router in the next-hop AS.
    (may be multiple links from current AS to next-hop-AS)
How does entry get in forwarding table?

Assume prefix is in another AS.

- Ties together hierarchical routing with BGP and OSPF.
- Provides nice overview of BGP!

**High-level overview**

1. Router becomes aware of IP prefix
2. Router determines the output port for the IP prefix
3. Router enters the prefix-port pair in forwarding table

![Diagram showing the process of adding an entry to the forwarding table](image)
BGP message contains “routes”
route = prefix + attributes: AS-PATH, NEXT-HOP,…
Example: route:
Prefix: 138.16.64/22; AS-PATH: AS3 AS131; NEXT-HOP: 201.44.13.125
Router may receive multiple routes

- Router may receive multiple routes for **same** destination prefix
- The router has to select one route
Select best BGP route to prefix

- Router selects route based on shortest AS-PATH

- Example:
  - AS2 AS17 to 138.16.64/22
  - AS3 AS131 AS201 to 138.16.64/22

- What if there is a tie? will come back to that!
Find best intra-route to BGP route

- Use selected route’s NEXT-HOP attribute
  - Route’s NEXT-HOP attribute is the IP address of the router interface that begins the AS PATH.

- Example:
  - AS-PATH: AS2 AS17; NEXT-HOP: 111.99.86.55

- Router uses OSPF to find shortest path from 1c to 111.99.86.55
Router identifies port for route

- Identifies port along the OSPF shortest path
- Adds prefix-port entry to its forwarding table:
  - (138.16.64/22 , port 4)
Hot Potato Routing

- If there exists two or more best inter-routes
- Then choose route with closest NEXT-HOP
  - use OSPF to determine which gateway is closest
  - Q: From 1c, chose AS3 AS131 or AS2 AS17?
  - A: route AS3 AS131 since it is closer
How does entry get in forwarding table?

Summary

1. Router becomes aware of prefix
   - via BGP route advertisements from other routers

2. Determine router output port for prefix
   - use BGP route selection to find best inter-AS route
   - use OSPF to find best intra-AS route leading to best inter-AS route
   - router identifies router port for that best route

3. Enter prefix-port entry in forwarding table
BGP Policy: how is it used in practice?

- **Objectives:** used by commercial ISPs to
  - fulfill bilateral agreements with other ISPs
  - minimize monetary costs (or maximize revenue)
  - ensure good performance for customers

- **Bilateral agreement between neighbor ISPs**
  - defines who will provide transit for what
  - depends on business relationships
    - Customer-provider relationship
    - Peer-to-peer relationship
Customers and Providers

- Customer pays provider for
  - access to the Internet and reachable from anyone

- Provider provides *transit service* for the customer
Nontransit vs. Transit ASes

- however, customer doesn’t allow traffic go through it
- NET A has two providers, called multi-homing
- traffic should NEVER flows from P1 through NET A to P2
- nontransit AS might be a corporate or campus network, or a “content provider”
Selective Transit

- NET A provides transit between B & C and C & D
- NET A DOES NOT provide transit Between D & B
- Most transit networks transit in a selective manner...
Customers Don’t Always Need BGP

- Static routing is the most common way of connecting an autonomous routing domain to the Internet.

**Diagram:**
- **Provider:**
  - Set routes 192.0.2.0/24 pointing to customer.

- **Customer:**
  - Set default routes 0.0.0.0/0 pointing to provider.

