Part I

End-to-End Routing Behavior in the Internet

Chapter 2

Overview of the Routing Study

The large-scale behavior of routing in the Internet has gone virtually without any formal study, the exception being Chinoy's analysis of the dynamics of Internet routing information [Ch93]. In this part of our thesis we analyze 40,000 route measurements conducted using repeated "traceroutes" between 37 Internet sites. The main questions we strive to answer are:

1. What sort of pathologies and failures occur in Internet routing?
2. Do routes remain stable over time or change frequently?
3. Do routes from A to B tend to be symmetric (the same in reverse) as routes from B to A?

Our framework for answering these questions is the measurement of a large sample of Internet routes between a number of geographically diverse hosts. We argue that the set of routes is representative of Internet routes in general, and analyze how the routes changed over time to assess how Internet routing in general changes over time.

We begin by giving an overview of the routing literature in general and, more specifically, how routing works in the Internet (Chapter 3). We find that while routing *protocols* (mechanisms) have been heavily studied, the literature offers very few *measurement* studies of how routing *behaves* in practice.

We then discuss our experimental methodology (Chapter 4). This includes our measurement apparatus, which is the `npd` "network probe daemon" and the `traceroute` utility for measuring Internet paths; the use of exponential sampling, which allows us to apply the *PASTA Principle* [Wo82] as the basis for the generalizations we derive from our measurements; and the use of the *Fisher's exact test* [Ri95] to test for significant differences between different sets of observations. We also discuss which aspects of our measurements are plausibly representative of Internet routing behavior in general (namely, aggregate observations of Internet paths), and which are not (those depending on the behavior of individual sites).

In Chapter 5 we give an overview of the 37 sites participating in the study, and details of the raw data and of the failures encountered when attempting to capture it. We also discuss how we assigned geographic locations to all of the 1,531 routers appearing in the paths we measured.

We performed two separate sets of measurements. The first, $\mathcal{R}_1$, consisted of 6,991 attempted measurements of 689 different paths through the Internet (i.e., distinct source/destination pairs). The $\mathcal{R}_1$ measurements were made with an average interval of 1-2 days between samples.

Upon analyzing the $\mathcal{R}_1$ data, we realized that we could not accurately answer crucial questions regarding routing stability without higher frequency sampling, nor could we unambiguously assess routing symmetry without simultaneous measurements of both directions of an Internet path. To resolve these difficulties, we conducted a second set of measurements, $\mathcal{R}_2$, consisting of 37,097 attempted measurements of 1,056 Internet paths. These measurements were made in two groups, one with an average interval of about 2.75 days between samples, and the other with an average measurement interval of 2 hours. The latter suffices for accurately assessing routing stability. We also *paired* the bulk (80%) of the measurements, conducting back-to-back measurements of the different directions of each Internet path. Pairing allows us to accurately assess routing asymmetries, and also to reduce a source of measurement error (§ 5.2).

Before analyzing the data for routing stability and symmetry, we needed to categorize any anomalies present in order to prevent them from skewing the analysis. In Chapter 6 we classify a number of routing pathologies:

- unresponsive routers, routing loops, routing changes in the middle of measurement, erroneous routes, omission of TTL decrement, and infrastructure failures, all of which were rare;
- host and stub network outages, which were fairly common (but for which our samples are probably not representative);
- and "fluttering," in which the path rapidly alternated between two different routes. In $\mathcal{R}_1$, fluttering was quite common, and sometimes had great impact on the routes of consecutive packets sent by a host. But, like outages, our samples are not persuasively representative, and fluttering was rare in $\mathcal{R}_2$.

Because $\mathcal{R}_1$ and $\mathcal{R}_2$ were made a year apart, we can analyze the relative prevalence of pathologies in each (§ 6.10). We find that the likelihood of encountering a major routing pathology more than doubled between the end of 1994 ($\mathcal{R}_1$) and the end of 1995 ($\mathcal{R}_2$), rising from 1.5% to 3.4%.

