Available at
www.ElsevierComputerScience.com
POWERED BY SCIENCE DIRECT®

Computer Communications 27 (2004) 557–567

computer communications

www.elsevier.com/locate/comcom

Issues with inferring Internet topological attributes

Lisa Amini[a,b,*], Anees Shaikh[a], Henning Schulzrinne[b]

[a]IBM Research, 19 Skyline Drive, Hawthorne, NY 10532, USA
[b]Department of Computer Science, Columbia University, 450 Computer Science Building, 1213 Amsterdam Avenue, New York, NY 10027-7003, USA

Received 8 August 2003; accepted 8 August 2003

Abstract

A number of recent studies of Internet network structure are based on data collected from inter-domain BGP routing tables and tools, such as `traceroute`, to probe end-to-end paths. The goal of these studies is often to infer Internet topological properties. There is growing evidence, however, that the amount and diversity of the data has a significant impact on the conclusions drawn about some of the structural properties. While systematic data collection from a number of network vantage points can reduce certain ambiguities, thus far, no methods have been reported for fully resolving these issues. The goal of our study was to quantify the effects of these anomalies on key Internet structural attributes. We report on our analysis of over 290,000 measurements from globally distributed sites. We contrast results obtained from router-level measurements with those obtained from BGP routing tables, and offer insights as to why certain inferred properties differ. We use multiple views of the same data to demonstrate that some topological attributes, such as the average path length, are relatively consistent across a variety of data sources. We also illustrate how using the same methodology to model other attributes, such as those based on the actual forwarding path between a pair of nodes, or the level of AS path asymmetry, can produce substantially misleading results.

Keywords: Internet mapping; Topology; Traceroute; Border Gateway Protocol; BGP

1. Introduction

The Internet's decentralized control and open nature have enabled it to evolve into an immense interconnection of millions of hosts, hundreds of thousands of address prefixes, and tens of thousands of separately administered routing domains. The structure of the network has significant implications on the performance and control of a number of existing and emerging network services, for example overlay and peer-to-peer networks, and content distribution architectures. Accurate measurement and characterization of the underlying network topology is critical to isolate faults and pathologies within the Internet, improve existing protocols, validate scalability of new services, and predict the future evolution of the Internet.

Developing a representative map of the Internet is challenging for a number of reasons. First, the structure of the Internet itself is not static—new nodes and edges are added daily. Second, as stated earlier, the number of nodes (hosts and routers) and edges (links) is enormous, and no single entity has complete knowledge of all Internet nodes and edges. Collaborative efforts, such as those by CAIDA [1], Oregon Route Views [2], and Looking Glass [3], have been established to acquire and share information about Internet structure. However, as more data is collected, it becomes obvious that data intended to represent the same properties, if collected with different tools or at different points within the network, can depict significantly different attributes. For example, a recent study of data collected from Internet routing tables led researchers to propose that the Internet's hierarchical structure can be represented by power laws [4]. However, the accuracy of this model is now a topic of debate [5], primarily because earlier findings were based on a limited view of the network.

In addition to the Internet's dynamic nature and decentralized administration, another complication is that the physical view of the of the topology is not necessarily the most relevant one. More specifically, policy-based routing creates a logical overlay on the Internet's physical structure that determines how packets are actually forwarded. Because routing policies reflect business relations,

* Corresponding author. Address: Department of Computer Science, Columbia University, 450 Computer Science Building, 1213 Amsterdam Avenue, New York, NY 10027-7003, USA.

0140-3664/$ - see front matter © 2003 Published by Elsevier B.V.
doi:10.1016/j.comcom.2003.08.021

network administrators frequently do not expose their routing policies.

Data currently being used for topology analysis is often obtained from the routing tables of the Internet's inter-domain routers, or by active, router-level probing techniques that generally rely on eliciting ICMP messages from remote routers or hosts [6]. A number of recent studies [5,7–9] have pointed out issues with utilizing this data to infer the Internet's structure. For example, since routers track reachability only from their location in the network, utilizing the routing tables of a single router, or a limited set of routers, will not provide a complete representation of Internet topology [5,9]. Router-level probing provides additional structural details, but if collected from a limited set of network vantage points, will also not provide a complete representation of the Internet's structure. Further, increasing the number of probing stations quickly reaches a point of diminishing returns [8]. While these studies have provided useful techniques to improve the data collection process, thus far no methods have been reported for fully resolving these issues.

In this study, we are interested in better quantifying the role that current sources of network metrics play in revealing Internet structural attributes. Our intent is not to fully resolve current Internet mapping ambiguities, but rather to gain insights into when representations based on various empirical data are sound, versus when they may be misleading. The remainder of this paper is structured as follows. We begin with background information to introduce potential causes of artifacts in topology data. In Section 3, we present our data collection methodology, and provide an overview of the datasets used in our study. We detail the results of our analysis in Section 4. We offer conclusions and describe plans for future work in Section 5.

2. Background

The Internet is a loose confederation of independent routing domains. Each domain is referred to as an Autonomous System (AS). This structure enables routing within a domain and between domains to be independently administered. Routing within a given domain is achieved using Interior Gateway Protocols (IGP), such as OSPF, IS-IS, and RIP. Routing between domains is achieved using Exterior Gateway Protocols (EGP)—BGP is the Internet's single de facto standard EGP [10,11]. Because our study is focused on the Internet's inter-domain topological attributes, we are most interested in the structure imposed by BGP routing policies. The two primary data collection methods used for inter-domain topology analysis are BGP routing tables and router-level probing. In this section, we discuss the major issues with each of these data sources.

For the remainder of this paper, we will refer to the network of interconnections between routers as a graph, $\mathcal{G}(\mathcal{V},\mathcal{E})$, where $\mathcal{V}$ is the set of nodes (or routers) and $\mathcal{E} \subseteq \mathcal{V} \times \mathcal{V}$ is the set of links, $deg(v)$ is the degree of node v. We use (u, v) to refer to the edge connecting nodes u and v, and Path (u, v) to refer to the sequence of edges traversed by a message flowing from node u to v.

2.1. BGP routing table issues

A BGP router learns of routes to remote networks from its BGP peers. Peers are routers with which a given router maintains a BGP session. Messages exchanged between BGP peers to convey route information are referred to as route advertisements. BGP routers maintain a routing table, with an entry for each remote network. An entry is keyed by the IP address prefix of the remote network, and includes an `AS_PATH` vector. The `AS_PATH` vector is intended to represent the sequence of ASes that would be traversed by a packet destined for the corresponding network, if that packet were forwarded to the next hop router specified in the entry. Each element of the `AS_PATH` is the uniquely assigned AS number of the AS to be traversed. Therefore, each `AS_PATH` provides a list of inter-AS edges. As the following subsections illustrate, there are a number of issues with using this information for inter-domain topology analysis.