- **Network:** 192.0.2.0/24
A multi-home with C and D, one of which is a backup.
The Peer-to-peer Relationship

- Peers provide transit between their respective customers
- don’t provide transit between peers
- often don’t pay each other (the relationship is settlement-free)
Peering Provides Shortcuts

Peering also allows connectivity between the customers of "Tier 1" providers.

peer
provider
customer
# Peering Dilemma

<table>
<thead>
<tr>
<th>To Peer</th>
<th>Not To Peer</th>
</tr>
</thead>
<tbody>
<tr>
<td>reduce upstream transit costs</td>
<td>you would rather have customers</td>
</tr>
<tr>
<td>improve end-to-end performance</td>
<td>peers are usually your competition</td>
</tr>
<tr>
<td>be the only way to connect customers to some part of the Internet (tier-1)</td>
<td>peering relationships may require periodic renegotiation</td>
</tr>
</tbody>
</table>

- Peering struggles are by far the most contentious issues in the ISP world!
- Peering agreements are often confidential.
MCI/Verizon free-peering requirements

Interconnection Requirements

1.1 Geographic Scope. The Requester shall operate facilities capable of terminating IP customer leased line connections onto a device in at least 50% of the geographic region in which the Verizon Business Internet Network with which it desires to interconnect operates such facilities. This currently equates to 25 states in the United States, 9 countries in Europe, or 3 countries in the Asia-Pacific region. The Requester also must have a geographically-dispersed network. In the United States, at a minimum, the Requester must have a backbone node in each of the following eight geographic regions: Northeast; Mid-Atlantic; Southeast; North Central; South Central; Northwest; Mid-Pacific; and Southwest.

1.2 Traffic Exchange Ratio. The ratio of the aggregate amount of traffic exchanged between the Requester and the Verizon Business Internet Network with which it seeks to interconnect shall be roughly balanced and shall not exceed 1.8:1.

1.3 Backbone Capacity. The Requester shall have a fully redundant backbone network, in which the majority of its inter-hub trunking links shall have a capacity of at least 9953 Mbps (OC-192) for interconnection with Verizon Business-US, 2488 Mbps (STM-16) for interconnection with Verizon Business-Europe, and 622 Mbps (OC-12) for interconnection with Verizon Business-ASPAC.

1.4 Traffic Volume. The aggregate amount of traffic exchanged in each direction over all interconnection links between the Requester and the Verizon Business Internet Network with which it desires to interconnect shall equal or exceed 1500 Mbps of traffic for Verizon Business-US, 150 Mbps of traffic for Verizon Business-Europe, and 30 Mbps of traffic for Verizon Business-ASPAC.

... for rest of it see http://www.verizonbusiness.com/uunet/peering/
Tier 1 Ases/ISPs

- Have access to the entire Internet only through its settlement-free peering links
- Top of the customer-provider hierarchy
- Typically large (inter)national backbones
- Have no upstream provider
- Peer with each other to form a full-mesh
- Around 10-12 Ases: AT&T, Sprint, Level 3
Other ASes

- Lower layer providers (tier-2, ...)
  - provide transit to downstream customers
    - but need at least one provider of their own
  - typically have national or regional scope
  - include a few thousand of ASes

- Stub Ases
  - do not provide transit service
  - connect to upstream provider(s)
  - most Ases (e.g., 85-90%)
  - e.g., NUS
Simplified logical model

Transit networks

Backbone service provider

“Consumer” ISP

Small corporation

Large corporation

“Consumer” ISP

“Consumer” ISP

Stub networks

Small corporation

Small corporation

Small corporation

Small corporation
More realistic competitive view

Multi-homing

- Large corporation
- "Consumer "ISP
- Small corporation
- "Consumer "ISP
- Backbone service provider

Peering point
AS Graphs Obscure Topology

The AS graph may look like this.

Reality may be closer to this...
CAIDA’s IPv4 & IPv6 AS Core
AS-level Internet Graph
Archipelago January 2014

This visualization illustrates the extensive geographical scope and rich interconnectivity of nodes participating in the global Internet routing system, and compares snapshots of macroscopic connectivity in the IPv4 and IPv6 address space.
At The Core
The top ASes ranked by customer cone size are displayed below.
For information about a specific AS, enter its AS name, its AS number, or the name of the Org of which the AS is a member.

