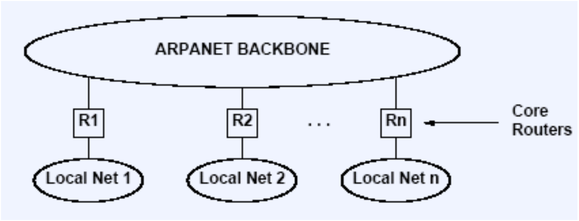
Routing Protocols

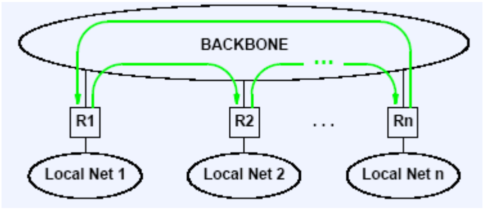
(1) Introduction

- Static routing versus dynamic routing
  - Static routing
    - Fixes routes at boot time
    - Useful only for simplest cases
  - Dynamic routing
    - Table initialized at boot time
    - Values inserted/updated by protocols that propagate route information
    - Necessary in large internets

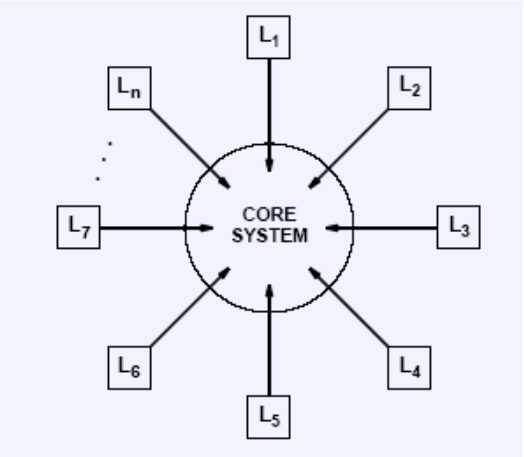
- Routing with partial information
  - The routing table in a given router contains partial information about possible destinations
  - For the unknown destinations, forward them to the *default router*.
  - Potential problem: some destinations might be unreachable.

- Original Internet and the problem if the core routers are allowed to have default routes.

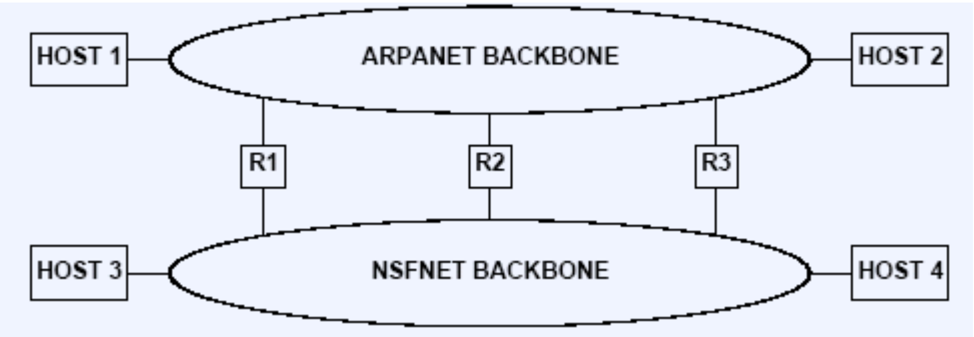




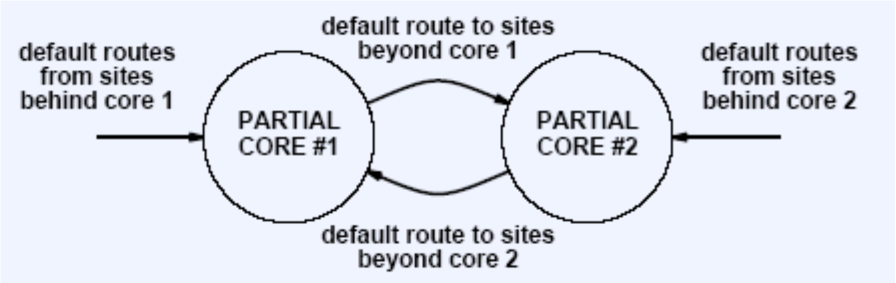
- Core routing architecture with single backbone.
  - Assumes a centralized set of routers that know *all possible destinations* in an internet.
  - Non-core routers use the core routers as their default routers.
  - Work best for internets that have a single, centrally managed backbone.
  - Inappropriate for multiple backbones.
  - Disadvantage
    - Central bottleneck for all traffic
    - Not every site could have a core router connected to the backbone: how do they get routing information?
    - No shortcut route possible: non-core routers always forward their traffic to the default routers even though another core router provides a better route. This is because the non-core routers do not know which one is better without full knowledge of all possible destinations.
    - Does not scale, because core routers must interact with each other.