2.1.1. Partial information

A router's BGP routing table provides complete information on the inter-AS edges that could be used to reach every known IP address from this router. However, not all routes received in advertisements are stored in the router's local routing table, nor are all of the stored routes advertised to peers. The router's input policy, route selection process, and output policy determine which routes are represented in the local routing table, which are used to forward packets, and which are advertised to peers [12].

Input policy. BGP routers receive advertisements from one or more peers. The router may be configured so that routing information received in certain advertisements is not used in the route selection process. Such routes are said to be filtered by the input policy. For example, advertisements from a given peer or with a given path attribute may be filtered. Routes filtered by the input policy, and thereby not used in the route selection process, are not reflected in the router's local routing table, nor are they advertised to peers.

Route selection process. A router is likely to receive advertisements for multiple routes to the same destination. The route selection process identifies the best routes to each destination, based on attributes such as AS path length or local preference. For example, consider the network depicted in Fig. 1. Because A is directly connected to C and D, edges (A, C) and (A, D) represent the shortest path and are likely to be selected as the best path to C and D, respectively. If so, the router at A would not have edge (C, D) in its routing table. In general, edges radiating outward from a router's vantage point toward remote networks tend

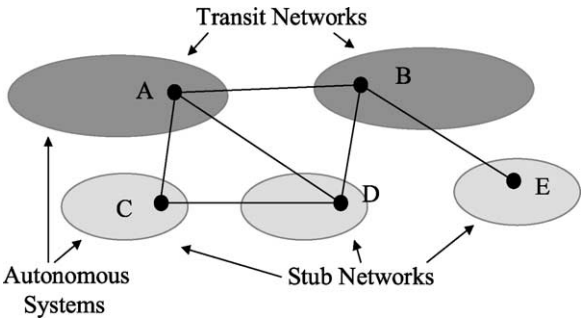

Fig. 1. Example of Internet routing between separately administered, or autonomous, domains. A domain is classified as stub if the path connecting two nodes traverses the domain only if one of the nodes is in the domain. A domain is classified as transit if it does not have this restriction. The practice of having certain domains provide transit for network traffic, while others do not, imposes a hierarchical structure.

to be well represented in its local routing table and advertisements. However, edges connecting the domains of different radiating paths typically are not.

Output policy. Not all routes in the local routing table are advertised to peers. Output policies may be configured on a per peer basis, and determine which of the routes in the local routing table are advertised to peers. Again we consider the topology in Fig. 1. *A* and *B* are providing transit services for *C* and *D*. As customers of *A* and *B*, the administrators of *C* and *D* are unlikely to want traffic from *A* to *B* (or vice versa) routed along path A–C–D–B. The output policy, in this case at *C* and *D*, would be configured to filter advertisement of routes utilizing (C, D) to *A* and *B*. To detect (C, D) in this scenario, a study would need to specifically include routing data collected from either *C* or *D*.

2.1.2. Intra versus inter-domain routing

BGP routing tables represent only the information that is propagated to BGP. Much of the routing policy within an AS is not propagated to BGP routing tables. The separation of intra-domain routing protocols and inter-domain routing protocols can result in an `AS_PATH` that does not reflect how packets are actually forwarded. The following are examples of intra-domain polices that may not be represented in BGP advertisements.

Static routes. Routes may be statically defined, as opposed to being learned from peers. Static routes are typically not advertised by BGP routers, but can be defined to direct, for example, all traffic to a given destination AS via a particular peer AS.

Source routing. The term source-routing is used to describe a policy in which a router forwards all packets with a given source IP address, to a specific next hop. For example, router *B* may be configured to forward all packets with a given source address to *A*.

Multi-hop BGP sessions. A pair of BGP routers may be logically, but not physically, adjacent. That is, a BGP session may be established across an intermediate AS. Packets will be forwarded to the link connecting the intermediate node, and therefore the intermediate node would be represented in the `traceroute` results. It is possible that the BGP advertisements generated at either of the endpoint nodes would not reflect the intermediate node.

2.1.3. AS_PATH manipulation

In addition to filtering route information, a BGP router may manipulate the attributes of the route. Minimally, the BGP router adds its AS number to the `AS_PATH` when it advertises a route to external peers. A router may be configured to manipulate the `AS_PATH` to affect inter-domain routing behavior. One common practice is for the router to repeat its own AS for a number of times, to increase the path length and thereby make a given path less likely to be selected. The `AS_PATH` can also be manipulated to include the AS of a different domain. However, this practice is less common as it can lead to routing loops or black holes.

2.2. Traceroute issues

An alternative to basing topological analysis on BGP routing tables is to actively probe end-to-end paths at the router-level. Router-level probing techniques, such as `traceroute`, generally rely on eliciting ICMP messages from all the packet routers along a network path. We will use the small network depicted in Fig. 2 throughout this section to illustrate the major issues with basing Internet structural analysis on data collected with `traceroute`.

We begin by reviewing the operational characteristics of `traceroute`. A `traceroute` command issued at *A* to probe the path to *D* begins by generating a UDP packet with a Time-to-live (TTL) field set to 1 hop. This packet would be forwarded to *B*, which would detect the expiration of the TTL, send an ICMP Time Exceeded message to *A*, and discard the UDP packet. Because the source address of the IP packet containing the ICMP message is set to *B*'s address, `traceroute` is able to determine that the first hop from *A* to *D* is *B*. `Traceroute` then increments the TTL and repeats the test until a response is received from *D*, or a specified maximum TTL value is reached. In the following subsections, we discuss issues with using `traceroute` results for topology analysis.

2.2.1. AS number resolution

The Internet's address space is allocated in a hierarchical manner. Blocks of the address space are allocated to Regional Internet Registries (RIRs), which, in turn, allocate

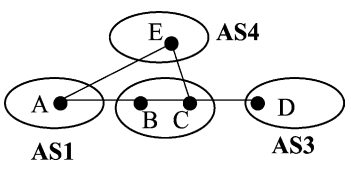

Fig. 2. Example inter-domain network.

address space to local registries who assign the addresses to end users [13]. Address space blocks are uniquely identified by IP address prefixes. Each AS is assigned a unique AS number, and an address space block. Each of the RIR's, ARIN, RIPE, and APNIC, maintain a database of the address space allocation for their region, and support database queries via the 'whois' protocol. Therefore, the AS number corresponding to a given IP address can be determined by querying the whois server of the appropriate registry. The Routing Arbiter Database (RADB) mirrors several regional databases, as well as other databases maintaining routing policy [14].