After removing anomalous measurements, we analyze the remainder to investigate routing stability and symmetry. This analysis is primarily done using the $\mathcal{R}_2$ data, for the reasons given above. We begin in Chapter 7 by reviewing the importance of routing stability for a variety of network applications. This review reveals that there are two distinct types of stability that are of interest. The first is *prevalence*: whether we are likely to observe the same route in the future as at the present. The second is *persistence*: whether the route we observe at the present is likely to remain *unchanged* for a considerable period of time.

We show that it is easy to assess routing prevalence, and find that Internet paths are strongly dominated by a single prevalent route. But routing persistence is much more difficult to assess, because we have no *a priori* reason for assuming that observing a route at time T_1 and then again at time T_2 tells us anything about whether it changed (and changed back) in between those two measurements.

We tackle this difficulty by first analyzing those measurements we made that were spaced only minutes apart. Doing so reveals that a minority of the paths have routes that persist for only tens of minutes, while the majority persist for significantly longer. After eliminating the quickly changing paths, we repeat the analysis at time scales of 1 hour, 6 hours, and days. We find that, at each time scale, some paths are prone to changes and others are not. Overall, about two thirds of the paths have routes persisting for days or weeks.

A final question concerning routing stability is how an endpoint can determine that its route has changed. We investigate a simple method based on observing changes in the Time To Live (TTL) field. We find that this method provides a useful heuristic, having an overall accuracy of about 95%, but is prone to false negatives (missing the fact that the route has changed), which limits its utility.

In Chapter 8 we turn to the question of routing symmetry. As with routing stability, we first discuss the importance of symmetry for a number of networking applications. We also look at different mechanisms that can introduce asymmetry into network routing. Of these, one in particular ("hot potato" routing between different Internet service providers) is expected to grow in the future, leading to a greater prevalence of routing asymmetry, and the differences in asymmetry between the $\mathcal{R}_1$ and $\mathcal{R}_2$ measurements suggest that this is happening.

Our first attempt at defining whether two routes are symmetric founders on the difficulties of determining whether two Internet addresses do indeed correspond to the same host. In the face of this problem, we revise our definition to consider two routes symmetric only if they visit exactly the same cities. If two routes are *asymmetric* according to this definition, then they visit at least one different city. Such asymmetries are *major* because they likely imply different path characteristics, such as propagation times and congestion levels.

We find that *half* of all Internet paths in $\mathcal{R}_2$ contained a major asymmetry, while only 30% in $\mathcal{R}_1$ did. About 20% of the $\mathcal{R}_2$ paths differed in two or more cities, and about 30% differed in the autonomous systems they visited.

The presence of pathologies, short-lived routes, and major asymmetries highlights the difficulties of providing a consistent topological view in an environment as large and diverse as the Internet. Furthermore, the findings that the prevalence of pathologies and asymmetries greatly increased during 1995 show in no uncertain terms that *Internet routing has become less predictable in major ways*.

A constant theme running through our study is that of widespread diversity. We repeatedly find that different sites or pairs of sites encounter very different routing characteristics. This finding matches that of our previous work [Pa94a], which emphasizes that the variations in Internet traffic characteristics between sites are significant to the point that there is no "typical" Internet site. Similarly, there is no "typical" Internet path. But we believe the scope of our measurements gives us a solid understanding of the breadth of behavior we might expect to encounter—and how, from an endpoint's view, routing in the Internet actually works.

Chapter 3

Related Research

The problem of routing traffic in communications networks has been studied for well over twenty years [Sc77, SS80]. The subject has matured to the point where a number of books have been written thoroughly examining the different issues and solutions [Pe92, St95, Hu95].

A key distinction we will make concerning the study of routing is that between routing *protocols*, by which we mean mechanisms for disseminating routing information within a network and the particulars of how to use that information to forward traffic, and routing *behavior*, meaning how in practice the routing algorithms perform. This distinction is important because, while routing protocols have been heavily studied, routing behavior has not.