- Multiple backbones
  - At beginning, NSFNET attached to the ARPANET backbone through a single router in Pittsburgh, routing is easy: routers inside NSFNET send all non-NSFNET traffic to ARPANET via the Pittsburgh router
  - Multiple connections were added later, and routing becomes complicated
    - Example: From host 3 to host 2, there are many possible routes, which one to choose?



- Partial cores are not a solution!
  - It is possible to have a single core system that spans multiple backbone networks.
  - It is not possible, however, to partition the core system into subsets that each keep partial information without losing functionality. The following figure illustrates the problem.


default route to sites beyond core 1
default routes from sites behind core 1
PARTIAL CORE #1
PARTIAL CORE #2
default routes from sites behind core 2
default route to sites beyond core 2

What We Need:

- Have a set of core routers know ***routes to all locations***
- Devise a mechanism that allows other routers to contact the core to learn routes (spread necessary routing information automatically)
- Continually update routing information

The Idea:

- The Autonomous System concept.

(2) EGP and BGP

- Autonomous System (AS)
  - Groups of networks under one administrative authority
  - Free to choose internal routing update mechanism
  - Connects to one or more other autonomous systems
  - AS number
    - ASes are assigned an AS number (ASN) by the Internet Corporation for Assigned Names and Numbers (ICANN).

- Different types of AS:
  - *Stub AS*: an AS that has only a single connection to one other AS. Naturally, a stub AS only carries local traffic.
  - *Multihomed AS*: an AS that has connections to more than one other AS, but refuses to carry transit traffic.
  - *Transit AS*: an AS that has connections to more than one other AS, and is designed (under certain policy restrictions) to carry both transit and local traffic.

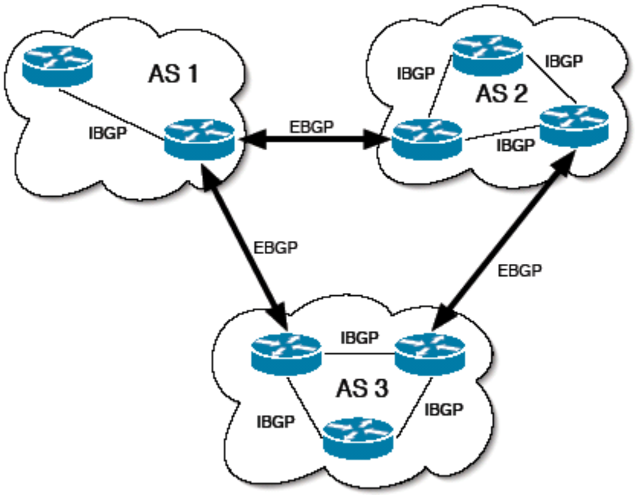
- EGP (External Gateway Protocol)
  - Originally a single protocol for communicating routes between two autonomous systems
  - Now refers to any exterior routing protocol

- BGP (Border Gateway Protocol)
  - The de facto standard of EGP in use in the Internet is BGP version 4.
  - **BGP** first became an Internet standard in 1989 and was originally defined in RFC 1105.
  - The current version, **BGP4**, was adopted in 1995 and is defined in RFC 1771 and its companion document RFC 1772.

- BGP Setup

- BGP *speaker:* a router running the BGP protocol is known as a BGP speaker. Each AS designates a border router to speak on its behalf. Some large ASs have several speakers.
- BGP *peering*:
  - BGP speakers communicate across TCP and become peers or neighbors. BGP uses TCP port 179 for establishing its connections.
  - Providers typically try to peer at multiple places. Either by peering with the same AS multiple times, or because some ASs are multi-homed, a typical network will have many candidate paths to a given prefix.
  - BGP peers are often directly connected at the IP layer; that is, there are no intermediate nodes between them. This is not necessary for operation, as peers can form a multi-hop session, where an intermediate router that does not run BGP passes protocol messages to the peer (this is a less commonly seen configuration).

- BGP peers and border routers.
  - BGP peers within the same AS are called internal peers; they communicate via Internal BGP (IBGP).
  - BGP peers from different ASes are called external peers; they communicate via External BGP (EBGP).
  - The routers that communicate using EBGP, which are connected to routers in different ASes, are called border routers.