A commonly used version of `traceroute`, referred to as the NANOG-traceroute [15], maps the IP address received in an ICMP probe reply to its corresponding AS number by querying the RADB. Specifically, it sends a query to whois.ra.net and parses the Origin field. An RADB response may list multiple Origins for a given query. In this case, the Origin associated with the most specific, i.e. longest, IP address prefix prevails.

Registry data. The method used to map the IP address in the ICMP message to an AS number represents the first issue. While some of the regional registries require network administrators to accurately maintain their registry data, others do not. Because changes are not automatically propagated to the registries, this information may be outdated, or simply incorrect.

Multiple AS numbers. A single service provider may register multiple AS numbers, but manage the corresponding address pools as a single routing domain. Suppose, for example, the administrator of the domain containing B and C in Fig. 2 had registered two AS numbers, AS2 and AS2′. B's interface to A, and C's interface to D could be assigned addresses from AS2, while B and C's internal interfaces could be assigned addresses from AS2′. A traceroute probe would indicate an AS path length of 4, whereas the BGP `AS_PATH` would indicate a length of 3.

2.2.2. ICMP message generation

ICMP message generation can also introduce issues. IP Version 4 router requirements [16] specify that the source address in an ICMP Time Exceeded message should be the address of the interface on which the ICMP response is sent. Compliance with this requirement can result in misleading `traceroute` results. Referring again to Fig. 2, suppose the path for packets sent from AS1 to AS3 traverse the path A–B–C–D. The corresponding AS path would thus be AS1, AS2, AS3. Due to ICMP message generation issues, it is possible for a `traceroute` invoked from within A and targeting D, to erroneously report an AS path of AS1, AS2, AS4, AS3, as follows.

The IP address assigned C's interface on edge (C, E) may be from the address pool assigned to either AS2 or AS4. If it was assigned from AS4, and C was configured to route packets to A via E, then the source address in an ICMP Time Exceeded message from C to A would be an address in AS4. In this case, a probe from A to D would indicate a path of AS1, AS2, AS4, AS3. Conversely, the BGP `AS_PATH` vector should correctly indicate a path of AS1, AS2, AS3.

As the issues described in this section illustrate, there are a number of difficulties in utilizing both BGP routing table data and end-to-end probing for Internet topology analysis. In this work, we do not develop or deploy new tools to address these issues, rather our goal is to quantify the effects of these issues on the resulting datasets, and to gain insights into which tools are more appropriate for various structural attributes. In Section 3, we describe the methodology used to collect data used for this experiment.

3. Experimental methodology

We collected four datasets representative of the types of measurements typically used for Internet topology analysis. The goal for our first dataset, $\mathscr{D}1$, was to collect AS path information representing the forward and reverse paths between pairs of nodes. We specifically sought a widely distributed, from both a topological and geographical sense, set of measurement servers from which we could issue `traceroute` commands, and query the local BGP router for corresponding `AS_PATH` data. We also focused on obtaining measurement samples that are representative of paths traversing the Internet core, using tools commonly used in Internet topological studies. In this section, we describe our collection methodology and summarize our evaluation of the datasets against these criteria.

We enlisted Looking Glass sites [3] distributed throughout the world for this purpose. Looking Glass sites provide an HTTP interface to invoke `traceroute` commands to specified IP addresses, and to query the site's local BGP router for the `AS_PATH` associated with an IP address. Many Looking Glass sites do not provide both functions, but we were able to identify 92 sites in which both the BGP query facility and the `traceroute` facility were active. Table 1 provides a list of the Looking Glass sites used in our study. They include sites in 28 countries, including 22 sites in North America, 57 in Europe, 12 in Asia-Pacific, and one in Africa. Additionally, the sites included a variety of Tier 1, 2 and 3 network providers (as classified in Ref. [19]). Thus the collections points were both geographically and topologically diverse.

Measurements were taken over an 18-day period, with each sample consisting of a BGP query and `traceroute` from both ends of a random pairing of Looking Glass servers. To ensure our random sampling did not introduce bias, we verified the distribution of samples to measurement sites. All sites were sampled between 1220 and 1241 times, with a mean of 1230 and standard deviation of 3.9.

Measurement invocations were spaced according to a Poisson distribution, with a mean of ten minutes between measurements from any given server. We chose this somewhat conservative interval between samples to

Table 1
Looking Glass sites used. For those Looking Glass sites that allow queries via different sites, the location represents the site selected

