Massachusetts Institute of Technology
Department of Electrical Engineering and Computer Science

6.829 Fall 2005 **Problem Set 2** September 22, 2005

This problem set has 3 questions, most with multiple parts. Answer them as clearly and concisely as possible. You may discuss ideas with others in the class, but your solutions and presentation must be your own (for each such discussion, please mention whom you collaborated with). Do not look at anyone else's solutions or copy them from anywhere (in particular, "bibles" are not allowed).

Turn in your solutions on **Thursday, September 29, 2005** in class.

1 Route reflection

The Border Gateway Protocol (BGP) has two modes of operation. eBGP runs between the border (or egress) routers of ASes to exchange reachability information between ASes. iBGP runs within an AS to disseminate the information about external destinations (learned through eBGP) amongst routers within an AS.

We define two correctness properties that are desirable in any intra-AS route dissemination mechanism.

P1 **Complete visibility:** The dissemination of information amongst the routers is "complete" in the sense that, for every external destination, each router picks[1] the same route that it would have picked had it seen the best routes from each eBGP router in the AS.

P2 **Loop-free forwarding:** After the dissemination of eBGP learned routes, the resulting routes (and the subsequent forwarding paths of packets sent along those routes) picked by all routers are free of forwarding loops.

1. To solve the intra-AS route dissemination problem, the designers of BGP proposed the use of a "full mesh" iBGP configuration. Here, each eBGP router in the AS establishes BGP sessions with all the other routers (both eBGP routers and internal routers) in the network. Show that the full mesh iBGP configuration always satisfies P1 and P2. (Assume there are no link or router failures in the network. Also assume that the underlying IGP implements shortest path routing.)

2. The full mesh iBGP does not scale well because it requires a quadratic number of iBGP sessions. Route reflection improves the scalability of intra-AS route dissemination. Refer L4 notes for information on how route reflection works.

   iBGP configurations with route reflectors do not necessarily satisfy P1 and P2. Consider the iBGP configuration shown in Figure 1. R1 and R2 are route reflectors and are iBGP peers of each other. C1 and C2 are route reflector clients of R1 and R2 respectively. The IGP costs of the network are marked on the links in the figure. Two routes to an external destination d, which are equally good with respect to the local preference, AS path length and MED

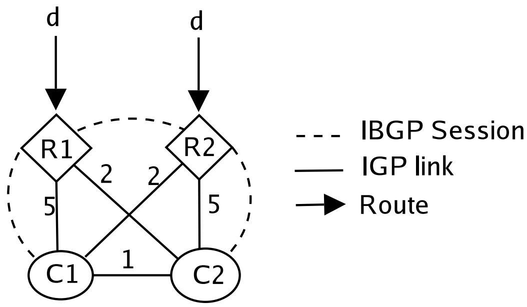

Figure 1: iBGP configuration with route reflectors

attributes, arrive at the routers R1 and R2. Under this situation, show that the given iBGP configuration violates properties P1 and P2.

3. Suppose the route reflectors R1 and R2 are modified so that they forward not just their best route but *all* the routes they hear to (and from) their clients. Show that, under this modification, this iBGP configuration satisfies the properties P1 and P2.

4. Consider an iBGP configuration which has the following property:[2] For every router A and every egress router B, one of the following conditions always holds (a) A is a route reflector and B is a client of A. (b) B is a route reflector and A is a client of B (c) A and B are normal iBGP peers (d) there exists a route reflector R on the shortest path between A and B such that both A and B are clients of R. Show that such an iBGP configuration satisfies P1 and P2.

2 Understanding BGP using table dumps

For this question, you will need to download the Routeviews routing table from `http://nms.csail.mit.edu/6.829/ps/ps2/oix-full-snapshot-2005-09-20-2200.dat.gz`

This file contains a Cisco BGP-4 routing table snapshot, taken at Oregon Route Views (`http://www.routeviews.org/`) on September 20, 2005. If you are curious about what other snapshots look like, you can find daily snapshots at `http://archive.routeviews.org/`. You can also find the peering structure of the Routeview router at `http://www.routeviews.org/peers/route-views.oregon-ix.net.txt`.

1. To start with, find the routing table entry for the MIT network (which corresponds to the prefix 18.0.0.0/8).

   (a) From the routing table file, what is the AS number for MIT?

   (b) What is the IP address of the best next hop from this router to MIT? How does this router know how to reach that next hop IP address?

[1]Refer to L4 notes for details on the BGP route selection process.