<table>
<thead>
<tr>
<th>AS rank</th>
<th>AS number</th>
<th>AS name</th>
<th>Org name</th>
<th>AS Type(s)</th>
<th>Number of ASes</th>
<th>IPv4 Prefixes</th>
<th>IPv4 Addresses</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>3356</td>
<td>LEVEL3</td>
<td>Level 3 Communications, Inc.</td>
<td>Tr Co</td>
<td>25,318</td>
<td>322,403</td>
<td>1,562,430,335</td>
</tr>
<tr>
<td>2</td>
<td>174</td>
<td>COGENT-174</td>
<td>Cogent Communications</td>
<td>Tr</td>
<td>17,484</td>
<td>208,063</td>
<td>744,220,957</td>
</tr>
<tr>
<td>3</td>
<td>3257</td>
<td>TINET-BACK...</td>
<td>Tinet SpA</td>
<td>Tr Co</td>
<td>15,623</td>
<td>222,392</td>
<td>846,663,937</td>
</tr>
<tr>
<td>4</td>
<td>1299</td>
<td>TELIANET</td>
<td>TeliaSonera International Carrier</td>
<td>Tr Co</td>
<td>15,178</td>
<td>228,540</td>
<td>785,632,128</td>
</tr>
<tr>
<td>5</td>
<td>2914</td>
<td>NTT-COMMUN...</td>
<td>NTT America, Inc.</td>
<td>Tr Co</td>
<td>14,876</td>
<td>224,278</td>
<td>929,277,565</td>
</tr>
<tr>
<td>6</td>
<td>3549</td>
<td>LVLT-3549</td>
<td>Level 3 Communications, Inc.</td>
<td>Tr Co</td>
<td>10,586</td>
<td>172,217</td>
<td>560,436,792</td>
</tr>
<tr>
<td>7</td>
<td>6453</td>
<td>AS6453</td>
<td>Tata Communications</td>
<td>Tr Co</td>
<td>10,229</td>
<td>167,716</td>
<td>610,754,120</td>
</tr>
<tr>
<td>8</td>
<td>6762</td>
<td>SEABONE-NET</td>
<td>TELECOM ITALIA SPARKLE S.p.A.</td>
<td>Tr Ao</td>
<td>9,904</td>
<td>129,816</td>
<td>405,609,356</td>
</tr>
<tr>
<td>9</td>
<td>6939</td>
<td>HURRICANE</td>
<td>Hurricane Electric, Inc.</td>
<td>Tr Co</td>
<td>6,240</td>
<td>73,271</td>
<td>288,745,110</td>
</tr>
<tr>
<td>10</td>
<td>1273</td>
<td>CW</td>
<td>Cable&amp;Wireless Worldwide</td>
<td>Tr</td>
<td>5,945</td>
<td>69,712</td>
<td>250,224,888</td>
</tr>
</tbody>
</table>

http://as-rank.caida.org/
The Great Peering War: Players

- **Level 3 (AS3356)**
  - also AS1 AS189 AS199 AS200 AS201 ...
  - ~49K on-net prefixes and 1325 BGP adjacencies
  - service provider to champions: Carrier’s Carrier

- **Cogent (AS174)**
  - also AS2149 AS4550 AS6259 AS6494 ...
  - ~11K on-net prefixes and 1332 BGP adjacencies
  - scrappy underdog, training hard, bulking up fast
The Timeline

- 31 Jul 2005: L3 Notifies Cogent of intent to disconnect. Both notify their sales departments; none notifies customers.
- 16 Aug 2005: Cogent begins massive sales, expecting Sept. 15 as depeering date.
- 31 Aug 2005: L3 Notifies Cogent again ...
- 7 Oct 2005 ~19:00: L3 reconnects Cogent.
The Event

- **Oct 5 between 9:00 and 11:00**
  - Cogent lost 5081 routes from L3.
  - L3 lost 2322 routes from Cogent.