Looking Glass site	Location	Looking Glass site	Location
212.49.128.150	Spain	noc.tele.dk	Denmark
alice.eng.level3.net	San Jose	noc.tele.net	Switzerland
as5388.net	United Kingdom	noc.toplink.net	Herrenberg
Bianaoh.cc.columbia.edu	New York	noc.urbanet.ch	Switzerland
dan.beesky.com	United States	noc.villageworld.com	New York
Debby.sunrise.ch	Switzerland	noc.wsisiz.edu.pl	Poland
doom.net	Massachussetts	onet.on.ca	Ontario
euro-guest.genuity.net	Frankfurt	pegas.carrier.kiev.ua	Kiev
home.mobikom.net	Bulgaria	probe.global-one.nl	Netherlands
kix.net	Seoul	ptlduh00.eli.net	Virginia
lg.1anetworks.net	United Kingdom	reporter.teleglobe.net	Palo Alto
lg.above.net	Atlanta	spirit.interware.hu	Hungary
lg.broadwing.net	Virginia	spnet.net	Bulgaria
lg.carrier1.net	London	stat.cybercity.dk	Denmark
lg.citec.net	South Africa	statistics.kpnqwest.net	Washington, D.C.
lg.cohaesio.net	Denmark	stats.deine.net	Paris
lg.conxion.net	California	stats.mia.net	Wisconsin
lg.drift.telia.dk	Denmark	stats.solnet.ch	Switzerland
lg.dtr.fr	Lyon	support.bbc.co.uk	London
lg.euronet.nl	Amsterdam	traceroute.nnit.dk	Denmark
lg.hea.net	Dublin	widell.net	Sweden
lg.lan.switch.ch	Zurich	www.ams-ix.net	Amsterdam
lg.lasting.ro	Romania	www.bbeng.gxn.net	New York
lg.noc.netscalibur.de	Germany	www.demos.net	New York
Looking-glass.in.bellnexxia.net	Toronto	www.doema.wirehub.nl	Netherlands
Looking-glass.optus.net.au	Australia	www.ebone.net	Bratislava
looking-glass.taide.net	Sweden	www.ecs-ip.net	Amsterdam
looking-glass.teaser.fr	France	www.eng.nac.net	London
lookingglass.tops.net	Bonn	www.gigapop.gen.tx.us	Texas
mail.kamp.net	Frankfurt	www.gitoyen.net	France
Neptune.dti.ad.jp	Japan	www.inetcomm.net	Russia
Netcollect.kpn.net	Amsterdam	www.inoc.imnet.ad.jp	Tokyo
netcon.internet.fo	Torshavn	www.intelideas.net	Madrid
Netmon.grnet.gr	Athens	www.macomnet.net	New Jersey
netstat.netone.com.tr	Istanbul	www.mediasat.ro	Romania
nms1.shinbiro.com	Korea	www.nat.bg	Bulgaria
noc.as8807.net	Denmark	www.noc.easynet.net	Delaware
noc.cetlink.net	Atlanta	www.noc.itgate.net	Milan
noc.colocall.net	Ukraine	www.nordu.net	Stockholm
noc.comstar.ru	Russia	www.opentransit.net	London
noc.datagrama.net	Spain	www.proxad.net	France
noc.kiev.sovam.com	Kiev	www.ripe.net	Amsterdam
noc.ngdc.net	Copenhagen	www.vianw.net	San Jose
noc.petrel.net	Switzerland	www.xmission.com	Utah
noc.runnet.ru	Russia	www.zimage.delbg.com	Bulgaria
noc.support.nl	Netherlands	www2.pt.lu	Luxembourg

ensure that we did not overload any single Looking Glass server. We also tried to reduce the possibility of routing dynamics affecting our measurements, by running the measurements simultaneously. Our measurement tool spawned four threads to collect a single measurement sample, where one thread initiated and processed the `traceroute` query to the source, while the second handled the BGP `AS_PATH` query to the source. In parallel, the remaining two threads collected the `traceroute` and BGP `AS_PATH` at the corresponding target. A sample was considered complete only if all four queries completed successfully.

Of the 116,302 attempted measurements of 8372 uniquely paired hosts, 12% failed due to an error at one of the hosts or an error in reaching one of the hosts. An additional 21% were incomplete due to the server failing before it delivered its results, inability of the probe to elicit a response for each node along the path, or inability of the probe to reach the requested target in one or both directions.

Additionally, we discarded all samples that met any of the criteria listed in Table 2, or for which we were unable to capture forward and reverse BGP and `traceroute` paths. We did not discard measurements with routing pathologies,

Table 2
Criteria for discarding measurements

Measurement discard criteria
Origin traceroute server not responding
Incomplete traceroute output
Node address in 10.x.x.x, 172.16.x.x-172.32.x.x, or 192.168.x.x range
Route did note terminate in target AS
Intermediate node did not respond to ICMP echo
No matching reverse probe for same time period

such as a routing loop, as long as each of the paired probes reached their target. If an `AS_PATH` had been manipulated to repeat an AS, we condensed the repeats to a single instance of the AS. Using these criteria, we were able to capture 26,978 fully paired `traceroute` and BGP queries representing 2840 unique route pairs.

The second dataset, $\mathscr{D}2$, was collected from the Oregon RouteViews website [2]. The website provides access to an archive of routing table snapshots for the Oregon RouteViews project's BGP router. The RouteViews router maintains multi-hop BGP sessions with 57 other BGP routers. These peered routers are widely distributed over the Internet, both from a geographic and network-layer topological perspective, and represent the networks of most large existing Internet Service Providers. For each IP address prefix, the RouteViews router maintains a list of all the AS paths it has learned from all of its peers. Therefore, retrieving the routing table of the RouteViews router provides all of AS path data advertised by all of its BGP peers. The $\mathscr{D}2$ dataset is a collection of all of the BGP snapshots for the same time period during which $\mathscr{D}1$ was collected. The snapshots were collected at 2-hour intervals.

Table 3 summarizes the characteristics of $\mathscr{D}1$ and $\mathscr{D}2$. While the number of unique ASes encountered in $\mathscr{D}1$ is small in comparison to the remaining datasets, we argue that the $\mathscr{D}1$ measurements are typical of routes traversing the Internet core. Our argument is not formal, but is based on recent studies [4,5,17] indicating the Internet core comprises a relatively small number of ASes that provide transit for the majority of the remaining ASes, which are stub networks. The $\mathscr{D}1$ dataset included 100% of the top 20 ASes in $\mathscr{D}2$ when ranked according to degree. Note that these top 20 nodes represent over 40% of the total edges discovered in $\mathscr{D}2$. Thus the sampled paths were representative of network paths that traverse the Internet core.

Table 3
Dataset summary. LG represents the Looking Glass sites listed in Table 1 and ORE represents the Oregon RouteView archive

Dataset	$\mathscr{D}1$	$\mathscr{D}2$	$\mathscr{D}3$	$\mathscr{D}4$
Date collected	3/2002	3/2002	4/2002	4/2002
Collection duration (days)	18	18	11	11
Data source	LG	ORE	LG	ORE
Number of nodes/AS's	337	13,054	7640	13,226
Number of edges	1937	53,816	25,812	55,410

The Looking Glass and Oregon RouteViews Web sites were also used to collect datasets $\mathscr{D}3$ and $\mathscr{D}4$, respectively. $\mathscr{D}4$ was collected in the same manner as $\mathscr{D}2$ but it represents the time period during which $\mathscr{D}3$ was collected. Unlike $\mathscr{D}1$ however, the goal of $\mathscr{D}3$ was not to collect forward and reverse paths between pairs of hosts. Instead, the goal was to capture path information for a large number of random destinations. We used the first routing table in $\mathscr{D}4$ to generate a list of IP address prefixes. Each measurement consisted of randomly selecting a Looking Glass server from the pool of servers listed in Table 1 and then executing a `traceroute` and BGP query. The target of the queries was generated by randomly selecting an IP address prefix from the $\mathscr{D}4$ routing tables, and then generating an IP host address from the corresponding netblock. While this meant that many of the probes were directed at nonexistent hosts, the valid IP prefix enabled the `traceroute` probe and BGP query to collect path information to the target network.

