Complexity of global routing policies

Andre Broido and kc claffy

Abstract—

In this paper we introduce a framework for analyzing BGP connectivity, and evaluate a number of new complexity measures for a union of core backbone BGP tables. Sensitive to engineering resource limitations of router memory and CPU cycles, we focus on techniques to estimate redundancy of the merged tables, in particular how many entries are essential for complete and correct routing.

We introduced the notion of *policy atoms* [BC01b] as part of a calculus in routing table analysis. We found that the number of atoms and individual counts of atoms with a given number of prefixes properly scale with the Internet's growth and with filtering of prefixes by length. We show that the use of atoms can potentially reduce the number of route announcements by a factor of two, with all routing policies being preserved. Atoms thus represent Internet properties in an accurate way, yet with much smaller complexity.

Several of our analysis results suggest that commonly held Internet engineering beliefs require re-consideration. We find that more specific routes had a relatively constant share of routes in backbone tables across 2000/2001. On the other hand, the churn of more specific routes was much larger than that of top prefixes. We also find that deaggregation of existing announcements is a second major source (beyond announcement of recently allocated address space) of new top (least specific) prefixes in global BGP tables. We also provide examples of misconfiguration and noise in BGP data, including multi-origin prefixes, AS paths with apparent routing loops (some of them due to typographical errors, other actual loops undetected by local BGP speakers), inadvertent transit through customer ASes.

I. Motivation

The aim of this paper is to classify changes in Internet routing characteristics over the last two years. We classify quantitative measures of the Internet's growth and complexity into extensive (volume and size) and intensive (relative and structural) metrics. Our observations confirm that many intensive quantities were invariant over the last two years, and that many extensive quantities were semi-invariant in that they scaled polynomially with the Internet's size growth. This latter scaling was quite often close to linear, i.e. the quantity grew in proportion with the Internet's size.

Global routing in today's Internet is negotiated among individually operated sets of networks known as Autonomous Systems (AS). An AS is an entity that

a) connects one or more networks to the Internet

b) applies its own policies to the exchange of traffic

c) has a globally recognized and unique identifier (AS number)

AS policy is used to control routing of traffic from and to certain networks via specific connections. These policies are articulated in router configuration language(s) and implemented by the Border Gateway Protocol (BGP) [RFC1771] [Stewart99].

A basic BGP exchange consists of a message regarding reachability of a single network via certain router. The reachability information includes an AS path, which is a sequence of ASes. BGP assumes that:

- this path is taken by the reachability message
- the advertised network can be reached via this path.

It is assumed that all ASes in the path forwarded the message deliberately, in accordance with their policies, and *ipso facto* agree to accept traffic destined to the advertised network.

A BGP table associates a network prefix identifier with the AS path through which the network is considered reachable. This table is an important ingredient of the packet forwarding process in a BGP-enabled router. In addition to the AS path, the table contains metrics associated with the paths, which are used for the best path selection in accordance with AS policies. Some of these metrics are specified locally; others are received from neighbors.

In addition to packet forwarding, information stored in a BGP table can be used to monitor aspects of the Internet's architectural evolution, since the data reflects consumption of vital and finite Internet resources [Houston01b]: IP addresses, AS numbers, network prefixes in routing table, CPU cycles in routers, bandwidth consumed by routing update traffic.

The number of networks in the table has bearing on both router memory and CPU cycles. Routing flaps, i.e., advertisements and withdrawals of prefixes, tend to increase with the number of prefixes in the table [Houston01b] [Doran01]. Reduction of prefix count is typically seen as beneficial to infrastructural integrity, and developing mechanisms to do so is thus important architectural research.

BGP inter-domain routing table data enables two different aggregations of IP address space:

1) from IP address (32-bit integers) to longest matching IPv4 network prefix, and

2) from network prefix to AS originating that prefix into the global routing mesh.

An *origin AS* is an AS that appears in the table at the end of an AS path for a network prefix, and is thus assumed to be the AS that originally advertised that prefix. In today's interdomain routing tables, almost 99% of prefixes consistently have a unique origin AS.

The mappings above are useful for converting IP addresses to 'home' prefixes and AS numbers. This usage of

Both authors are with CAIDA, San Diego Supercomputer Center, University of California, San Diego. E-mail: {broido, kc}@caida.org.

Support for this work is provided by the Defense Advanced Research Project Agency (DARPA), through its Next Generation Internet program, and by the National Science Foundation (NSF). CAIDA is a collaborative organization supporting cooperative efforts among the commercial, government and research communities aimed at promoting a scalable, robust Internet infrastructure. CAIDA is based at the University of California's San Diego Supercomputer Center (SDSC). More information is available at the CAIDA website.

BGP data is helpful for tasks such as visualization of Internet AS topology [HBCFKLM00] or for analysis of trends in routability of IP space given by registries to Internet service providers (ISPs) and customers [BC01a].

Analyzing BGP tables can support studies of many questions regarding Internet properties discussed in the Internet community (e.g., NANOG, IETF):

1. How much IP address space is routable?
2. Are allocated IP addresses actually being routed?
3. Which ASes are the most important?
4. How many ASes are single-/multihomed?
5. How many prefixes do ASes advertise?
6. Does the core table grow due to multihoming?
7. Are more-specific prefixes driving up the table size?

In conjunction with connectivity data [BC01a] it can also answer the questions as:

8. Are IP addresses within an AS topologically connected?
9. What is the IP hop diameter of an AS?
10. Can Internet topology, especially particularly highly connected, central, or vulnerable points, be inferred from BGP tables?

A BGP table is useful for answering questions about Internet engineering, when other data sets are prohibitive in size, diversity, or aggregatability. BGP data also has an advantage that its two most basic concepts, *prefix* and *AS* are close approximations to real-life notions of *network* and *administrative domain*.

AS interconnections given by BGP AS paths represent contractual relations between autonomous systems, from which one can infer business models and relations between ISPs [Gao00], as well as assess statistics of inter-AS connectivity (peering sessions) [Faloutsos99]. These uses are theoretical in nature since the presence of an AS in a path does not guarantee that the traffic will be actually carried by this AS. BGP connectivity is declared rather than observed. Nonetheless, such analyses still provide considerable, and in some cases otherwise unavailable, insight into the global routing mesh, especially when taken in conjunction with other types of data.

In this paper we describe a calculus for analyzing complexity of global routing policies. We define and evaluate complexity measures, e.g, the number of *policy atoms*, for a union of core backbone BGP tables. With resource limitations of router memory and CPU cycles in mind, we focus on techniques to estimate redundancy of the merged tables, mainly how many entries are essential for complete and correct routing. These complexity measures can answer questions such as:

1. How many network prefixes does it take to cover the entire routable address space?
2. What is the complexity of the system of AS paths associated with individual network prefixes?
3. How many network prefixes are globally distinguishable with respect to routing policies?
4. How many routing policies are applied by the Internet to addresses originated by one AS?

We will also discuss per-line compression of routing table, in particular

1. How many bits does it take to encode an AS number?
2. How many bits does it take to encode an AS path?

We will also try to illustrate counterintuitive properties of BGP data of paramount engineering relevance:

- presence of prefixes with multiple origin AS;
- apparent routing loops in individual paths and entangled atoms
- existence of different AS paths that ramify (branch) at an AS while reaching the same destination prefix.

In addition to BGP table size reduction, another application for finding aggregate units of routing is the design of Internet active measurement systems. A system that covers many BGP prefixes is more likely to provide relevant information, e.g. about network topology, than a system that covers just a few. However, measurement resource limitations (bandwidth and CPU) require coarser probing frequency for a larger set of monitored objects. A system that covers routes known to be distinct could eliminate redundancy in path probing, maximizing both coverage and sampling frequency.

Significant contribution to the analysis of BGP tables was made by Geoff Houston [Houston 2001a,b], with whom we share many motivations. Our results differ in two ways: we analyze a larger set of tables (up to 33), not just one; and we take only prefixes common to backbone ASes, which avoids the bias of locally maintained routes. We thus capture a more complete inventory of distinct units of routing policy than any previous study. The increased resolution is moderate in numerical terms (a factor of about $\sqrt{2}$, compared to using one table) but significant in a policy context.

II. BGP data availability.

There are several publicly available sources of raw and processed BGP data. Summary plots of several metrics of interest for Australian ISP Telstra's BGP routing table are updated daily [Houston 2001a].

Samples of BGP data from diverse sources are available via a variety of *looking glasses* [CAIDA01], a globally distributed, independent set of servers that support examination of BGP AS paths for individual IP addresses. Looking glasses are primarily intended as debugging tool and cannot provide a full BGP table upon request.

RIPE NCC (Reseaux IP Europeens) [RIPE01] recently started collecting over 70 BGP route views, mostly from European ISPs accredited at LINX (London Internet Exchange). We intend to analyze RIPE's BGP tables in the near future.

Another source of data on inter-AS relations, though not on AS paths, is the `whois` service provided by RIPE, APNIC and some other registries [IRR 2001]. These registries store AS traffic policies in succinct database entries. These data allow analysis of policies as relations between triples of AS: the third AS is announced by the second to the first with an intent to invite the first to use the second AS for transit to the third AS.[1]

[1]We analyze this data in a companion paper where we introduce and study the *constrained dual AS graph*, a policy-oriented representation of inter-AS connectivity in the form of a graph where nodes

Another relevant public collection of data is the records of reservations (to a country), allocations (transfers to ISPs) and assignments (transfers to customers) of address blocks by three authoritative Internet registries, ARIN, RIPE and APNIC [ARIN 2001]. We analyzed this data in [BC 2001a].