3.1 Studies of routing protocols

The literature contains many studies of routing protocols. In addition to the books cited above, see, for example, McQuillan et al.'s discussion of the initial ARPANET routing algorithm [MFR78] and the algorithms that replaced it [MRR80, KZ89]; the Exterior Gateway Protocol used in the NSFNET [Ro82, Re89], and the Border Gateway Protocol that replaced it [RL95, RG95, Tr95a, Tr95b]; the related work by Estrin et al on routing between administrative domains [BE90, ERH92]; Awerbuch's technique of reducing asynchronous networks to synchronous ones to simplify routing algorithms [Aw90]; Perlman and Varghese's discussion of difficulties in designing routing algorithms [PV88]; Deering and Cheriton's seminal work on multicast routing [DC90]; Perlman's comparison of the popular OSPF and IS-IS protocols [Pe91]; and Baransel et al.'s survey of routing techniques for very high speed networks [BDG95].

3.2 Studies of routing behavior

For routing behavior, however, the literature contains considerably fewer studies. Some of these studies are based on pure analysis, such as Bertsekas' study of routing dynamics for different topologies [Be82]; or on simulation, such as Zaumen and Garcia-Luna Aceves' studies of routing behavior on several different wide-area topologies [ZG-LA91, ZG-LA92], and Sidhu et al.'s simulation of OSPF [SFANC93]. In only a few studies do measurements play a significant role: Rekhter and Chinoy's trace-driven simulation of the tradeoffs in using inter-autonomous system routing information to optimize routing within a single autonomous system [RC92]; Chinoy's study of the dynamics of routing information propagated inside the NSFNET infrastructure [Ch93]; and Floyd and Jacobson's analysis of how periodicity in routing messages can lead to global synchronization among the routers [FJ94].

This is not to say that studies of routing protocols ignore routing behavior. But the presentation of routing behavior in the protocol studies is almost always qualitative, such as the discussion of the poor performance of the original ARPANET routing algorithm [MFR78] or the tendency for the revised algorithm to oscillate under heavy load [KZ89]. Even [MRR80], which presents the revised algorithm, and notes that to test it the authors subjected the network during off-hours to a greater volume of test traffic than users generated during peak hours, discuss this stress-testing in general terms, rather than delving into any measurement specifics.

Of the measurement studies mentioned above, [RC92] and [FJ94] are both devoted to examining a tightly focussed question. Only Chinoy's study is devoted to characterizing routing behavior in-the-large, and it remains the only formal measurement study of routing in wide-area networks of which we are aware.[1]

Chinoy found wide ranges in the dynamics of routing information: For those routers that send updates periodically regardless of whether any connectivity information has changed, the vast majority of the updates contain no new information. Most routing changes occur at the edges of the network and not along its "backbone." Outages during which a network is unreachable from the backbone span a large range of time, from a few minutes to a number of hours. Finally, most networks are nearly quiescent, while a few exhibit frequent connectivity transitions.

3.3 End-to-end routing dynamics

Chinoy's study concerns how routing information propagates *inside* the network. It is not obvious, though, how these dynamics translate into the routing dynamics seen by an end user. One of the areas noted by Chinoy as ripe for further study is "the end-to-end dynamics of routing information."

We will use the term *path* to denote the network-level abstraction of a "virtual link" between two Internet hosts. For example, when Internet host A wishes to establish a network-level connection to host B, A need not have any knowledge of the routing infrastructure upon which the Internet is built. As far as A is concerned, the network layer provides it with a link, or *path*, directly to B. Similarly, B has a *path* to A. We will sometimes abbreviate the notion of the path from A to B as $A \Rightarrow B$.

At any given instant in time, the path $A \Rightarrow B$ is realized at the network layer by a single *route*, which is a sequence of Internet routers along which packets sent by A and destined for B are forwarded. We will refer to a single *hop* of a particular route for the path as $R_1 \rightarrow R_2$, indicating that after arriving at router R_1, packets are next forwarded to R_2.