$\mathscr{D}3$ measurements were also made at Poisson intervals with a mean of 10 minutes between measurements initiated at the same host. Of the 62,645 attempted measurements, 14% failed due to an error in reaching or querying the Looking Glass host. An additional 23% were incomplete due to reasons similar to those mentioned above for the $\mathscr{D}1$ dataset. Ultimately, we were able to successfully collect `traceroute` and corresponding BGP path data for 27,185 unique paths.

4. Results

In Section 3, we described the methodology used to collect datasets from a wide variety of vantage points within the Internet. The goal of collecting both BGP routing tables and router-level probes was to capture data reflecting both the advertised portion of routing policy, and the packet forwarding behavior of the corresponding paths. We have pointed out a number of issues with current data sources for Internet topology analysis—issues from which our datasets also suffer. In this section, we attempt to quantify the effect these measurement artifacts have on key topological attributes.

4.1. AS path asymmetry

We begin our evaluation with an analysis of a well-known property of Internet topologies—AS path asymmetry. Paxson [7] defined route asymmetry as the property of having $\text{Path}(u,v) \neq \text{Path}(v,u)$ for any $u,v \in \mathscr{V}$, and used `traceroute` probing to show that nearly half of all Internet paths included a major asymmetry. That is, about 20% of the end-to-end paths differed in two or more of the cities visited and about 30% differed in the ASes visited. Paxson's study was conducted on data collected in 1995.

We used the $\mathscr{D}1$ dataset for this particular analysis. For those IP addresses that were not mapped to an AS

number by the traceroute output, we resolved the AS number by querying the routing registry in a manner similar to the NANOG-traceroute described in Section 2.2. However, if the whois.ra.net query failed, the following order was used to query regional databases: whois.ripe.net, whois.apnic.net, whois.nic.mil, whois.arin.net. If multiple Origin records were reported for the longest corresponding IP address prefix, the most recent record was used. If the response included the AS name assigned the IP address range of the address in question, but not the AS number, we considered all addresses within the specified IP address range to be in the named AS.

Not surprisingly, we found the AS path asymmetry based on data collected in 2002 was significantly higher than what had been reported for 1995 data. At 69.8% AS path asymmetry, the size of the increase was nevertheless surprising. However, several factors may be artificially inflating this statistic, including ICMP message generation issues, outdated routing registry records, and multiple ASes managed as a single domain.

Our first target was the procedure for mapping IP addresses to AS numbers. To gain insights on this issue, we re-evaluated the asymmetry for $\mathscr{D}1$, this time using $\mathscr{D}2$ to perform IP address to AS number resolution. Specifically, we retrieved the AS_PATH for the IP address's network prefix from $\mathscr{D}1$, and used the final AS number in the AS_PATH as the AS number for the corresponding IP address. We refer to this method as BGP-AS-resolution. Using BGP-AS-resolution, we calculated AS asymmetry of 61.4%. While BGP-AS-resolution is based on up-to-date information, it does not reduce potential inaccuracies due to multiple ASes managed as a single domain or ICMP message generation. It may also introduce an issue of hiding ASes due to route aggregation.

We investigated the ICMP message generation issue first. Recall from Section 2.2, if the source IP address for the ICMP Time Exceeded reply is set to the interface on which the ICMP message is sent and this interface's address was assigned from a neighboring AS's pool, traceroute would incorrectly indicate the neighboring AS was traversed. We began by checking the source code of the IP stack for AIX, FreeBSD, and Linux. The FreeBSD and AIX implementations set the source IP address to the address of the interface on which the traceroute probe is received, and therefore would not suffer from incorrectly reporting the neighboring AS. Though this does not adhere to the requirement stated in [16], the specification governs router behavior and not specifically host behavior—a role in which this software is often deployed. We also tested the behavior on three Cisco router models (7500, 6500, 2651), an Extreme Networks (6800) router, and a Windows 2000 server in our lab. All of the Cisco routers, as well as the Extreme Networks router and the Windows 2000 system, set the source IP address to the interface on which the traceroute probe was received. However, the Linux IP stack sets the IP source address to the interface on which the ICMP message is sent. We are also aware of tests of at least one router (Cisco 3660) on which the source address of ICMP Time Exceeded messages are set to the outgoing interface address.

Without a mechanism to determine the IP implementation of intermediate routers, we were unable to isolate which routes erroneously included neighboring ASes. Therefore, our next test was to evaluate asymmetry without using traceroute. Specifically, we calculated asymmetry for the same routes using the BGP AS_PATH of $\mathscr{D}2$. Because the BGP AS_PATH was collected from the router local to the Looking Glass host originating the corresponding traceroute, it should predict the AS path to be followed by the traceroute probe—with the exception of the anomalies listed in Section 2.1.

Many of the BGP routers responded with multiple AS_PATH's for a given destination, and labeled one of the paths as 'best.' The 'best' qualification may be assigned based on attributes such as AS path length or a configured policy. We considered only the BGP path labeled 'best' in calculating AS path asymmetry.

When calculated using AS_PATH, the asymmetry for the same route set was 56.3%. Fig. 3 compares the hop difference distribution for each of the three data sources. While the distribution of routes with one or more AS-hop differences was not significantly different for each of the methods, the nearly 15% difference in number of fully symmetric paths under BGP AS_PATH, as opposed to traceroute with RIR AS resolution, clearly indicates a difference in the traceroute path and that predicted by AS_PATH. Our next challenge was to better quantify this difference on a per route basis.

4.2. BGP AS_PATH prediction of traceroute AS path

In this section, we compare the AS path reported by traceroute with that predicted by the BGP AS_PATH of

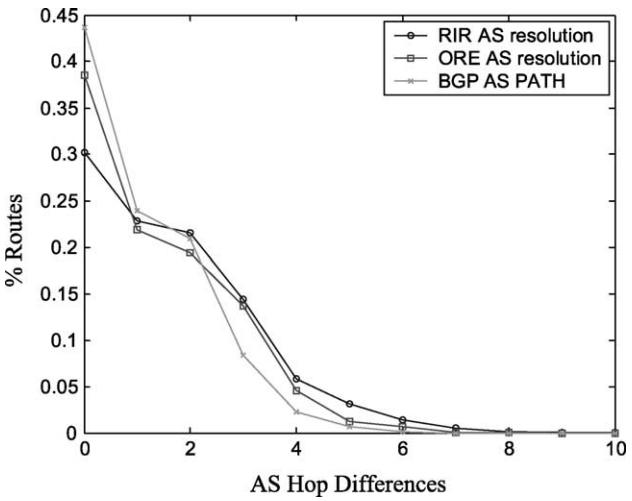

Fig. 3. Comparison of AS asymmetry when calculated using traceroute with RIR resolution of IP addresses to AS numbers, with Oregon Route View resolution of AS numbers, and using BGP AS_PATH data.