A. University of Oregon RouteViews data

The data used for this paper comes from a public source based at the University of Oregon's Advanced Network Technology Center [Meyer 2001a]. RouteViews data is a union of several dozen unpruned backbone BGP tables [Meyer 2001b]. Note that there are other mechanisms to obtain this and similar data, including the RIPE registry database project described above.

Participating peer ASes often contribute more than one routing table view to RouteViews. It is an unsolved question which selection of views yields the best picture of global routing. The question is important since tables are 12M gzipped (over 250M uncompressed) as of July 2001, rapidly consuming disk space if more than one snapshot a day is taken. One of our goals is to find metrics that can help with peer selection when one cannot handle processing data from all peers, such as the marginal utility of adding another peer after N peers have already been merged.

RouteViews daily sampled tables stored at Hans-Werner Braun's server `moat.nlanr.net` [NLANR 2001] start at 1997-11-08 and end in March 2001. PCH [PCH 2001] began storing daily snapshots on 2001-02-19, and RouteViews itself started storing samples every 2 hours from 2001-04-20 [Meyer 2001b].

For this paper, we processed RouteViews BGP tables 1999-04-30, 2000-05-02, 2001-05-01 and 2001-05-03. The dates differ slightly so as to use the largest table within a few days of May 01 of the respective year. To check for five months trends, we also processed 2000-11-28, 2000-11-29 Route View tables.

RouteViews peer participation has grown since inception. All except a few peers contribute a default-free backbone routing table. Remaining peers contribute a table less than half size, usually a few thousand prefixes.

The major difficulty in analyzing trends in RouteViews data is that growth of individual tables is accompanied by a corresponding growth in the number of contributing peers. Continuous additions, drops and changes of peer IP addresses make it almost impossible to find a representative collection of peers with large enough tables present for long enough to do trend analysis. For example, there are only six backbone peers in common in the three tables of 1999-2001. Other fluctuations come from policy changes. Filtering of prefixes shorter than allocated address blocks, enforced by some ISPs, currently causes a relatively large (15%) gap between their table sizes and the rest of the contributors, which was not present in mid-1999 or mid-2000. Diversity measures such as *atoms* discussed in section VII are influenced by the number and choice of peer tables, masking trends one intends to analyze. We thus devote the section IV-D to a discussion of selection of peer tables for analysis.

are ordered pairs of ASes and nodes AB and BC are linked when B was observed announcing C's prefixes to A.

III. BGP tables: properties and caveats

An entry in a BGP table looks like

Network	Next Hop	Mc	Path
12.0.0.0	204.29.239.1		6066 3549 7018
12.0.48.0/20	204.29.239.1		6066 3549 209 1742
	213.200.87.254	20	3257 13646 1742

The first field in the above table entry is the target network prefix; the second entry (an IP address) is the peer who contributed the view. The third is the metric (which is usually multi-exit Discriminator, see [HM 2000]) and the fourth is AS path. We did not show four other parameters that have missing, null or almost constant values, and we will not analyze the metric value either.

The table is ordered numerically by network prefix. Repeated prefixes are not shown. Networks that align on classful boundaries (/8 in Class A, /16 in Class B and /24 in Class C space) are shown without their prefix mask length.

In the example above, the address block 12.0.48.0/20 is originated by AS 1742. Two peers advertise reachablility to it via two AS paths. One of them contains 4 ASes (3 hops) and another 3 ASses (2 hops). Note that this block is a subset (or *more specific*) of address block 12.0.0.0/8; nonetheless, their origin ASes are different and the AS paths diverge a few hops away from a common peer.

A. AS path length

The AS path length is the fundamental metric of a BGP routing table. Internet providers generally over-provision capacity for bandwidth and packet processing in their networks [Atkinson 2001]. This engineering framework implies that packet loss is most likely to occur on AS boundaries, either when passing packets to the customer or to another provider, rather than internal to a backbone itself. If packet loss events on the boundaries are independent, the probability of receiving a packet through an AS path is the product of probabilities of crossing each AS boundary without packet loss. In the absence of reasonable estimates, one can assume probability of such loss-free crossing to be a constant p. The probability of getting a packet through an AS path with n AS crossings would then be equal to p^n. Since $p < 1$, this expression decreases as n grows. It therefore makes sense to minimize AS path length n in order to minimize the probability of systemic packet loss.

And in fact, BGP does. Although it is possible to override this metric on the basis of vendor-specific weights [Cisco 2001] or local preference [Stewart1999], when several paths with equal local preference exist for the same prefix, BGP selects the shortest AS path. This algorithm ostensibly minimizes packet loss properties and also theoretically minimizes hassles involved when something goes wrong between two arbitrary ASes.

The fundamental nature of the AS path length metric in the BGP decision algorithm gives rise to a common practice of prepending extra copies of an AS number to the be-

ginning of the path, to reduce the likelihood of selection of that (now longer) AS path for forwarding traffic. Over 10% of the AS paths in the RouteView table currently exhibit prepending. Prepending can even be explicitly articulated as policy, e.g., in a line from RIPE's `whois` database:

```
Community Definition
NNNN:3062 To LINX PEERS prepend NNNN NNNN
```

where NNNN is an AS number.[2] The practice of prepending results in mistyped lines in most RouteViews tables, which we discuss below.

After path selection occurs, repeated AS names are no longer relevant; we remove redundant prepended instances of an AS in our table before analysis. As far as we can tell, existing analyses of BGP data do not show any sign of awareness of these and some other idiosyncrasies in the RouteViews data.

Selection of shortest AS path is not a policy; it is part of the BGP standard. But there are several mechanisms that can override this standard; we will assume these *are* policies:

1. prepending (own) AS number more than once;
2. use of BGP attributes that take precedence over AS path length
3. traffic engineering tools that update the output of BGP selection process [Meyer 2001d];

We can study the use of these policies and their ramifications by analyzing BGP AS paths.

IV. Other caveats of BGP data.

There are several other caveats in dealing with analysis of BGP tables, the most prominent being the wide fluctuations of more specific prefixes from peer to peer and from day to day, and occasional rises in the use of multiple origin AS prefixes. In section IX we will discuss other caveats, including AS loops and tangles.

The wide fluctuations of table size from peer to peer are mainly due to locally carried customer routes. For address space analysis it is important to note that some of /8s (16.8M addresses) are periodically announced and withdrawn [Houston01a], which represents a substantial fraction (1/3) of new allocations for a year.

Other anomalies include the presence of private ASes, announcements of RFC1918 space (e.g. 10.0.0.0/8), and announcements of default route 0.0.0.0/8.

A salient property of BGP data used for inferring AS connectivity is that it does not guarantee correctness of the announced AS paths. We discuss connectivity distortions by BGP in [BG01c]. A counterintuitive consequence of AS granularity being too coarse is the multitude of ramifications (see section IX) in same-prefix path systems, including *tangles* (AS loops arising from paths passing two ASes in opposite directions while reaching same prefix). Apparent routing loops arise from typos in prepended sequences configured by hand. These typos are consistently present in RouteViews tables. (On 28 June 2001, 7 prefixes originated by one AS are misconfigured by a prepending typo.)

[2] A *community* is an BGP attribute associated with a set of network prefixes.

The ASes involved are (slowly) changing. There are also real routing loops, undetected by the local BGP process. For example, in 02 May 2000 data, 43 prefixes have a loop in combination A B A present at the end of AS A and 5 customer AS announcements, and one prefix by prepending. We analyzed one such annoncement in March 2001 and found that a double-homed AS got its route announced to one upstream back from the regional Internet exchange and then announced it to another upstream without checking for a loop. Misconfigurations of that type appear less frequently so far in 2001 (they were present in March, but none in April-June.) For obtaining correct BGP AS graph, we currently shrink the portion of AS path between first and last instance of repeated ASes, if the AS is the same in both instances. If there are two different ASes in first and last repeated instances, we exclude the whole route announcement from analysis.

BGP data also presents challenges caused by its incompleteness and uncertainty. Some ASes are transit-only and do not advertise their networks; other ASes advertise only part of their blocks, and others either advertise networks that they do not own, or truncate AS paths so that other ASes appear to do the same when many origins are found in the table(s) for the same network. Occasional spikes in the use of multi-origin announcements are not an uncommon feature of BGP tables.

A. Multi-origin announcements.

In our dataset, 27 large (table over 98K) peers for the 03 May 2001 merged table contains 117029 prefixes. 1110 of these have multiple origins. The whole set contains the following number of prefixes with different origins:

origins	1	2	3	4	5	6	Total
prefixes	115919	1080	20	6	2	2	117029

Of 97250 prefixes in the intersection of 27 large tables (global prefixes), 1067 have multiple origin AS. There are 621 different groups of ASes that appear as multiple origins, These groups include 768 ASes out of 10937. The largest number of prefixes multiply originated by an AS is 93.

These numbers differ slightly for the full (37 peer) table, in that there are 1156 prefixes, 642 AS groups, 798 ASes in multiple origin groups, and at most 104 prefixes multiply originated by one AS.

1049 prefixes among 1067 have two origins and 15 prefixes have three origins. 3 prefixes have four origins; they are originated with AS numbers given to European branches (DE, NL, CH and "Europe") of a large US provider.