[2]Do not worry about how the iBGP should be configured so that this property is satisfied.

(c) How many AS's must a packet traverse between the time it leaves the router and the time that it arrives at MIT?

(d) Use `traceroute` now to trace the route from MIT to the router that took the snapshot. Is the current route from MIT to the router the same as the reverse route in the trace data?

(e) On September 20, 2005 the AS path to `route-views.oregon-ix.net` from MIT was `10578 11537 4600 3582`. Why is this path not simply the reverse of the path advertised by MIT to Routeviews? Why does this traceroute (which was run at the same time), not match the AS path?

```
traceroute to route-views.oregon-ix.net (128.223.60.103), 30 hops max, 38 byte packets
 1  legacy31.default.csail.mit.edu (18.31.0.1) [AS3]  0.430 ms  0.371 ms  0.370 ms
 2  kalgan.trantor.csail.mit.edu (128.30.0.245) [AS40] 0.401 ms  0.408 ms  0.400 ms
 3  B24-RTR-2-CSAIL.MIT.EDU (18.4.7.1)  7.190 ms [AS3] 0.830 ms  1.092 ms
 4  EXTERNAL-RTR-1-BACKBONE.MIT.EDU (18.168.0.18) [AS3] 3.714 ms  0.942 ms  0.764 ms
 5  EXTERNAL-RTR-2-BACKBONE.MIT.EDU (18.168.0.27) [AS3] 0.829 ms  0.692 ms  0.755 ms
 6  nox230gw1-Vl-526-NoX-MIT.nox.org (192.5.89.89) [(null)] 5.305 ms  0.783 ms  0.609 ms
 7  nox230gw1-PEER-NoX-NOX-192-5-89-10.nox.org (192.5.89.10) [(null)] 6.119 ms  9.434 ms  6.08
 8  chinng-nycmng.abilene.ucaid.edu (198.32.8.82) [(null)] 26.016 ms  39.404 ms  26.142 ms
 9  iplsng-chinng.abilene.ucaid.edu (198.32.8.77) [(null)] 270.931 ms  281.590 ms *
10  kscyng-iplsng.abilene.ucaid.edu (198.32.8.81) [(null)] 50.490 ms  39.170 ms  39.077 ms
11  dnvrng-kscyng.abilene.ucaid.edu (198.32.8.13) [(null)] 49.970 ms  49.781 ms  49.803 ms
12  snvang-dnvrng.abilene.ucaid.edu (198.32.8.1) [(null)] 74.424 ms  74.614 ms  74.453 ms
13  pos-1-0.core0.eug.oregon-gigapop.net (198.32.163.17) [AS4600] 86.717 ms  86.726 ms  86.792
14  uo-0.eug.oregon-gigapop.net (198.32.163.147)  89.333 ms [AS4600] 110.026 ms  102.835 ms
15  ge-5-1.uonet1-gw.uoregon.edu (128.223.2.1) [AS3582] 87.760 ms 86.942 ms ge-5-1.uonet2-gw.u
16  g0-1.route-views.routeviews.org (128.223.60.103) [AS3582] 87.468 ms *  87.431 ms
```

(f) How many routes are there to get from this router to MIT?

(g) From the routing table, what is the best route to MIT? Why was this route selected as the best route?

(h) Using the data in the Routeviews table and using `dig` (for DNS lookups), infer which companies are the likely ISPs for MIT.

2. Several of the IP prefixes in the table are formatted as $w.x.y.z/m$. The mask field, m, specifies the length of the network mask to use when matching input destination addresses to entries in the table.

(a) Write down an expression using bit-wise operations to determine whether a destination address, A_i, matches a prefix A/m in the routing table. A_i and A are 32 bits each.

(b) Find the first "Class C" CIDR address in the table (address prefix $\geq$ 192.0.0.0). How many class C networks does this address correspond to? What is the maximum number of routing table entries that this single CIDR address saves? Why is it that we can only infer the maximum, and not the actual, number of addresses that this CIDR address saves?

(c) In the table, there are examples of groups of prefixes that have the same advertised AS path, but show up as separate entries in the routing table.[3]

[3] For both parts of this problem, it's sufficient to find the existence of one AS path that is advertised more than once. It is *not* necessary to find two prefixes for which *all* advertised paths are the same.

(i) Provide an example of non-contiguous prefixes (and the corresponding AS path) for which this is true. Why might non-contiguous prefixes have the same AS path?