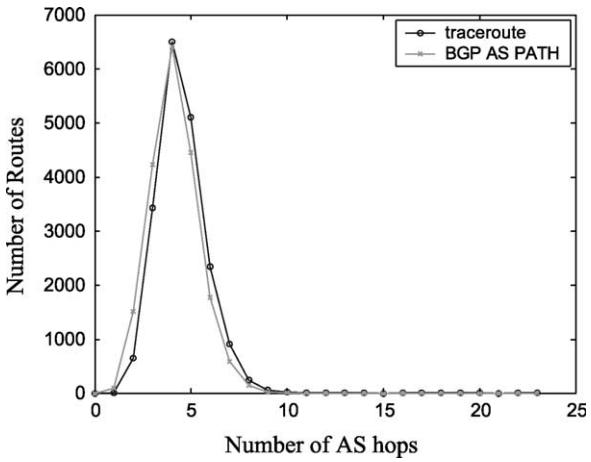

Fig. 4. Comparison of AS path length distribution for traceroute data and BGP AS_PATH data.

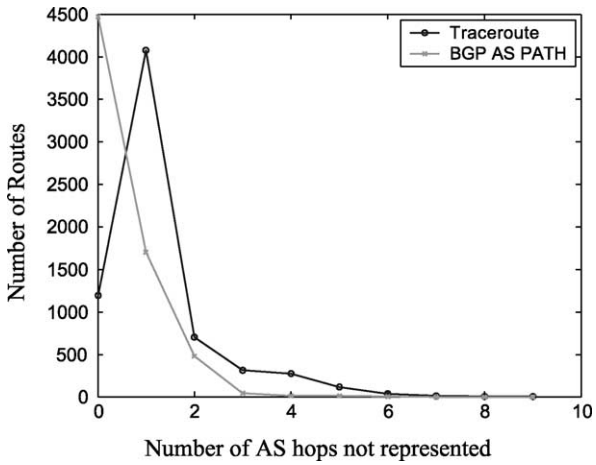

Fig. 5. Comparison of AS hops not represented the corresponding traceroute/BGP AS path. Only paths in which the traceroute path did not match the BGP AS_PATH are represented.

the Looking Glass router local to the node performing the `traceroute`. We started with a relatively simple attribute of the Internet's AS topology—AS path length. For this particular test, we used the $\mathscr{D}3$ dataset so paths would not be limited to those between Looking Glass sites. As described in Section 4.1, we resolved IP addresses to AS numbers by using the RouteViews router's BGP tables for the corresponding time period ($\mathscr{D}4$). We discarded those routes that did not reach at least one node in the target AS.

We found the average AS path length was 4.49 when calculated using `traceroute` probe data, and 4.15 when calculated using BGP `AS_PATH` for the corresponding measurement. Likewise, the AS path length distributions, plotted in Fig. 4, showed little difference when calculated with either `traceroute`, or corresponding BGP `AS_PATH` data.

However, when the data was compared on a per-route basis, the differences were more significant. For example, we found that 32.7% of all measured `traceroute` paths indicated a different path length than the corresponding BGP `AS_PATH`. Recall from Section 2.1, issues with ICMP message generation can cause `traceroute` paths to incorrectly include ASes not actually traversed—thus inflating path length when compared to the corresponding BGP `AS_PATH`. Similarly, BGP-related issues are likely to result in `AS_PATH`s that also differ from actual forwarding behavior—in which case, the `traceroute` path lengths are not strictly longer, nor shorter, than the corresponding `AS_PATH`.

Fig. 5 plots the distribution of routes according to the number of ASes that are not represented in the corresponding `traceroute` or BGP `AS_PATH`. For example, there were approximately 500 routes in which the BGP `AS_PATH` included two ASes that were not represented in the `traceroute` results, and there were approximately 700 routes in which the `traceroute` results included two ASes that were not in the BGP `AS_PATH`. The format of this graph was chosen to illustrate that the `traceroute` paths were not strictly longer, nor shorter, than the `AS_PATH`'s, and vice versa. Additionally, the number of `traceroute` paths that exhibited two or more nodes not in the corresponding BGP path was similar to the number of BGP paths exhibiting the reverse. The results for 0 or 1 hops, however, were quite different. More precisely, 74% of the routes with different path lengths were different because the `traceroute` path included a single additional node that was not included in the BGP `AS_PATH`, and the corresponding BGP `AS_PATH` contained no nodes that were not represented in the `traceroute` path.

We did a visual inspection of the paths with a single additional node and found that in many cases, the single additional node was an exchange point. Exchange points, or Network Access Points (NAPs), provide a collocation site where Internet Service Providers (ISPs) can connect routers to a common network in order to exchange inter-ISP traffic. IP addresses are assigned to ISP routers installed at exchange points from the address space of the exchange point's AS. Because peering arrangements can be directly between ISPs, it is possible that the AS of the exchange point would not be represented in the BGP `AS_PATH`. A `traceroute`, however, is likely to elicit a response from the interface assigned an address in the exchange point's AS. Thus the `traceroute` is likely to include the exchange point's AS.

Another potential cause for a single missing node in BGP `AS_PATH` is route aggregation. For example, consider a scenario with m hops, where the m^{th} hop (AS_m) is a customer of the $(m-1)^{\text{th}}$ hop (AS_{m-1}). As a service provider for AS_m, AS_{m-1} may aggregate network address prefixes, and advertise BGP `AS_PATH`s with AS_{m-1} as the last hop AS for network prefixes belonging to AS_m. The `traceroute`, because it elicits a response from the target interface, which is in AS_m, would include a single additional node, namely AS_m.

4.3. AS degree

A number of recent studies have focused on characterizing the Internet's inter-domain topology according to the distribution of edges per AS node, or AS degree. Representing the Internet's degree distribution with a purely mathematical formulation, as opposed to routing policies and hierarchies, would significantly simplify network analysis. In this section, we use the $\mathscr{D}3$ and $\mathscr{D}4$ datasets, along with the insights gained from the asymmetry and path prediction experiments, to investigate differences in characterizing AS degree.