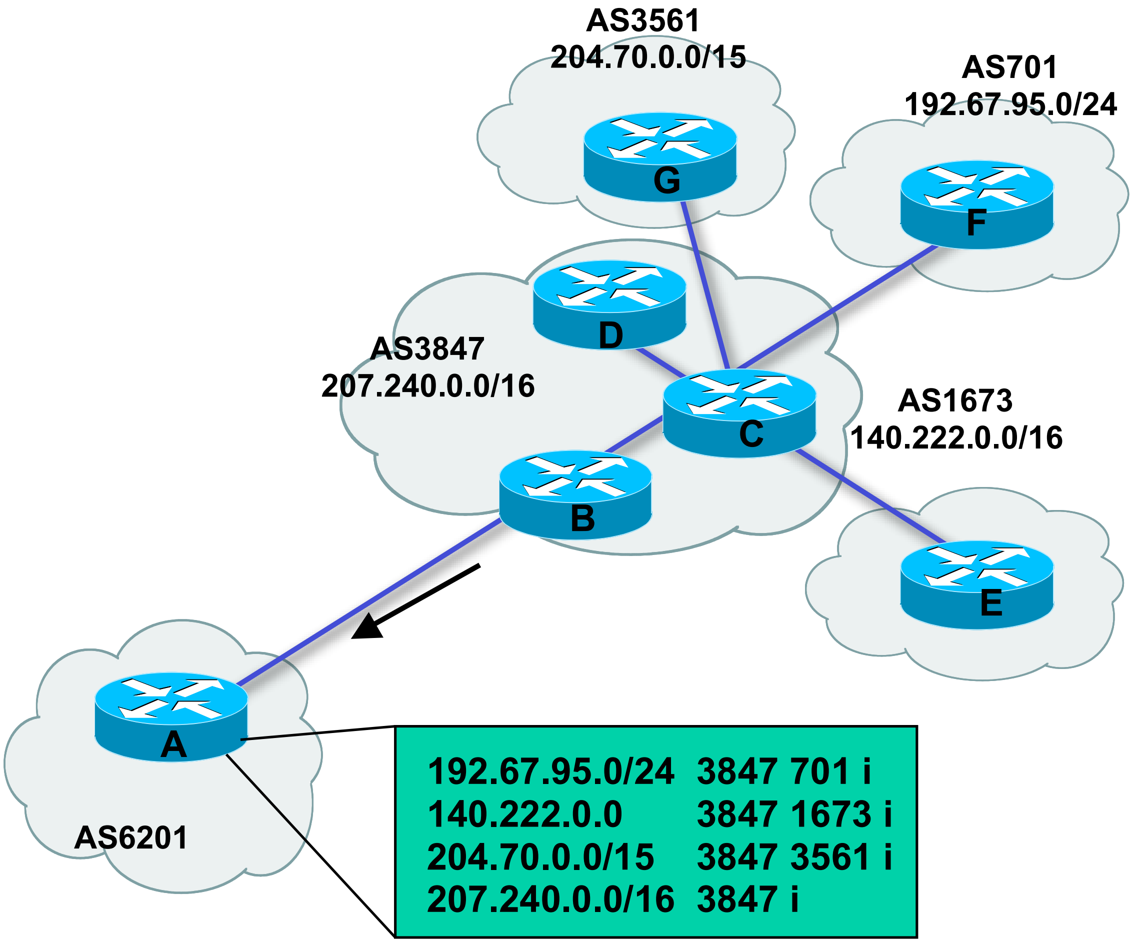
- BGP Aggregation
  - Routes can be aggregated
  - For example, a BGP speaker at the border of an autonomous system (or group of autonomous systems) must be able to generate an aggregated route for a whole set of destination IP addresses over which it has administrative control (including those addresses it has delegated), even when not all of them are reachable at the same time.

- BGP statistics in the BGP table of AS4637 (Reach) on November 30, 2005
  - AS4637 is a large AS.
  - 20946: Number of ASes in routing system
  - 173244: Number of network prefixes
    - 8700: Number of ASes announcing only one prefix
    - 1458: Largest number of prefixes announced by an AS: AS7018 (AT &T WorldNet Services)

- 91316736: Largest address span announced by an AS: AS721 (DoD Network Information Center).

- BGP routing
  - Each AS originates one or more prefixes representing the addresses assigned to hosts and devices within its network.
  - CIDR representation: prefix / (# most significant bits). For example, 192.68.0.0/16.
  - BGP peers constantly exchange the set of known prefixes and paths for all destinations in the Internet via UPDATE messages.
  - Each AS advertises the prefixes it is originating to its peers.
  - All ASes update their routing tables based on their neighbors' reachability information, and forward the received information to each of their other neighbors.

- AS_path
  - ASes establish a AS path for each advertised prefix
  - The paths are vectors of ASes that packets must traverse to reach the originating AS.
  - Path vectors are stored in a routing table and shared with neighbors via BGP.
  - As a BGP route travels from AS to AS, the AS number of each AS is stamped on it when it leaves that AS.
  - An example:

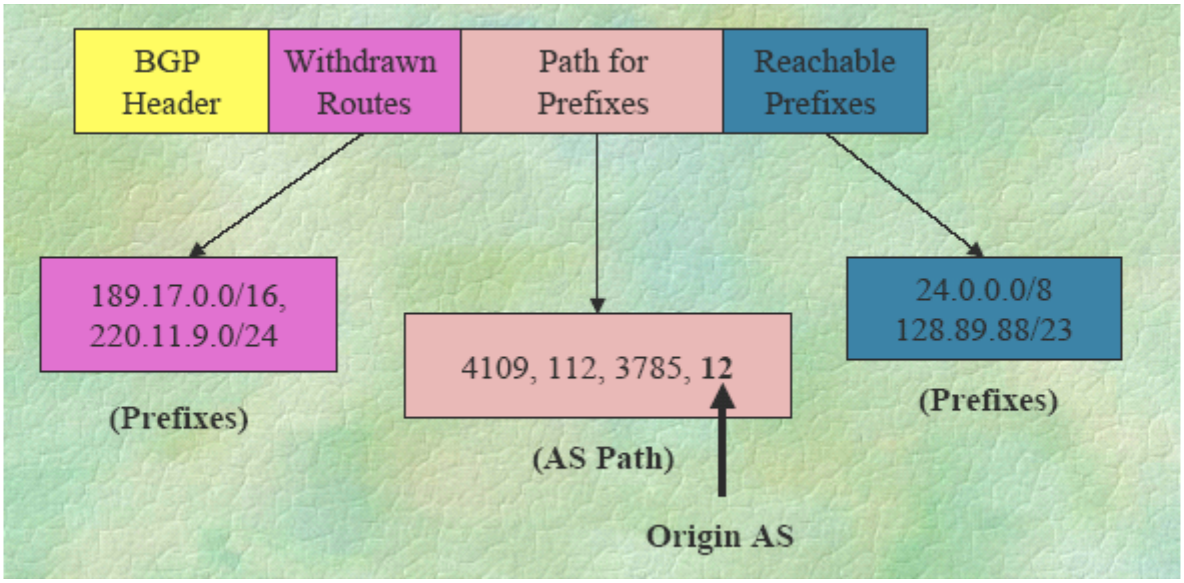


192.67.95.0/24	3847 701 i
140.222.0.0	3847 1673 i
204.70.0.0/15	3847 3561 i
207.240.0.0/16	3847 i

- BGP Update Message Fields
  - BGP packets in which the type field in the header identifies the packet to be a BGP update message packet include the following fields. Upon receiving an update message packet, routers will be able to add or delete specific entries from their routing tables to ensure accuracy. Update messages consist of the following packets:

  - *Withdrawn Routes*---Contains a list of IP address prefixes for routes being withdrawn from service.
  - *Path Attributes*---Describes the characteristics of the advertised path. The following are possible attributes for a path:
    - Origin: Mandatory attribute that defines the origin of the path information
    - AS Path: Mandatory attribute composed of a sequence of autonomous system path segments
    - Next Hop: Mandatory attribute that defines the IP address of the border router that should be used as the next hop to destinations listed in the network layer reachability information field
    - Multi-Exit Discriminator: Optional attribute used to discriminate between multiple exit points to a neighboring autonomous system
    - Local Preference: Discretionary attribute used to specify the degree of preference for an advertised route
    - Atomic Aggregate: Discretionary attribute used to disclose information about route selections
    - Aggregator: Optional attribute that contains information about aggregate routes
  - *Network Layer Reachability Information*---Contains a list of IP address prefixes for the advertised routes

- An example: A simplified BGP UPDATE message:



- Routing Policy
  - BGP enforces routing policies, such as the ability to forward data only for paying customers through a number of protocol features.
  - Routing policies are related to political, security, or economic considerations.
    - A multihomed AS can refuse to act as a transit AS for other AS's. (It does so by only advertising routes to destinations internal to the AS.)
    - A multihomed AS can become a transit AS for a restricted set of adjacent AS's, i.e., some, but not all, AS's can use the multihomed AS as a transit AS. (It does so by advertising its routing information to this set of AS's.)
    - An AS can favor or disfavor the use of certain AS's for carrying transit traffic from itself.
  - BGP uses the attribute values in UPDATE messages to help enforce policies.
  - Policies configured in a BGP router allow it to do the following:
    - Filter the routes received from each of its peers
    - Filter the routes advertises to its peers
    - Select routes based on desired criteria
    - Forward traffic based on those routes
  - Setting policy often involves techniques to bias BGP's route selection algorithm.

- Multiple Path
  - BGP could possibly receive multiple advertisements for the same route from multiple sources.
  - BGP selects only one path as the best path.
  - When the path is selected, BGP puts the selected path in the IP routing table and propagates the path to its neighbors.

- BGP Path Selection.
  - One of the major tasks of a BGP speaker is to evaluate different paths from itself to a set of destination covered by an address prefix, select the best one, apply appropriate policy constraints, and then advertise it to all of its BGP neighbors.
  - Metric
    - Each AS can use its own routing protocol
    - Metrics differ (hop count, delay, etc)
    - BGP does not communicate or interpret distance metrics.
    - The only interpretation is the following: "My AS provides a path to this network".
  - Where there are more than one feasible paths to a destination, all feasible paths should be maintained.
    - Each feasible path is assigned a preference value.
    - The process of assigning a degree of preference to a path can be based on several sources of information:
      - Information explicitly present in the full AS path.
      - A combination of information that can be derived from the full AS path and information outside the scope of BGP (e.g., policy routing constraints provided as configuration information).
  - Possible criteria for assigning a degree of preference to a path are
    - Prefer the path with the largest *weight* (weight is defined by Cisco, and is local to a router).
    - *Local preference*: prefer an exit point from the local AS. The local preference attribute is propagated throughout the local AS.
    - *Shortest* AS_path (the AS count).