- **Oct 7 around 19:00**
  - Cogent regained 4070 routes from L3.
  - L3 regained 2210 routes from Cogent.
The Damage

- 4.3% of prefixes in the global table were isolated from each other
- ~1% of globally visible ASes were affected
- Single-homed Victims
  - Cogent: 15299 Columbia Management, 18714 Perry Capital, 19040 FirstMerit N.A, 22288 Republic First Bancorp, 26264 Millennium Bank, N.A., 33378 Cathay Financial, 20330 New York State Unified CourtSystem
BGP Routing Information Bases

- **What is a route in a BGP speaker?**
  - route = prefix + attributes = NLRI + Path Attributes

- **How about all the routes in a BGP speaker?**
  - Routing Information Bases (RIBs)
  - RIBs = Adj-RIBs-In + Loc-RIB + Adj-RIBs-Out
    - Adj-RIBs-In: unprocessed routes from peers via inbound UPDATE; input for decision making
    - Loc-RIB: selected local routes used by the router
    - Adj-RIBs-Out: selected for advertisement to peers
BGP Decision Process: Overview

BGP provides policy-based routing

Best Route Selection

Inbound UPDATE

Adj-RIBs-In

Loc-RIB

Adj-RIBs-Out

Outbound UPDATE

Apply Export Policies

IP Forwarding Table

Install forwarding entries for best routes

Apply Import Policies

Apply Policies

BGP provides policy-based routing
BGP: applying policy to routes

- **Import policy**
  - filter unwanted routes from neighbor
    - e.g., prefix that your customer does not own
  - used to rank customer routes over peer routes
  - manipulate attributes to influence path selection
    - e.g., assign local preference to favored routes

- **Export policy**
  - filter routes you don’t want to tell your neighbor
    - E.g., export only customer routes to peers & providers
  - manipulate attribute to control what they see
    - e.g., make paths look artificially longer (AS prepending)
Customer-Provider Relationship

- Customer pays provider for access to Internet
  - provider exports customer’s routes to everybody
  - customer exports provider’s routes to customers

Traffic to the customer

Traffic from the customer

advertisements

traffic

NUS

dest

Singtel

NUS

Singtel
Peer-to-Peer Relationship

- Peers exchange traffic between customers
  - AS exports *only* customer routes to a peer
  - AS exports a peer’s routes *only* to its customers

Traffic to/from the peer and its customers

Diagram :

- Singtel
- Comcast
- NUS
- Princeton

Traffic connections and advertisements.
BGP routing policy

- A, B, C are provider networks
- X, W, Y are customer (of provider networks)
- X is dual-homed: attached to two networks
  - X does not want to route from B via X to C
  - .. so X will not advertise to B a route to C
BGP routing policy

- A advertises path AW to B
- B advertises path BAW to X
- Should B advertise path BAW to C?
  - no way! B gets no “revenue” for routing CBAW since neither W nor C are B’s customers
  - B wants to force C to route to w via A
  - B wants to route only to/from its customers!
Sibling to Sibling Relationship

- When exporting to a sibling
  - an AS exports its routes, routes of its customers, and also its provider or peer routes

- Models multiple Ases that belong to the same commercial organization, which owns multiple ASNs

- Defined in L. Gao’s work:

  On inferring autonomous system relationships in the Internet, IEEE/ACM Transactions on Networking, 9(6), 2001
Valley Free Property (by Gao)

- **Typical valid AS paths** (you might see from BGP routing tables)
  - single peak (uphill + downhill)
  - single flat top (uphill + 1 peering + downhill)
  - any sub-paths of the above are valid

- **Invalid patterns**
  - provider ➔ customer ➔ peering
  - provider ➔ customer ➔ provider
  - peering ➔ peering
  - peering ➔ provider
Valley-free AS paths
BGP best route selection

1. Calculation of degree of preference
   - If the route is learned from an internal peer, use **LOCAL_PREF** attribute or preconfigured policy
   - Otherwise, use preconfigured policy