We calculated the AS degree for nodes represented in $\mathscr{D}3$'s traceroute results, $\mathscr{D}3$'s BGP `AS_PATH` results, and

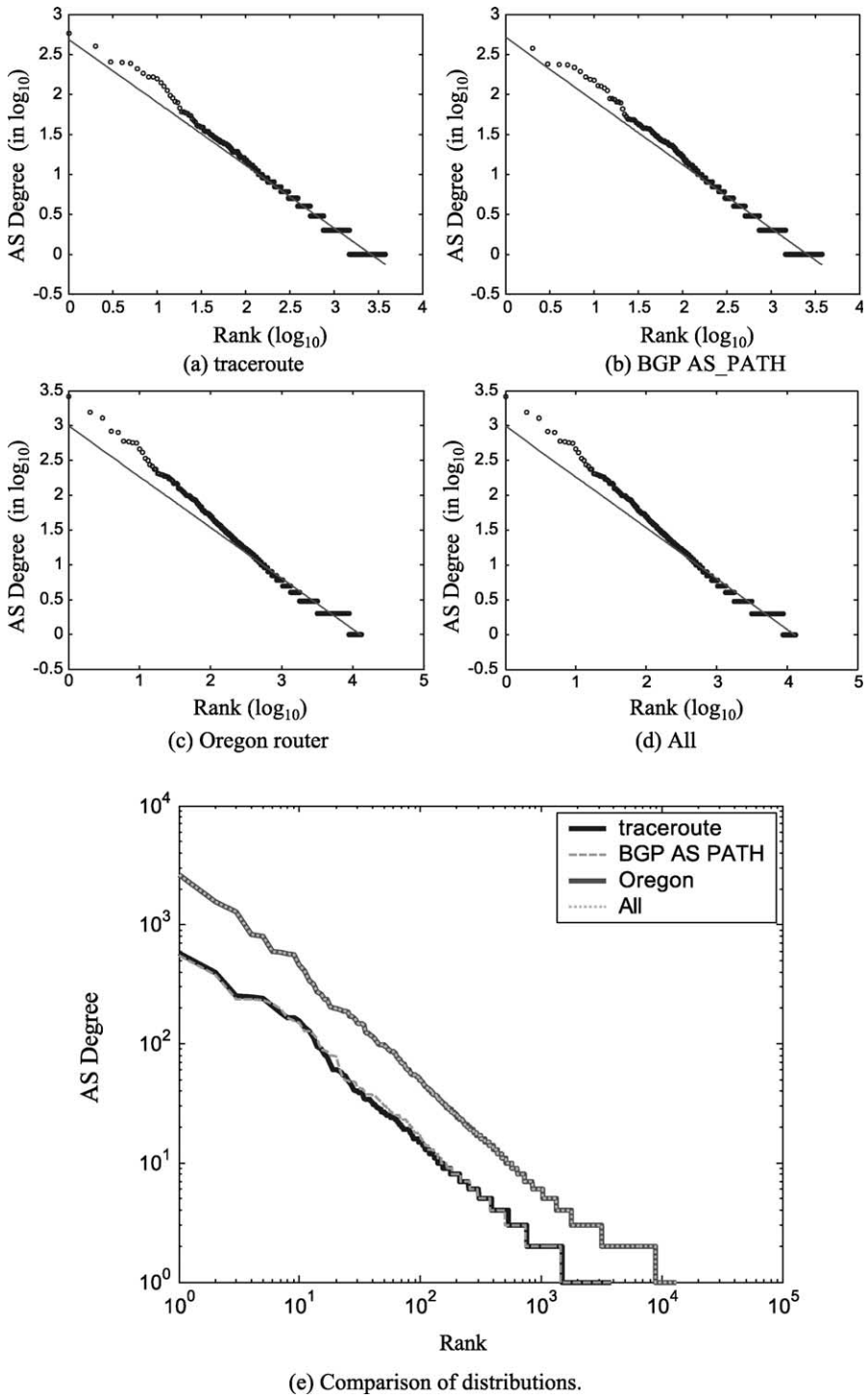

Fig. 6. AS edge degree (y-axis) versus rank (x-axis) in sequence of decreasing degree. Graphs (a), (b) and (c) represent the degree distributions based on the traceroute and BGP AS_PATH of $\mathscr{D}$ and the Oregon routing tables of $\mathscr{D}$, respectively. The distribution in graph (d) represents the aggregation of edges discovered in any of these three datasets. (e) Provides a side-by-side comparison. The solid line in (a)–(e) represents the best exponential fit for the distribution.

$\mathscr{D}4$'s RouteViews routing tables. For each of these datasets, we ranked the ASes in descending order of AS degree and plotted their distribution in Fig. 6(a)–(c), respectively. We also created a third dataset, labeled 'All' in Fig. 6(d), to represent the aggregate of all nodes and edges discovered by all three methods. The first four plots include an overlay of the best fit exponential curve for the distribution, to make the visual comparison amongst the plots easier. The best fit is calculated as a polynomial of degree 2; we chose a curve that gives the minimum deviation from all points to the fitted curve for the distribution. The correlation coefficient for the 'traceroute', 'AS_PATH', 'Oregon', and 'All' distributions and their corresponding best exponential fit were 0.97, 0.97, 0.96, and 0.96, respectively. These high correlations were achieved without discarding any outliers.

The comparison in Fig. 6(e) makes it easier to see several important features of the distributions. First, the distribution obtained from $\mathscr{D}3$'s traceroute data and BGP AS_PATH data are almost completely overlapped. The number of nodes discovered with traceroute was slightly more than discovered with AS_PATH–traceroute data included 18 nodes not represented in the AS_PATH data. This supported earlier findings that, of the paths that differed, the majority differed because the traceroute path included a single additional node. It was, at first, surprising that the AS_PATH data included nearly 200 edges that were not represented in the traceroute data. We did a visual inspection of many of the AS_PATH routes that produced edges that were not represented in traceroute. Of those routes we inspected, most followed the pattern depicted in Fig. 7. That is, the traceroute included a single additional node, and that node was the AS of an exchange point. Moreover, it was the edge directly connecting the ASes immediately before and after the exchange point in the traceroute that was represented in the BGP AS_PATH data and not the traceroute data.