        - Lowest *origin* type: A path learned entirely from BGP (i.e., whose endpoint is internal to the last AS on the path) is generally better than one for which part of the path was learned via EGP or some other means.
        - Policy considerations
        - Presence or absence of a certain AS or AS's in the path
        - Link dynamics: Stable paths should be preferred over unstable ones.

- The length of an AS path vector
    - This is one of the most significant criteria BGP uses for path selection
    - This length can be modified by an organization repeatedly adding its AS number to a path, in order to discourage its use (a technique known as *padding* or *prepending*).

- Third party restriction
    - EGP restricts a (noncore) router to advertise only those networks reachable entirely from within its autonomous systems.

---

(3) Case Study: Syracuse University

- Announced Prefixes (http://www.cidr-report.org/cgi-bin/as-report?as=AS11872)
    - The data is collected from AS4637 on November 30, 2005.

```
        AS11872: SYRACUSE-UNIVERSITY - Syracuse University


              Prefix               (AS Path)
          128.230.0.0/16       4637 6395 11872
          149.119.0.0/16       4637 6395 11872
          192.155.14.0/24      4637 6395 11872
          192.155.16.0/24      4637 6395 11872
Note:     AS6395 is BROADWING - Broadwing Communications Services, Inc.
```

- Address Space
    - AS11872 (Syracuse University)
        - Originate Address Space: 131584/14.99
        - Transit Address Space: 0 (i.e., AS11872 does not provide transit service).
    - AS6395 (Broadwing, a Tier-I ISP)
        - Originate Address Space: 1365504/11.62
        - Transit Address Space: 7683584/9.13 (i.e., it does provide transit service)
    - AS4637 (Reach)
        - Originate Address Space: 314880/13.74
        - Transit Address Space 1713352013/1.33 (the transit coverage is larger than AS6395)

- Partial Topology (using `traceroute` and `whois`)
    - AS11872 (Syracuse University) peers with AS6395 (Broadwing), AS4323 (Time Warner), AS1785 (NYSERNet R&E Network), ….
    - NYSERNet only provides backbone in New York State. It does not carry commercial traffic.

- NYSERNet buys the Internet service from Broadwing, a Tier-I ISP (and maybe others). Therefore, commercial traffic goes to those backbone.
- NYSERNet connects to the Abilene (AS11537). The Abilene Network is an Internet2 high-performance backbone network that enables the development of advanced Internet applications and the deployment of leading-edge network services to Internet2 universities and research labs across the country. The network has become the most advanced native IP backbone network available to universities participating in Internet2.
- If we `traceroute` to an Internet2 universities, most likely the traffic goes through the Abilene backbone (abilene.ucaid.edu).
- If we `traceroute` to a company (e.g., yahoo.com), most likely the traffic goes through the Broadwing backbone.

---

(4) Attacking BGP

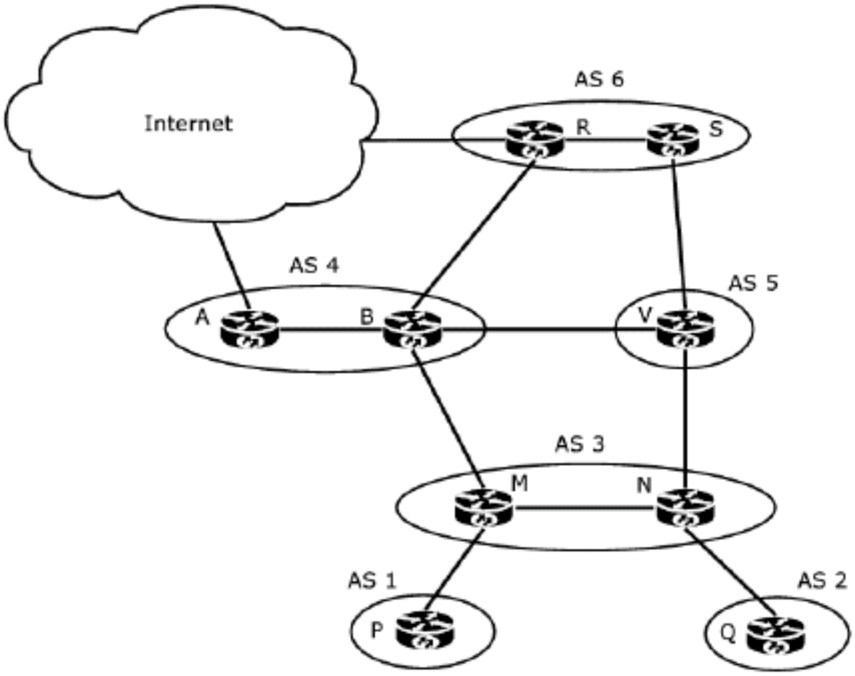
- Misconfigurations
  - Misconfigurations are quite common in practice, and they can cause the same problems that an attack could cause.
  - April 25 1997: AS7007 flooded the Internet with incorrect advertisements, announcing AS7007 as the origin of the best path to essentially the entire Internet.
  - April 7 1998: AS8584 announced about 10,000 prefixes it did not own.
  - April 6 2001: AS15412 announced about 5,000 prefixes it did not own.