2. Route selection (recommended process)
   - Highest degree of **LOCAL_PREF** (or the only route to the destination), and then tie breaking conditions on:
     - Smallest number of AS numbers in **AS_PATH** attribute
     - Lowest origin number in **ORIGIN** attribute
     - Most preferred **MULTI_EXIT_DISC** attribute
     - Routes from eBGP are preferred (over iBGP)
     - Lowest interior cost based on **NEXT_HOP** attribute
LOCAL_PREF attribute

- **LOCAL_PREF:**
  - 4-byte unsigned integer (default value 100)
  - for a BGP speaker to inform its other internal peers of its degree of preference for a route
  - should include in UPDATE messages that are sent to internal peers; should not send to external peers

![Routing Diagram]

- AS 200
- AS 400
- AS 100
- AS 300
- LOCAL_PREF 500
- LOCAL_PREF 800
- 160.10.0.0/16 LOCAL_PREF 500 > 160.10.0.0/16 LOCAL_PREF 800
MULTI_EXIT_DISC attribute

- **MULTI_EXIT_DISC (MED):**
  - 4-byte unsigned integer (default value 0)
  - for a BGP speaker to discriminate among multiple entry points to a neighboring AS to control inbound traffic
  - if received over eBGP, may be propagated over iBGP, but must not be further propagated to neighboring ASes

![Diagram showing MED attributes in AS 201 and AS 202 with MED values 2000 and 1000 respectively, illustrating the MED comparison for routes 120.68.1.0/24]
COMMUNITY attribute

- Described in RFC 1997
- 4-byte integer value
- Used to group destinations
  - Each destination could be member of multiple communities
- Very useful in applying policies within and between Ases
  - import and export policies based on the COMMUNITY attributes
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
BGP Subprefix Hijacking

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

- 18:47:00, 24 Feb 2008, Pakistan Telecom (AS 17557) began advertising 208.65.153.0/24, a more specific route of the prefix 208.65.152.0/22 used by YouTube (AS 36561)

- found 20 mins later and took ~2 hours to restore
  http://research.dyn.com/2008/02/pakistan-hijacks-youtube-1/

- can be visualized by BGPlay
  https://stat.ripe.net/special/bgplay
18:47:45  1st hijacked route propagated in Asia, AS path 3491 17557
18:48:00  9 big trans-Pacific providers carrying hijacked route
18:48:30  47 DFZ providers now carrying the bad route
18:49:00  most of the DFZ now carrying the bad route (93 ASNs)
18:49:30  all who will carry the hijacked route have it (97 ASNs)

20:07:25  AS 36561 advertises the hijacked /24 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  40 providers have stopped using the hijacked route
20:18:43  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 see the routes advertised at 20:07
20:50:59  attempted prepending, AS path was 3491 17557 17557
20:59:39  hijacked prefix is withdrawn by 3491, disconnected 17557
Preventing (Sub)Prefix Hijacking

- **Best common practice for route filtering**
  - each AS filters routes announced by customers
  - e.g., based on the prefixes the customer owns

- **But not everyone applies these practices**
  - hard to filter routes initiated from far away
  - so, BGP remains very vulnerable to hijacks

- **Other techniques**
  - secure extensions to BGP (e.g., S-BGP, soBGP)
  - anomaly detection of suspected hijacks
How is BGP used in practice?

- Three classes of “knobs”
  - preference: add/delete/modify attributes
  - filtering: inbound/outbound filtering
  - tagging: e.g., COMMUNITY attribute

- Applications
  - business relationships
    - influencing the decision process (LOCAL_PREF)
    - controlling route export (COMMUNITY)
  - traffic engineering
    - inbound traffic control (MED, AS prepending)
    - outbound traffic control (LOCAL_PREF, IGP cost)
    - remote control (COMMUNITY)