Another interesting observation regarding the distribution for traceroute and AS_PATH data, is that it appears to flatten slightly where the highest degree edges are represented, and taper slightly where the lowest degree edges are represented. This is not surprising and, in comparison with the Oregon RouteViews data, indicates not all of the stub nodes were discovered. It is reasonable to assume that discovering these additional stub nodes would lead to discovering additional edges for the highest degree nodes.

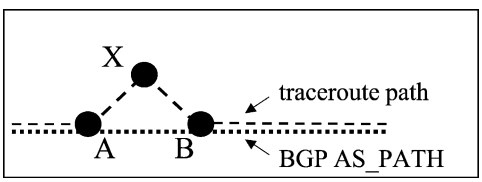

Fig. 7. BGP AS_PATHs with at least one edge (A,B) not found in traceroute dataset, the corresponding traceroute path included an additional node (X).

Fig. 6 also demonstrates that multiple views of the same data can enable a more complete perspective on some topological attributes. That is, the number of nodes discovered using traceroute and BGP AS_PATH queries is substantially less than the number discovered in the Oregon RouteViews data. It is also interesting to note that the distributions for the Oregon RouteViews data, and the aggregation of all three sources were almost completely overlapped. That is, of the more than 3700 nodes discovered in the traceroute and BGP AS_PATH data, only 35 were not already represented in the Oregon RouteViews data.

5. Conclusions and future work

A faithful representation of Internet topological attributes would significantly improve our ability to model, design, deploy, and control the performance of network services. This realization has led to a great deal of effort in measuring and quantifying the structural attributes of the Internet. Analyses are generally based on BGP routing tables and traceroute probing. In this report, we identified a number of issues with the current data sources used for Internet topological analysis and reported on the effects these issues introduce in modeling key topological attributes. Our analysis is based on data collected from a wide variety of vantage points within the Internet, distributed both geographically and topologically. An important feature of our study was the collection of both BGP routing tables and router-level probes, thus ensuring that the data represents both the advertised portion of routing policy, and the packet forwarding behavior of the corresponding paths.

Our experiments demonstrate that for some attributes, such as the average AS path length and path length distribution, representation is relatively consistent across multiple data sources. For other attributes, such as the AS degree distribution, a combination of different data types, taken from multiple vantage points, is necessary to obtain a more accurate view. The advantage of combining multiple views has also been demonstrated by recent work on mapping ISP topologies, which uses BGP data to guide traceroute probing, resulting in improved relevance and reduced network probe overhead [18]. Additionally, we show that even when using multiple views of the network, modeling certain other attributes, such as the actual forwarding path between a pair of nodes and the level of AS path asymmetry, the results can still be misleading.

There remain a number of interesting questions regarding the differences between router and inter-domain level data in representing the Internet structure. Though we collected data for two relatively long periods, our focus was on quantifying differences between data sources, so we did not collect data to allow us to directly compare the two views over time. This extension to our study would address

the question of whether differences between the views are relatively stable, or changing over time. We also anticipate that incorporating the role of an AS may help in resolving some of the AS path anomalies. For example, in Section 4.2 we speculated on potential causes of AS path anomalies, including the presence of exchange points and route aggregation by network service providers. Applying techniques such as those in [19] to classify the role of an AS, and then using it to help resolve ambiguities may prove to be an effective tool in improving the fidelity of Internet mapping.

Acknowledgements

We thank administrators of the Looking Glass servers, whose public service enabled this work. We also thank Alan Crosswell, Bartlomiej Solarz-Niesluchowski, Avi Freedman, Jennifer Rexford, and Morley Mao for valuable discussions on router configuration and Internet Service Provider routing policies, and Bill Rippon and Jisoo Lee for configuring the Cisco routers used in the ICMP message handling experiments. Finally, we are grateful to the anonymous reviewers for their useful and extensive feedback.

References

[1] Cooperative Association for Internet Data Analysis (CAIDA). http://www.caida.org
[2] D. Meyer, University of Oregon Route Views Project. http://www.antc.uoregon.edu/route-views
[3] T. Kernen, Traceroute Organization, http://www.traceroute.org
[4] M. Faloutsos, P. Faloutsos, C. Faloutsos, On the power–law relationships of the internet topology, Proceedings of ACM SIGCOMM, September, 1999.
[5] Q. Chen, H. Chang, R. Govindan, S. Jamin, S. Shenker, W. Willinger, The origin of power laws in internet topologies revisited, Proceedings of IEEE Infocom, June, 2002.
[6] V. Jacobson, Traceroute software, ftp://ftp.ee.lbl.gov/traceroute.tar.gz, 1989.
[7] V. Paxson, End-to-end routing behavior in the internet, IEEE/ACM Transactions on Networking 5 (5) (1997) 601–615.
[8] P. Barford, A. Bestavros, J. Byers, M. Crovella, On the marginal utility of network topology measurements, ACM SIGCOMM Internet Measurement Workshop, November, 2001.
[9] A. Broido, K. Claffy, Internet topology: connectivity of IP graphs, SPIE International Symposium on Convergence of IT and Communication, August, 2001.
[10] J. Stewart, BGP4: inter-domain routing in the internet, Addison-Wesley, Reading, MA, 1998.
[11] Y. Rekhter, T. Li, A border gateway protocol 4 (BGP-4), Internet Engineering Task Force Request for Comment 1771, March, 1995.
[12] S. Halabi, D. McPherson, Internet routing architectures, second ed., Cisco Press, Indianapolis, IN, 2000.
[13] D. Karrenberg, G. Ross, P. Wilson, L. Nobile, Development of the Regional Internet Registry System, The Internet Protocol Journal, http://www.cisco.com/warp/public/759/ipj_4-4/ipj_4-4_regional.html
[14] Routing Arbiter Database. http://www.radb.net
[15] NANOG-traceroute. ftp://ftp.login.com/pub/software/traceroute
[16] F. Baker. Requirements for IP Version 4 Routers, Internet Engineering Task Force Request for Comment 1812, September, 1981.
[17] K. Calvert, M. Doar, E.W. Zegura, Modeling internet topology, IEEE Communications Magazine June (1997).
[18] N. Spring, R. Mahajan, D. Wetherall, Measuring ISP topologies with rocketfuel, Proceedings of ACM SIGCOMM, August, 2002.
[19] L. Subramanian, S. Agarwal, J. Rexford, R.H. Katz, Characterizing the Internet hierarchy from multiple vantage points, Proceedings of IEEE INFOCOM, June, 2002.