The following table shows that among 2-origin prefixes, there is often imbalance between number of peers seeing one or another origin.

Min.AS cn.	1	2	3	4	5	6	7	8	9	10	11	12	13
2-or.pref.	185	98	162	78	71	57	65	55	43	52	49	78	56

In about 42.4% cases, the less frequent AS is seen by one, two or three peers. Between 4 and 13 peers the distribution is roughly uniform: if a less frequent origin AS appears in more than three peer tables, chances are about the same that its peer count will be anywhere between 4 and 13. (13

is maximum that the smaller summand of 27 can be.) The data has no gap between prefixes with preferred origin and prefixes for which the origin is undecided, This means that assigning an origin by peer "majority vote" cannot be data-driven without making an arbitrary cutoff choice. Even with such a choice, a non-negligible number of prefixes will remain undecided ("inconsistent", in BGP parlance.)

In a set of May 2000 data that we analyzed, about 10% of the multiple origin AS prefixes arose from truncation of an AS path at the end, i.e. a prefix was advertised both by customer and upstream provider.

As of 1 June 2001, the *union* of the 32 tables contains 2315 multiple origin AS prefixes, 693 groups, and 876 ASes in groups. The maximum number of multiply originated prefixes, 953 and 951, is advertised by two ASes belonging to the same provider mentioned above. (These two AS numbers are used for US networks.) 930 prefixes are advertised by exactly these two ASes. These multiple origin prefixes, almost all of which are /24 subnets of a class A block, are seen by only four peers. Three of these peers see prefixes with one origin (which has much smaller AS number) and one peer with another origin AS.

These fluctuations will be filtered when only prefixes common to chosen peers are analyzed, as we mostly do in the rest of the paper.

B. Private ASes

AS numbers between 64512-65525 are sometimes assigned by a provider to a customer who wants to speak BGP but does not qualify for a legitimate AS number from a registry. Another reason for their appearance is subdivision of an AS into a confederation [HM00] [Stewart99]. Like private addresses, private ASes should not be leaked beyond boundaries of their consensual use.

For 03 May 2000 RouteViews data, out of 335606 AS paths in the union of 27 tables, 32 paths (one in 10K) contain a private AS that was not truncated by upstream. These paths carry 41 prefix announcements. 19 paths out of 31 contain a private AS between two legitimate AS numbers (one of which is the origin), both under 750.

AS sets were designed for use with routes that aggregate less specific and more specific announcements having different AS paths. The motivation was to enable a check for an AS loop in the case where there are more ASes on the line than any single path would contain. Fortunately, this feature is no longer significantly present in BGP tables.[3]

C. Inadvertent transit.

There are signs of inadvertent transit through customer ASes. When the BGP graph is stripped by removing successive transit levels [BC01a], only a small portion, (3%) of the ASes remain. The AS graph of 28 June 2001 obtained from all 41 backbone peer tables (87K prefixes and over) has 11408 AS nodes and 24495 links. There are 333 ASes (2.92%) and 2656 links in the core. The stripping has increased the link/node ratio from 2.15 to 7.98. However, 180 ASes in the core (54%) have outdegree 1. These are most likely customers found in AS paths connecting their upstream providers due to a common BGP misconfiguration: a customer announcing its upstream's routes to another upstream.

[3]There are 125 AS set tokens (instances of an AS set) in the union of 27 tables of 2001-05-03, and only 7 different AS sets. Compared with 10937 Ases, or 3.3M lines in the table, this is not much.

D. Peer selection among RouteViews

We need to find the set of prefixes generally agreed as routable. This set will serve as an input to path identification algorithms, which analyze routing diversity of the globally visible prefixes. To capture this diversity, we need the maximum number of complete peer views. However, as the number of views increases, the number of prefixes shared among can decrease, complicating analysis.

Another problem is that peers contribute (full or partial) tables with sizes varying from a few thousand prefixes to 108K and more. We need to make peer contributions comparable. Otherwise, conclusions will depend upon a mix of data with varying levels of statistical legitimacy. Another problem is that each table contains a mix of globally routed and locally carried prefixes, and proportion of these may vary. It seems reasonable to choose a fixed cutoff for table sizes as a fraction of the maximum size. However, an arbitrary threshold may cut the sample between relatively close sizes. As is often the case with Internet data, the table size spectrum contains several large gaps. The data classifies itself (self-tiers) into the intervals bounded by these gaps. Choosing a cutoff at the gap rather than at a fraction of maximum has an additional advantage of being more robust against an occasional occurrence of inflated tables which leave everyone else behind.

The counts of BGP tables with a given number of prefixes of length /24 or shorter in the 03 May 2001 Route View data are shown below. Table sizes are binned by 1K intervals (1000); e.g. 98K entry counts tables with 98,000-98,999 prefixes. We list only non-zero counts.

#pref,K	105	103	101	100	99	98	92	85	1-8
#peers	3	2	1	6	13	2	2	4	4

The table shows that as of 03 May 2001, there are three different groups of backbone tables with 98-105K, 92K and 85K entries. They likely arise from three different types of route selection policies. The 85K table is obtained by filtering on prefix length according to Regional Internet Registries' (RIR) allocation boundaries (/19 and /20 in class A, /16 in class B) [Houston 2001a]. Filtering policy may be in some cases moderated by the goal of preserving reachability, as can be inferred from data shown in [RBB01]. This is accomplished by leaving least specific (top) prefixes in the table, even if they are longer than RIR boundaries. The most common table size is 99K. An 98K cutoff leaves 27 peer tables. An 85K prefix cutoff leaves 33 peers. We will mainly discuss 27 tables, since this results in the most homogeneous sample.

V. More specific prefixes

A. Internet address blocks

An Internet address *block* is a contiguous interval of integers. Such an interval is competely specified by its lowest end (we call it *base*) and its size. It is referred to as a *CIDR block* if the size is a power of 2 and the size divides the base.

The system of CIDR addressing blocks can be abstracted into a binary tree, in which each node is a bit string of length between 0 and 32, and two nodes are connected if one is a substring of another which is one bit shorter. For example, 192.168.0.0/16 represents a bit string of length 16 for address block which will never have a specific owner and is free for use by anyone [RFC1918]. Its parent node is 192.168.0.0/15.

An individual IP address is given by a 32-bit string. An *address block* consists of leaves of a subtree rooted at a node.

Proposition. *Two CIDR address blocks have an IP address in common, if and only if one is a subset of another.*

Proof. Follows from the uniqueness of path between any two nodes in a tree. Both nodes representing CIDR blocks belong to a common path that starts at the root of CIDR tree (empty string 0.0.0.0/0) and terminates at their common IP address. Thus, one node is a parent of another in the tree.

Definition. An announced address block is called *more specific* if it is a subset of another announced block. This latter block is called *less specific.* An address block A is *immediate* less specific of block B if there is no intermediate block C, which is a subset of A and a superset of B (more specific than A and less specific than B.)

A block is called *top* or *least specific* if it is not more specific of any other block. A top block is called *root* if it has more specifics, and *standalone* or *non-specific* otherwise.

B. Specificity as business relation

The premise that packets are forwarded according to the route of the most specific announcement implies a degree of cooperation from the owners of the less specific block. If anything goes wrong and the more specific announcement will be withdrawn, the immediate less specific receives all traffic destined to the more specific block. However, this is not always a customer-provider relation. A larger ISP can use a subblock out of a smaller ISP's allocation.

C. More specifics in Route View tables

Let us examine how many more specifics exist in the table. The table of 03 May 2001 contains 118379 prefixes in the union of 37 peer views. We select prefixes common to 19 peer tables. Of those, 102095 are common to six or more peers. Among those prefixes, more specifics make up 55210, or 54% of all common prefixes. Another 41731 prefixes are non-specifics. Their address space is covered by no other prefix. 5154 prefixes are *roots* which have more specific prefixes in the intersection of 19 tables. The set of more specific prefixes of a root has the structure of a tree, which is induced by the similar structure of CIDR subdivisions of the whole IPv4 address space. Each tree can be specified by its height and its number of nodes. Height one corresponds to a tree with only one level of hierarchy, i.e. a prefix and one or more immediate more specifics. The distribution of tree height is as follows:

Max chain, pf	1	2	3	4	5
Top prefixes	42188	4237	670	109	9

The number of nodes in the trees (excluding the root node) varies from 1 (1611 trees) and 2 (954 trees) to large unique trees of 1132 and 1243 nodes (with tree height 2) and 1324 nodes (tree height 4). We conclude that there are prefixes with up to 5 less specifics, a fact rarely mentioned in the literature, even though it is easy to observe (we saw it first in May 2000; it occurs in the table of 2001-04-03, but not of 2001-04-20 and 2001-04-29, where the maximum height is 4.)

The distribution of prefixes by specificity level for 03 May 2001 data is given in the table below. The count for level 0 includes roots and non-specifics; the count for level 1 reflects immediate more-specifics of the roots; the count for level 2 includes more-specifics of the latter, and so on.

Level	0	1	2	3	4	5
prefixes	46885	42758	11366	1013	64	9

The address space consumption by prefixes of different specificity level is given in the following table:

Level	0	1	2	3	4	5
IPs	1122M	106M	9.44M	0.67M	24200	2304

Each level of specificity requires at least one decimal order of magnitude less address space than the previous level. In particular, most IP addresses are covered by only one prefix.[4] The extended version of this report describes further implications of specificity [BC2001b].