(ii) Provide an example of contiguous prefixes (and the corresponding AS path) for which this is true. This practice is often called *deaggregation*. Why might this be done?

3. Ben Bitdiddle is interested in studying the characteristics of the Internet using routing table snapshots. The Oregon Exchange has agreed to give Ben Bitdiddle some partial routing table snapshots from 1995 to the current day, including some snapshots from before the upgrade to BGP-4. They will give him snapshots containing the following:

   (a) Only the destination network address/mask.

   (b) Only the lines marked `*>`.

   (c) Only the paths, with best next-hops marked.

   Ben doubts that these partial snapshots could tell him anything interesting, but you disagree. What information about the evolution of the Internet could you infer from each type of partial snapshot?

3 Inferring AS Relationships

As you know, the Internet is composed of about 20,000 distinct origin ASes that exchange routes to establish global connectivity, and that business relationships determine which routes are exchanged between each pair of ASes.

Recall that one network will re-advertise its customer routes to its peers and providers, but will not re-advertise routes heard from a peer to other peers or providers. With the knowledge of these "rules" and a view of a default-free routing table (or multiple tables), one can deduce relationships between AS pairs based on links that exist in the AS graph.[4]

In *On Inferring Autonomous System Relationships in the Internet*[5], Lixin Gao observes that, because of these constraints, AS paths must adhere to one of the following patterns:

1. a series of customer-provider links (an *uphill path*)
2. a series of provider customer links (a *downhill path*)
3. an uphill path followed by a downhill path
4. an uphill path followed by a peering link
5. a peering edge followed by a downhill path
6. an uphill path followed by a peering link, followed by a downhill path

[4]There rules are sometimes violated in practice, but for the purpose of this problem we will assume they hold.

[5]You can find a copy of this paper at `http://www-unix.ecs.umass.edu/~lgao/ton.ps`. While you don't need to read the paper to solve this problem, you may find it helpful and interesting.

This is called the "valley free" property of AS paths. The hard question, of course, is: where is the "top of the hill"? Gao suggests using the AS in the path that contains the largest degree: that is, the AS that connects to the most other ASes.

We have provided a Routeviews routing table for you at `http://nms.csail.mit.edu/6.829/ps/ps2/oix-full-snapshot-2005-09-20-2200.dat.gz`. Your task is to produce a good guess about relationship between each AS pair in the table.

1. Produce the CCDF (Complementary Cumulative Distribution Function) of AS degree (i.e., plot the fraction of AS's that have degree of $\geq n$, for all $n > 0$) on a log-log scale. Also include a table of the "top 10" AS's for degree and the value of their degrees. Do not count a link from an AS to itself as an edge. Also, consider *all* AS paths that are given in the table, not just the best path for each prefix.

2. For each of the following AS paths, list the transit relationships inferred for each pair, based on that path. *This is a two-step process.*

   First, for each AS path, note the transit relationships. For example, for the path $ABCD$, if C were the AS with the highest degree, you would write "Transit relationships: $A \rightarrow B$, $B \rightarrow C$, $D \rightarrow C$". This will give you a list of AS pairs that have transit relationships.

   Once you have scanned all AS paths, you may find that you have a commutative transit relationship: i.e., A transits B *and* B transits A. This is called a sibling relationship. For all pairs in the following paths, note which AS transits for the other, or if the two pairs have a sibling relationship.

   (a) 16150 1239 701 703 80
   (b) 7660 2516 1239 7018 2386
   (c) 3277 3267 3343 1299 3549 206
   (d) 4513 7911 3320 8551

3. Finding the "top of the hill" by using the AS with the highest degree sometimes produces the wrong answer. Another way to do this is to view the AS paths from one vantage point as a directed graph, and using a reverse pruning algorithm to the AS graph in order to assign ranks to each AS.

   First, leaf nodes of the AS graph are assigned the lowest rank. Then, these nodes and their incident edges are removed from the graph. The nodes that are leaves in this new graph are assigned the next highest rank. The process repeats until the graph is strongly-connected (i.e., there are no leaves); each node in the strongly-connected component of the graph receives the highest rank. AS relationships are inferred by comparing rankings from AS graphs as visible from *multiple vantage points*[6].

   (a) What are advantages of using this type of ranking scheme over a power-law based scheme? What are the disadvantages?
   (b) Why does this scheme require multiple vantage points to be effective?

[6] You can find the paper that describes this algorithm in detail at `http://www.cs.berkeley.edu/~sagarwal/research/BGP-hierarchy/infocom02.pdf`.