- Attacking assumptions
  - Attackers have already compromised and taken complete control of one or more BGP speakers.
- Objectives of an attacker
  - **Blackholing**: occurs when a prefix is unreachable from a large portion of the Internet.
    - Intentional blackhole routing is used to enforce private and non-allocated IP ranges.
    - Malicious blackholing refers to false route advertisements that aim to attract traffic to a particular router and then drop it.
  - Redirection: occurs when traffic flowing to a particular network is forced to take a different path and to reach an incorrect, potentially also compromised, destination.
  - Subversion: is a special case of redirection in which the attacker forces the traffic to pass through a certain link with the objective of eavesdropping or modifying the data.
  - Instability: can be caused by successive advertisements and withdrawals for the same network.

- Fraudulent Origin Attacks
  - A malicious AS can advertise incorrect information through BGP UPDATE messages passed to routers in neighboring ASes.
  - *Prefix hijacking*: A malicious AS can advertise a prefix originated from another AS and claim that is the originator.
  - *Prefix deaggregation*: This occurs when the announcement of a large prefix is fragmented or duplicated by a collection of announcements for smaller prefixes.
    - BGP performs *longest prefix matching*, whereby the longest mask associated with a prefix will be the one chosen for routing purposes.

  - For example, if the prefixes 12.0.0.0/8 and 12.0.0.0/16 are advertised, the latter prefix, which corresponds to a more specific portion of the address block, will be chosen.
  - If an AS falsely claims to be the origin of a prefix and the update has a longer prefix than others currently in the global routing table, it will have fully hijacked that prefix. The false updates will eventually be propagated throughout the Internet.

- Subversion of Path Information
  - A malicious AS can tamper with the path attributes of an UPDATE message.
  - Recall: BGP uses path vector; routing to destinations is performed based by sending packets through the series of ASes denoted in the path string.
  - An AS can modify the path it receives from other ASes by
    - Inserting or deleting ASes from the path vector
    - Changing the order of the ASes
    - Altering attributes in an UPDATE message, such as the multi-exit discriminator (used to suggest a preferred route into an AS to an external AS) or the community attribute (used to group routes with common routing policies)



- Setup of the above figure
  - AS1 and AS2 are stub networks that have been assigned address blocks from their provider AS3.
  - All ASes provide transit service to their customers, which reside at the lower levels of the diagram.
  - The horizontal lines (e.g. between routers B-V) represent backup links and non-transit relations between the corresponding ASes.

- Attacking Scenarios:
  - <u>Router B wants to subvert traffic destined to AS2</u>:

        - It could announce a fake route, announcing that it has a direct connection to AS2.
        - It could also claim ownership of the address blocks originated by AS2. Routers A and R would then forward traffic destined to AS2 to B.
    - B can de-aggregate the prefix announced by AS2 to two prefixes that are longer by one bit, while keeping the AS-PATH to AS2 the same. In that case, traffic originating anywhere in the Internet, except in AS2, and destined to AS2 would be forwarded towards router B.
    - If AS2 owned a prefix that was aggregated with other prefixes by the provider AS2, then B could simply announce the original AS2 prefix.
    - Note that a compromised BGP speaker can use de-aggregation to blackhole a victim network anywhere in the Internet, regardless of the proximity between the two.
    - Redirect traffic:
        - Normally, B should announce the AS1 route that goes through {AS1, AS3, AS4}.
        - Instead, B can propagate that route only to A indicating that it should not be announced any further, and announce the padded route that goes through AS5 to R.
    - Update modifications
        - Suppose that AS3 uses the link V-N only for backup purposes because it is cheaper to use link B-M instead. To achieve this, router N can pad the UPDATEs going to V, making the corresponding AS-PATH longer.
        - Assume that R is compromised, and that it wants to redirect traffic to AS3 through the more expensive link V-N.
        - R can drop the padding in the route that includes the {AS5, AS3} link, and instead pad the route that includes the {AS4, AS3} link (or simply not announce it). This would force traffic for AS3 to take the more costly V-N route.

- Route flapping and Route dampening
    - If a router goes offline frequently, the routes it advertises will disappear and reappear in peer routing tables. This is called *route flapping*.
    - In order to lower burden, unstable routes are often penalized through a process called *route dampening*.
    - Neighboring routers will ignore advertisements from the router for an increasing amount of time, depending on how often the route flapping occurs.