VI. Evolution of prefix set

The large number of more specific prefixes is generally thought of as contributing to routing table growth. It is thus important to understand whether more specifics are routed differently than their less specific counterparts. The results should also help explain why the use of more specifics is popular.

We examine the distribution of more specifics through atoms. There are 97250 prefixes common to 27 peers in the table of 03 May 2001, out of a total 117K in the union of 27 tables. The total number of more specifics in the union of tables is 70K, of which 51K (52.3% of 97K) are in the atoms, i.e. are carried by all 27 peers. Of the remaining 19K, 2.9K are longer than /24, and 16.2K are shorter than /24, yet not globally visible. (The union also contains 67 top (least specific) non-global prefixes longer than /24.) The remaining part of the section deals only with prefixes of lengths /8-/24.

Among the globally visible more-specifics, 14.84% are in the same atom as their immediate less specific, and for 86.46%, all their less specifics are in another atom. An unexpected group of 355 prefixes (0.70% of 51K) consists of "patches", which are in the same atom with some of their

[4] Some addresses have 6-fold coverage, which is probably as safe as buying five insurance policies for the same type of accident.

less specifics, yet their *immediate* less specific is in another atom. These prefixes revert the policy hole punched by their immediate less specific, making it to conform back to that of a larger aggregate.

Another unexpected group of prefixes consists of non-global less specifics, which however have some globally visible more specifics. Among those, 71 are root prefixes (do not have less specifics themselves) and 86 are more specific. The most prominent of those are 63.64.0.0/11 (651 more-specifics), 63.96.0.0/11 (548), 62.0.0.0/8 (419); the rest have 68 or less more-specifics.

For the 33 peers with over 85K entries in RouteViews[5], the result does not significantly change despite a large drop in the count of more specifics, to 38K, or 45.3% out of 83.7K common prefixes. The number of more specifics present in their atom together with a less specific prefix is 13.64% of 38K, and the number of those that have a less specific in the same atom is 86.45%.

It is clear that more specific prefixes are introduced into the routing table for the purpose of expressing routing policies differing from those carried by larger aggregates, as previously suggested in [Houston 2001b]. Our result also shows strong quantitative agreement with the data given by Houston for January 2001, where he states that among 37.5K more specific announcements, 30K, or 80% of them use different AS paths from their corresponding aggregate and thus are introduced to express different routing policies from that of the larger aggregates.

To analyze medium-range trends in the dynamics of the routing table, we compare the growth of non-specifics, roots and more specifics across six months. We use 28 Nov 2000 RouteViews table with 90K cutoff and 26 peers. We initially tried the 20 November 2000 table until we discovered that one large peer dropped 4K prefixes between 28 and 29 Nov. This results in 4.4K prefixes seen by only 25 peers on 29 Nov, and 3.5K drop in prefixes common to all 26 peers. (For comparison, in the 03 May 2001 table only 760 prefixes have a next-to-largest peer count; for 28 Nov 2000 it is 611.) As we show below, the count of root prefixes (as an example) changed less than that between December and May. Hence, the improper choice of a day in RouteViews can bring our "signal-to-noise ratio " well under 1.

The number of peers common among large peer portions of 2000-11-28 and 2001-05-03 tables is 19.

Prefixes seen by n peers are few when n is over 10 and under $m-3$, where m is the number of large peers with an unfiltered table, which allows for classification of prefixes into locally and globally reachable. It is sometimes the case that a peer does not see a prefix otherwise shared by all large peers. It might therefore make sense to include prefixes whose large peer count is one or two under the maximum. However, to maximize comparability, we will take 19 large peers common in both the 28 Nov 2000 and 03 May 2001 Route View tables, and compare the sets of prefixes common to all 19 peers, which disappeared, emerged, or changed. We will present the same comparison between the set of global prefixes common to all 26 large peers on 28 Nov 2000 and the set of prefixes common to all 27 large peers on 03 May 2001.

In the tables below, "Vac" stands for vacuum (no prefix), counting cases when prefixes emerged, i.e. were not present before (shown as transition from vacuum to matter) or disappeared (transition from matter to vacuum.) "M.sp" stands for more specifics, and "St.al" for standalones; "p" for previous (28 Nov) count, i.e. row sum for more specifics, roots and standalones, and "c" for current (03 May) sum, i.e. sum of column values.

The row index (Root, M.sp etc) denotes the prefix's status on 2000-11-28; the column index its status on 2001-05-03.

Global 19 peers prefixes, Nov.28, vs. May 05.

	St.al	Root	M.sp	Vac	Sum	Out
St.al	30900	740	888	3195	35723	13.50
Root	447	3452	86	352	4337	20.41
M.sp	3247	190	34464	12025	49926	30.97
Vac	7594	643	15250	0	23487	100.00
In	31.60	36.27	32.50	66.30	89986p	
Sum	42188	5025	50688	15572	97901c	

In the next table, we mark prefixes which disappeared from the table in five months as "Out" and those which are new as "In".

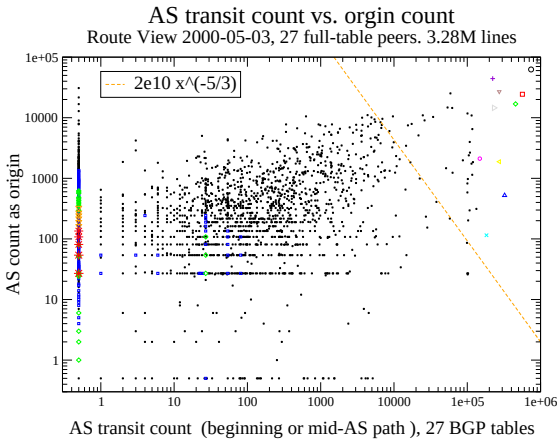
	St.al	Root	M.sp	Vac
Out	13.50	20.41	30.97	100.00
In	31.60	36.27	32.50	66.30
Net	18.10	15.86	1.53	−33.70

Our data does not support the claim that more specifics are driving up the size of BGP table. In the five months between December 2000 and May 2001, the change in the amount of more specifics is insignificant, relative to other sources of variation. Top prefixes have increased by 16 to 18 percent. The difference in changes of root prefixes and standalone blocks (non-specifics) is too small to distinguish. Also, the margin of error here may be as large as 4%, which is the amount of positive change for more specifics when using a table from 29 Nov rather than 28 Nov 2000. Recall that the former table is missing 3.5K global prefixes due to a 4K drop at one large peer.

There is, however, a property that distinguishes more specifics from the rest of the address blocks: they churn (turn over) much faster than top prefixes. Although the rate of addition of new more specifics was comparable with the same rate for non-specifics, the rate of their depletion was 2.5 times as high, resulting in an approximately zero net change. (In a slightly different comparison setup, it is 0.68%). On that scale, roots had intermediate turnaround, which fits their role as middle ground objects in the infrastructure (they are both specifics and top blocks.) This is consistent with more specifics being typically assigned to small ISPs or customer companies that have more volatility.[6]. In general, more dependency causes more churn, and

[5]This table size results from applying a specific common filtering policy [RBB01].

[6]Many .com's jumped on the Net in 1999/2000, had to sign off later in 2000/2001. Such prefixes would wear and tear much faster than other prefixes, pretty much like circulating currency of lower denominations



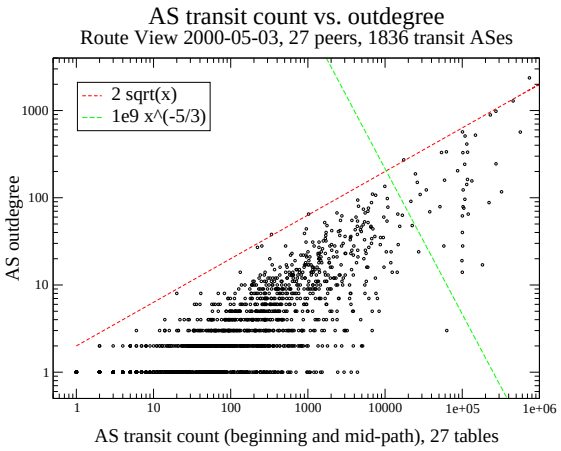


Fig. 1.

the maximal churn is for more specifics (net increase almost zero, but with 30% churn).

A. *Evolution of top prefixes*

The increase in top prefixes comes from four sources: allocation, deaggregation, expansion and aggregation. We next analyze the relative importance of these causes, comparing top prefixes present in the intersection of 19 peers tables on 28 Nov 2000 and top prefixes common to the same set of peers in the 03 May 2001 table.

There are 8237 new top global prefixes in the table of 03 May 2001, compared to the table of global 19-peer prefixes for 28 Nov 2000.[7]

4525 prefixes (55%) covering 38.75M addresses are completely new. They have no related prefixes (more or less specifics) in 28 Nov table. We will label them "allocation". Actual allocation of the address block by registries [ARIN 2001] [BC 2001a] may have occurred long before that.

3712 prefixes (45%) have related prefixes in Nov.28 table. Of those, 3306 (40%) are fully covered by global prefixes from 28 Nov table and 406 (5%) are partly covered.