The path $A \Rightarrow B$ may oscillate very rapidly between different routes, or it may be quite stable (an issue we explore in Chapter 7). So Chinoy's suggested research area can be viewed as:

[1] Since publishing some of the results from this part of our thesis [Pa96b], we have learned of a very interesting study of Internet routing, similar in spirit to that of Chinoy's, by Jahanian, Labovitz and Malan [JLM97]. We will discuss this new work in the version of [Pa96b] presently undergoing revision for publication in *IEEE/ACM Transactions on Networking*. We unfortunately learned of the work too late to include discussion of it here.

given two hosts A and B at the edges of the network, how does the path $A \Rightarrow B$ between them behave over time? This is the question we attempt to answer in our study.

3.4 Routing in the Internet

For routing purposes, the Internet is partitioned into a disjoint set of *autonomous systems* (AS's), a notion first introduced in [Ro82]. Originally, an AS was a collection of routers and hosts unified by running a single "interior gateway protocol." Over time, the notion has evolved to be essentially synonymous with that of *administrative domain* [HK89], in which the routers and hosts are unified by a single administrative authority. Within the domain or AS are one or more *routing domains*, which are hosts and routers that communicate using the same routing protocol.

Routing between autonomous systems provides the highest-level of Internet interconnection. RFC 1126 [Li89] outlines the goals and requirements for inter-AS routing (of particular interest for our study are the goals of infrequent loops and stable routes). [Re95] gives an overview of how inter-AS routing has evolved.

When the NSFNET formed the "backbone" of the Internet, inter-AS routing was done using the Exterior Gateway Protocol (EGP) [Ro82, Re89]. A major constraint of EGP, however, is that it requires a tree-like topology between the AS's (with the NSFNET backbone at the root), and, if the topology is violated, loops can result. EGP has since been replaced with the Border Gateway Protocol (BGP), currently in its fourth version [RL95, RG95]. BGP is now used between all significant AS's [Tr95a]. BGP removes the EGP topology restrictions, allowing arbitrary interconnection topologies between AS's. It also provides a mechanism for preventing routing loops between AS's, which we discuss in § 6.3.1 and § 6.3.3.

The key to whether use of BGP will scale to a very large Internet lies in the *stability* of inter-AS routing [Tr95b]. If routes between AS's vary frequently—a phenomenon termed "flapping" [Do95]—then the BGP routers will spend a great deal of their time updating their routing tables and propagating the routing changes. Daily statistics concerning routing flapping are available from [Me95b] (see also [Co91-95]).

It is important to note that stable inter-AS routing does *not* guarantee stable end-to-end routing, because AS's are large entities capable of significant internal instabilities. In our study we focus on end-to-end routing behavior at the granularity of individual routers, though we also note where appropriate how the behavior changes when the granularity is shifted to that of autonomous systems (where the route for the path $A \Rightarrow B$ is viewed as a sequence of AS's rather than a sequence of routers).

One final note: since the publication of Chinoy's study, the Internet has undergone a major topological and administrative change. Inter-AS routing now uses BGP rather than EGP, as discussed above; and the network topology is no longer constrained to a tree with the NSFNET backbone at the root, but has switched to a number of commercial network service providers supporting a potentially arbitrary interconnection topology. Our measurements spanned this transition, with the first dataset taken at the end of 1994, before the NSFNET backbone was decommissioned in Spring 1995, while the second was taken at the end of 1995. Thus, our measurements give us an opportunity to determine whether Internet routing changed significantly during the year separating them. As discussed in § 6.10 and § 8.5, we find significant increases in the prevalence of routing "pathologies" and in routing asymmetry. These changes are not, however, necessarily due to the

NSFNET transition; in particular, two thirds of the routes measured in the first dataset already did not transit the NSFNET, traversing instead commercial providers such as SprintLink and MCINET, or networks outside the U.S.