- Attacks using route flapping
    - Can be used to trigger route dampening for a victim network at an upstream router.
    - This can be done by withdrawing and re-announcing the target routes at a sufficiently high rate that the neighboring BGP speakers dampen those routes.
    - A dampened route would force the traffic to the victim AS to take a different path, enabling traffic redirection.
    - The dampening can be triggered when a single route flap forces BGP peers to consider several backup paths, causing a large number of additional withdrawals and announcements.

- Congestion-induced BGP session failures
    - When the BGP peers are under heavy congestion, the TCP-based BGP sessions can be so slow that they are eventually aborted, causing thousands of routes to be withdrawn.
    - When BGP sessions are brought up again, routers must exchange full routing tables, creating large spikes of BGP traffic and significant routing convergence delays.
    - For example, studies have shown that during the adverse effects of the Code Read and Nimda worms of 2001, BGP traffic "exploded" by a factor of 25 (later, another study has shown that over 40% of the observed BGP updates are due to other reasons).

- Other Denial of Service Attacks

- TCP RST attacks
- SYN flooding attacks
- ICMP attacks

(5) Securing BGP

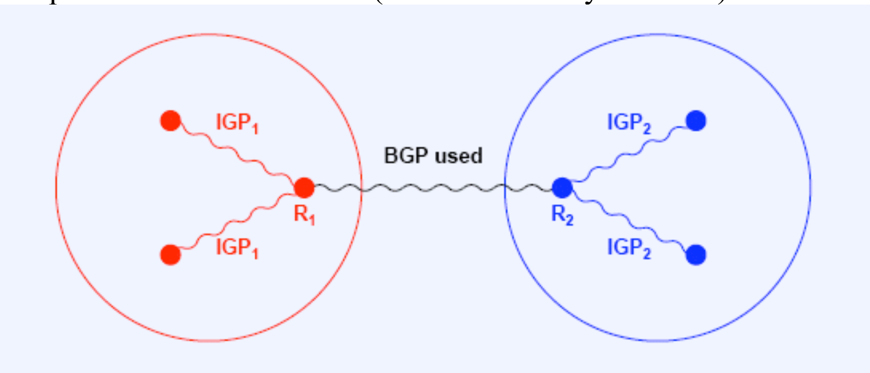
- S-BGP: Secure BGP
  - Designed by researchers at BBN with the objective to protect BGP from erroneous or malicious UPDATEs.
  - S-BGP makes three major additions to BGP
    - It introduces a Public Key Infrastructure (PKI) in the interdomain routing infrastructure to authorize prefix ownership and validate routes
    - A new transitive attribute is introduce to BGP updates. That attribute ensures the authorization of routing UPDATEs, and prevents route modification from intermediate S-BGP speakers
    - All routing message can be secured using IPSec, if routing confidentiality is a requirement.
  - Address Attestations (AAs)
    - Issued by the owner of one or more prefixes, to identify the first AS authorized to advertise the prefixes.
  - Route Attestations (RAs)
    - Issued by a router on behalf of an AS (ISP), to authorized neighbor ASes to use the route in the UPDATE containing the RA.

- The Protocol Operation:
  - When generating an UPDATE, a router generates a new RA that encompasses the path and prefixes plus the AS # of the neighbor AS
  - When receiving an UPDATE from a neighbor, it
    - Verifies that its AS # is in the first RA
    - Validates the signature on each RA in the UPDATE, verifying that the signer represents the AS # in the path
    - Checks the corresponding AA to verify that the origin AS was authorized to advertise the prefix by the prefix “owner”

- Limitations of S-BGP
  - Require the presence of a hierarchical PKI infrastructure and distribution system, trusted by all participating ISPs.
  - S-BGP is quite cryptographically intensive
  - Routers may need a large memory space (about 20MB per peer) to store the public keys. The space requirement can be significant for a speaker with tens of peers
  - Aggregation is an additional problem for S-BGP
  - S-BGP cannot prevent “collusion attacks” (or the wormhole attack). Such attacks are possible when two compromised routers fake the presence of a direct link between them. For the rest of the Internet, it then appears as if those two ASes are connected.