Of 3306 fully covered, 2941 (35.7%) have one less specific in the table. 150 prefixes (1.8%) have two less specifics, and 0 prefixes have 3 or more less specifics. These prefixes are obtained by *deaggregation.*

215 prefixes (2.6%) out of 3305 fully covered do not have less specifics. These are obtained by *aggregation.* Those can also be augmented by 5% of partly covered prefixes, making total of 7.6%.

The breakdown of sources of the new top prefixes is:

Allocation	55%
Deaggregation	37.5%
Expansion	5%
Aggregation	2.6%

We note that among 5% of prefixes that were partly covered in 2000-11-28 table (which we count as expansion) expansion of just one prefix makes 2.5% and another 2.5% is expansion of more than one partly covering prefix to a larger single block.

[7]The number of new top global prefixes is closer to 8100 when the set of global 19 peer prefixes in 03 May table is compared to the union of all peer tables for 28 Nov.

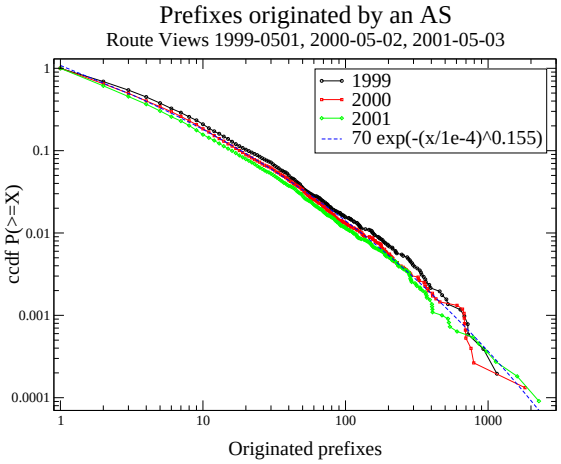


Fig. 3.

VII. BGP atoms and classification of routing policies.

In this section, we explore a new way of grouping addresses on the basis of their global routing properties.

In a single AS setup, we can group prefixes (address blocks) according to the AS path to which they map in the BGP table. The number of groups is then equal to the number of different AS paths. All path counts refer to reduced (non-prepended) paths with no repeated adjacent ASes. Geoff Houston's report [Houston2001a] implements this grouping. It finds 14081 AS paths to 108K prefixes carried by Telstra on 2 May 2001.

We seek a more effective way to group prefixes, which reflects properties of the global routing system rather than its single AS view. Otherwise we risk policy biases inherent in any one specific view, including (but not limited to) strong preference for some AS paths or an excessive number of local prefixes. Moreover, a single AS may not see a policy difference in the routing of two prefixes, simply because they are routed through the same path from it – the projection of data to the 'observation plane' of this AS loses important detail.

We offer a generalization of prefix grouping by AS path:

Definition. Two prefixes are said to be *path equivalent* if we cannot find a BGP peer who sees them with different AS paths. An equivalence class of this relation is called a

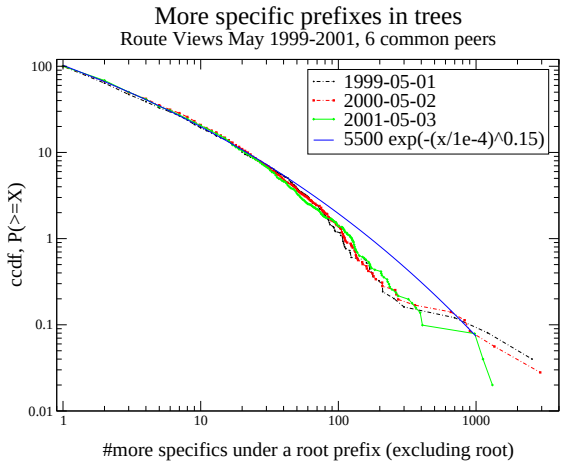


Fig. 4.

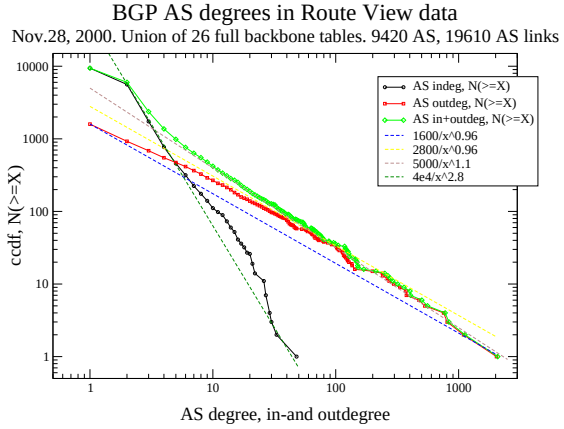


Fig. 5.

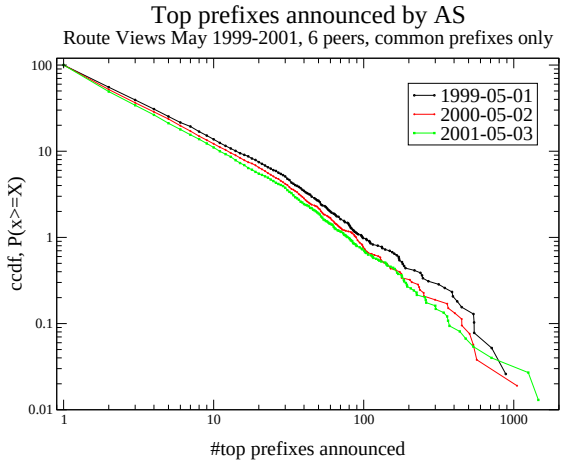


Fig. 6.

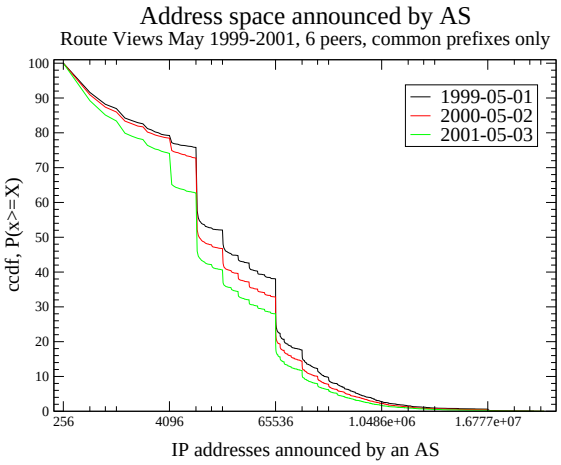


Fig. 7.

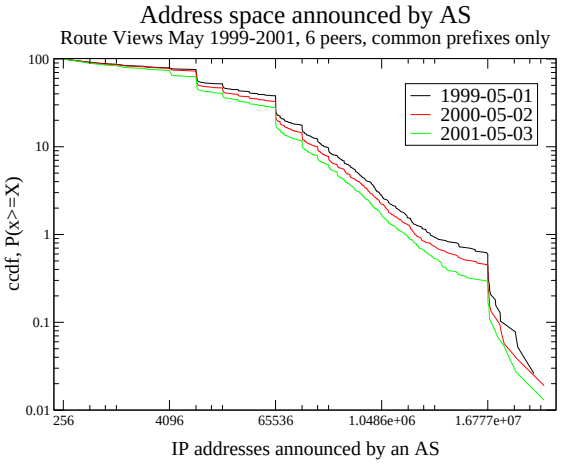


Fig. 8.

BGP atom.

Convenience associated with description of global routing in terms of atoms derives from the fact that an atom captures the part of routing policy relevant to AS paths and applicable to many prefixes at once. Intuitively, an atom can be conceptualized as a double-sided object like a coin, with a system of AS paths on its head and set of prefixes on its tail.

The algorithm for constructing BGP atoms is:

1. Find all prefixes common to a chosen system of peers.

2. Associate a system of peer AS paths with each prefix.

3. For each system of AS paths, find all prefixes that share this system of paths.

The following table shows relevant statistics for the three May tables from 1999-2001 (using May 01 table for 2001) choosing the 6 peers present throughout this period: AS 1, 7018, 3561, 2828 (US) 1755, 3333 (Europe).

Year	1999	2000	2001
AS count	4893	7482	10832
common prefixes	57720	75174	99009
AS paths, max	6603	9859	14207
atoms	8615	12327	17474
atoms/paths	1.30	1.25	1.23
atoms with 1 prefix	3912	5814	8582
largest atom, pf.	1152	1799	2290
crown atoms	4697	6684	9465

The AS and prefix counts in the table differ slightly from those given elsewhere in the paper since we used only six peers. Prefix counts refer to the number of prefixes common to all peers, which were used to compute atoms.

The table shows that the growth in the number of atoms seen by six peers closely follows the growth in the number of ASes, prefixes and AS paths. All four numbers approximately doubled from 1999 to 2001.

The table suggests strong potential for atomic reduction of routing tables. Atoms generalize the grouping of prefixes by AS paths; yet as time goes by, their counts approach the maximum AS path counts (which for this peer selection are consistently from the same European AS.) This evolution may result in the diversity of AS paths seen by one peer eventually becoming close to what can be found by analysis of global collection including dozens of them. However, currently this is not yet the case, as we will find when analyzing tables with 27 and 33 peers.

Approximately 50% of atoms contain only one prefix, though there are also atoms with many prefixes; this is where reduction of the BGP table occurs. The number of one-prefix atoms and the maximum prefix count for an atom have doubled in 1999-2001, matching the evolution of the table size. The cumulative distribution of atoms by size (Figure 9) is close to a Weibull curve [Extreme 2000]

$$P\{n > x\} = \exp(-ax^b)$$

where n is the number of prefixes in an atom, with $b \approx 0.15$ for $x \leq 100$ prefixes. Frequencies of individual counts satisfy the relation

$$P[n = x] \leq cx^d, \qquad x \leq 100$$

with $d \approx 1.8$. It is also possible to approximate $P\{n \geq x\}$ with power function $cx^{-1.3}$ [Faloutsos 1999].

With regard to address space content, sizes under /16 accumulate about 5%. Atoms of total address range of a /16 (65.5K) contribute about 10% of the address space, and a comparable number of atoms. Atoms larger that /16, but smaller than /8, accumulate address space in a logarithmically uniform way, i.e. each binary order of magnitude size contributes approximately an equal number of addresses, for a total of 45% of routable address space. Another 40% of the routable address space is contributed by atoms containing large blocks (/8s, i.e. 16.8M, and more.)

Most atoms are small in terms of address size: more than 60% are under 8192 addresses (/19) and about 75% have sizes under 65K. The only size in which both many atoms exist and a significant portion of address space is covered is

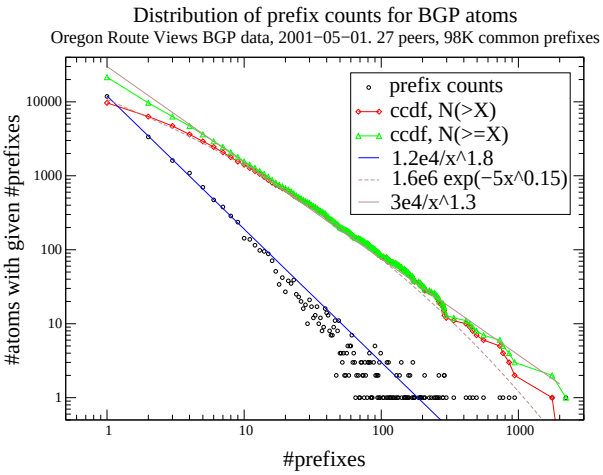


Fig. 9.

/16. Note also that both distributions have convex curves after spikes at powers of 2, where atoms with a large prefix and one or more smaller blocks are accumulated.

We note that the data can easily be identified with a power function rather than Weibull, a fairly common phenomenon in Internet topology analysis. A power function predicts a much heavier tail than Weibull distribution, and has greater deviation from the observed ccdf (complementary cumulative distribution function), especially for large object sizes. The AS degree distribution in the RouteViews table was observed by [Faloutsos 1999] to be close to a power function. Barabasi *et al.* [AJB 2000] used this result to infer far-reaching conclusions regarding resilience of the Internet to directed attacks upon AS nodes, despite the fact that removing a large AS node implies disabling hundreds or thousands of routers at once. In reality, more carefully actively measured data [BC01c] suggests that most Internet object (e.g., IP, prefix, AS node) size distributions, while amenable to rough approximation by a power function (with relative accuracy between 30% and 120%), fit better to a Weibull distribution, often with relative accuracy of 7-15%. A possible explanation of this phenomenon is that Weibull is an *extreme value distribution* [Extreme 2000], which limits the minimum of many positive variables, and Internet object sizes are usually constrained by various resource limitations. It may also be the case that Weibull is a better fit to this data because it has more parameters (degrees of freedom) [Willinger 2001].

The fact that some of the Route Views have BGP AS degree distributions which are close to power functions is therefore an exception, rather than a rule in Internet statistics, even when restricted exclusively to graphs' degrees. For example, comparison of Figure 11 with Figure 5 shows that even different daily samples of BGP data can have very different AS degree distributions, close to power function in one day and further away from it in another.

VIII. Crown atoms

Despite growth in multihoming, a significant number (close to 40%) of ASes are still singly homed, at least as observed in RouteViews data.

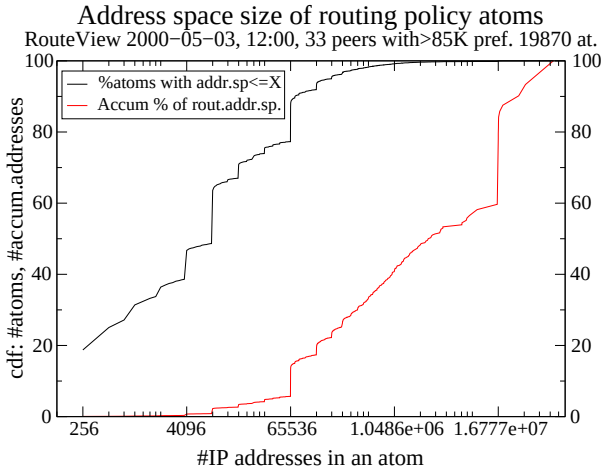


Fig. 10.

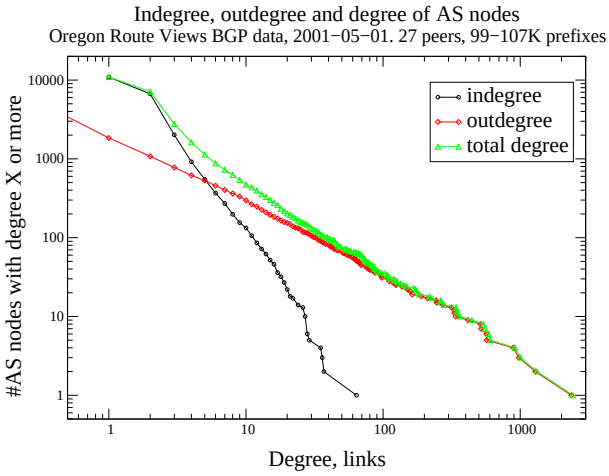


Fig. 11.

From the global routing standpoint, if all paths to AS B pass through AS A, it would make no difference if B's networks were advertised by A. This circumstance is one reason for the existence of multi-origin prefixes. One necessary condition for truncation of paths at B is that it occur in paths consistently always before A and other ASes closer to A than B (this is not always the case, but exceptions are rare, not more than a dozen atoms.)

DEFINITION. An AS B is called a *focal point* for an atom **A** if B is present in all paths and every other AS consistently either follows B or precedes B in all paths in which this AS appears.

A *crown point* is an AS that has the largest number of following ASes among all focal points.

A *crown atom* is an atom whose system of AS paths has been truncated at the crown point.

Note that if an atom has a unique origin AS, this origin is a focal point (with 0 ASes following), and therefore most atoms have an associated crown atom. If there is no focal point, we will let the crown atom coincide with the original atom by definition.

When AS paths are truncated, systems of paths that differ only in the truncated part may become equal, which will reduce the number of atoms (prefix sets with identical AS path systems will be merged into larger atoms.) For the 01 May 2001 data, the number of crown atoms equals 15174, which is 30% less than the number of atoms, and about 7 times less than the number of prefixes.

The reduction is much greater when the number of peers is small. For example, for the 8 peers with the largest number of AS paths, the number of atoms is 20109 and the number of crown atoms is 12452. For the 6 peers present in all data sets, it is 9465. The magnitude of these reductions is not unexptected. The growth in the number of peers, on the other hand, results in growth in the number of observed AS links and in a decrease in single-homing. Crown atoms are likely to lose their advantage as the number of peers grows large. Nonetheless, single-homed networks do exist and for these, truncation to crown atoms will always reduce the length of their AS paths. The problem is, however, that truncated atoms might not have existed, and thus truncation of some crowns may not reduce the number of existing atoms.

A. Dependence of atoms on peer choice

An inherent limitation of our definition of atoms is that it relies on a specific collection of available peer views, and the results will differ using different views. We need to show that for a sufficiently complete set of peers, this dependence diminishes. Ideally, the addition of new peers should not change the system of atoms after we have 'complete' coverage; the equivalence relation is maximally refined, and the set of classes converged to its *projective limit*.[8]

For the data of May 01, 2001, we analyzed other choices of peers than selecting all large peers as we did above:

A. Top 8 peers by AS path count

B. Top 8 peers by prefix count (without AS repetitions)

C. 8 non-US peers

From (C) we ask what is thae cost in coverage if all US peers were removed from our data set?

The next table shows the dependence of atom counts on the choice of peers, in terms of peer IP addresses and peer AS.

Selection	IPs	AS	Pref.	AS pth	Atoms
many prefixes	27	24	97940	14566	21512
many paths	8	8	99140	14566	20109
many prefixes	8	8	99256	14566	19368
non-US peers	9	8	98764	14207	18376

The number of atoms using by 27 big RouteViews peers is 90% reached with the 8 peers with the largest prefix and/or path count. International providers have somewhat smaller resolution, 85.4% of the maximum for 27 peers.

B. Distinct routing policies for an AS

How many different routing policies can exist in the Internet with regard to address blocks in the same AS?

The following table shows data for May 03, 2001, computed with 27 peers seeing a total of 21573 atoms.

[8]A projective limit is the limit of subdivision of existing objects; its dual is the *injective limit*, which is the limit of adding completely new objects [Lang 1992].

Atoms	1	2	3	4	5	6	7	8-65
#AS	6851	2188	856	430	217	150	89	262

In this data, there are 36 different atom counts that can occur for an AS. The maximum number of routing policies applied in the Internet to prefixes from the same AS is therefore 65. The last atom count assumed by two different AS is 30 (counts of 32,33,41,43,46,52,54 and 65 are assumed by one AS each.) The last count with over 10 AS sharing it is 14 atoms.)