(6) Within an Autonomous Systems (RIP, OSPF)

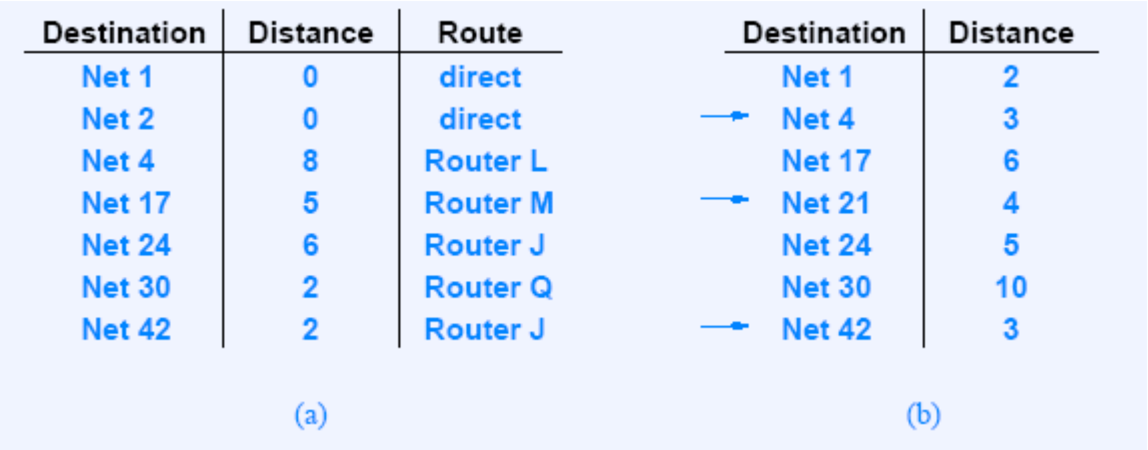
- Relationship between BGP and IGP (Interior Gateway Protocol)



- IGP Protocols
  - There is no single standard for IGP.
  - Examples of IGP: RIP, HELLO, OSPF

- Distance-Vector Routing
  - Each entry in the table identifies a destination network and gives the distance to that network, usually measured in hops
  - Initially, a router initializes its routing table to contain an entry for each directly connected network.
  - Periodically, each router sends a copy of its routing table to any other router it can reach directly. When a report arrives at router K from router J, K replaces its table entry under the following conditions:
    - If J knows a shorter way to reach a destination
    - If J lists a destination that K does not have
    - If K currently routers to a destination through J and J's distance to that destination changes.

- Distance-Vector Routing Example
  - (a) is an existing route table for router K
  - (b) is an incoming routing update message from router J. The marked entries will be used to update existing entries or add new entries to K's routing table.



Destination	Distance	Route
Net 1	0	direct
Net 2	0	direct
Net 4	8	Router L
Net 17	5	Router M
Net 24	6	Router J
Net 30	2	Router Q
Net 42	2	Router J

(a)

	Destination	Distance
	Net 1	2
→	Net 4	3
	Net 17	6
→	Net 21	4
	Net 24	5
	Net 30	10
→	Net 42	3

(b)

- RIP: Routing Information Protocol
  - Implemented by UNIX program `routed`
  - RIP operates on UDP port 520
  - Distance-Vector protocol
  - Uses hop count metric (16 is infinity)
  - Relies on broadcast
  - Current standard is RIP2

- Two modes of RIP
  - Active mode:
    - Broadcast a message every 30 seconds.
    - The message contains information taken from the router’s current routing database.
    - Each message consists of pairs, where each pair contains (IP, hop count)
    - Only routers can run RIP in active mode.
  - Passive mode
    - Listen and update their routing tables.
    - Both host and router can run in passive mode.

- Link-State Routing (Shortest Path First, or SPF)
  - Participating routers learn internet topology
  - Think of routers as nodes in a graph, and networks connecting them as edges or links
  - Pairs of directly-connected routers periodically
    - Test link between them
    - Propagate (broadcast) status of link
  - All routers
    - Receive link status messages
    - Recompute routes from their local copy of information using the well-known *Dijkstra shortest path algorithm*. Note that Dijkstra’s algorithm computes the shortest paths to all destinations from a single source.

- OSPF: Open SPF
  - Includes *type of service routing*. Multiple routes to a given destination can be installed, one for each type of service.
  - Provides *load balancing*.
  - Partition networks into subsets called *areas*.

  - Require message authentication.
  - Support network-specific, subnet-specific, host-specific, and CIDR routes.

- OSPF authentication
  - Simple Authentication
    - A password (key) is configured on each router and is included in plaintext in each OSPF packet originated by that router.
    - It is not secure.
  - MD5 Authentication
    - It is based on shared secret keys that are configured in all routers in the area.
    - Each router computes an MD5 hash for each packet based on the content of the packet and the configured secret key. Then it includes the resulting hash value in the OSPF packet.
    - The receiving router, using the pre-configured secret key, will compute an MD5 hash of the packet and compare it with the hash value that the packet carries thus verifying its authenticity.
    - Sequence numbers are also employed with MD5 authentication to protect against replay attacks.

(7) References

1. Comer's TCP/IP Slides
2. Bellovin's slides (2003) http://www.cs.columbia.edu/~smb/talks/routesec.pdf
3. Mao. http://www.eecs.umich.edu/~zmao/eecs589/notes/lec3_6.pdf
4. Ola Nordstrom and Constantions Dovrolis, Beware of BGP Attacks.
5. K. Butler, T. Farley, P. McDaniel, and J. Rexford. A Survey of BGP Security