These results change slightly with 85K prefixes cutoff which enables taking 33 peers:

Atoms	1	2	3	4	5	6	7	8-60
#AS	6491	1905	738	391	185	146	85	214

Using this method, the atomic reduction of prefix set is huge, almost five-fold, which has compelling implications for the state of Internet engineering.

IX. Ramified atoms and locality of AS policies

A fundamental premise of BGP is that an AS shares only its best paths with its neighbors. If an AS implements uniform routing policies throughout its networks, then packets to a given network from anywhere in that AS will always follow the same AS path to the destination. There would be at most one outgoing link at any node of the atom's AS graph, and the graph would be an (inbound) tree.

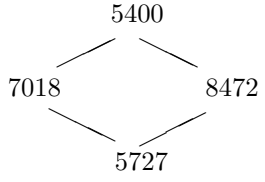
In reality, an AS can have both geographical and logical spread, and different networks within an AS may implement varying policies due to their local conditions, differences in peer interconnection, or loose intra-AS coupling (e.g. companies acquired by a large ISP may still follow their old policies [McCreary 2001].) This policy dependence on local properties of the AS that propagated the connectivity message, though hidden in addressing by AS numbers, can be observed through *ramification* of atoms.

Ramification is a phenomenon where several paths destined for the same prefix can branch or loop at an AS, resulting in the system of paths whose span is a graph with positive outdegrees and/or positive cyclomatic number (defined below.)

For the BGP table of 03 May 2001 (with selection of 27 peers with over 98K prefixes), 7532 atoms (35%) are ramified. This ramification is mainly observed at a handful of ASes, which either participate in RouteViews, or are known to be large providers, or both. ASes which are ramified in the largest number of atoms have counts as large as 3295, 1990, 1533, 1047, 538, 481, 135 and 124. All other ASes are ramified in 35 atoms or less.

For example, one ramified atom in the table for May 01[9] consisted of 7 prefixes with origin AS 8472, with two distinct AS paths:

(1) 7018-5400-8472
(2) 7018-5727-8472



[9]his atom is not found in the table on May 03; prefixes from AS 8472 are no longer common to all 27 peers.

Path 1 is seen by AS 1740 and 3 other peers. Path 2 is seen by AS 7018 and 5 other peers.

AS 7018 is a global provider, present almost everywhere though more dense on the US East Coast, whereas AS 1740 is present only in California. 'Hot potato' routing dictates that an AS should send traffic destined for a non-customer via the point closest to where that traffic entered its network. Consistent with this policy, AS 1740 reaches AS 8472 via a different intermediate provider, and most likely using different parts of AS 7018's infrastructure than those peering with 7018 at the East Coast.

The rest of the path system for this atom is comprised of 40 other ASes that form one or more inbound (fan-in) trees, i.e. graphs with all nodes of outdegree 1 except one (root) node with outdegree 0. We provide a quantitative analysis of atoms' cyclic complexity below.

The next example represents an atom with two cycles in undirected graph. Here, ramified AS paths have different lengths:

1-1239-852-838 6453-1239-852-838
1-701-1691-852-838 6453-701-1691-852-838

We call an atom a *tangle* if it has a directed loop. The table of May 03, 2001 contains 9 tangles, 5 of them the result of a typo in prepended sequence. The following example of a tangle is most likely a result of traffic engineering in two cross-continental backbones. In an atom with cyclomatic number (see below) 7, paths 8 and 9 contain a directed loop between AS 1239 and AS 6461.

8) 6453-1239-2914-11908-6461-4926-4270-4387
9) 6461-1239-5511-4000-7303-4270-4387

There are several causes of ramification. An AS may have a valid engineering reason, (e.g. load balancing) to announce different AS paths to different peers, even if peerings occur within geographical and topological proximity of each other. An AS may not announce to neighbours the best available path or announce the best path to only some of them. Another possibility is that an AS announces a path that has nothing to do with where the traffic is being sent, i.e. to keep knowledge of business relations private. All these scenarios run contrary to the basic assumptions of BGP, but they are allowed under BGP. Such flexibility is often consciously leveraged by those who know exactly what they are doing.

This suggests that ramified atoms are a rule rather than exception. The more observation points we establish, the higher the chance that some paths will enter and exit infrastructure of some global provider at different points, which will manifest as a ramification in the atom's path system. Ramification is then associated not so much with individual atoms, but with multihomed transit ASes with highly diverse policies. It is also an indication that an atom's view provided by contributing peers is reasonably complete.

A. Cyclomatic numbers

An internal structure of the AS graph spanned by an atom can be simplified by recursive removal of nodes with indegree 0 and outdegree 1. ([BC01a].) This algorithm can then be repeated, removing inbound and outbound trees

(second run removes nodes with outdegree 0 and indegree 1.) What is left (if it is non-empty) represents a graph whose symmetrization will contain cycles.

The most common type of cyclic remainder in the current RouteViews inventory of ramified atoms is a diamond-shaped configuration, like that shown above. This remainder has just one cycle when viewed as a non-directed graph.

To measure complexity of ramification, we use the notion of *cyclomatic number* [Harary 1975]:

DEFINITION. The cyclomatic number of a graph equals its count of links minus the count of nodes plus the count of connected components.

Cyclomatic number is a convenient parameter of the graph since it does not change when attached trees are removed. The cyclomatic number of a tree is 0, so it can be computed with or without preliminary stripping of trees off the graph. It measures the dimension of the first homology group of the graph viewed as a simplicial complex [Mac Lane 1995], so that a graph with cyclomatic number c has 2^c different cycles and combinations of cycles.

For 2001-05-03 dataset with the same set of 27 peers as that used on 2001-05-01 (with 90% or more of the maximum prefix count, i.e. 99K or more) have the set of prefixes split into 21573 atoms (61 more than on May 01). The number of ramified atoms is 7532 (27 more). The whole set of atoms has the following distribution of cyclomatic number:

Cyc.num.	0	1	2	3	4	5	6-10
#atoms	14152	5678	1568	145	16	9	5

X. NETWORK AGGREGATION BY ROUTING POLICY.

In this final section we show how to reduce the size of a complete backbone BGP table by a factor of two, preserving all globally visible routing policies, including relations of more- and less-specificity, and using exactly as many IP addresses.

The BGP standard encourages carrying multiple prefixes in an UPDATE message [Rekhter, Li 1995]: BGP is able to carry atoms in its messages at the connection setup when the whole table is exchanged. The standard size of a TCP data segment, 1460 bytes, will have space for at least 300 prefixes, that is, for every atom out of 21.6K except 13 (99% of atoms have 50 or less prefixes.)

This approach could potentially provides a substantial savings on processing overhead and on network bandwidth. If the atoms are agreed upon in advance, – e.g. through the use of globally defined communities or a new transitive attribute, the savings could be leveraged across the global Internet without the need to collect prefixes with equal AS paths each time updates are sent (as suggested by BGP specification.) Assembling updates is even possible with an ASIC or content-addressable memory. However, as prefixes change status, individual advertisements are required. Savings in BGP transmission and processing may be more difficult to realize at that stage.

Recall that BGP encourages advertising aggregates that may be only partly reachable, so as to reduce route flaps. The BGP specification introduces aggregation algorithms that, as we have shown, are not currently used due to conceptual loopholes in the (otherwise streamlined) BGP design. BGP supports aggregation of a route to a more specific prefix with a (different) route for a less specific, creating an AS path that contains AS sets. This solution is intended to throttle explosive growth of more specifics. It creates, however, an ambiguity in the AS path (leaving, at best, only as much information as necessary to detect AS loops, if the path is not truncated when aggregated) which runs contrary to the logic of BGP.

Atoms offer the global routing community a better option: merge together prefixes that share routing policy, and redistribute addresses into smaller number of CIDR blocks.

DEFINITION. Two prefixes belong to the same *routing policy group*, if they have equal

a) Level in the hierarchy of prefixes (counting roots and non-specifics as level 0)

b) atoms (i.e. same AS path from each BGP peer)

c) height of the tree of subordinated more-specifics

d) routing policy group of their immediate less specifics [10]

For the purposes of redivision of IP space into CIDR blocks, which will be done recursively for each level (starting from 0), we will also require addresses in prefixes from one group to be covered by

e) equal CIDR block of the (new) subdivision of address space contained in their less specifics.

It is possible to prove (by induction) that IP addresses that belong to prefixes in one group can be reordered and split into new CIDR blocks with all properties (a)-(e) preserved.

Some prefixes visible to all peers may have less specifics in the union of peer tables that are not globally visible themselves. To avoid underestimating the minimum number of CIDR blocks, (i.e. to avoid being too optimistic about possible table reduction) we will make each non-specific or less specific prefix of length /24 or shorter into a pseudo-atom, for the purpose of computing prefix groups only. This approach increases the reduced table size in two ways, through addition of these prefixes and through avoidance of merging prefixes with different less specifics, when those non-global blocks are included.

The minimum number of CIDR blocks that can accommodate addresses covered by a set of prefixes depends only on address count:

LEMMA. *The minimum number of CIDR blocks containing a (renumbered union of a) set of prefixes that cover n different IP addresses equals the number of 1's in the binary expansion of n.*

Splitting less-specific address space into new CIDR blocks leaves some freedom as to which more specifics end up in which block. We have chosen to proceed from the largest CIDR block (high-order bit in n's binary expansion), filling each CIDR block by more-specifics in the order of decreasing size, and relaying overflow more-specifics to the next CIDR block. This may not be exactly the optimal

[10] d) holds automatically for level 0 prefixes, because their less specifics make up an empty set.

approach with respect to the total number of blocks, but examples show that the gain obtained by different orderings is small.

Each prefix (starting from level 0) is assigned a tag ("first name") containing items from the list above: a) level b) atom c) height e) CIDR block size.

We prepend this with a "last name", which is the full name assigned to the group to which its less specific prefix belongs, including the length of a particular CIDR block to which the less specific group's address space is subdivided. Note that if the subdivision is done according to the lemma, each length is represented by at most one block.

An example illustrates this strategy. The table of 03 May 2001 contains 27 peers who carry over 90% of the maximum prefix count. The number of atoms for this system of peers is 21570, 22.2% of the common prefix count, which is 97250. More detailed data is given in the section which compares evolution of more specifics and top prefixes in one of the previous sections.

When the naming algorithm is applied, the number of CIDR blocks to which the global prefixes are split, is 47589, or 49% of all global prefixes.[11] This shows that the current number of globally routable objects can be reduced by a factor of two with preservation of all routing policies, including subset relations between prefixes.

XI. Conclusion

We have described a framework for analyzing BGP tables. Several topics are expanded more completely in the extended version of this report [BC 2001b]. We have covered idiosyncracies and architecturally revelant trends in current core BGP tables. Some of the analyzed metrics have had sustained growth over two years. Many of them had variable rates, mostly slowing down in 2000/2001. The number of ASes grew fastest, more than doubling over two years.

Our results differ from those we hear from Internet engineers. We find that more specific routes had a relatively constant share of routes in backbone tables across 2000/2001. On the other hand, the churn of more specific routes was much larger than that of top prefixes. We also find that deaggregation of existing announcements is a second major source (beyond announcement of recently allocated address space) of new top (least specific) prefixes in global BGP tables.

We have also listed manifestations of misconfiguration and noise in BGP data, including multi-origin prefixes, AS paths with apparent routing loops (some of them due to typos, some other being true loops undetected by local BGP speakers), inadvertent transit through customer ASes, leaking announcements of private (RFC1918) space and private ASes. Other sources of uncertainty in BGP data include transit-only (non-origin) ASes, allocated but unannounced address blocks, locally carried prefixes, global prefixes missed by a few peers, prefix length filters, partial tables, and possibility of AS path truncation, tampering, or mismatch with the actual traffic path.

We introduced the notion of *policy atoms* [BC01b] as part of a calculus in routing table analysis. We found that the number of atoms and individual counts of atoms with a given number of prefixes properly scale with both the Internet's growth and with filtering of prefixes by length. We also found that atoms' AS path systems can have rich internal structure, and that complex routing policies used by major backbone networks result in many ramified atoms (those with non-tree AS graphs), some of them even with directed cycles. We have shown that the use of atoms can potentially reduce the number of route announcements by a factor of two, with all routing policies being preserved. Atoms thus represent Internet properties in an accurate way, yet with much smaller complexity. We continue to investigate the use of atoms as a framework for analysis of Internet routing in the next decade.

We recognize that any attempt to capture an 'Internet route map' in its entirety will inevitably produce noise-like phenomena that render parts of the data irregular, incomprehensible or chaotic. Neither common-sense assumptions nor specifications and standards should be taken at their face value. The more we study the routing, the more astounded we are that a system with such diversity tends to work reasonably well so much of the time.

[11] If we forget about CIDR blocks, the number of prefix groups (each of which can be merged into one interval of IP space) is 33894.

XII. Acknowledgments

Many thanks to David Meyer of U.Oregon, Sean McCreary of Packets Clearing House, Brad Huffaker, David Moore, and Daniel Plummer of CAIDA and Geoff Houston of Telstra for their helpful feedback and guidance.

References

[AJB00] R.Albert, J.Jeong, A.-L.Barabasi. Error and attack tolerance of complex networks. Nature, v.405, 27 July 2000, 378-381

[ARIN01] American Registry for Internet Numbers. ftp://ftp.arin.net/pub/stats/

[Atkinson01] R.J.Atkinson, end2end-interest posting, Apr.24, 2001

[BC01a] Broido Andre and k claffy, 'Analysis of available IPv4 address space allocation, assignment, routing data', http://www.caida.org/~broido/addr/addr.html

[BC01b] A.Broido, kc claffy, Analysis of Route Views BGP data: policy atoms. Proceeedings of the Network-Related Data Management workshop, Santa Barbara, May 23, 2001.

[BC01c] A.Broido, kc claffy. Internet topology: connectivity of IP graphs. Proceedings of the SPIE conference, Denver, Colorado, August 2001.

[BC01d] A.Broido, kc claffy. Internet topology, 30 pp., in preparation.

[CAIDA01] http://www.caida.org/tools/measurement/reversetrace/

[CR00] E.Chen, Ya.Rekhter. BGP support for four-octet AS number space. draft-chen-as4bytes-00.txt, November 2000. http://www.ietf.cnri.reston.va.us/internet-drafts/

[Cisco01] BGP Best Path Selection Algorithm. http://www.cisco.com/warp/public/459/25.shtml

[Doran01] S.Doran. Routing System Scaling – Disaster Looming, but Medium-Term Fixes Known. Posted to nanog@merit.edu, 2 Apr 2001.

[Extreme00] Extreme value distributions. In: Engineering statistics handbook, Ch.8. National Institute of Standards, 2000. http://www.itl.nist.gov/div898/handbook/apr/section1/apr163.htm

[Faloutsos99] M.Faloutsos, P.Faloutsos, Ch.Faloutsos, "On Power-Law Relationships of the Internet Topology", ACM SIGCOMM'99.

[FLS89] R.Feynman, R.Leighton, M.Sands. The Feynman lectures on physics, v.3. Quantum mechanics. Addison-Wesley, 1989.

[Gao00] L.Gao. On Inferring Automonous System Relationships in the Internet. IEEE Global Internet, Nov 2000. http://www-unix.ecs.umass.edu/∼lgao/ globalinternet.ps

[HM00] S.Halabi, D.McPherson. Internet Routing Architecturess, 2nd ed, Cisco Press, 2000, 498 p.

[Harary75] F.Harary. Graph Theory. Addison Wesley, 1975.

[Houston01a] G.Houston, BGP routing table statistics, updated daily. http://www.telstra.net/ops/bgp/

[Houston01b] Geoff Houston, 'Analyzing the Internet's BGP Routing Table', *The Internet Protocol Journal,* Volume 4, Number 1, March 2001. http://www.telstra.net/gih/papers/ipj/4-1-bgp.pdf

[Huffaker01] B.Huffaker, CAIDA internal presentation, June 2001.

[HMC01] B.Huffaker, D.Moore, kc claffy, in preparation.

[HBCFKLM00] B.Huffaker, A.Broido, kc claffy, M.Fomenkov, K.Keys, E.Lagache, D.Moore, Skitter AS Internet Graph. Published by CAIDA, 2000.

[IRR01] Internet Routing Registry. List of routing registries. http://www.irr.net/docs/list.html

[Knuth97] Donald Knuth, 'The Art of Computer Programming: Seminumerical Algorithms (Vol 2, 3rd Ed)', Addison Wesley, 1997.

[Lang92] Serge Lang, Algebra, 3rd edition, Addison-Wesley, 1992.

[LV97] M.Li, P.Vitanyi. An introduction to Kolmogorov complexity. 2nd ed., New York, Springer, 1997, 637 p.

[Mac Lane 95] Saunders Mac Lane. Homology. Classics in Mathematics, Springer-Verlag, 1995

[McCreary00] Sean McCreary, private communication, 2000.

[Meyer01a] U. Oregon's Advanced Network Technology Center http://www.antc.uoregon.edu/

[Meyer01b] RouteViews, U. Oregon's RouteViews project, http://www.antc.uoregon.edu/route-views/

[Meyer01c] RouteViews, daily updates http://archive.routeviews.org/bgp

[Meyer01d] David Meyer, private communication, 2001

[NLANR97] http://moat.nlanr.net/Routing/rawdata/

[PCH01] Sean McCreary, Bill Woodcock. PCH RouteViews archive. http://www.pch.net/documents/data/routing-tables/

[RBB01] J.Rexford, R.Bush, S.Bellovin. Some Initial Measurements of Prefix Length Philtres. Presentation at NANOG, Scotsdale, AZ, May 21, 2001. http://research.att.com/∼jrex/nanog/lost.html, http://psg.com/∼randy/010521.nanog/

[RFC1771] Y.Rekhter, T.Li. A Border Gateway Protocol 4 (BGP-4) RFC 1771, March 1995. ftp://ftp.isi.edu/in-notes/rfc1771.txt

[RIPE01] BGP data. http://www.ripe.net

[Shannon49] Shannon, Claude. The mathematical theory of communication. Urbana, Univ.Illinois Press, 1949, 117p.

[Stewart99] J.W.Stewart III. BGP4: Inter-Domain routing in the Internet. Addison-Wesley, 1999, 137 p.

[Willinger2001] W.Willinger, private communication, March 2001.